2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

PyraMoT: A Novel Framework for Enhanced Facial Thermal Landmarks Detection

Kais Riani , Salem Sharak and Mohamed Abouelenien
Computer and Information Science, University of Michigan, Dearborn, USA
kriani@umich.edu sharak@umich.edu zmohamed@umich.edu

***Abstract*— Facial analysis and recognition are vital components in many real-world applications, such as driver safety monitoring, security systems, and healthcare. Despite the need for thermal images in many of these applications, there has been limited research conducted on facial landmark recognition for thermal images. This scarcity can be partly attributed to a lack of thermal datasets for comprehensive analysis. In this paper, we make three major contributions to the field of thermal facial landmark detection. First, we present D5050, a novel thermal face dataset that includes 166 video sequences and 5050 annotated thermal images from 163 different subjects, which is, to our knowledge, the largest dataset with comprehensive landmark annotation in terms of the number of subjects. Second, we propose PyraMoT, an innovative thermal facial landmark detection framework that combines a customized encoder-decoder structure and a Feature Pyramid Network (FPN) with the efficiency of MobileNetV2 and the incorporation of the anisotropic loss function. Third, we conduct a thorough comparative analysis with existing approaches, including six methods, three datasets, and four color palettes for thermal facial landmark detection. This study employs measures such as Normalized Mean Error (NME) and Failure Rate (FR) to determine and comparatively evaluate the accuracy and reliability of the various detection methods. Overall, our proposed approach outperforms other methods on different datasets and contributes to the field of thermal image processing, proposing three main advancements that address current limitations in the field.**

I. INTRODUCTION

Facial landmark localization plays a pivotal role in computer vision, providing a framework for numerous advanced applications. It enables precise face recognition by identifying key facial features, used in security systems and personal identification methodologies [38]. Age estimation, an emerging application, leverages these landmarks to accurately infer age, which has implications in sectors like surveillance, investigation, and social entertainment [15]. In healthcare, facial landmark localization is instrumental in vital sign monitoring, facilitating non-intrusive observation of heart rate and respiration [19]. Additionally, this technology is crucial in emotion recognition, where it aids in interpreting human emotions by analyzing facial expressions, thereby contributing significantly to human-computer interaction and psychological studies [1], [31].

Thermal cameras offer distinct advantages in various fields, particularly in their ability to operate effectively in unlit or low illumination conditions [11]. Unlike RGB cameras that rely on visible light, thermal cameras detect infrared radiation. Thermal cameras are employed in various applications, including drivers' alertness detection [26], circadian rhythm detection [28], and monitoring human vital signals [21], [27]. However, the domain of thermal facial landmark detection, despite its potential, remains relatively unexplored when compared to landmark detection using RGB cameras. Thermal imaging requires specialized approaches different from traditional visual landmark detection [24] due to the major differences in the pixel format and the representation of different objects as temperatures, which in return loses some of the sharp edges and boundaries found in RGB images. Deep learning techniques, though promising, face hurdles due to the distinct nature of thermal imagery [6] and the significant lack of proper annotated datasets in the field. This limited research availability highlights the need for dedicated efforts in exploring the use of thermal imaging to detect facial landmarks.

The development of effective models for facial landmark detection with thermal imaging is significantly hindered by the lack of annotated datasets. Most existing datasets with thermal images are either too small to train robust models or lack the comprehensive annotations necessary for detailed analysis [13]. This minimal availability of quality data leads to models that are not sufficiently generalized, limiting their applicability in real-world scenarios. In addition, the lack of diverse datasets encompassing various demographics and environmental conditions hampers the creation of universally effective models [3].

In this paper, we make three major contributions, including a new annotated thermal dataset, an innovative thermal facial landmark detection approach that can be applied directly to radiometric thermal data without the need for RGB or grayscale formats, and an extensive comparison with existing approaches. First, we introduce a novel thermal face dataset, which encompasses an extensive collection of thermal facial images featuring 163 distinct subjects. Second, we have developed a novel approach for detecting facial landmarks in these images, tailored to the unique characteristics of thermal imaging, providing a novel method for situations where traditional visual data and techniques prove inadequate. The process of training and evaluating models using this dataset is comprehensively outlined. In particular, we employ a hybrid approach that leverages a pre-trained model, Mobilenetv2, in conjunction with a Convolutional Neural Network (CNN) encoder-decoder including a Feature Pyramid Network (FPN). This combination aims to enhance the performance of facial landmark detection using thermal

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3503-9494-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/FG59268.2024.10581870

images. Additionally, we explore the use of anisotropic loss function, an aspect of our methodology that contributes to the effectiveness of our approach. Third, we conduct a comprehensive comparison between our proposed model and other existing methods applied to thermal datasets. This comparative analysis aims to assess our approach in relation to established techniques.

II. RELATED WORK

Several studies have created thermal datasets to facilitate research and development in the field of computer vision and facial landmark detection. These datasets serve as valuable resources for training and evaluating landmark detection models specifically designed for thermal imagery, which is a relatively new research area. In this section, we explore the datasets utilized for training and evaluating landmark detection models, along with the approaches employed in literature to tackle this challenging task.

