4-12-2021

Visible-to-Thermal Transfer Learning for Facial Landmark Detection

Domenick D. Poster

Shuowen Hu

Nathan J. Short

Benjamin S. Riggan

Nasser M. Nasrabadi

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 22, 2021, accepted March 23, 2021, date of publication March 31, 2021, date of current version April 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3070233

Visible-to-Thermal Transfer Learning for Facial Landmark Detection

DOMENICK D. POSTER[1], (Member, IEEE), SHUOWEN HU[2], NATHAN J. SHORT[3], BENJAMIN S. RIGGAN[4], (Member, IEEE), AND NASSER M. NASRABADI[1], (Fellow, IEEE)

[1]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA
[2]U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory, Adelphi, MD 20783, USA
[3]Booz Allen Hamilton, Laurel, MD 20707, USA
[4]Electrical and Computer Engineering Department, University of Nebraska-Lincoln, Lincoln, NE 68508, USA

Corresponding author: Domenick D. Poster (dposter@mix.wvu.edu)

ABSTRACT There has been increasing interest in face recognition in the thermal infrared spectrum. A critical step in this process is face landmark detection. However, landmark detection in the thermal spectrum presents a unique set of challenges compared to in the visible spectrum: inherently lower spatial resolution due to longer wavelength, differences in phenomenology, and limited availability of labeled thermal face imagery for algorithm development and training. Thermal infrared imaging does have the advantage of being able to passively acquire facial heat signatures without the need for active or ambient illumination in low light and nighttime environments. In such scenarios, thermal imaging must operate by itself without corresponding/paired visible imagery. Mindful of this constraint, we propose visible-to-thermal parameter transfer learning using a coupled convolutional network architecture as a means to leverage visible face data when training a model for thermal-only face landmark detection. This differentiates our approach from models trained either solely on thermal images or models which require a fusion of visible and thermal images at test time. In this work, we implement and analyze four types of parameter transfer learning methods in the context of thermal face landmark detection: Siamese (shared) layers, Linear Layer Regularization (LLR), Linear Kernel Regularization (LKR), and Residual Parameter Transformations (RPT). These transfer learning approaches are compared against a baseline version of the network and an Active Appearance Model (AAM), both of which are trained only on thermal data. We achieve a 6.5% - 9.5% improvement on the DEVCOM ARL Multi-modal Thermal Face Dataset and a 4% improvement on the RWTH Aachen University Thermal Face Dataset over the baseline model. We show that LLR, LKR, and RPT all result in improved thermal face landmark detection performance compared to the baseline and AAM, demonstrating that transfer learning leveraging visible spectrum data improves thermal face landmarking.

INDEX TERMS Biometrics, face recognition, infrared imaging, landmark detection, thermal sensors.

I. INTRODUCTION

Landmark detection is a critical component for facial analysis applications, including face recognition, 3D modeling, and expression classification. Precise and accurate detection of facial landmarks enable faces to be registered (or aligned) to a common frame of reference, often referred to as canonical coordinates. Face registration may be performed using a variety of approaches such as similarity transformations [1], affine transformations [2], projective transformations [3], or frontalization methods [4].

A significant amount of landmark detection research has been performed on visible spectrum imagery under a wide array of conditions, such as variable pose, illumination, expression, and occlusion, driven by applications in the commercial and government sectors. There has been substantially less landmark detection research for thermal infrared imagery. The primary advantage of thermal imagery is it can be captured by a passive system requiring no illumination. It has been shown that the fusion of thermal and visible face images can lead to increased facial recognition performance [5], [6]. However, for surveillance in low-light/nighttime settings without active illumination, thermal imaging is oftentimes used alone. Yet, landmark-annotated face datasets in the thermal spectrum contain substantially

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

fewer subjects than those in the visible spectrum. For example, the UMDFaces [7] and the CelebA+ [8], [9] visible face datasets contain over 8,000 subjects, whereas some larger, commonly used thermal face datasets [10]–[13] contain between 94 to 238 subjects. Therefore, it would be desirable if the vast amount of visible face data available could be leveraged to help train a thermal face landmark detection system.

The key assumption of this work is that visible domain data is only available during training. At deployment/test time, the model must operate only on thermal data. As such, we do not investigate multi-modal fusion techniques because they require both visible and thermal data at test time to perform inference. Using transfer learning, we leverage visible domain data in conjunction with thermal data during training. Although visible data is used during training, the resulting thermal face landmark detection model is intended for deployment on thermal domain data (i.e. in low-light or nighttime scenarios).

Given that faces captured in either domain represent the same physical entity and share the same underlying geometric structure, it is reasonable to assume there is knowledge to be leveraged from the visible domain which can bolster the performance of a thermal-only model. The human ability to identify facial landmarks in thermal imagery, even for individuals with little prior experience with thermal imagery, is an inspiration for seeking a means of cross-domain knowledge transference in automated facial landmark detection.

Transfer learning aims to translate the knowledge embedded in a model trained on a source domain to a target domain model that typically has less available labelled data. For example, Siamese networks [14] learn domain-invariant features through weight sharing and have been shown to be effective for heterogeneous face recognition in the case of matching target domain probes to a gallery of source domain images [15]. An alternative to weight sharing is to establish linear or non-linear relationships between the weights of the two networks [16], [17]. This strategy learns to translate the feature detectors from one domain to another, as opposed to aligning the data distributions of the domains through the extraction of domain-invariant features.

The objective of this work is to assess the potential of visible-to-thermal transfer learning for the purpose of enhancing thermal face landmark detection. Specifically, the contributions of this paper are:

- The first investigation into the use of parameter transfer learning for enhancing thermal face landmark detection by leveraging visible face data during training.
- The implementation and comparative analysis of four parameter transfer learning techniques (Siamese networks [14], Linear Layer Regularization (LLR) [16], our proposed Linear Kernel Regularization (LKR), and Residual Parameter Transformation (RPT) [17]) versus a baseline network and an Active Appearance Model (AAM).
- A series of ablation studies informing how to generate thermal-visible image pairs by assessing the impact of image alignment, data augmentation, and dataset sub-sampling on transfer learning.
- A thorough evaluation of the models on the DEVCOM ARL Multimodal and RWTH Aachen University Thermal Face Datasets, as well as a comparison to human level performance.

The remainder of the paper is organized as follows. Section II reviews related thermal face landmark detection and transfer learning methods. Section III describes the baseline and transfer learning models in greater detail. Section IV presents and discusses the experimental results. Section V concludes this paper with a summary of contributions and results as well as brief thoughts on future work.

II. RELATED WORK

A. THERMAL LANDMARK DETECTION

Much of the prior research in thermal landmark detection has focused on certain facial sub-regions, in particular the eyes and nose, as these landmarks are often integral components of a larger task. Bourlai and Jafri [18] utilized the pronounced appearance of hair in medium wave infrared (MWIR) to detect eyebrows from horizontal integral projections and thereby limit the search space of a template-based matcher. Working with long wave infrared (LWIR), Wang *et al.* [19] also used integral projections to detect eyeglasses and estimate pupil locations. If eyeglasses are not detected, then a Support Vector Machine classified eye regions from Haar-like feature. Hussien *et al.* [20] found Histogram of Gradients (HoG) to outperform Local Binary Patterns and Haar features when training a cascade classifier to detect eye regions. Tzeng *et al.* [21] localized nose regions in video frames based on the temperature fluctuations caused by breathing.

More recent efforts are geared towards detecting a holistic face shape comprised of 5 to 68 landmarks. Kopaczka *et al.* [22] used Active Appearance Models (AAM) trained on HoG and Scale-Invariant Feature Transform (SIFT) features to perform landmark tracking in thermal videos. Avoiding landmark detection altogether, Sun and Zheng [23] aligned visible and thermal image pairs by iterative point-to-point matching with Canny edge maps. Poster *et al.* [24] compared the performance of three deep-learning based landmark detection methods on thermal face data, including the Multi-Task Convolutional Neural Network (MTCNN) [25] and Deep Alignment Network (DAN) [26] architectures, achieving the best performance with DAN. Kopaczka *et al.* [12] found DAN to outperform AAMs in both landmark detection precision and speed.

In these prior works, only thermal face images were used to train the models. The recent success of deep learning based face landmark detection methods on visible spectrum data is due in part to the increased availability of annotated visible

face datasets. In this paper, we focus on assessing the benefit of using transfer learning to draw upon the information provided by visible face data.

B. TRANSFER LEARNING

A key objective of transfer learning is to improve performance on a target domain by leveraging data from a source domain. Typically, the source domain contains ample training data as compared to the target domain. We use the term source model or target model to refer to models based on the domain from which they receive their input.

While it seems intuitive that there is useful and relatable information to be shared between the visible and thermal domains, it is less clear exactly what or how knowledge should be shared. To the best of our knowledge, the only research on transfer learning for the purpose of landmark detection on thermal imagery was conducted by Wu and Ji [27], who proposed a constrained Deep Boltzman Machine for learning a common feature subspace between the thermal and visible domains, framing the problem of landmark detection as a classification task. However, existing DBM architectures are limited in terms of their scalability and training time compared to Multi-layer Perceptrons (MLP) or Convolutional Neural Networks (CNNs), which have seen more widespread success [28].

Instance, feature, parameter, and relational knowledge are four different categories of transfer learning techniques [29], with feature and parameter transfer learning being the most relevant to landmark detection. A common form of feature transfer involves learning a shared representational subspace wherein extracted image features from two distinct domains exhibit the same distribution. The Maximum Mean Discrepancy (MMD) [30] and the Kullback-Leibler Divergence [31] are two commonly used measures for minimizing the difference between the global statistics of the feature representations. However, these measures are typically used when the feature representations are optimized for classification tasks.

Alternatively, parameter transfer learning seeks to learn a beneficial relationship between the parameters of a model ingesting data from a source domain and the parameters of a model operating on a target domain. Parameter transference binds the models together by their feature extraction and transformation processes instead of by the literal features the models produce. For example, in the case of convolutional layers, parameter transfer learning models the relationship between the weights of the convolutional kernels of the source and target networks, whereas feature transfer learning models the relationship between the output feature maps. As MMD and KL-Divergence are computed from the global statistics of the feature maps, we instead focus on parameter transfer learning techniques which can more easily take into account the spatial information of the convolutional kernels.

A straight-forward method of parameter transfer learning is a Siamese network [14], [32] wherein two networks share the same weights. Rather than sharing the model parameters, Rosantsev *et al.* [16] penalized weights in the target network for which a linear transformation of the corresponding weights in the source network could not be learned, in addition to minimizing the MMD between the feature representations output by the source and target networks. Our work omits the MMD loss term because of its ambiguous applicability in regressing landmark locations from the feature representations, as opposed to classifying images. Further relaxing the constraints on the model parameters, Rosantsev *et al.* [17] learned a residual transformation from the weights in the source network to the corresponding weights in the target network, again for the goal of image classification. Instead of MMD, they employ an auxiliary discriminator network to classify the domains of input samples, which is trained in an alternating fashion with the main network. This auxiliary domain classifier is also omitted from our version of the model in order to examine the parameter transfer approach in isolation.

III. VISIBLE-TO-THERMAL TRANSFER LEARNING

In this section, we formalize the problem of face landmark detection within a single domain, namely thermal imagery, using a baseline CNN architecture. We show how the baseline architecture is then extended into the proposed coupled network framework to facilitate transfer learning. Four different methods for performing visible-to-thermal parameter transfer learning are presented in detail.

A. BASELINE ARCHITECTURE

The baseline face landmark detection architecture utilized for this study is a single stage version of the Deep Alignment Network (DAN) [26]. The DAN is a VGG-like [33] CNN, composed primarily of eight convolutional layers and two fully-connected layers, that regresses a set of fiducial landmark coordinates from an input face image. The DAN was originally designed to detect landmarks in visible face imagery. However, [12] and [24] have demonstrated successful landmark detection to a certain degree when trained using only thermal imagery. A more detailed breakdown of the layers of the baseline network is given in Table 1. This architecture serves as the basis for the coupled network framework depicted in Figure 1.

More specifically, the network regresses a set of offset values that, when summed together with a mean face shape computed from the training data, yields the final predicted landmark locations. Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$ be the sets of input images and ground truth landmark coordinates respectively, and let $\Theta = \{\theta_h\}_{h=1}^{\Omega}$ be the parameters of each network layer. Given L landmarks, let $\bar{\mathbf{y}} \in \mathbb{R}^{L\times 2}$ be the mean face shape computed from the training data. In practice, we use $L = 5$ corresponding to the centers of the eyes, the base of the nose, and the corners of the mouth.

The output of the network is an update $\Delta\mathbf{y}_i \in \mathbb{R}^{L\times 2}$, which is then added to the mean shape such that the predicted landmarks $\hat{\mathbf{y}}_i \in \mathbb{R}^{L\times 2}$ are given by

$$\hat{\mathbf{y}}_i(\mathbf{x}_i|\Theta, \bar{\mathbf{y}}) = \bar{\mathbf{y}} + \Delta\mathbf{y}_i(\mathbf{x}_i|\Theta). \tag{1}$$

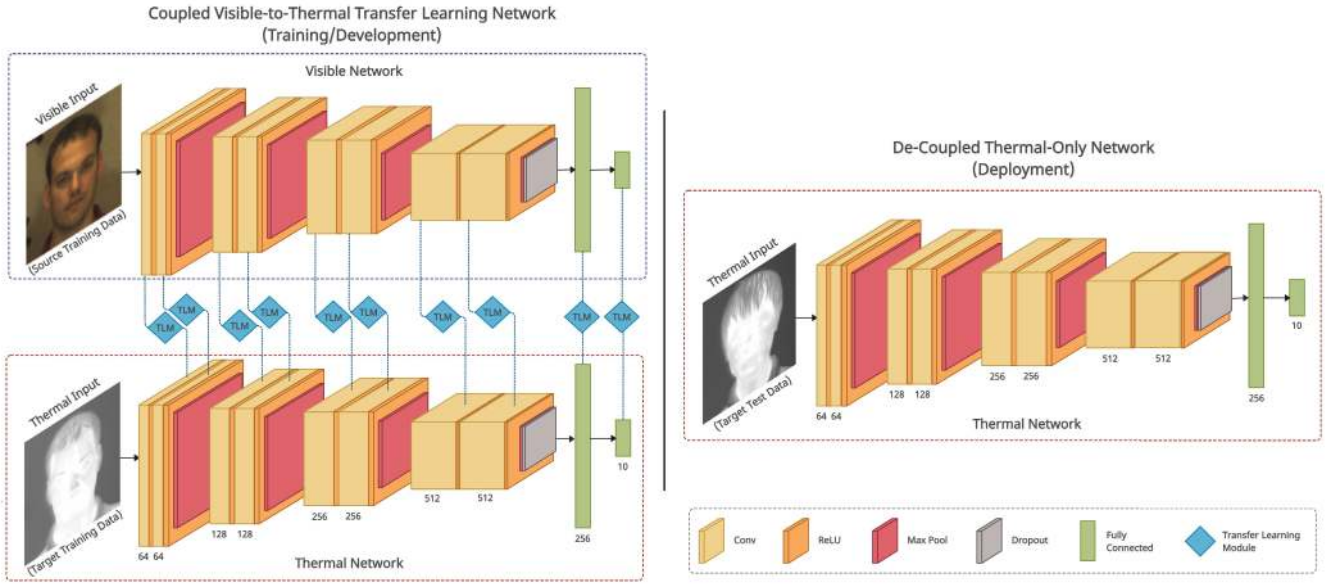

FIGURE 1. The proposed coupled network architecture (left) with Transfer Learning Modules (TLM) between the pairs of convolutional and fully-connected layers. A TLM may be a Siamese Layer, Linear Layer Regularizer, Linear Kernel Regularizer, or Residual Parameter Transformer. After the coupled network has been trained, the thermal domain sub-network is de-coupled for evaluation on the test data (right). The thermal sub-network is architecturally identical to the baseline network but has learned a different set of weights via the transfer learning process. Best viewed in color.

TABLE 1. DAN-based architecture. Kernels are described as height × width × depth, stride.

Layer	Output Shape	Kernel
conv1	$112 \times 112 \times 64$	$3 \times 3 \times 1, 1$
conv2	$112 \times 112 \times 64$	$3 \times 3 \times 64, 1$
pool1	$56 \times 56 \times 64$	$2 \times 2 \times 1, 2$
conv3	$56 \times 56 \times 128$	$3 \times 3 \times 64, 1$
conv4	$56 \times 56 \times 128$	$3 \times 3 \times 128, 1$
pool2	$28 \times 28 \times 128$	$2 \times 2 \times 1, 2$
conv5	$28 \times 28 \times 256$	$3 \times 3 \times 128, 1$
conv6	$28 \times 28 \times 256$	$3 \times 3 \times 256, 1$
pool3	$14 \times 14 \times 256$	$2 \times 2 \times 1, 2$
conv7	$14 \times 14 \times 512$	$3 \times 3 \times 256, 1$
conv8	$14 \times 14 \times 512$	$3 \times 3 \times 512, 1$
pool4	$7 \times 7 \times 512$	$2 \times 2 \times 1, 2$
dropout	$7 \times 7 \times 512$	-
fc1	$1 \times 1 \times 256$	$7 \times 7 \times 512, 1$
fc2	$1 \times 1 \times 10$	$1 \times 1 \times 256, 1$

To train the network, the following objective function is minimized via gradient descent:

$$\mathcal{L}(\Theta|\mathbf{X}, \mathbf{Y}, \overline{\mathbf{y}}) = \lambda_d \mathcal{L}_d(\Theta|\mathbf{X}, \mathbf{Y}, \overline{\mathbf{y}}) + \lambda_w \mathcal{L}_w(\Theta), \quad (2)$$

$$\mathcal{L}_d(\Theta|\mathbf{X}, \mathbf{Y}, \overline{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^{N} e(\Theta|\mathbf{x}_i, \mathbf{y}_i, \overline{\mathbf{y}}), \quad (3)$$

$$\mathcal{L}_w(\Theta) = \sum_{h=1}^{\Omega} \|\theta_h\|_2^2, \quad (4)$$

where $\mathcal{L}_d(\cdot)$ is the mean landmark detection loss and $\mathcal{L}_w(\cdot)$ is a L2 weight decay term. λ_d and λ_w are coefficients modulating the influence of the two aforementioned loss terms.