A. Datasets

In the context of thermal facial landmark detection, a significant challenge lies in the limited availability of comprehensive thermal datasets. These datasets are pivotal for driving research progress, enabling model development and rigorous evaluation. They provide diversity in subject demographics, facial expressions, and head positions. Due to their instrumental role, researchers have curated datasets to facilitate model development and evaluation in thermal facial landmark detection. The RWTH-Aachen University Database offers 2,935 thermal images from 90 subjects, featuring diverse facial expressions and 68 manually annotated landmark points, making it a valuable resource for studying facial recognition in thermal imagery [10]. The ARL-VTF (DEVCOM Army Research Laboratory Visible-Thermal Face Dataset) stands out with an extensive collection of 500,000 images from 395 subjects, captured concurrently in thermal and visible spectrums, complete with annotations for facial bounding boxes and six key facial landmarks including the left eye center, right eye center, base of nose, left mouth corner, right mouth corner, and center of mouth [23]. The TFW (Thermal Faces in the Wild) dataset presents thermal images from diverse indoor and outdoor environments. It includes 9,982 frames and 16,509 labeled faces from 147 subjects, with challenging scenarios and conditions [12]. The SF-TL54 Dataset, derived from the SpeakingFaces dataset, provides 2,556 thermal images of faces from 142 participants recorded in multiple positions, it emphasizes controlled conditions and diverse facial expressions [13]. The Charlotte-ThermalFace: UNC Charlotte Thermal Face Database offers images in a 16-bit TIFF format and documents controlled thermal variation, environmental properties, and facial landmarks [3].

Comprehensive facial landmark datasets, aside from Charlotte, are generally available only in RGB or Grayscale format. Despite representing thermal images, these datasets introduce additional challenges and require extra processing to build models that handle these converted formats. A model could be simplified by interacting directly with thermal frames, which contain temperature values rather than pixel values. This temperature data further expands the applicability of the data in different real-world applications, which may utilize temperature data directly instead of other formats.

B. Thermal Facial Landmark Detection Methods

Several approaches have been used to leverage the potential of thermal face biometrics for diverse applications, encompassing detection, monitoring, recognition, and identification [11]. Within this context, landmark detection in thermal imagery emerges as a pivotal technique for accurate and comprehensive facial feature analysis.

In earlier studies, Kopaczka et al. [9] developed an LWIR face tracking method based on Active Appearance Models (AAM), addressing challenges in LWIR facial landmark detection. Their work stressed the need for a dedicated thermal face database and employed dense image features with contrast-enhancing preprocessing for AAM training. In a related study [10], the authors introduced a head pose estimation method using random forest regression to improve landmark detection for non-frontal faces in the thermal spectrum. Furthermore, they evaluated the effectiveness of Deep Alignment Networks (DANs) for thermal facial landmark detection and tracking. Poster et al. [22], [23], employed DAN, in addition to multi-task convolutional neural network (MTCNN) and the Multi-Class Patch-Based Classifier (PBC), as part of their research on landmark detection in thermal imagery. Chu et al. [6], developed a neural network for facial landmark detection and emotion recognition in thermal face images. They used a U-Net structure, trained in two stages, with multi-task learning. Mallat et al. [16], conducted training using both Active Appearance Models (AAMs) and a Deep Alignment Network (DAN) on pseudo-thermal images. However, AAM cannot be used in real-time situations since the computation required by the AAM approach is costly [18].

More recently, YOLOv5-Face was used by kuzdeuov et al. [12], which is specifically designed for facial detection in images, providing outputs in the form of bounding boxes and the coordinates of five crucial facial landmarks. In a similar vein, Anghelone et al. [2], proposed TFLD which is based on YOLOv5, which consists of two consecutive modules: a face detection module and a landmark detection module. Furthermore, kuzdeuov et al. [13] trained two models for the facial landmark detection task to show the efficacy of their dataset. The first model is a classic machine learning model, dlib, based on an ensemble of regression trees. In contrast, the second model employs a deep learning architecture using the U-Net design.

III. DATASET DESCRIPTION

In this section, we present an overview of our novel thermal dataset, D5050, for facial landmark detection. We provide insights into the dataset's acquisition, composition, annotation methodology, and inter-annotator agreement, which are crucial for landmark detection and related research.

A. Data Acquisition

The thermal dataset was acquired using a FLIR SC6700 camera, capable of capturing high-resolution thermal images at 640x512 pixels with a substantial capacity of 7.2 million electrons. To ensure the capture of dynamic facial expressions and movements, the camera was configured to record at a frame rate of 100 frames per second.

The dataset encompasses 166 video sequences, featuring a total of 163 unique subjects. From this collection, 5050 thermal images were selected and annotated for facial landmark detection, capturing a diverse array of scenarios. Among these images, 130 include subjects wearing glasses, comprising five individuals. Furthermore, the dataset comprises subjects captured at different times. The dataset's core comprises 157 subjects, each providing 30 annotated images for analysis. We also captured 160 images of the same subjects at different times, with two of the subjects providing 30 images in the morning and 30 images in the evening, and another subject contributed only 10 images in the evening of the following day. Three other subjects were represented by 60 images each, with 30 images having no occlusions and 30 with occlusions using glasses. A noteworthy aspect of this dataset is its balanced gender representation, with 70 of the subjects, about 43%, being female, further enhancing its diversity in terms of gender characteristics.

B. Data Annotation

To annotate this comprehensive dataset, we employed a multi-step approach that combined computer vision models with manual correction. Initially, we utilized face detection and facial landmark prediction models from the dlib library as a starting point for the annotations. However, due to variations in camera specifications and image quality, as well as the fact that dlib was not designed to be run directly on the temperature values of thermal images, but on thermal images in grayscale format, we observed inaccuracies in the predicted facial landmarks, as expected. To address this, two annotators individually examined each annotation and corrected and refined landmarks as needed.

The annotation scheme for this dataset is based on the SF-TL54 annotation [13], which, in turn, relies on the Multi-PIE annotation scheme and the Labeled Face Parts in the Wild (LFPW) dataset [5], [29]. It focuses on 54 facial landmarks, including key areas such as the chin, eyebrows, nose bridge, nose tip, eyes, and lips, with specific landmarks as follows: chin (1-17), left eyebrow (18-22), right eyebrow (23-27), nose bridge (28-31), nose tip (32-36), left eye (37-42), right eye (43-48), and lips (49-54). While Multi-PIE typically employs a 68-point landmark set according to the IBUG standard, our adaptation aligns with the SF-TL54 scheme's 54-point landmark set.