The landmark detection loss $\mathcal{L}_d$ is measured using the Normalized Root Mean Square Error (NME), calculated as

$$e(\Theta|\mathbf{x}_i, \mathbf{y}_i, \overline{\mathbf{y}}) = \frac{\left\|\hat{\mathbf{y}}_i(\Theta|\mathbf{x}_i, \overline{\mathbf{y}}) - \mathbf{y}_i\right\|_2}{d_i}, \quad (5)$$

where $d_i \in \mathbb{R}$ is the inter-pupil distance (IPD). Specifically, the IPD is the Euclidean distance between the left and right eye center landmarks.

Ultimately, the input image $\mathbf{x}_i \in \mathbb{R}^{112 \times 112 \times 1}$ is mapped to a $7 \times 7 \times 512$ tensor via four max pooling layers. The 3×3 kernels in the final convolutional layer have a receptive field of 68×68 pixels.

B. COUPLED ARCHITECTURE

Parameter transfer learning relates the weights of a source network to the weights of a target network. By modeling this relationship, knowledge can be transferred from one domain to the other, allowing a target network to effectively learn from data in the source domain via the parameters in the source network. Typically, this relationship is learned by applying a constraint or penalty term to the weights of the networks during training. For the purposes of this work, we consider the source domain to be visible spectrum imagery and the target domain to be thermal infrared imagery.

The goal is to utilize visible domain data during training in order to improve the performance of the thermal face landmark detection network beyond what it could achieve if it were only trained on thermal data. In order to conduct visible-to-thermal transfer learning, we specify a dual network architecture with one network trained to perform face landmark detection on visible domain data, while the other is trained on

thermal data. These two networks have the same architecture as the baseline network.

Coupling the pairs of layers in the visible and thermal networks are Transfer Learning Modules (TLM). TLMs are generic, abstract architectural components which facilitate the parameter transfer learning process. In other words, a TLM represents an implementation of a parameter transfer learning method on a pair of visible and thermal network layers. Figure 1 depicts the proposed coupled architecture.

We consider four methods of parameter transfer learning: Siamese (shared) Layers [14], Linear Layer Regularization (LLR) [16], Residual Parameter Transformation (RPT) [17], and our customized version of LLR entitled Linear Kernel Regularization (LKR). The methods differ by the manner in which the parameters of the source and target networks are related.

Ultimately, four different versions of the coupled architecture are compared - one for each of the transfer learning methods examined. Important to note is that the Siamese TLM is employed for all fully-connected layers as we found this configuration to yield better performance. For example, the Siamese model uses the Siamese TLM for all of its coupled layers. However, the RPT model uses the RPT TLM for all its convolutional layers, but the Siamese TLM for its fully-connected layers. The remainder of this section discusses the details of each of the transfer learning methods.

C. SIAMESE LAYER

A Siamese network [14], [32] encourages the learning of parameters capable of meaningfully processing input from both source and target domains. This is also sometimes referred to as weight sharing. A Siamese Transfer Learning Module imposes the constraint that the weights of the paired visible and thermal layers be equal. Structurally, the Siamese and baseline networks are identical. The input and output signature of the Siamese network is the same as that of the baseline network. Given an input image $\mathbf{x}_i \in \mathbb{R}^{112\times112\times1}$ from either the visible or thermal domain, the network outputs a shape update $\Delta\mathbf{y}_i \in \mathbb{R}^{L\times2}$.

Let $\mathbf{X}^v = \{\mathbf{x}_i^v\}_{i=1}^{N^v}$ and $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^{N^t}$ be the sets of visible and thermal training images, respectively. Similarly, let $\bar{\mathbf{y}}^\mathbf{v}$ and $\bar{\mathbf{y}}^\mathbf{t}$ be the mean shapes of the visible and thermal training data. The predicted landmarks are given by

$$\hat{\mathbf{y}}_i^v(\mathbf{x}_i^v|\Theta, \bar{\mathbf{y}}^\mathbf{v}) = \bar{\mathbf{y}}^\mathbf{v} + \Delta\mathbf{y}_i(\mathbf{x}_i^v|\Theta), \tag{6}$$

and

$$\hat{\mathbf{y}}_i^t(\mathbf{x}_i^t|\Theta, \bar{\mathbf{y}}^\mathbf{t}) = \bar{\mathbf{y}}^\mathbf{t} + \Delta\mathbf{y}_i(\mathbf{x}_i^t|\Theta), \tag{7}$$

yielded by the same set of network parameters Θ.

The Siamese network is trained on batches of images from both visible and thermal domains but is otherwise optimized in a similar fashion to the baseline network. Let $\mathbf{Y}^v = \{\mathbf{y}_i^v\}_{i=1}^{N^v}$ and $\mathbf{Y}^t = \{\mathbf{y}_i^t\}_{i=1}^{N^t}$ be the sets of visible and thermal ground truth landmark coordinates. The objective function for the Siamese network is

$$\mathcal{L} = \lambda_v\mathcal{L}_v + \lambda_t\mathcal{L}_t + \lambda_w\mathcal{L}_w, \tag{8}$$

$$\mathcal{L}_v(\Theta|\mathbf{X}^v, \mathbf{Y}^v, \bar{\mathbf{y}}^\mathbf{v}) = \frac{1}{N^v}\sum_{i=1}^{N^v} e(\Theta|\mathbf{x}_i^v, \mathbf{y}_i^v, \bar{\mathbf{y}}^\mathbf{v}), \tag{9}$$

$$\mathcal{L}_t(\Theta|\mathbf{X}^t, \mathbf{Y}^t, \bar{\mathbf{y}}^\mathbf{t}) = \frac{1}{N^t}\sum_{i=1}^{N^t} e(\Theta|\mathbf{x}_i^t, \mathbf{y}_i^t, \bar{\mathbf{y}}^\mathbf{t}), \tag{10}$$

where $\mathcal{L}_v$ and $\mathcal{L}_t$ are the landmark detection loss terms for the visible and thermal streams, while λ_v and λ_t are their respective coefficients.

This strategy does not add additional network parameters. However, if the two domains are sufficiently dissimilar, a Siamese network may be too limited in its capacity to sufficiently exploit the common information.

D. LINEAR LAYER REGULARIZATION

Given two separate but structurally identical networks, Linear Layer Regularization (LLR) [16] encourages a linear relationship to form between the corresponding sets of weights of the two networks. Let $\Theta^v = \{\theta_h^v\}_{h=1}^{\Omega}$ and $\Theta^t = \{\theta_h^t\}_{h=1}^{\Omega}$ be the parameters of the visible and thermal networks, respectively, where Ω is the number of layers in the baseline network. Given $j \subseteq \Omega$ sets of paired layers, let $\mathbf{a} \in \mathbb{R}^j$ and $\mathbf{b} \in \mathbb{R}^j$ be the scalar and bias terms learned by minimizing the LLR constraint

$$\mathcal{L}_{LLR} = \sum_{j\subseteq\Omega} \left\| a_j\theta_j^v + b_j - \theta_j^t \right\|_2^2. \tag{11}$$

This constraint applies a penalty to the loss when the network fails to learn a linear transformation of the visible weights to the thermal weights. The full objective function being minimized is

$$\mathcal{L} = \lambda_v\mathcal{L}_v + \lambda_t\mathcal{L}_t + \lambda_w\mathcal{L}_w + \lambda_{LLR}\mathcal{L}_{LLR}(\Theta^v|\mathbf{X}^v, \Theta^t|\mathbf{X}^t, \mathbf{a}, \mathbf{b}), \tag{12}$$

with the coefficient λ_{LLR} regulating the influence of the $\mathcal{L}_{LLR}$ loss. Like the Siamese network, training requires input batches composed of visible and thermal data.

Outside of the training process, however, the networks can be effectively "de-coupled," as indicated by their respective, mutually-independent predictive functions,

$$\hat{\mathbf{y}}_i^v(\mathbf{x}_i^v|\Theta^v, \bar{\mathbf{y}}^\mathbf{v}) = \bar{\mathbf{y}}^\mathbf{v} + \Delta\mathbf{y}_i(\mathbf{x}_i^v|\Theta^v), \tag{13}$$

$$\hat{\mathbf{y}}_i^t(\mathbf{x}_i^t|\Theta^t, \bar{\mathbf{y}}^\mathbf{t}) = \bar{\mathbf{y}}^\mathbf{t} + \Delta\mathbf{y}_i(\mathbf{x}_i^t|\Theta^t). \tag{14}$$

Therefore, while this method more than doubles the number of learnable parameters as compared to the baseline network during training, the resultant thermal network is the same size as the baseline network when evaluating on thermal test data (e.g. the deployment phase).

E. LINEAR KERNEL REGULARIZATION

The application of a single scalar and bias term to all the kernels of a layer may be too constrained to relate knowledge

between the visible and thermal domains. As an alternative approach, individual linear transformations may be learned for each kernel, thereby expanding the network's flexibility in a semantic, piecemeal fashion.

Given the visible parameters $\boldsymbol{\theta}_j^v \in \mathbb{R}^{C_j \times K_j}$ and thermal parameters $\boldsymbol{\theta}_j^t \in \mathbb{R}^{C_j \times K_j}$, let $(\mathbf{A}_j)^T = [\mathbf{a}_j \dots \mathbf{a}_j] \in \mathbb{R}^{K_j \times C_j}$ and $(\mathbf{B}_j)^T = [\mathbf{b}_j \dots \mathbf{b}_j] \in \mathbb{R}^{K_j \times C_j}$, where $\mathbf{a}_j \in \mathbb{R}^{K_j}$ and $\mathbf{b}_j \in \mathbb{R}^{K_j}$ are sets of learnable regularization parameters replicated C_j times. C_j and K_j are the sizes of the inputs and outputs to the j-th layer, respectively. In the case of a convolutional layer, C_j can be considered to be the height $\times$ width $\times$ depth of the kernel and K_j the number of kernels. The kernels can be independently regularized using a loss term of the form

$$\mathcal{L}_{LKR} = \sum_{j \subseteq \Omega} \left\| \mathbf{A}_j \odot \boldsymbol{\theta}_j^v + \mathbf{B}_j - \boldsymbol{\theta}_j^t \right\|_2^2. \tag{15}$$

where $\odot$ is the Hadamard product function. This loss term replaces the $\mathcal{L}_{LLR}$ loss term in the objective function

$$\mathcal{L} = \lambda_v \mathcal{L}_v + \lambda_t \mathcal{L}_t + \lambda_w \mathcal{L}_w + \lambda_{LKR} \mathcal{L}_{LKR}. \tag{16}$$

F. RESIDUAL PARAMETER TRANSFORMATION

Depending on how drastic the difference in appearance between the two domains is, a linear transformation of the weights may be too restrictive to translate the underlying phenomenological similarities. Residual Parameter Transformation (RPT) allows greater flexibility in modeling the similarities and differences of the respective domains at the cost of more learnable parameters.

Unlike LLR and LKR, when using RPT, the thermal network parameters are not learnable parameters in and of themselves. Instead, they are the output of two-layer residual networks, as illustrated in Fig. 2. Given encoding matrices $\mathbf{A}_j^e \in \mathbb{R}^{P_j \times C_j}$ and $\mathbf{B}_j^e \in \mathbb{R}^{K_j \times Q_j}$, decoding matrices $\mathbf{A}_j^d \in \mathbb{R}^{C_j \times P_j}$ and $\mathbf{B}_j^d \in \mathbb{R}^{Q_j \times K_j}$, and bias matrix $\mathbf{U}_j \in \mathbb{R}^{P_j \times Q_j}$, the visible parameters $\boldsymbol{\theta}_j^v \in \mathbb{R}^{C_j \times K_j}$ are transformed into the thermal parameters $\boldsymbol{\theta}_j^t \in \mathbb{R}^{C_j \times K_j}$ according to the equations

$$\boldsymbol{\theta}_j^t(\boldsymbol{\theta}_j^v, \mathbf{A}_j^e, \mathbf{B}_j^e, \mathbf{U}_j, \mathbf{A}_j^d, \mathbf{B}_j^d) = \boldsymbol{\theta}_j^v + \Delta\boldsymbol{\theta}_j^v, \tag{17}$$

$$\Delta\boldsymbol{\theta}_j^v = \mathbf{A}_j^d \sigma(\mathbf{A}_j^e \boldsymbol{\theta}_j^v \mathbf{B}_j^e + \mathbf{U}_j)\mathbf{B}_j^d, \tag{18}$$

where σ is the ReLU activation function, C_j and K_j are the input and output dimensions of the j-th layer parameters, P_j is the size to which the input dimension is reduced, and Q_j is the size to which the output dimension is reduced. See Table 2 for the actual sizes of the RPT subnets used in the experiments.

The network is tasked with learning an update $\Delta\boldsymbol{\theta}_j^v$ to be combined with the input visible weights $\boldsymbol{\theta}_j^v$. To this end, an additional loss term is added to the objective function of the form

$$\mathcal{L}_{RPT} = \mathcal{L}_r - \ln(\mathcal{L}_r), \tag{19}$$

$$\mathcal{L}_r = \sum_{j \subseteq \Omega} \left\| \Delta\boldsymbol{\theta}_j^v \right\|_{Fro}^2. \tag{20}$$

The loss term $\mathcal{L}_{RPT}$ is smallest when $\mathcal{L}_r = 1$. The natural log barrier function prevents learning the trivial transformation $\mathcal{L}_r = 0$ and discourages $\mathcal{L}_r$ from being very large in magnitude. However, it also has the side-effect of pressuring the total magnitude of the $\Delta\boldsymbol{\Theta}^v$ updates to be 1, which is a highly arbitrary constraint. The final objective function is

$$\mathcal{L} = \lambda_v \mathcal{L}_v + \lambda_t \mathcal{L}_t + \lambda_w \mathcal{L}_w + \lambda_{RPT} \mathcal{L}_{RPT}. \tag{21}$$

Even though the thermal network parameters Θ^t are a function of Θ^v, once the networks have been trained, the parameters are frozen. Consequently, the thermal network can be de-coupled from the visible network and residual transformation modules.

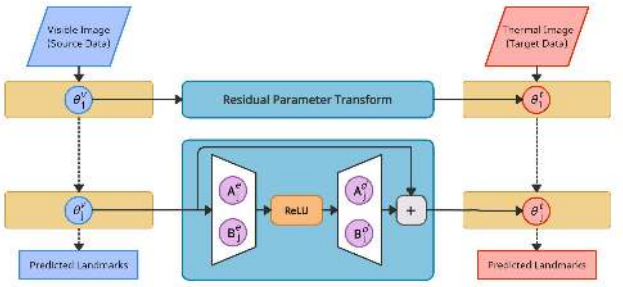

FIGURE 2. Residual transformation of the visible network layer parameters θ_j^v into thermal parameters θ_j^t via a learnable encoder-decoder sub-network with parameters $\mathbf{A}_j^e$, $\mathbf{B}_j^e$, $\mathbf{A}_j^d$, and $\mathbf{B}_j^d$ (bias term $\mathbf{U}_j$ not shown).

TABLE 2. Parameter sizes of the RPT subnets. * indicates no encoding/decoding performed on this dimension.

Layer j	$C_j \times K_j$	P_j	Q_j
conv1	9×64	$-*$	32
conv2	576×64	32	32
conv3	576×128	32	64
conv4	1152×128	64	64
conv5	1152×256	64	128
conv6	2304×256	128	128
conv7	2304×512	128	256
conv8	4608×512	256	256

IV. EXPERIMENTS AND RESULTS

In this section, we present and discuss the experimental results. To do so, we first describe the two datasets used for evaluating the landmark detection models, as well as the implementation details and preprocessing steps. Also, we discuss our experimental results which studies the effects of dataset sub-sampling and image and subject alignment when constructing input image pairs. The algorithms are also assessed in terms of their landmark detection accuracy compared to human level performance as well as their generalizability to a realistic, unobserved dataset.

As discussed in the previous sections, when evaluating the models (the deployment phase), visible data is not expected to be present. In cases where both a source and target network are learned, such as with LLR and RPT, the source network and the intermediary parameters bridging the two networks are not needed during the evaluation of the target network.

Therefore, all models, including those trained via transfer learning, are architecturally identical to the baseline network during evaluation on thermal domain test data and do not require data from the visible domain once deployed.

A. DATASETS

1) DEVCOM ARL POLARIMETRIC THERMAL FACE DATASET

This study uses Volumes 1 and 2 of the DEVCOM ARL Polarimetric Thermal Face Dataset released in [10] and extended in [11]. Subjects have been captured using thermal and visible light cameras. The dataset includes manually annotated landmarks (left eye center, right eye center, base of the nose, left mouth corner, and right mouth corner) for both domains.

Data from Volume 1 [10] contains 60 subjects captured at 2.5, 5, and 7.5 meters. The subjects were recorded displaying a baseline neutral facial expression and then were instructed to count out loud from one to ten in order to record a series of non-neutral expressions. The thermal imagery was recorded using a LWIR camera at 60 frames per second with a 640 $\times$ 480 pixel-sized Stirling-cooled mercury cadmium telluride focal plane array. Each subject has 96 frames of baseline thermal images and 288 frames of expression thermal images per range. However, each frame is highly correlated to its temporal neighbors. The average inter-pupil distance (IPD) of the thermal imagery at each range is 93 pixels, 48 pixels, and 33 pixels respectively. The visible spectrum imagery was captured by an array of four Basler Scout series cameras (two monochrome cameras and two color cameras), with resolutions of 640×480 and 640×492 respectively and focal lengths ranging from 4.5mm to 50mm. At the time of publishing, annotations for one monochrome and one color/RGB camera were available. Each subject has 4 annotated images per visible camera per range.

Volume 2 contains 51 subjects [11] captured at 2.5 meters using the same LWIR and color camera employed in Volume 1. The average IPD of the thermal imagery is 84 pixels. There is 1 neutral and 30 non-neutral expression visible images per subject. There are eight times as many thermal image frames recorded within the same period of time (although they are highly correlated in the same way as Volume 1), yielding 8 neutral and 240 non-neutral expression thermal images per subject.

The dataset is divided into Protocols 1 and 2 as described in [11]. Protocol 1 contains images from Volume 1 with subjects evenly split into training and testing sets. Protocol 2 is a composite of Volumes 1 and 2 with 85 subjects in the training set and 26 subjects sequestered for testing. Due to their highly correlated nature, we select a subset of thermal images from each subject by sampling every other frame. All results have been averaged over five random folds of subject-disjoint training and testing sets as per [24].