C. Inter-Annotator Agreement

In our evaluation, we initially focused on metrics that assess annotator agreement, providing insights into the consistency and reliability of our annotations. We employed Pearson's Correlation Coefficient (CC) to quantify the linear relationship between the landmark coordinates ($\hat{l}$) generated by two annotators [20], [17].

Moreover, we utilized the Concordance Correlation Coefficient (CCC) to extend our evaluation, combining CC with measures of agreement in means and variances between the landmark coordinates generated by the two annotator [36].

These metrics capture both correlation and agreement between the annotations made by the two annotators and provide a robust assessment of the consistency and reliability of the annotations.

When assessing inter-annotator agreement for our dataset, we achieved results showing a CCC of 0.921, a CC of 0.967, and a normalized Root Mean Square Error (nRMSE) of 0.022. The nRMSE was calculated by normalizing the Euclidean distance between two points based on the diagonal of the bounding box encompassing a face. This bounding box was determined by finding the maximum and minimum coordinates of the annotated landmarks [4], [30]. This compares well to Bandini et al. [4], where the nRMSE (normalized Root Mean Square Error) was lower than 5%, and Sagonas et al. [30], with a normalized mean euclidean distance of 0.0262.

In summary, our dataset serves as a valuable resource for advancing the field of automatic thermal facial landmark detection. It features high-resolution imaging, a diverse set of subjects, and comprehensive annotated facial landmarks. Employing combined annotations from the two annotators as the ground truth enhances the dataset's overall quality and reliability, establishing it as an asset for future research.

IV. METHODOLOGY

In this section, we elaborate on the key components of our methodology, addressing the critical aspects of loss function selection and providing a detailed insight into the deep learning model architecture used for thermal facial landmark detection.

A. Loss function

The choice of an appropriate loss function is critical for achieving optimal model performance. Previous research [34], [33] has commonly employed widely recognized L1 and L2 loss functions for facial landmark localization models. However, these traditional loss functions were outperformed by Wing-Loss introduced in [7].

In our research, we adopt the anisotropic loss function as outlined in [25]. The choice is motivated by the need to account for likely uncertainties in training point positioning. Rather than employing Euclidean distances and treating all points uniformly, we leverage the Mahalanobis distance with anisotropic covariance matrices. This choice allows each facial landmark point to be assigned a distinct weight, enhancing the model's adaptability to variations in point significance. The anisotropic covariance matrices play a crucial role in penalizing movements away from boundaries, particularly for points situated on edges. This deliberate construction aims to address the challenges posed by uncertain training point positions, prioritizing accuracy

in proximity to edges. Additionally, landmarks with well-defined x and y directions, such as the corners of the eyes or mouth, are assigned unit covariance matrices. This relatively simple change not only improves the overall accuracy but also allows for faster convergence during training.

Suppose we have two shapes, each containing n points, represented by $\{x_i\}$ and $\{z_i\}$. We assume a covariance matrix, S_i, at each point $\{z_i\}$. The cost function is defined as:

$$Loss = \sum_{i=1}^{n} (x_i - z_i)^T W_i (x_i - z_i) \tag{1}$$

where $W_i = S_i^{-1}$ is the weight matrix at each point. To represent anisotropic distributions, a covariance matrix S is constructed with a variance of a^2 along direction u ($|u| = 1$) and b^2 along the orthogonal direction:

$$S = a^2 uu^T + b^2 (I - uu^T) \tag{2}$$

This results in the weight matrix:

$$W = S^{-1} = \frac{1}{a^2} uu^T + \frac{1}{b^2} (I - uu^T) \tag{3}$$

Assuming the points $\{z_i\}$ are roughly equally spaced along a curve, the tangent at point i is given by:

$$t_i = \frac{z_{i+1} - z_{i-1}}{|z_{i+1} - z_{i-1}|} \tag{4}$$

similar to the one outlined in [13] For points at the ends of an open curve, $t_1 = \frac{z_2 - z_1}{|z_2 - z_1|}$ and $t_n = \frac{z_n - z_{n-1}}{|z_n - z_{n-1}|}$. The weight matrices are then set up as:

$$W_i = S_i^{-1} = \frac{1}{a^2} t_i t_i^T + \frac{1}{b^2} (I - t_i t_i^T) \tag{5}$$

where $a > b$, providing more freedom to slide along the curve than normal to it. For our paper, based on experimental results conducted on a validation set, we choose $a = 4$ and $b = 2$. At well defined corners we use Wi = I, where I is the identity matrix. The cost function exhibits more rapid convergence, as observed in practical applications [25].

B. Pyramid MobileNet Thermal (PyraMoT) architecture

The architecture of our model is illustrated in Fig. 1. Our training methodology follows a three-stage training approach, similar to the one outlined in [13], which consists of mask extraction, fully connected layer pre-training, and full training.

In the first stage of our methodology, we present a hybrid architecture that combines the efficiency of MobileNetV2 with a custom encoder-decoder structure. The initial stage consists of using MobileNetV2 as the base model, which is a neural network architecture known for its efficiency, speed, and suitability for real-time applications on resource-constrained devices [32]. We extracted the bounding boxes for our dataset by padding landmarks, which was used to crop the images to the face. Following that, the resizing of the input image and passing it through a 1x1 convolution to align with the MobileNetV2 input shape occurred simultaneously with the encoder block. This block consists of a series of convolutional layers with 3x3 filter size and varying strides. These layers progressively downsample the input image, capturing diverse features at different scales. Leaky ReLU and ReLU activation functions are applied to introduce non-linearity and enhance feature representation. Following the encoder, we integrate Feature Pyramid Networks (FPN) into our model architecture [14]. To generate feature maps at various scales, we utilize both upsampling and lateral connections. These connections play a vital role in enhancing the fusion of features from different stages, thereby improving the overall feature representation. Resizing MobileNetV2 features to align with the encoder's output size is followed by a fusion layer that concatenates the extracted features with features from the preceding stage, establishing a connection between the MobileNetV2 base and the custom encoder. The decoder block comprises convolutional transpose layers with dropout to reconstruct the image. The upsampling and concatenation steps mirror the FPN-like structure, allowing the model to recover spatial information at different scales. Dropout is incorporated for regularization to prevent overfitting. The final part of the model consists of a series of convolutional transpose layers that gradually upsample the features. The output layer, using a 1x1 convolution with tanh activation, produces the final output. The model is further detailed in Table I.

TABLE I

OVERVIEW OF THE PYRAMOT MODEL ARCHITECTURE, DETAILING LAYER SPECIFICATIONS INCLUDING KERNEL SIZES, STRIDES, ACTIVATION FUNCTIONS, AND DROPOUT RATES.

Layer	Kernels	Stride	Activation	Dropout
encoder	64 × (3 × 3)	2 × 2	LeakyReLU	-
encoder	128 × (3 × 3)	1 × 1	LeakyReLU	-
encoder	128 × (3 × 3)	2 × 2	relu	-
encoder	256 × (3 × 3)	2 × 2	relu	-
encoder	256 × (3 × 3)	1 × 1	relu	-
encoder	512 × (3 × 3)	2 × 2	relu	-
encoder	256 × (3 × 3)	2 × 2	relu	-
p5	256 × (1 × 1)	-	relu	-
p4	-	UpSampling2D	-	-
p3	-	UpSampling2D	-	-
p3	256 × (1 × 1)	-	relu	-
p4	256 × (1 × 1)	-	relu	-
p5	256 × (1 × 1)	-	relu	-
fusion	-	-	-	-
decoder	512 × (3 × 3)	2 × 2	relu	25%
fusion	-	-	-	-
decoder	256 × (3 × 3)	2 × 2	relu	25%
fusion	-	-	-	-
decoder	128 × (3 × 3)	2 × 2	relu	25%
decoder	64 × (3 × 3)	2 × 2	relu	25%
decoder	64 × (3 × 3)	1 × 1	relu	-
decoder	32 × (3 × 3)	2 × 2	relu	-
output_layer	1 × (1 × 1)	-	tanh	-

In Fig. 2, we illustrate the predicted mask obtained in the initial step of our facial landmark detection model, giving a straightforward representation of the model's performance in identifying facial landmarks.

Following the mask extraction, the second stage involves the introduction of three fully connected layers connected to the output layer from the PyraMoT model. The first and second layers contain 2,048 and 512 neurons, respectively,

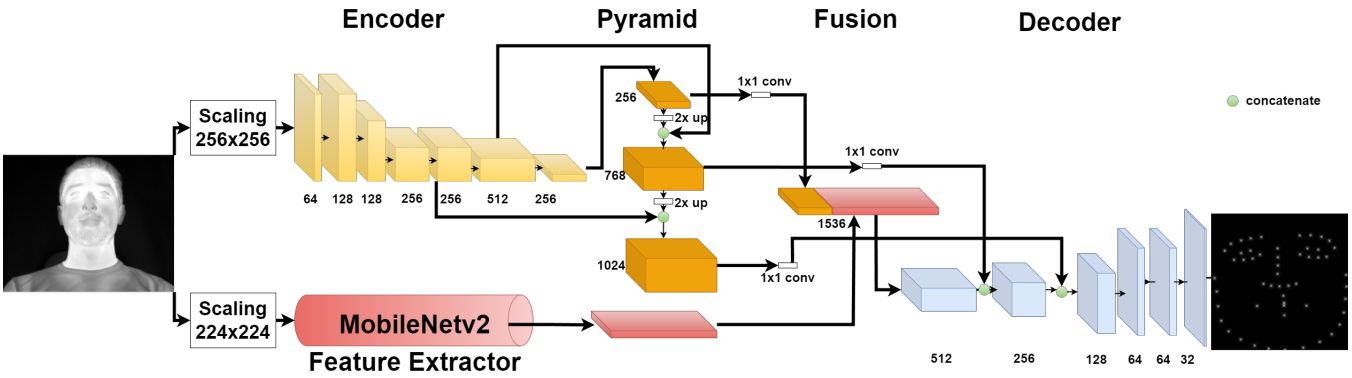

Fig. 1. Overview of the first stage of PyraMoT architecture

with a 30% dropout. The final layer has 108 neurons as the number of x, y coordinates is 2 × K, where K represents the number of landmarks. We freeze the trainable parameters of the PyraMoT model during this stage, aiming to pre-train the final connected layers without altering the weights of the PyraMoT model. The training objective is set to minimize the anisotropic loss, emphasizing the importance of accurate predictions for facial landmark coordinates.

The third and final stage is a comprehensive training process where the frozen layers of the model are released, enabling the complete network to be fine-tuned. The objective is to adjust the weights of both the PyraMoT model and the last fully connected layers, enhancing the overall performance of the model in facial landmark detection. This three-stage approach draws inspiration from the literature [13] and aims to optimize the model efficiency and accuracy.

Fig. 2. a) Input face image. b) The ground-truth mask. c) The predicted mask

V. EXPERIMENTAL RESULTS

In this section, we present the evaluation metrics employed in our assessment, along with a discussion of the results for the various methods used in our experiments.

A. Evaluation Metrics

1) Normalized Mean Error (NME): In evaluating the precision of our facial landmark detection models, we assessed accuracy using the Normalized Mean Error (NME) metric [7], [13], [29], which measures the average of the squared differences between predicted and actual landmark coordinates, normalized by the interocular distance. The NME is calculated as:

$$\text{NME} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\hat{l}_i - l_i\|}{K \cdot D_i} \tag{6}$$

Where $\hat{l}_i$ and l_i represent the predicted and ground-truth landmark coordinates for the i-th landmark point, K is the total number of facial landmarks, D_i is the interocular distance, typically defined as the distance between the outer corners of the eyes for the i-th image [29], [37], and N represents the total number of images.