2) RWTH AACHEN UNIVERSITY THERMAL FACE DATASET

We also evaluate on the RWTH Aachen University Thermal Face Dataset [12] containing 2,935 images from 90 subjects. Subjects were filmed in four different sequences related to a variety of poses and expressions. The four sequences captured subjects performing: 1) an s-shaped head movement pattern, 2) seven fundamental facial movements, referred to as action units (AUs) [34], 3) neutral, happy, sad, and surprised emotions, 4) arbitrary head movements and expressions.

The images were recorded with an Infratec HD820 thermal infrared camera at 30 frames per second with a 1024 $\times$ 768 pixel-sized uncooled microbolometer focal plane array and a thermal resolution of 0.03 K at 30 $^\circ$ C and equipped with a 30-mm f/1.0 prime lens. The subjects were seated 0.9 meters from the camera. The resulting spatial resolution of the face was approximately 0.5 mm per pixel. The average IPD is 147 pixels.

A subset of the frames were manually annotated with a 68-point landmark scheme Since we train and evaluate using a 5-point landmark scheme, we calculate the eye center landmarks as the mean of the landmarks around each eye. The landmarks for the base of the nose and corners of the mouth are directly taken from the 68-point set.

B. IMPLEMENTATION DETAILS

As mentioned in Section III-B, all of the transfer learning models utilize the Siamese TLM for the fully-connected layers (e.g. the weights of the fully-connected layers are always shared).

The models were implemented using Python 3.6 and Tensorflow 1.14. Training and evaluation was done on machines equipped with dual RTX 2080TI GPUs. The code for the AAM was provided by the authors [12], using HoG features and the Wiberg Inverse-Compositional optimization algorithm.

The hyperparameter and training settings for the CNN models are given in Table 3. The optimizer, learning rate, cost coefficients, and number of training steps were experimentally tuned to report the highest possible performance for each model. For all models, the moving average momentum of batch normalization is set to 0.8 and the dropout rate to 50%. A decay of 0.9 and momentum of 0.0 is used for the RMSProp optimizer. For models using the Adam optimizer, the beta1 and beta2 parameters are 0.9 and 0.999 respectively.

The two GPUs were utilized in a data-distributed fashion such that each device received half of the batch size. We used a batch size of 128 thermal images for the baseline model and 128 thermal-visible image pairs for the transfer learning models. The baseline, Siamese, LLR, and LKR models were trained for 10,000 steps (approximately 74 epochs of Protocol 1 data and 40 epochs of Protocol 2 data). The RPT network is trained for 25,000 steps (about 185 epochs) due to its increased number of network parameters and lower learning rate.

C. PREPROCESSING

The input images are face crops obtained by extending the bounds of the ground truth face shape by 70%. Each image is independently normalized by subtracting its mean value

TABLE 3. Hyperparameter settings for CNN models. λ_t, λ_v, and λ_w are the cost coefficients for the thermal domain NME, visible domain NME, and L2 weight regularization respectively. α is the learning rate and Steps refers to the number of training steps.

Model	Optimizer	α	λ_t	λ_v	λ_w	Steps
Baseline	RMSProp	1e−3	1.0	0.0	0.0	10k
Siamese	Adam	5e−4	5.0	0.0	0.0	10k
LLR	Adam	5e−4	5.0	1.0	5e−4	10k
LKR	Adam	5e−4	8.0	1.0	5e−4	10k
RPT	Adam	1e−4	10.0	1.0	1e−3	25k

TABLE 4. Due to more thermal images per subject than visible in the ARL dataset, image pairs are created by cycling through the available visible data.

Thermal Files	Visible Files
(thermal_1.png,	visible_1.png)
(thermal_3.png,	visible_2.png)
(thermal_5.png,	visible_3.png)
(thermal_7.png,	visible_4.png)
(thermal_9.png,	*visible_1.png*)
(thermal_11.png,	*visible_2.png*)

divided by the standard deviation. All images are converted to grayscale and resized to 112×112. In every training batch, each thermal image has a 50% chance of being horizontally flipped. For the transfer learning models, the thermal images in the ARL datasets are paired with the visible images of the same subject captured at the same range. Random horizontal flipping is coordinated such that if a thermal image is flipped, so too is the visible image.

D. CONSTRUCTING IMAGE PAIRS

The ARL dataset contains more thermal than visible images. For each subject, we sub-sample every other frame of thermal data. An equal number of visible images is collected by cycling through all of the subject's visible data, as exemplified in Table 4.

Protocol 1 data is captured at three ranges from one thermal camera and two visible cameras. Thermal data from each range is paired with visible data from the camera and range that is most similar in terms of mean IPD.

E. EVALUATION METRICS

The standard face landmark detection evaluation metric [26], [35], [36] is the mean Normalized Root Mean Square Error (NME), defined as the Euclidean distance between the predicted landmarks $\hat{\mathbf{y}}_i$ and ground truth landmarks $\mathbf{y}_i$, normalized by the inter-pupil distance (IPD) d_i:

$$\text{NME} = \frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{L}\sum_{j=1}^{L}\left\|\hat{\mathbf{y}}_{i,j}-\mathbf{y}_{i,j}\right\|_2}{d_i}, \tag{22}$$

where N is the number of images in the test set and L is the number of landmarks. The notation NME (%) indicates the values have been scaled by a factor of 10^2. We consider any image with a NME (%) equal to or greater than 10% to be a failed detection. As per [37], in addition to the NME metric, the Standard Deviation (Std), Median, Median Absolute Deviation (MAD), and Maximum Error statistics are also reported.

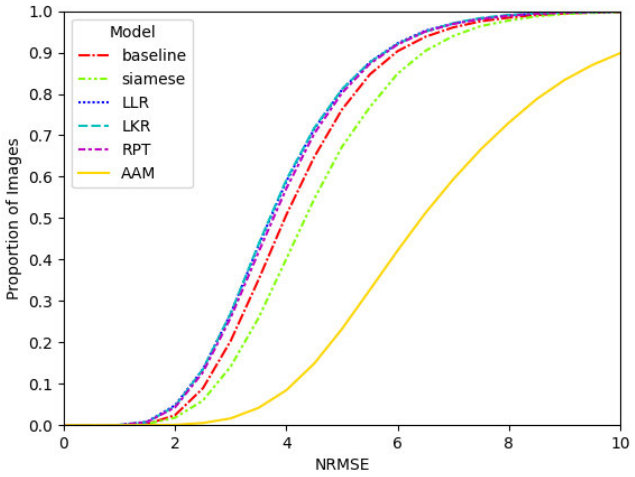

FIGURE 3. CED curve in terms of NME(%) on Protocol 1.

We provide the Cumulative Error Distribution (CED) curve, displaying the proportion of images with NMEs falling below threshold values ranging from 0% to 10%, as well as the Area Under the CED Curve ($\text{AUC}_{10\%}$) and the Failure Rate at 10% ($\text{FR}_{10\%}$), defined as the proportion of test images with a NME (%) greater than 10%.

F. EVALUATION ON ARL DATASET

All of the models learned using transfer learning techniques outperform the baseline model with the exception of the siamese network. The tabulated results are given in Tables 5 and 6. Fig. 3 displays the CED curve for Protocol 1. LKR performs best on Protocol 1 with a NME (%) of 3.90, followed closely by LLR at 3.91. Although, as shown in Table 9 and further discussed in Section IV-G, the best performance on Protocol 1 is actually achieved by RPT with a NME (%) of 3.82 due to more rigorous sub-sampling of the highly correlated thermal data. The RPT model, with its increased number of parameters during training, is more prone to over-fitting but has a greater capacity to relate knowledge between the source and target domains. This is further demonstrated on Protocol 2, where the RPT model performs best, able to take advantage of the increased amount of training data. We can see the siamese network is unable to beat the baseline network, indicating the non-trivial differences between the visible and thermal domains.

Inspection of the qualitative examples from the ARL dataset in Fig. 4 indicates the high level of accuracy of the neural networks. That the LLR, LKR, and RPT approaches achieve similar error rates may be an indication that the models are approaching a soft limit in terms of performance on the ARL Dataset. One may also observe that the human annotated ground-truth landmarks are not always perfect (e.g., row 4 right eye in Fig. 4), further placing a soft cap on performance.

TABLE 5. Landmark detection performance statistics in terms of NME(%) on Protocol 1 of the ARL Dataset.

Model	NME(%)	Std	Median	MAD	Max Error	$AUC_{10\%}$	$FR_{10\%}$
DAN [24]	5.6	-	-	-	-	-	-
AAM [22]	6.81	2.38	6.43	1.86	22.15	0.34	10.14
baseline	4.17	1.42	3.97	1.09	16.50	0.58	0.25
siamese	4.51	1.48	4.33	1.16	13.82	0.55	0.24
LLR	3.91	**1.38**	3.71	**1.07**	13.86	**0.61**	0.15
LKR	**3.90**	**1.38**	**3.70**	**1.07**	14.34	**0.61**	**0.12**
RPT	3.95	**1.38**	3.76	**1.07**	**13.11**	0.60	0.14

TABLE 6. Landmark detection performance statistics in terms of NME(%) on Protocol 2 of the ARL Dataset.