2) Failure Rate (FR): In addition to NME, we also present the Failure Rate (FR) metric, which serves as an indicator of localization quality. The FR metric designates predictions as failures when the NME (Normalized Mean Error) surpasses a predefined threshold (t). In our case the threshold is set to t=10% [7], [35].

3) Cumulative Error Distribution (CED) Curve: Another supplementary metric we employed is the Cumulative Error Distribution (CED) curve [25]. The CED curve illustrates the proportion of examples with NME values below a given threshold (t), providing a comprehensive view of prediction quality. A higher Area Under Curve (AUC) in the CED curve signifies superior performance. These combined metrics offer a robust assessment of our landmark detection models' accuracy and localization quality, as used in literature.

B. Comparative Analysis of Methods Performance

1) Data Split: In our analysis we report the metrics using the three available datasets, which are the RWTH-Aachen University Dataset, the SF-TL54 Dataset, and our own D5050 Dataset. Table III provides an overview of each dataset, detailing the sizes of the training, validation, and test image sets, as well as the number of subjects involved, the presence of occlusions, the color palette, and the number of landmarks. The color palette in this table corresponds to the different format used to represent the thermal images, including gray, iron, negative, and radiometric temperatures (Thermal). The Charlotte-ThermalFace Dataset, which contains over 10,000 images, was also considered in our experiments; several analyses, tests, and models were performed and run on the dataset's thermal (gray and radiometric temperature)

TABLE II

COMPARISON OF NME FOR DIFFERENT METHODS ON THE THERMAL DATASETS

Dataset	D5050			SF-TL54			RWTH-Aachen
Cobebase/Type	Gray	Iron	Thermal	Gray	Iron	Negative	Gray
U-Net	0.0335	0.035	0.0365	0.0358	0.0368	0.0364	0.0371
MobileNetV2	0.0391	0.0456	0.0863	0.0422	0.0415	0.0407	0.0372
ResNet50	0.0626	0.0544	0.0712	0.128	0.133	0.113	0.0811
PyraMoT	**0.0329**	**0.0337**	**0.035**	**0.034**	**0.0354**	**0.0358**	**0.0353**
YOLOv5	0.045	0.054	0.041	0.046	0.064	0.046	0.078
AAM	0.0562	0.0559	0.0586	0.0737	0.0724	0.0728	0.044

images. Due to some deficiencies in the dataset, including widespread missing or erroneous ground truth landmarks, the dataset and our results on it were removed from the paper. Our attempts at filtering Charlotte's data were not successful due to the randomness of the deficiencies, which would only be correctable with a significant manual effort to certify or recreate the ground truth of the dataset. Although the dataset contains a diverse range of images, to the best of our knowledge, no previous studies on landmark detection have been conducted on it.

TABLE III

DATASETS OVERVIEW

Dataset	D5050	SFTL54	RWTH-Aachen
Subjects	163	142	94
Color palette	3	3	1
Number of landmarks	54	54	68
occlusions	Yes	Yes	No
Train $\|\|subject$	3850 $\|\|126$	1800 $\|\|100$	2042 $\|\|58$
Validation $\|\|subject$	600 $\|\|20$	180 $\|\|10$	222 $\|\|8$
Test $\|\|subject$	600 $\|\|17$	576 $\|\|32$	671 $\|\|24$

2) Performance Analysis: We thoroughly evaluated the performance of various methods, including AAM [10], YOLOv5 [2], MobileNetV2 [32], ResNet50 [8], PyraMoT (our proposed method), and U-Net [13]. Furthermore, we explored the potential of dlib; however, because we used it in our semi-manual annotation process and to maintain impartiality, we decided not to include the results obtained with dlib in the paper. The Active Appearance Models (AAM) method was implemented using the Menpo library for facial landmark detection in thermal images. Key preprocessing steps included converting landmarks to a standard PTS format and adjusting them to the thermal image dimensions. The AAM, configured with HolisticAAM and Image Gradient Orientation (IGO) feature, was trained on a subset of images. LucasKanadeAAMFitter was also implemented to the fitting process.

In terms of the YOLOv5 implementation, the lack of specified parameters in the paper, particularly the width and height of the bounding boxes, presented a challenge. Given the lack of clear guidance on these specific parameters, we defined our own settings in accordance with the author's methodology. We calculated the width (w) and height (h) of the bounding boxes as proportional to the inter-eyes distance (IED), using the formulas $w = IED * k$ and $h = IED * k$, where k is a scaling factor less than one.

The implementation of MobileNetV2 and ResNet50 involved modifying the last layer to align with the specific number of classes corresponding to our landmarks. Subsequently, the models were trained directly on the datasets with this adjustment.

For U-Net, we followed the exact steps provided in the paper, including the application of the Adam optimizer and a dropout rate of 0.5 for regularization. The Wing loss function with parameters ($w = 10$, $\epsilon = 2.0$) was also employed.

For all datasets used in our research, we implemented data augmentation by horizontal flipping, thereby doubling the dataset size for training, validation, and testing. This method was chosen to allow for a direct comparison with the results reported in this paper [13]. Our goal was to maintain consistency in the augmentation strategy while assuring a fair and meaningful evaluation for a thorough comparative analysis.