Model	NME(%)	Std	Median	MAD	Max Error	$AUC_{10\%}$	$FR_{10\%}$
baseline	4.07	**1.34**	3.88	1.05	**15.00**	0.59	**0.08**
siamese	4.45	1.50	4.26	1.16	18.97	0.55	0.23
LLR	3.76	1.42	3.52	1.10	15.78	0.62	0.15
LKR	3.77	1.40	3.53	1.09	16.04	0.62	0.16
RPT	**3.72**	1.39	**3.49**	**1.07**	15.31	**0.63**	0.13

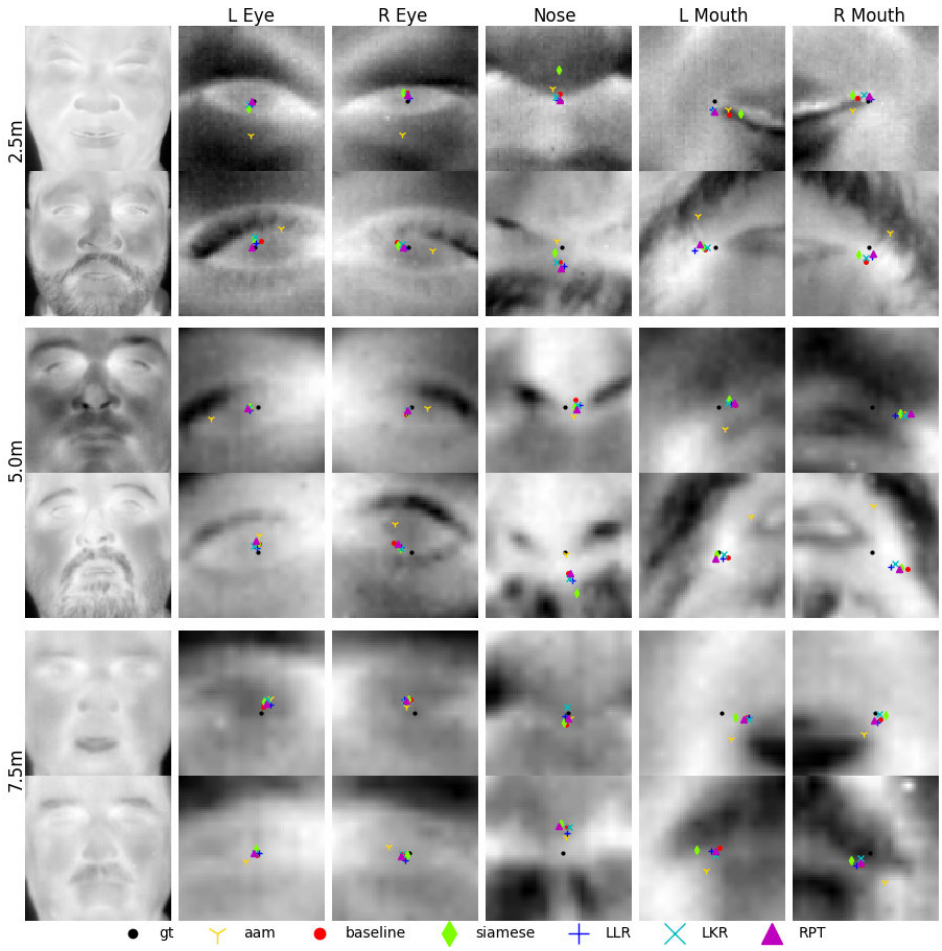

FIGURE 4. Qualitative results on ARL Dataset by landmark region. Region crops are centered around the ground truth landmark.

Table 7 presents the average performance on each of the five landmark locations: left eye center (LEC), right eye center (REC), base of nose (BoN), left mouth corner (LMC), right mouth corner (RMC). These results are obtained from the models trained on Protocol 1 with ground truth bounding boxes.

TABLE 7. NME(%) per individual facial landmark.

Model	LEC	REC	BoN	LMC	RMC
baseline	3.14	2.82	3.98	5.29	5.60
siamese	3.38	2.99	4.72	5.36	6.12
LLR	3.06	**2.53**	**3.60**	4.96	5.37
LKR	**3.02**	2.61	3.63	**4.90**	**5.34**
RPT	3.06	2.67	3.72	4.93	5.38

TABLE 8. NME(%) per camera range.

Model	2.5m	5.0m	7.5m
DAN [24]	5.00	5.60	6.10
baseline	3.56	4.18	4.75
siamese	4.17	4.44	4.93
LLR	**3.19**	3.95	4.58
LKR	3.23	**3.94**	**4.53**
RPT	3.26	4.00	4.60

Unsurprisingly, all models detected the eyes and nose more accurately than the corners of the mouth. It is less clear why the left mouth corner has consistently lower error rates than the right mouth corner. Given the overall imbalance in confidence for the different landmark locations, it may be prudent to consider a weighted approach when calculating a transformation to be used for image alignment.

Table 8 presents the average performance based on the subject's distance from the camera. The results are obtained from Protocol 1 using ground truth bounding boxes.

Both LLR and LKR perform similarly well, with LLR having an advantage at the 2.5m range and LKR having an advantage at the 7.5m range.

G. SUB-SAMPLING AND THERMAL-TO-VISIBLE RATIO

There are only 4 visible images available per camera per range per subject versus 384 thermal images per range per subject in Protocol 1. However, the thermal data was recorded at 60 frames per second, resulting in many images being highly correlated. This ablation study examines the effect of sub-sampling the thermal image frames at a variety of rates to obtain different ratios of thermal-to-visible image pairs.

The coupled models are trained on six different ratios of thermal-to-visible data, as shown in Table 9. In each case, the baseline model is trained on the same set of thermal images as the coupled models. The models are evaluated on the same sets of test data used in Subsection D. Due to the small training set sizes when sampling at ratios of 1:2 and 1:4, the number of training steps are reduced to 5,000 for all models.

Interestingly, the coupled models perform better when sampling thermal images at ratios of 8:1, 4:1, and 2:1 compared to when using the default ratio of 48:1 when sampling every other thermal frame. This may be due to an increased likelihood at the 48:1 ratio that training batches contain more highly correlated images, leading to overfitting. The highest performance achieved on Protocol 1 is with RPT when using an 8:1 thermal-to-visible ratio. With more parameters than the other networks, RPT is more susceptible to overfitting, and therefore benefits from careful data sampling.

TABLE 9. NME(%) by ratio of unique thermal to visible training images.

Model	8:1	4:1	2:1	1:1	1:2	1:4
baseline	4.26	4.25	4.22	4.22	4.35	4.49
siamese	4.43	4.30	4.43	4.51	4.70	4.93
LLR	3.84	**3.83**	**3.83**	**3.91**	**4.06**	4.67
LKR	3.87	3.89	3.88	4.15	4.23	5.86
RPT	**3.82**	3.85	3.92	4.00	4.15	**4.34**

H. IMAGE PAIR ALIGNMENT

In this ablation study, we examine the importance of well-aligned imagery in the training of the transfer learning models. The results given in Section IV-F are obtained from transfer learning models trained with same-subject image pairs which were relatively well-aligned by virtue of using the ground-truth landmarks to calculate the face crop.

Two means of inducing image misalignment are considered: random-subject image pairing, and random image augmentations. In random-subject image pairing, the same-subject constraint is removed. In order to introduce a misalignment between same-subject image pairs, we consider two scenarios in which random data augmentation is applied to the pair of images either jointly or independently. In addition to horizontal flipping, random translations and scaling are applied to the face bounding boxes. In the joint augmentation case, horizontal flipping of the thermal and visible image is synchronized and the bounding box coordinates are randomly offset by the same values. In the independent augmentation case, random horizontal flipping and the random bounding box offset values are independently determined for each image. The $(\Delta x, \Delta y)$ offset values for the top-left and bottom-right corners of the bounding box are obtained by uniformly sampling four integers in the ranges of $[-5, 5]$, $[-10, 10]$, and $[-15, 15]$. During evaluation, no random augmentation was performed.

Table 10 presents the results of the ablation study on Protocol 1. The use of a same-subject image pairing strategy with joint random augmentations results in better performance. However, image misalignment induced by independent perturbations to the face bounding boxes do not drastically reduce the performance of the transfer learning models, even when the variance is increased to ± 15 pixels. Compared to transfer learning strategies such as Maximum Mean Discrepancy (MMD) which apply a constraint to the outputs of corresponding layers, LLR, LKR, RPT, and siamese networks instead place their constraints on the network parameters themselves. We reason that this important difference makes parameter transfer learning methods less reliant on well-aligned, time-synchronized paired data—a useful property as time-synchronized datasets are limited in number and more difficult to collect. These results, along with those from the previous Section IV-G, indicate special attention should be paid to how the data is sampled and paired.