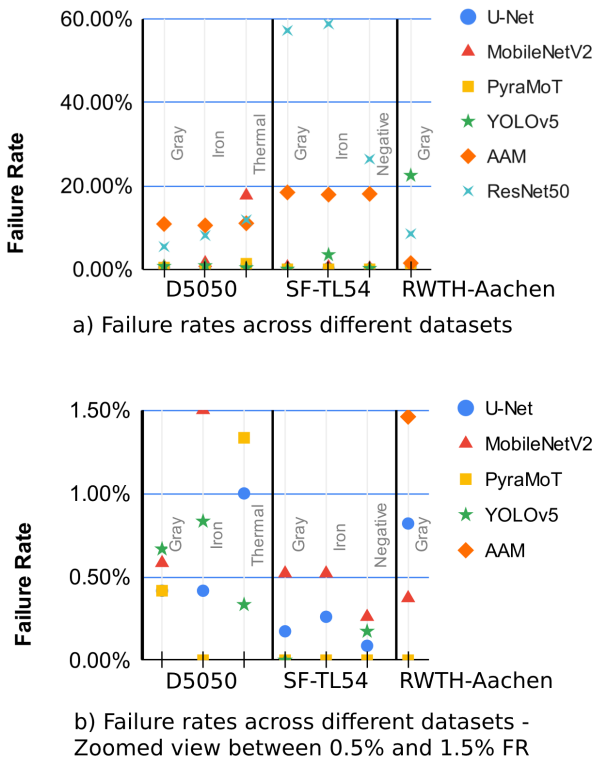

Fig. 3. Comparison of Failure Rate for Different Methods on the Thermal Datasets

Table II presents the NME reported for all methods across

diverse datasets. The D5050 dataset achieved the lowest errors compared to the same image palette in the other datasets. Among the methods, U-Net reached an NME of 0.0335, 0.0358, and 0.0371 on the D5050, SF-TL54, and RWTH-Aachen datasets using the gray-scale images, respectively. However, PyraMoT outperformed U-Net with an NME of 0.0329 on D5050, and 0.035 and 0.0353 on SF-TL54 and RWTH-Aachen using the gray-scale images, respectively. One possible explanation for PyraMoT's higher performance is its architecture that integrates MobileNetV2 as feature extractor and Feature Pyramid Network as multi-scale features, which employs hierarchical lateral connections for object detection tasks within thermal images, while U-Net utilizes symmetrical skip connections that focuses on preserving fine-grained spatial information for semantic segmentation purposes. The combination of low level features with high level features allows the model to capture both fine and coarse details in the image. This feature fusion results in a comprehensive representation of the input image at multiple scales, enhancing object localization, particularly for small objects, through the provision of feature maps at different resolutions.

PyraMoT also outperformed all methods, including U-Net which follows the three-stage training approach. In addition, U-Net demands more resources and time for training, as it has around 34 million trainable parameters compared to PyraMoT's approximate 16 million trainable parameters for the first stage of training.

MobileNetV2 and ResNet50 are primarily designed and trained for RGB images and, in particular, image classification tasks. Radiometric thermal image performance issues may surface as a result of RGB channel optimization, necessitating potential adaptation or transfer learning for better results on such datasets. This may provide one explanation of the relatively poorer performance seen on MobileNetV2 on the radiometric thermal D5050 dataset, with 0.0863 NME, compared to the gray and iron palettes that align more closely with the RGB images, at 0.0391 and 0.0456 respectively. ResNet50, when compared against MobileNetV2, performed relatively worse on SF-TL54 and RWTH-Aachen than on D5050. One possible reason for that is ResNet's deeper architecture, which typically requires a larger dataset to achieve better performance.

Aside from the effect of the characteristics of the different models on their performance, we observed that all methods performed better on average on the D5050 dataset when compared to the SF-TL54 and RWTH-Aachen datasets. This may be due to the extensive number of images and their high resolution in the D5050 dataset.

While the YOLOv5 architecture is capable of effectively detecting objects and their bounding boxes, it isn't inherently optimized for the detection of facial landmarks, particularly in the thermal color palettes. Furthermore, a requirement for YOLOv5 is that the input needs to be converted into an image first, meaning the model is not able to process the raw thermal data stored as numpy arrays. Similarly, the AAM model's architecture utilizes shape and texture in

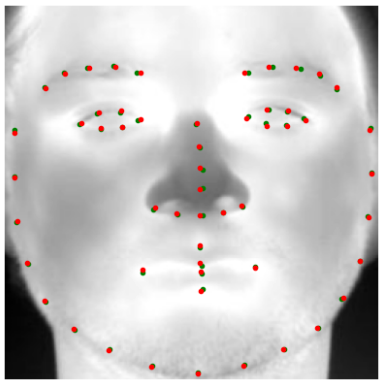

Fig. 4. Comparison of Predicted Landmarks (in red) and Ground Truth (in green).

defining facial landmarks, both of which are not as apparent in thermal images as compared to RGB images. Compared to U-Net and PyraMoT, these two methods had a higher error across all datasets.

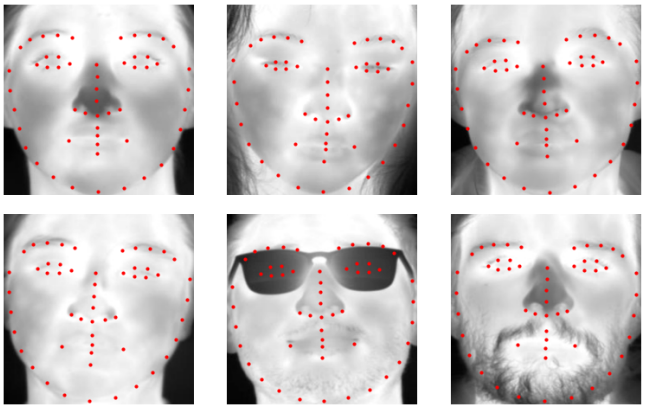

Fig. 5. Facial landmark predictions in diverse images, encompassing different subjects, poses, and occlusions.