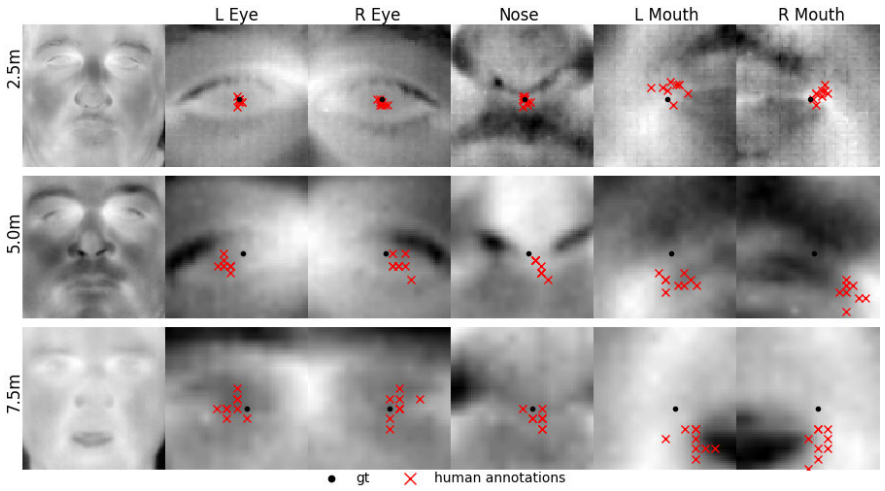

FIGURE 5. The variance of the human annotations.

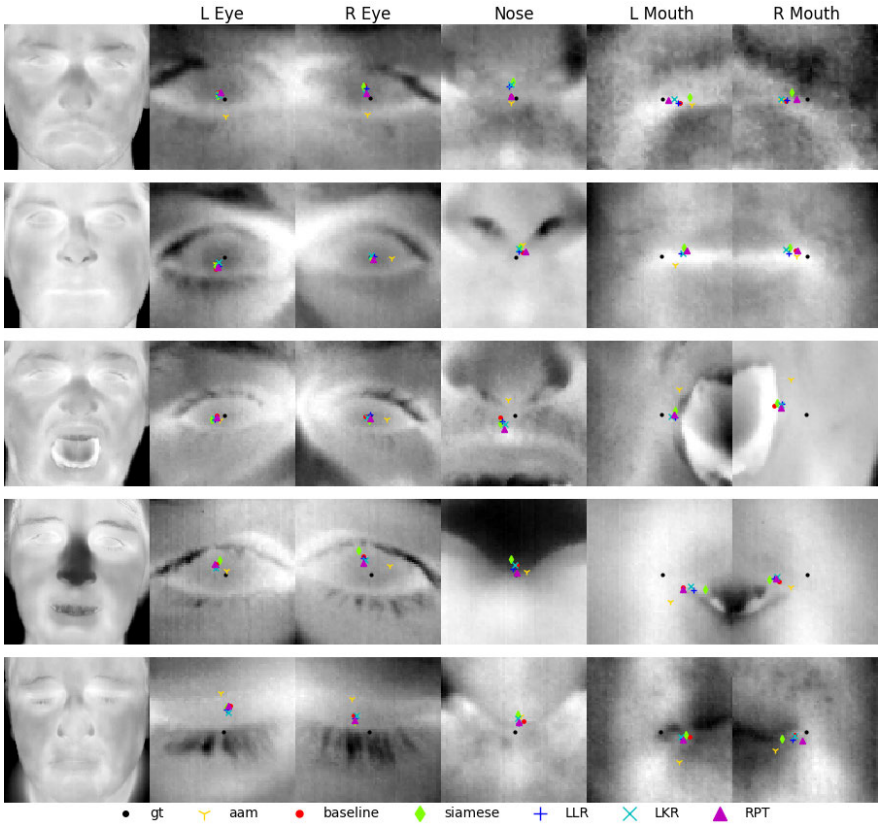

FIGURE 6. Qualitative results on RWTH Aachen dataset by landmark region. Region crops are centered around the ground truth landmark.

I. HUMAN LEVEL BENCHMARK

To establish a point of reference for performance, we asked 10 human participants each manually annotate 20 images representing 20 different subjects across Volumes 1 and 2. All participants received the same set of 20 images to annotate, the same annotation software, and the same instructions, including how to zoom in and out.

For a given model trained on a fold of the data, for each landmark in a given image, we measure the Mahalanobis distance from a model's prediction to the cluster of human annotations. Mahalanobis distance measures the distance from a point to a distribution by taking into account the covariance of the distribution. For this reason, it is widely used for outlier detection. We assume the distribution of landmark

TABLE 10. Impact of image (mis)alignment on NME(%). Image pairs are generated using either a Same-Subject (SS) or Random-Subject (RS) pairing policy. Misalignment is also induced with Independent (I) and Joint (J) augmentation strategies per range of random bounding box offset values (e.g., [−5, 5]).

Model	±0		±5		±10		±15	
	I+RS	J+SS	I+SS	J+SS	I+SS	J+SS	I+SS	J+SS
baseline	-	4.17	5.03	-	5.38	-	5.76	-
siamese	4.54	4.51	5.32	5.20	5.77	5.73	7.22	6.85
LLR	**3.91**	**3.91**	4.25	**4.24**	**4.49**	**4.49**	**4.73**	**4.73**
LKR	3.97	3.90	4.35	4.32	4.60	4.58	4.78	4.74
RPT	4.02	3.95	4.64	4.69	4.76	4.72	4.83	4.78

TABLE 11. Percentage of outliers by confidence level.

Model	Protocol 1		Protocol 2	
	95%	99%	95%	99%
AAM [22]	28.78	13.17	19.29	5.71
baseline	9.27	2.44	9.29	2.14
siamese	17.07	5.37	19.29	5.71
LLR	**8.29**	2.44	10.71	0.71
LKR	8.78	**1.95**	**7.86**	**0.0**
RPT	8.78	2.44	**7.86**	**0.0**

TABLE 12. Percentage of outliers by range on Protocol 1 ($\alpha = 0.05$).

Model	2.5m	5.0m	7.5m
AAM [22]	20.49	6.83	1.46
baseline	2.44	5.85	**0.98**
siamese	6.83	7.8	2.44
LLR	2.44	**4.88**	**0.98**
LKR	2.44	5.37	**0.98**
RPT	**1.95**	5.85	**0.98**

annotations follow a normal distribution and is therefore equivalent to the chi-square distribution with 2 degrees of freedom. Predictions are considered outliers if their Mahalanobis distance exceeds critical values calculated at confidences of 95% and 99%. Table 11 reports the percentage of outlying predictions for each model across the five folds of both Protocols 1 and 2, while Table 12 accounts for the outlier percentages by camera range. Fig. 5 depicts the variance of the human annotations at the three different camera ranges.

The low percentage of outliers among the LLR, LKR, and RPT models indicate they are reaching human-level performance. This could explain why, of the three aforementioned models, it is difficult for one to consistently outperform the others, thus supporting the reasoning that they are achieving results that are as good as can be expected given the imperfect nature of the ground truths. The majority of outlying predictions occur at the 5.0m range. At the 7.5m range, the models perform nearly as well as the human annotators.

J. GENERALIZATION TO RWTH AACHEN DATASET

It is not possible to train the transfer learning networks on the RWTH Aachen dataset because it lacks any corresponding visible imagery. Thus, we evaluate on the entire Aachen dataset using models trained on Protocols 1 and 2 of the ARL dataset. The results are presented in Table 13.

TABLE 13. NME(%) on RWTH Aachen university thermal face dataset. Models were trained on Protocols 1 and 2.

Model	Protocol 1	Protocol 2
AAM [22]	8.46	8.12
baseline	6.38	5.93
siamese	7.77	7.35
LLR	6.36	**5.77**
LKR	6.36	5.83
RPT	**6.31**	5.80

TABLE 14. NME(%) for RWTH Aachen video sequences A (Off-Pose), B, C, and D (Off-Pose). Models trained on Protocol 1.

Model	Seq A	Seq B	Seq C	Seq D
AAM [22]	13.4	7.4	7.1	9.0
baseline	**9.4**	5.0	5.6	**7.9**
siamese	10.2	6.3	7.1	9.5
LLR	9.6	4.8	5.5	8.1
LKR	9.6	4.8	5.5	8.1
RPT	**9.4**	**4.8**	**5.4**	**7.9**

The LLR, LKR, and RPT outperform both the AAM and baseline models. Of the models trained on Protocol 1, RPT is the best performing, whereas of the models trained on Protocol 2, LLR is the best performing. This differs from the evaluation results on the ARL Dataset, where LKR performed best on Protocol 1 and RPT on Protocol 2.

Of the 2,935 images evaluated, 546 images are from the s-shaped head movement video sequence (Sequence A), 377 are from the facial action units (AUs) sequence (Sequence B), 1,782 are from the emotions sequence (Sequence C), and 230 are from the arbitrary head movement and expression sequence (Sequence D). Table 14 shows the results on the individual video sequences. The models perform poorly on Sequences A and D, wherein a majority of faces are at an off angle from the camera. This drop in performance is expected because this type of variance is not expressed in the training data. However, the qualitative results for upright and frontal faces, shown in Fig. 6, demonstrate the generalizability of the neural network models to an unseen dataset that has similar variability in facial pose and expression to that of the ARL Dataset.

In terms of landmark detection speed, we observed an AAM fitting time per image of 5.1s, which is similar to the 7.31s reported in [12]. For the CNN-based models, we timed a detection speed of 0.02 seconds per image compared with 0.03 seconds in [12]. The implementation of the AAM algorithm used in both this study and [12] does not take advantage of GPU-accelerated parallel processing, resulting in expectedly slower speeds compared to the CNN models.