In addition to the NME results, we present the FR results in Fig. 3. The y-axis, with a threshold of 10%, shows the percentage of points that have an NME higher than 0.1. Fig. 3a shows that ResNet50 and AAM have very high FR. Fig. 3b, which is a zoomed view of the Failure rates between 0% and 1.5%, shows lower FR for U-Net and PyraMoT. Both methods exhibit lower FR percent across the majority of datasets, with only 1.33% and 1% Failure Rate (FR) for PyraMoT and U-Net, respectively, on the temperature scale. This further proves the effectiveness of both methods for this task compared to the other methods. Of note is that YOLOv5 resulted in low FR, in part due to NME being calculated for the predicted landmarks, as YOLOv5 fails to consistently predict all landmarks, which resulted in low errors. Nevertheless, the average NME remains higher than that of PyraMoT.

To assess PyraMoT's performance more comprehensively, Fig. 4 visualizes the error by comparing predicted landmarks in red to ground truth landmarks in green. The visual

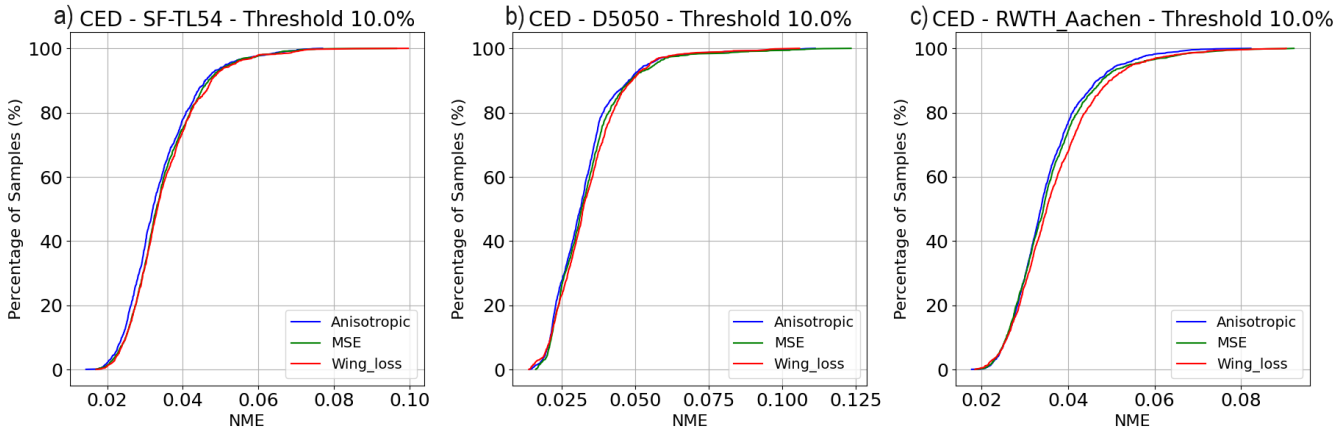

Fig. 6. Cumulative Error Distribution (CED) for Different Datasets and Loss Functions

disparity between the two sets allows a quick assessment of the model's accuracy.

Fig. 5 extends this evaluation by showcasing PyraMoT's predicted landmarks across various images, including occlusions. While our model currently demonstrates proficiency in handling glasses as an occlusion, addressing potential unknown occlusions like face masks or caps requires ongoing research for thermal landamrks detection. The model maintains reliable landmark placement around the subject's face, even in challenging scenarios that include glasses and beards, scoring an NME of 0.0335. This indicates PyraMoT's potential reliability for future facial segmentation tasks on thermal images.

Following that, we delve into the analysis of the loss functions performance across the three datasets in Fig. 6. We observe that MSE performs better on the RWTH-Aachen dataset compared to the Wing loss function, while for the other datasets, Wing loss shows slightly better results. The anisotropic loss function outperformed both loss functions across all three datasets. The approach of penalizing movements away from boundaries proves effective, as explained earlier, especially for points on edges, as we assign distinct weights to each facial landmark point.

VI. CONCLUSION

In this paper, we presented novel contributions to the field of thermal facial landmark detection, a domain that, despite its potential in security, autonomous vehicles, and health domains, has been relatively under-explored. We introduced a comprehensive new thermal dataset, a novel thermal facial landmark detection framework, PyraMoT, along with an extensive comparative analysis against existing methodologies across three datasets and four color palettes, including gray, iron, negative, and radiometric thermal. The D5050 dataset, featuring a diverse collection of thermal facial images from 163 distinct subjects, including instances with occlusions, addressed a critical gap in the availability of quality thermal image data with comprehensive facial landmark annotations through manual annotation.

By integrating the efficient MobileNetV2 architecture with a custom encoder-decoder structure, including a Feature Pyramid Network, PyraMoT showed better performance in extracting and utilizing features from thermal images. Our comparative analysis against U-Net, MobileNetV2, ResNet50, YOLOv5, and AAM showed that our proposed approach outperformed other approaches for all of the datasets and scenarios using the Normalized Mean Error (NME) and Failure Rate (FR) to quantitatively and comparatively evaluate the accuracy and reliability of various detection models. On the D5050-gray dataset, PyraMoT achieved a lower Normalized Mean Error (NME) of 0.0329 compared to U-Net's 0.0335, and continued this trend across SF-TL54 and RWTH-Aachen datasets with NMEs of 0.034 and 0.0353, respectively. In particular, PyraMoT achieved the lowest NME of 0.035 on the D5050 radiometric thermal images compared to all the other methods. The PyraMoT architecture, which consists of MobileNetV2 as a feature extractor and FPN for multi-scale features, improves object localization, particularly for small objects, by providing feature maps at different levels. This allowed the model to capture both fine-grained and coarse-grained details in the image, making it capable of handling objects of different sizes, such as facial landmarks. Unlike models optimized for RGB images, such as MobileNetV2 and ResNet50, PyraMoT proved better at thermal image processing, evidenced by lower failure rates and higher accuracy in challenging conditions, including occlusions. This is complemented by an advanced loss function analysis, where PyraMoT's use of an anisotropic loss function proved more effective across datasets than traditional MSE and Wing loss functions, particularly in penalizing deviations for edge points. Overall, PyraMoT's combination of lower error rates, efficient training, and robust performance in diverse conditions suggests it as an effective approach for thermal landmark detection.

We believe that our findings will serve as a foundation for future research and development, driving advancements in both the theoretical and practical aspects of thermal image processing for facial landmark detection.

REFERENCES

[1] F. Abdat, C. Maaoui, and A. Pruski. Human-computer interaction using emotion recognition from facial expression. In *2011 UKSim 5th European Symposium on Computer Modeling and Simulation*, pages 196–201. IEEE, 2011.

[2] D. Anghelone, S. Lannes, V. Strizhkova, P. Faure, C. Chen, and A. Dantcheva. Tfld: Thermal face and landmark detection for unconstrained cross-spectral face recognition. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2022.

[3] R. Ashrafi, M. Azarbayjani, and H. Tabkhi. Charlotte-thermalface: A fully annotated thermal infrared face dataset with various environmental conditions and distances. *Infrared Physics & Technology*, 124:104209, 2022.

[4] A. Bandini, S. Rezaei, D. L. Guarín, M. Kulkarni, D. Lim, M. I. Boulos, L. Zinman, Y. Yunusova, and B. Taati. A new dataset for facial motion analysis in individuals with neurological disorders. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1111–1119, 2020.

[5] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[6] W.-T. Chu and Y.-H. Liu. Thermal facial landmark detection by deep multi-task learning. In *2019 IEEE 21st international workshop on multimedia signal processing (MMSP)*, pages 1–6. IEEE, 2019.

[7] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2235–2245, 2018.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] M. Kopaczka, K. Acar, and D. Merhof. Robust facial landmark detection and face tracking in thermal infrared images using active appearance models. In *VISIGRAPP (4: VISAPP)*, pages 150–158, 2016.

[10] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof. A thermal infrared face database with facial landmarks and emotion labels. *IEEE Transactions on Instrumentation and Measurement*, 68(5):1389–1401, 2018.

[11] M. Krišto and M. Ivasic-Kos. An overview of thermal face recognition methods. In *2018 41St international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1098–1103. IEEE, 2018.

[12] A. Kuzdeuov, D. Aubakirova, D. Koishigarina, and H. A. Varol. Tfw: Annotated thermal faces in the wild dataset. *IEEE Transactions on Information Forensics and Security*, 17:2084–2094, 2022.

[13] A. Kuzdeuov, D. Koishigarina, D. Aubakirova, S. Abushakimova, and H. A. Varol. Sf-tl54: A thermal facial landmark dataset with visual pairs. In *2022 IEEE/SICE International Symposium on System Integration (SII)*, pages 748–753. IEEE, 2022.

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[15] X. Liu, Y. Zou, H. Kuang, and X. Ma. Face image age estimation based on data augmentation and lightweight convolutional neural network. *Symmetry*, 12(1):146, 2020.

[16] K. Mallat and J.-L. Dugelay. Facial landmark detection on thermal data via fully annotated visible-to-thermal data synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020.

[17] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[18] K. Nagumo, T. Kobayashi, K. Oiwa, and A. Nozawa. Face alignment in thermal infrared images using cascaded shape regression. *International Journal of Environmental Research and Public Health*, 18(4):1776, 2021.

[19] T. Negishi, S. Abe, T. Matsui, H. Liu, M. Kurosawa, T. Kirimoto, and G. Sun. Contactless vital signs measurement system using rgb-thermal image sensors and its clinical screening test on patients with seasonal influenza. *Sensors*, 20(8):2171, 2020.

[20] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[21] D. Perpetuini, C. Filippini, D. Cardone, and A. Merla. An overview of thermal infrared imaging-based screenings during pandemic emergencies. *International Journal of Environmental Research and Public Health*, 18(6):3286, 2021.

[22] D. Poster, S. Hu, N. Nasrabadi, and B. Riggan. An examination of deep-learning based landmark detection methods on thermal face imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[23] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman, X. Di, C. N. Fondje, V. M. Patel, N. J. Short, B. S. Riggan, N. M. Nasrabadi, et al. A large-scale, time-synchronized visible and thermal face dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1559–1568, 2021.

[24] D. D. Poster, S. Hu, N. J. Short, B. S. Riggan, and N. M. Nasrabadi. Visible-to-thermal transfer learning for facial landmark detection. *IEEE Access*, 9:52759–52772, 2021.

[25] F. Rayhan, A. Galata, and T. F. Cootes. Not all points are created equal-an anisotropic cost function for facial landmark location. In *BMVC*, 2020.

[26] K. Riani, M. Papakostas, H. Kokash, M. Abouelenien, M. Burzo, and R. Mihalcea. Towards detecting levels of alertness in drivers using multiple modalities. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–9, 2020.

[27] K. Riani, S. Sharak, M. Abouelenien, M. Burzo, R. Mihalcea, J. Elson, C. Maranville, K. Prakah-Asante, and W. Manzoor. Non-contact based modeling of enervation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.

[28] K. Riani, S. Sharak, K. Das, M. Abouelenien, M. Burzo, R. Mihalcea, J. Elson, C. Maranville, K. Prakah-Asante, and W. Manzoor. Towards classifying human circadian rhythm using multiple modalities. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

[29] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013.

[30] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903, 2013.

[31] A. Samara, L. Galway, R. Bond, and H. Wang. Affective state detection via facial expression analysis within a human–computer interaction context. *Journal of Ambient Intelligence and Humanized Computing*, 10:2175–2184, 2019.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[33] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 824–832, 2015.

[34] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[35] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[36] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016.

[37] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018.

[38] Y. Wu and Q. Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127:115–142, 2019.