V. CONCLUSION

This work demonstrates that thermal face landmark detection can be improved via transfer learning with visible data. Except for the siamese network, all of the parameter

transfer learning methods yield improvement over the baseline network. The Residual Parameter Transfer (RPT) model achieved the best performance on Protocol 1 (when using an 8:1 thermal-to-visible image ratio) and Protocol 2 of the DEVCOM ARL Dataset. The Linear Layer Regularization (LLR) model performed best on the RWTH Aachen University Thermal Face Dataset, showcasing its ability to generalize to an unseen dataset, while RPT performed second best, indicating a susceptibility for overfitting. Additionally, we demonstrate the resilience of LLR, LKR, and RPT to image pair misalignment. As such, these methods are well suited to multi-spectral datasets which lack precisely aligned and time-synchronized data.

A point of future work is to extend the scope of this study to training on paired thermal and visible datasets that include variations in pose or occlusion. Quantitative and qualitative results suggest we may have reached a soft cap on the performance attainable on Volumes 1 and 2 of the ARL Dataset, limiting efforts to study the relative capabilities of the transfer learning approaches. The use of training data that does not rely on same-subject image pairs is also an avenue of exploration as it would potentially allow for a much larger body of visible data to be leveraged for transfer learning.

ACKNOWLEDGMENT

The authors would like to thank Ms. Michelle Giorgilli and Tom Cantwell of the Defense Forensics and Biometrics Agency (DFBA) for their guidance and discussions on this work.

REFERENCES

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[2] D. A. Socolinsky and A. Selinger, "Thermal face recognition in an operational scenario," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2004, p. 2.

[3] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[5] O. Arandjelovic, R. Hammoud, and R. Cipolla, "On person authentication by fusing visual and thermal face biometrics," in *Proc. IEEE Int. Conf. Video Signal Based Surveill.*, Nov. 2006, p. 50.

[6] G. Bebis, A. Gyaourova, S. Singh, and I. Pavlidis, "Face recognition by fusing thermal infrared and visible imagery," *Image Vis. Comput.*, vol. 24, no. 7, pp. 727–742, Jul. 2006.

[7] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "UMDFaces: An annotated face dataset for training deep networks," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 464–473.

[8] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[10] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan, "A polarimetric thermal database for face recognition research," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 119–126.

[11] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 845–862, Jun. 2019, doi: 10.1007/s11263-019-01175-3.

[12] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof, "A thermal infrared face database with facial landmarks and emotion labels," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 5, pp. 1389–1401, May 2019.

[13] R. S. Ghiass, H. Bendada, and X. Maldague, "Université laval face motion and time-lapse video database (UL-FMTV)," Université Laval, Québec, QC, USA, Tech. Rep., 2018. [Online]. Available: https://www.qirt2018.de/portals/qirt18/doc/We.4.B.3.pdf

[14] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.

[15] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 54–62.

[16] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, Apr. 2019.

[17] A. Rozantsev, M. Salzmann, and P. Fua, "Residual parameter transfer for deep domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4339–4348.

[18] T. Bourlai and Z. Jafri, "Eye detection in the middle-wave infrared spectrum: Towards recognition in the dark," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, Nov. 2011, pp. 1–6.

[19] S. Wang, Z. Liu, P. Shen, and Q. Ji, "Eye localization from thermal infrared images," *Pattern Recognit.*, vol. 46, no. 10, pp. 2613–2621, Oct. 2013.

[20] M. N. Hussien, M.-H. Lye, M. F. A. Fauzi, T. C. Seong, and S. Mansor, "Comparative analysis of eyes detection on face thermal images," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Sep. 2017, pp. 385–389.

[21] H.-W. Tzeng, H.-C. Lee, and M.-Y. Chen, "The design of isotherm face recognition technique based on nostril localization," in *Proc. Int. Conf. Syst. Sci. Eng.*, Jun. 2011, pp. 82–86.

[22] M. Kopaczka, K. Acar, and D. Merhof, "Robust facial landmark detection and face tracking in thermal infrared images using active appearance models," in *Proc. 11th Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, Feb. 2016, pp. 150–158.

[23] L. Sun and Z. Zheng, "Thermal-to-visible face alignment on edge map," *IEEE Access*, vol. 5, pp. 11215–11227, 2017.

[24] D. Poster, S. Hu, N. Nasrabadi, and B. Riggan, "An examination of deep-learning based landmark detection methods on thermal face imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 980–987.

[25] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 964–975, Feb. 2018.

[26] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 88–97.

[27] Y. Wu and Q. Ji, "Constrained deep transfer feature learning and its applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5101–5109.

[28] Y. Bengio, I. Goodfellow, and A. Courville, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2017.

[29] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[30] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 513–520.

[31] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[32] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[34] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. London, U.K.: Oxford Univ. Press, 1997.
[35] Z. Liu, X. Zhu, G. Hu, H. Guo, M. Tang, Z. Lei, N. M. Robertson, and J. Wang, "Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3467–3476.
[36] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 115–142, Feb. 2019.
[37] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2116–2125.

DOMENICK D. POSTER (Member, IEEE) received the B.A. degree in international studies and the M.S. degree in computer science from West Virginia University, in 2008 and 2015, respectively. His past research has focused on spoof detection for face and iris recognition. As an Oak Ridge Associated Universities (ORAU) Journeyman Fellow at the DEVCOM Army Research Laboratory, from 2018 to 2020, he studied face recognition and landmark detection in the thermal spectrum.

SHUOWEN (SEAN) HU received the B.S. degree in electrical and computer engineering from Cornell University, in 2005, and the Ph.D. degree in electrical and computer engineering from Purdue University, in 2009. He was awarded the Andrews Fellowship to study at Purdue University, conducting research in biomedical signal processing. Following graduation from Purdue University, he joined the DEVCOM Army Research Laboratory (ARL) as an Electronics Engineer with the Image Processing Branch. He has more than 50 conference and journal publications. His current research interests include cross-spectrum face recognition as well as on object detection and classification.

NATHAN J. SHORT received the M.S. and Ph.D. degrees in computer engineering from Virginia Tech, in 2012. He is currently the Senior Lead Scientist of Booz Allen Hamilton. He conducts research and development in computer vision, image processing, and machine learning, supporting government organizations within DOD, DHS, DOJ, and the IC. His experience includes research and development of imaging systems for unmanned vehicles, multi-biometric solutions for mobile and traditional assets, as well as developing next-generation human identification technology to support forensic and ISR applications. He has several peer-reviewed publications in the fields of biometrics, computer vision, and optics; and has co-organized multiple technical workshops and tutorials covering cutting-edge technology in these areas. He holds two patents related to biometric recognition and security. He also supports the Intelligence Advanced Research Projects Activity (IARPA) Odin Program, providing technical subject matter expertise on hardening biometric security systems; and the U.S. Army Research Laboratory, developing custom face recognition technology. His publications have received multiple awards and recognition, including Best Paper Awards at highly competitive, academic-level biometrics, and computer vision conferences. His latest work was featured in the Optical Society's Optics and Photonics News Special Issue "Optics in 2015," which highlights "the most exciting research in optics" conducted over the year.

BENJAMIN S. RIGGAN (Member, IEEE) received the B.S. degree in computer engineering and the M.S. and Ph.D. degrees in electrical engineering from North Carolina State University, in 2009, 2011, and 2014, respectively. He is currently an Assistant Professor with the Electrical and Computer Engineering Department, University of Nebraska-Lincoln. Prior to joining the Electrical and Computer Engineering Department, UNL, he has worked with the U.S. Army Research Laboratory's Image Processing and Networked Sensing and Fusion Branches, where he focused on domain adaptation, cross-spectrum facial recognition, and multi-modal analytics. His research interests include computer vision, image and signal processing, and biometrics/forensics that are related to multi-modal analytics and machine learning. He has received the Best Paper Award at the IEEE Winter Conference on Applications of Computer Vision (WACV) in 2016, the Runner-Up Best Paper Award at the IEEE International Conference on Biometrics: Theory, Applications, and Systems in 2016, and the Best Paper Award at IEEE WACV in 2018. He has also helped to organize workshops and tutorials at the IEEE International Conference on Automatic Face and Gesture Recognition (FG) in 2017, IEEE/IAPR Joint Conference on Biometrics (IJCB) in 2017, and IEEE WACV in 2018 and 2019.

NASSER M. NASRABADI (Fellow, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the Imperial College of Science and Technology, University of London, London, U.K., in 1980 and 1984, respectively. He was a member of the Technical Staff at the Phillips Research Laboratory, NY, USA, as an Assistant Professor with the Department of Electrical Engineering, Worcester Polytechnic Institute, an Associate Professor with the Department of Electrical and Computer Engineering, University at Buffalo, and the Senior Research Scientist of the U.S. Army Research Laboratory. In 2016, he joined West Virginia University, where he is currently a Professor with the Lane Department of Computer Science and Electrical Engineering. He founded the Biometrics and Identity, Innovation Center, and he is also the Director of the Cognitive Computing Laboratory (CCL). His research and teaching interests include image processing, computer vision, machine learning, deep neural networks, biometrics, and sparsity theory. His current research interests include image processing, computer vision, biometrics, deep learning, statistical machine learning theory, sparsity, robotics, and neural networks applications to image processing. He has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS, SYSTEMS AND VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS.