2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

PyraSegNet: A Novel Framework for Thermal Facial Image Segmentation

Kais Riani[1] , Mohamed Abouelenien[1] and Oumaima Jouiri[2]
[1]Computer and Information Science, University of Michigan, Dearborn, USA
[2]Concordia University, Montreal, Quebec, Canada

kriani@umich.edu[1] zmohamed@umich.edu[1] Oumaima.jouiri@mail.concordia.ca[2]

2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3315-5341-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/FG61629.2025.11099379

Abstract— **Thermal facial region segmentation is critical for applications such as physiological monitoring, stress detection, and human-computer interaction. Despite its importance, progress in this domain has been hindered by challenges such as occlusions, the lack of thermal annotated datasets and the absence of tailored, high-performing models on thermal images. In this paper, we address these challenges through four key contributions. First, we introduce the TFR dataset, a novel thermal face dataset that includes 5050 annotated thermal images extracted from 166 video sequences using 163 different subjects, which is, to our knowledge, the largest thermal dataset with comprehensive annotated regions in terms of the number of subjects, captured under diverse conditions, including variations in lighting, occlusions, and demographics, ensuring robustness and adaptability. Second, we propose PyraSegNet, an advanced deep learning architecture combining a customized encoder-decoder structure with a Feature Pyramid Network (FPN) to enhance multi-scale feature learning. Third, we compare PyraSegNet to other state-of-the-art methods. Finally, we perform an in-depth comparative analysis of loss functions, including Dice Loss, Tversky Loss, and Jaccard Loss, providing actionable guidance on optimizing thermal facial segmentation performance. Our findings underscore the significance of the TFR dataset, the effectiveness of PyraSegNet, and provide insights into optimal loss function selection for thermal facial region segmentation. This research paves the way for enhanced accuracy and reliability in applications leveraging thermal facial analysis.**

I. INTRODUCTION

Facial analysis and recognition have become crucial tools in various fields, including healthcare, security, and human-computer interaction [1], [2]. Facial features, as a medium of non-verbal communication, offer significant insights into identity, emotional states, and physiological conditions [9]. While traditional approaches to facial analysis rely heavily on RGB imagery, thermal imaging has recently emerged as a compelling alternative due to its ability to capture physiological information, such as heat distribution, under low-light or visually challenging conditions [5]. These advantages have made thermal imaging particularly relevant for applications such as stress detection, vital sign monitoring, and surveillance in low-visibility environments [7], [32]. Furthermore, thermal imaging provides the added benefit of being less intrusive, making it ideal for continuous monitoring in both clinical and non-clinical settings.

Despite its promise, the development of robust thermal facial analysis models is hindered by several challenges. A primary limitation is the scarcity of large, annotated datasets specifically designed for thermal images [3]. In contrast to RGB datasets, which are abundant and well-documented, thermal datasets remain limited in both size and diversity, constraining the development of machine learning models tailored to thermal data. Additionally, although thermal images present an important alternative to RGB images, especially in the conditions previously mentioned, the unique properties of thermal images, such as low contrast, indistinct boundaries, and increased noise, present significant hurdles for precise facial region segmentation and landmark detection [35]. These challenges underscore the need for more advanced and targeted approaches in thermal image processing.

Advances in deep learning have shown potential in addressing some of these challenges. Semantic segmentation models like U-Net and DeepLabV3+ have been adapted for thermal imaging with varying success [18], [14]. However, achieving optimal performance often requires customized architectural and algorithmic approaches. For instance, feature extraction mechanisms must account for the reduced visual contrast in thermal data, and loss functions must effectively address the uneven representation of facial regions, such as the smaller pixel density of areas, such as the eyes compared to the cheeks. This highlights the need for specialized models that can better exploit the unique characteristics of thermal images [16], [23].

In this work, we introduce PyraSegNet, a novel deep learning architecture explicitly designed for thermal facial region segmentation. PyraSegNet leverages the capabilities of InceptionV4 as a feature extractor and incorporates a custom encoder-decoder structure to enable accurate segmentation of key facial regions, including the full face, eyes, forehead, cheeks, and nose. By employing a hybrid architecture and integrating different loss functions, such as DiceLoss and TverskyLoss, PyraSegNet achieves state-of-the-art segmentation accuracy for thermal facial regions.

This paper makes the following contributions:

- TFR Dataset: Introduction of the TFR dataset, a robust annotated thermal facial dataset of the whole face, forehead, eyes, cheeks, and nose, designed to advance thermal facial region segmentation research.
- Model Innovation: Development of the PyraSegNet architecture, incorporating multi-scale feature extraction to address the unique challenges of thermal imaging.
- Loss Function Comparison: Comparison of several loss

functions, including Dice Loss, Tversky Loss, and Jaccard Loss, to evaluate their effectiveness in addressing challenges, such as uneven region distribution and low contrast in thermal image segmentation.
- Comprehensive Evaluation: Detailed experiments comparing PyraSegNet with state-of-the-art methods, including DeepLabV3+, FPN, U-Net, U-Net++, PAN, MAnet, PSPNet, and LinkNet showcasing its performance in segmentation accuracy, robustness, and generalizability across diverse thermal image conditions.

By addressing the gaps in current thermal facial segmentation methodologies, this work provides a foundation for advancing applications in physiological monitoring, stress detection, and beyond. Moreover, the insights gained from this research could have significant implications for improving human-computer interaction in environments where traditional visual analysis is limited or impractical.

II. Related Work

Various studies have explored the challenge of detecting regions of interest (ROIs) in thermal images, which is crucial for applications such as physiological monitoring and stress detection [6], [25], as well as extracting physiological signals related to enervation from facial thermal data [29]. This section reviews the datasets and methods employed for ROI segmentation in thermal imaging, highlighting methods that address these challenges and improve detection accuracy.

A. Thermal Datasets

The lack of large-scale annotated datasets for thermal facial analysis has been a barrier to the advancement of this research. To overcome this, several key datasets have been developed recently. For instance, a custom dataset created using the FLIR SC655 thermal camera, which captured thermal images of 23 subjects for medical thermography [8]. Additionally, efforts in facial skin temperature distribution for the estimation of psychophysiological state have been aided by a thermal face-tracking dataset designed for segmentation and tracking techniques [11].

Further developments include a dataset aimed at non-invasive forehead segmentation in thermal imaging systems, featuring more than 18,000 images with variations such as glasses and beards [30]. Multiclass face analysis in thermal infrared images has been supported by another dataset, captured with a FLIR AX5 bolometer camera [24]. The effectiveness of thermal image semantic segmentation has also been demonstrated using datasets such as Segment Objects in Day and Night (SODA), MFNet, and SCUT-Seg [20], [10], [36], which focus on general scene segmentation tasks rather than facial region analysis. These datasets contribute to thermal segmentation techniques [26].

Efforts in facial image segmentation also benefit from the Thermal Face Database, which, while lacking region annotations, includes manually annotated landmarks across 2,935 images. These annotations have been used to train novel frameworks such as the Self-Adversarial Multi-scale Contrastive Learning (SAM-CL) [14]. In addition, advances in facial region segmentation associated with breathing have been made using thermal images captured from various angles [19]. To predict thermal comfort by extracting skin temperature from facial components, the Thermal Face Database and Charlotte-Thermal Face Dataset were used [13].

In stress detection, datasets capturing temperature variations across facial regions (nose, cheeks, forehead) have been particularly useful [12]. The ApeTI dataset, which segments ape faces to retrieve physiological signals, introduces additional insights into the broader application of thermal imaging [22]. Other notable data sets include DeepBreath, which captures breathing patterns for stress detection [6], and UBComfort, which collects physiological metrics in semi-ecological car environments [25]. One of the largest datasets, ARL-VTF, contains 500,000 images of 395 subjects annotated with facial landmarks [27], while TFW (Thermal Faces in the Wild) offers 16,509 labeled faces in varied environments [17]. The SpeakingFaces dataset provides more than 2,500 thermal images from 142 participants with facial landmark annotations [18], and the D5050 dataset adds another valuable resource with 5,050 annotated images from 163 subjects [28]. These datasets are annotated with facial landmarks but lack region segmentation annotations, a problem we are addressing in this paper.

B. Semantic Segmentation methods

Adapting semantic segmentation techniques from RGB to thermal images presents challenges due to low contrast and unclear boundaries. Despite these obstacles, deep learning models such as U-Net and DeepLabV3+ have been successfully adapted for thermal image segmentation. These models rely on encoder-decoder architectures that help retain spatial details and improve pixel-level accuracy in thermal data [14], [16].

To enhance segmentation performance, recent approaches have incorporated novel architectural modifications tailored to thermal images. For example, FTNet introduces multiscale feature extraction and edge-guidance mechanisms to focus on object boundaries, which are often less defined in thermal images. This attention to boundaries enhances the precision of segmentation in thermal datasets [26]. Additionally, the Self-Adversarial Multiscale Contrastive Learning Framework (SAM-CL) applies contrastive learning and domain-specific augmentation to address the unique characteristics of thermal images, improving segmentation in facial regions, particularly for physiological monitoring tasks [14].

Multiclass segmentation techniques have also been adapted for thermal images, where each pixel is assigned a label corresponding to a specific facial region. This approach has been effective in medical and physiological analysis, enabling more accurate segmentation of areas such as the nose, forehead, and cheeks. Such segmentation is crucial for applications, such as stress detection and health monitoring [24]. Non-contact breathing monitoring has also benefited from these techniques, as precise segmentation of the nose region allows for more stable extraction of breathing signals from different angles [19].

Semantic segmentation has also been applied in practical domains like thermal comfort prediction. By segmenting facial regions and extracting skin temperature, researchers can gain insights into indoor comfort levels [13]. However, the lack of large, annotated thermal datasets remains a limitation, constraining further progress in improving the accuracy and generalizability of these methods [28].

III. DATASET

This section outlines the key aspects of the Thermal Facial Regions (TFR) dataset, which was created for the purpose of facial region segmentation. We describe the dataset's acquisition, its structure, and the annotation process, emphasizing its importance for region-specific facial analysis.

A. Data Acquisition

The thermal images in this dataset were captured using a FLIR SC6700 camera, known for its ability to record high-resolution thermal images at a resolution of 640×512 pixels and a sensitivity of 7.2 million electrons. The camera was set up to capture dynamic facial expressions and movements at a frame rate of 100 frames per second.

The dataset consists of 166 video sequences, which involve 163 unique individuals. From these sequences, we selected 5050 thermal images for annotation. These images cover a wide array of scenarios, including variations in time of day and the presence of occlusions. Notably, 130 of these images feature subjects wearing glasses, with five individuals contributing to this subset. The core of the dataset includes 157 subjects, each providing 30 annotated images. In addition, we gathered 160 images from the same set of subjects at different times, with two subjects contributing 30 images in the morning and another 30 in the evening, while a third subject provided 10 images in the evening of the next day. To account for occlusions, three subjects contributed 60 images each, half of which were taken without occlusions and the other half with occlusions, such as glasses. The dataset is gender-balanced, with 43% of the subjects (70 individuals) being female, ensuring diversity.

B. Data Annotation

Five facial regions were annotated within the dataset: the whole face, the forehead, the eyes, the cheeks, and the nose. The annotations were made using the Supervisely Unified OS/Platform for computer vision tasks [33]. Initially, an experienced annotator carried out and verified the annotations, which were then reviewed and adjusted by another annotator to ensure consistency and accuracy. Each facial region is represented as a polygon, providing precise delineation for region-based analysis tasks. The iterative nature of the annotation process, including several rounds of review, ensured the creation of a high-quality ground truth for facial region segmentation.

Moreover, the dataset includes detailed facial landmark annotations, based on the SF-TL54 annotation [18], which incorporates 54 key facial points, such as the chin, eyebrows, nose bridge, nose tip, eyes, and lips. To ensure accuracy, the initial landmark annotations were generated using a face detection and facial landmark prediction model from the dlib library, followed by manual correction and refinement by two annotators. This multi-step annotation process improved the reliability of the landmarks. When evaluating inter-annotator agreement for our dataset, we obtained a Concordance Correlation Coefficient (CCC) of 0.921, a Pearson's Correlation Coefficient (CC) of 0.967, and a normalized Root Mean Square Error (nRMSE) of 0.022.

In conclusion, TFR is a comprehensive resource for advancing thermal imaging research, particularly for automatic facial region segmentation. It features high-resolution thermal images, a diverse subject pool, and annotations for key facial regions. These characteristics make it a significant contribution to the fields of computer vision and thermal analysis.

IV. METHODOLOGY

In this section, we describe the key components of our methodology, focusing on the architecture of the deep learning model used for the thermal segmentation of the facial region and the choice of loss functions.

A. PyraSegNet Architecture for Thermal Facial Region Segmentation

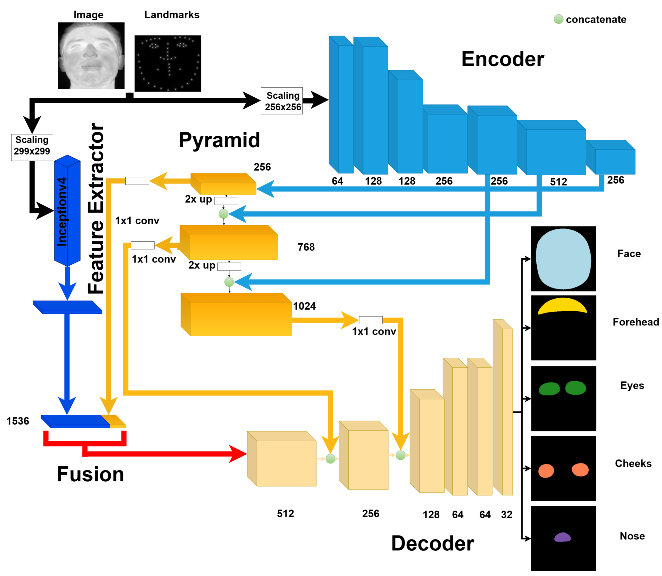


Fig. 1: Overview of PyraSegNet architecture

The architecture of our model is illustrated in Fig. 1. In this work, we used a single deep learning model for thermal facial region segmentation, building on a hybrid architecture that combines the efficiency of InceptionV4 with a custom encoder-decoder structure. InceptionV4 was chosen for its feature extraction capabilities, particularly in handling multi-scale features and providing detailed spatial information. Two types of input configurations were used in this work. The first configuration involved using only the thermal facial images as input to the model, providing a baseline for segmentation performance. Thermal images were chosen for their ability to capture temperature variations across the face, where each pixel represents the heat emitted from the surface, i.e., the absolute temperature value assigned to each pixel. This temperature data is useful for potential applications that rely on precise thermal measurements, however first, the challenge lies in accurately segmenting the facial regions to extract meaningful temperature information. The second configuration augmented the thermal facial images with our annotated facial landmarks, represented as an additional input channel. This additional information was expected to enhance the model's ability to localize and segment facial regions that require finer spatial precision, such as the eyes and nose. By comparing these configurations, we evaluated the impact of incorporating landmark information on segmentation accuracy and robustness. The input images are resized and passed through a 1x1 convolution to align with the InceptionV4 input shape, as seen in Fig. 1.

The encoder utilizes a series of convolutional layers with 3x3 kernels and varying strides to progressively downsample the input while capturing features at multiple scales. The varying strides play a crucial role in multi-scale feature extraction by enabling the model to process fine-grained details and broader contextual information simultaneously. For segmentation tasks, this capability is vital as it ensures the network can distinguish subtle differences in facial regions, such as the edges of the nose or the contours of the eyes, while still retaining global structural context. This approach improves the network's ability to produce accurate and coherent segmentation maps, particularly for complex or overlapping regions. LeakyReLU and ReLU activation functions are used to introduce non-linearity, ensuring robust feature extraction. Specifically, the encoder includes layers with increasing kernel numbers, as detailed in Table I. The architecture starts with 64 filters and progresses to 512 filters, balancing computational efficiency with feature richness. Dropout is excluded from this stage to maximize feature retention and ensure that the extracted features remain comprehensive.

To enhance multi-scale feature fusion, Feature Pyramid Networks (FPN) [21] are incorporated after the encoder. The FPN structure aggregates features from different encoder stages, enabling finer localization and semantic segmentation. The pyramid levels (P3, P4, P5) include convolutional layers and upsampling operations, merging features from high-resolution and low-resolution layers. The fusion layer integrates the outputs of the encoder and FPN by combining multi-scale features obtained from both components. Specifically, it employs a series of concatenation and 1x1 convolutional layers to merge the high-resolution and low-resolution features, ensuring seamless integration of detailed spatial information with broader contextual patterns. The fusion process ensures that the decoder receives a balanced combination of semantic and spatial features, crucial for segmenting intricate facial regions like the eyes and nose. This step ensures that features extracted at different scales are effectively combined to capture both global and local patterns. The decoder block then processes these features through convolutional transpose layers, progressively upsampling to reconstruct the spatial resolution of the input image. Dropout layers with a rate of 25% are applied in the decoder to prevent overfitting, particularly in the higher-capacity layers.

TABLE I: Overview of the PyraSegNet model architecture, detailing layer specifications including kernel sizes, strides, activation functions, and dropout rates.

Layer	Kernels	Stride	Activation	Dropout
encoder	$64 \times (3 \times 3)$	2×2	LeakyReLU	-
encoder	$128 \times (3 \times 3)$	1×1	LeakyReLU	-
encoder	$128 \times (3 \times 3)$	2×2	Relu	-
encoder	$256 \times (3 \times 3)$	2×2	Relu	-
encoder	$256 \times (3 \times 3)$	1×1	Relu	-
encoder	$512 \times (3 \times 3)$	2×2	Relu	-
encoder	$256 \times (3 \times 3)$	2×2	Relu	-
p5	$256 \times (1 \times 1)$	-	Relu	-
p4	-	UpSampling2D	-	-
p3	-	UpSampling2D	-	-
p3	$256 \times (1 \times 1)$	-	Relu	-
p4	$256 \times (1 \times 1)$	-	Relu	-
p5	$256 \times (1 \times 1)$	-	Relu	-
fusion	-	-	-	-
decoder	$512 \times (3 \times 3)$	2×2	Relu	25%
fusion	-	-	-	-
decoder	$256 \times (3 \times 3)$	2×2	Relu	25%
fusion	-	-	-	-
decoder	$128 \times (3 \times 3)$	2×2	Relu	25%
decoder	$64 \times (3 \times 3)$	2×2	Relu	25%
decoder	$64 \times (3 \times 3)$	1×1	Relu	-
decoder	$32 \times (3 \times 3)$	2×2	Relu	-
output_layer	$5 \times (1 \times 1)$	-	Sigmoid	-

The final output layer applies a 1x1 convolution followed by a sigmoid activation function to produce segmentation maps for five distinct facial regions: the whole face, forehead, eyes, cheeks, and nose. This structure allows the network to simultaneously handle multi-label segmentation tasks.

B. Loss Function

To optimize the model for accurate segmentation of the thermal facial region, we performed an evaluation of various loss functions. After thorough experimentation, we identified three loss functions that outperformed others in a preliminary comparison of segmenting specific facial regions such as the face, forehead, eyes, cheeks, and nose.

Dice Loss proved particularly effective for larger regions like the face and forehead, where maximizing overlap between predicted and ground truth regions is crucial. Its ability to measure overlap efficiently made it an ideal choice for these areas [4]. The Dice Loss is calculated as:

$$L_{\text{Dice}} = 1 - \frac{2\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i} \tag{1}$$

where y_i is the ground truth label for pixel i, $\hat{y}_i$ is the predicted probability for pixel i, and the summation is over all pixels.

Jaccard Loss demonstrated better performance in handling regions with complex boundaries, notably the eyes and nose. By evaluating the intersection over the union of predicted and true regions, it helped to achieve clearer distinctions in challenging areas [34]. The Jaccard Loss is given by:

$$L_{\text{Jaccard}} = 1 - \frac{\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i - \sum_i y_i \hat{y}_i} \quad (2)$$

Tversky Loss was invaluable for addressing the pixel-wise region imbalance by allowing fine-tuned weighting between false positives and false negatives. This was especially beneficial for smaller, underrepresented regions, such as the eyes and nose, which contain fewer pixels compared to larger regions such as the whole face. This ensured that the model properly attended to these areas despite their smaller size [31]. The Tversky Loss is defined as:

$$L_{\text{Tversky}} = 1 - \frac{\sum_i y_i \hat{y}_i}{\sum_i y_i \hat{y}_i + \alpha \sum_i y_i (1 - \hat{y}_i) + \beta \sum_i (1 - y_i) \hat{y}_i} \quad (3)$$

where $\alpha = 0.5$ and $\beta = 0.5$ in the multilabel setting. Here, α and β control the relative importance of false positives and false negatives, respectively.

In addition to the aforementioned loss functions, we explored the suitability of other metrics [15] , including RMI Loss, Focal Loss, and Lovasz Loss, but found them to be less effective for our specific thermal facial region segmentation task. In particular, the Focal Loss did not outperform the Tversky Loss in our scenario, especially for smaller facial regions.

V. EXPERIMENTAL RESULTS

A. Experimental Setup and Evaluation

1) Implementation Details: Our experimental framework for facial region segmentation utilizes a well-structured pipeline. The dataset, consisting of images with a resolution of 640×512 pixels, was pre-processed to extract face bounding boxes. These boxes were then cropped and resized to ensure consistent input dimensions for the model. The data was split into 3850 training images, 600 validation images, and 600 test images, spread across 126 subjects for training, 20 for validation, and 17 for testing. To improve generalization, data augmentation was applied exclusively to the training set. The augmentations included Gaussian Blur (with a kernel size of 5) to simulate varying lighting conditions, Vertical Flip (VFlip) to account for pose variability, and Horizontal Flip (HFlip) through landmark mirroring to enhance the model's robustness to changes in facial orientation.

The models were trained on a NVIDIA RTX 6000 Ada Generation GPU. The training process employed the Adam optimizer with a learning rate of 0.0001. A ReduceLROnPlateau learning rate scheduler was used with a factor of 0.5 and a patience of 5 epochs, reducing the learning rate when the validation loss plateaus. Early stopping was applied with a patience of 10 epochs to prevent overfitting, ensuring that the model would halt training once the model performance stops improving.

2) Evaluation Metrics: To evaluate the performance of our facial region segmentation model, we focus on several key metrics. Mean Intersection over Union (mIoU) is used to measure the overlap between the predicted regions and the ground truth, providing an indication of the model's segmentation accuracy. The F1 score (Dice coefficient) is also employed to balance Precision and Recall, giving an overall assessment of segmentation quality. Precision reflects how accurately the model detects facial regions without misclassifying background pixels, while Recall measures the model's ability to capture all relevant facial regions, ensuring no important features are overlooked. Together, these metrics offer a comprehensive view of the model's performance in terms of both accuracy and completeness.

B. Performance Comparison

In this section, we compare the performance of our proposed model, PyraSegNet, against several state-of-the-art models for facial region segmentation. In addition, we compare different variants of our model architecture.

1) Comparison with SOTA Methods: We compared the performance of PyraSegNet, which uses InceptionV4 as the feature extractor, against several state-of-the-art models for thermal facial region segmentation, including DeepLabV3+, FPN, U-Net, U-Net++, PAN, MAnet, PSPNet and LinkNet. The results presented in Table II show that PyraSegNet outperforms other methods consistently across most key evaluation metrics for results obtained without using landmarks as input.

PyraSegNet achieved a mIoU of 92.87% and an F1 score of 96.30%, demonstrating its ability to accurately segment facial regions with minimal overlap errors between predicted and ground truth masks. Additionally, it achieved a Precision of 95.59%, indicating its effectiveness at reducing false positives. The Recall of 97.04% suggests that PyraSegNet is effective at detecting nearly all relevant facial regions, ensuring comprehensive segmentation.

Other models, such as DeepLabV3+ and U-Net, showed lower performance on these metrics. DeepLabV3+ achieved a mIoU of 80.16% and an F1 score of 88.87%, while U-Net had a mIoU of 74.72% and an F1 score of 85.47%. While U-Net++ demonstrated a high Recall of 98.86%, its low Precision of 47.34% indicates it is not capable of balancing between capturing all regions and minimizing false positives. The comparison highlights PyraSegNet's balanced performance, achieving both high Precision and Recall, making it an effective model for thermal facial region segmentation.

The relatively lower performance of other models, especially in terms of Precision and mIoU, suggests that PyraSegNet's architecture, specifically its use of the InceptionV4 feature extractor, allows it to better handle the complexities of thermal imagery. The InceptionV4 network potentially aids in capturing multi-scale features, which is important for accurately detecting the facial regions, even under challenging conditions like low contrast and noise in thermal images.

2) Impact of Landmark Masks: We assessed the effect of incorporating landmark masks alongside thermal images as input. The results, shown in Table III, demonstrate the improvement in performance with the inclusion of landmark masks for all methods. For PyraSegNet, adding landmark

TABLE II: Performance Comparison of PyraSegNet and Other Models with DiceLoss

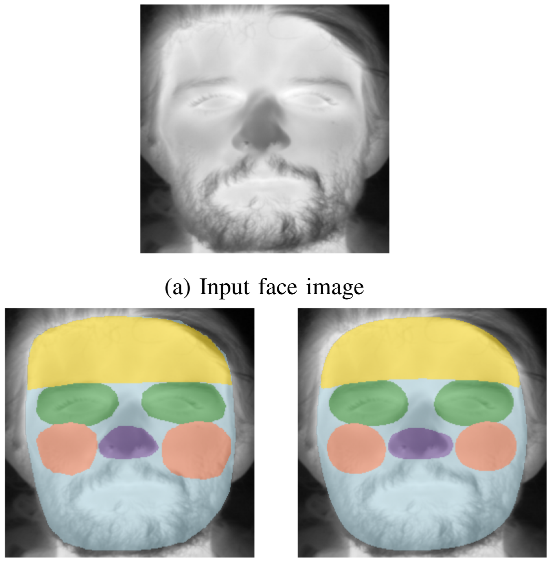
Model	mIoU(%)	F1	Precision	Recall
PyraSegNet	**92.87**	**96.30**	**95.59**	97.04
DeepLabV3+	80.16	88.87	85.99	92.10
FPN	76.68	86.73	87.09	86.48
U-Net	74.72	85.47	76.28	**97.29**
MAnet	74.15	85.08	81.45	89.11
PAN	71.54	83.32	87.88	79.36
LinkNet	68.10	80.96	70.39	95.29
PSPNet	53.44	69.64	79.80	61.84
U-Net++	47.09	63.98	47.34	98.86

masks improved the mIoU from 92.87% to 93.14% and the F1 score from 96.30% to 96.45%, suggesting that the additional landmarks provide more precise localization of facial regions. Precision increased from 95.59% to 96.30%, while Recall decreased slightly from 97.04% to 96.60%, though this slight reduction in Recall does not undermine the overall improvement in segmentation quality.

Other models, such as DeepLabV3+ and U-Net, also present significant performance improvements with the addition of landmark masks. DeepLabV3+ achieved an increase in mIoU from 80.16% to 89.92%, and its F1 score improved from 88.87% to 94.69%. U-Net also benefited, with its mIoU improving from 74.72% to 88.08%, and its F1 score rising from 85.47% to 93.66%. These improvements suggest that landmark masks assist in localizing facial features more effectively, helping the models to better capture the facial regions. These models depend on texture and intensity variations, whereas thermal images represent temperature, making it more difficult to differentiate regions without additional structural guidance from landmarks. In summary, adding landmark masks enhances segmentation performance by providing additional spatial and structural context and helping the model more accurately localize and segment facial regions, which appear less distinct due to thermal noise or low contrast. While the inclusion of landmarks generally results in better segmentation, the extent of the improvement varies depending on the model's ability to leverage this additional information.

Fig. 2 shows the segmented regions predicted by PyraSegNet for a typical thermal face image. Subfigure (a) shows the input image, (b) shows the ground-truth segmentation, and (c) shows the predicted regions, which align closely with the ground-truth.

In the following figure, Fig. 3, we show a case where occlusions, specifically glasses, are present on the face. The image in this case contains the predicted eyes and nose segmented regions overlaid on the input image. Despite the presence of glasses, which partially obscure the eyes and nose, the model continues to segment the facial regions with high accuracy. The overlay of the predicted regions demonstrates the model's resilience to the occlusions caused by glasses, maintaining an effective segmentation of the face. It should be noted that in the training set there were images with glasses annotated for the eyes region to aid the model in detecting those instances.



(a) Input face image

(b) The ground truth (c) The predicted regions

Fig. 2: Comparison of the input face image, the ground truth segmentation regions, and the predicted segmentations.

Fig. 3: Predicted Regions in Presence of Occlusions

The results in Table IV demonstrate that while PyraSegNet achieves slightly better performance compared to other models, several architectures, such as DeepLabV3+ and FPN perform remarkably well on grayscale images, achieving mIoU scores of 92.76% and 92.71%, respectively, with F1-scores close to PyraSegNet's performance. This indicates that these models, originally designed for RGB inputs, can effectively generalize to grayscale images. However, when comparing these results to those in Tables II and III, it is evident that these models struggled on raw thermal images, with mIoU scores dropping significantly—for instance, DeepLabV3+ achieved only 80.16%, and FPN scored 76.68%. This demonstrates that PyraSegNet consistently outperforms other methods for thermal images and is slightly better when it comes to grayscale images. The hybrid encoder-decoder structure, along with its tailored feature extraction approach, enables PyraSegNet to better capture the distinct characteristics of thermal images, such as lower contrast and higher noise, while still performing well on grayscale images. Furthermore, the table confirms the previous finding that using landmarks as input consistently improves results across all models.

TABLE III: Performance Comparison of PyraSegNet and Other Models with DiceLoss (Including Landmarks as Input)

Model	mIoU(%)	F1	Precision	Recall
PyraSegNet	**93.14**	**96.45**	**96.30**	**96.60**
DeepLabV3+	89.92	94.69	94.89	94.52
FPN	89.51	94.46	93.41	95.56
U-Net	88.08	93.66	94.31	93.05
MAnet	84.60	91.65	88.38	95.20
PAN	80.57	89.23	91.01	87.56
LinkNet	79.86	88.79	90.69	87.02
PSPNet	55.31	71.20	63.96	80.38
U-Net++	52.10	68.49	74.53	63.40

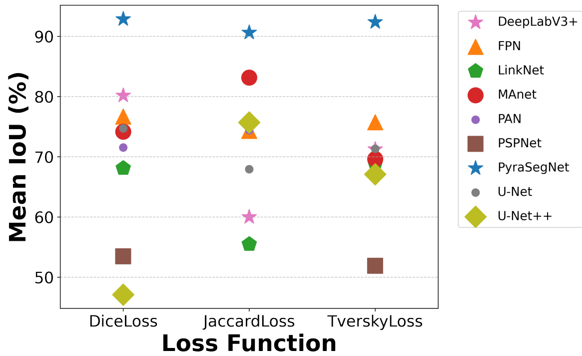
TABLE IV: Performance Comparison of PyraSegNet and Other Models with DiceLoss on Gray scale Images

Model	Including Landmarks		Excluding Landmarks	
	mIoU (%)	F1	mIoU (%)	F1
PyraSegNet	**92.89**	**96.31**	**92.73**	**96.22**
DeepLabV3+	92.76	96.24	92.47	96.08
FPN	92.71	96.22	92.15	95.91
U-Net	92.64	96.18	91.61	95.59
MAnet	92.26	95.97	91.58	95.57
PAN	92.70	96.21	92.43	96.06
LinkNet	92.39	96.04	91.83	95.71
PSPNet	91.78	95.70	90.65	95.05
U-Net++	92.69	96.20	92.07	95.85

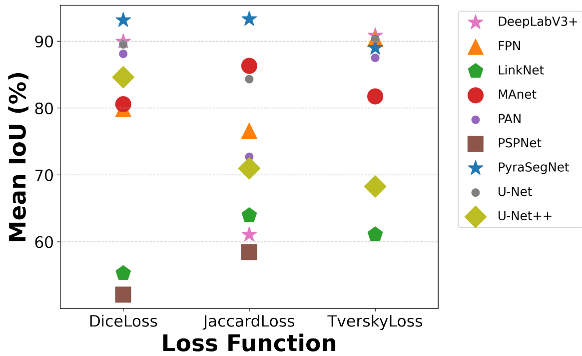
3) Loss Function Comparison: This section evaluates the impact of loss functions on segmentation performance, focusing on DiceLoss, TverskyLoss, and JaccardLoss. The results, depicted in Figure 4, report the mean Jaccard index as a percentage to emphasize differences between the loss functions.

As shown in Fig.4a, where thermal images without landmarks were used as input, DiceLoss achieved the highest performance with PyraSegNet, reaching a mean Jaccard index of 92.87%. TverskyLoss followed closely at 92.36%, while JaccardLoss had a slightly lower value of 90.62%. These results indicate that both DiceLoss and TverskyLoss are highly effective for thermal image segmentation, potentially due to their ability to handle segmented region size imbalances in foreground and background regions, which is a common characteristic of thermal images.

In Fig.4b, where facial landmarks were integrated with thermal images, performance improved across most loss functions. PyraSegNet with JaccardLoss achieved the highest mean Jaccard index of 93.30%, followed closely by DiceLoss at 93.14%, while TverskyLoss achieved 88.97%. The addition of facial landmarks provided valuable spatial context, leading to improved segmentation results for JaccardLoss and DiceLoss, although TverskyLoss showed lower performance with PyraSegNet. The improvement in JaccardLoss can be explained by how the loss function works: it measures the overlap between predicted and true regions. By adding facial landmarks, the model was able to align predicted boundaries with the ground truth, which increased the overlap and reduced the total area being considered. These findings highlight the importance of selecting the proper loss function and show that using complementary data, such as facial landmarks, can improve segmentation performance by providing additional spatial information.



(a) Mean IoU Comparison Across Loss Functions



(b) Mean IoU Comparison Across Loss Functions Including Landmarks

Fig. 4: Comparative Analysis of Loss Functions with and without Landmarks

C. Ablation Study

The ablation study investigates the contributions of different components of PyraSegNet to its performance. Specifically, the study examines the effectiveness of the pyramid structure, the feature extractor, and various backbone architectures. The results are summarized in Fig. 5, which highlights the impact of these components.

1) Model Performance Across Backbones: To investigate the impact of various backbone architectures on segmentation quality, we compared PyraSegNet with ResNet, EfficientNet, and SE-ResNet backbones. PyraSegNet's default backbone (InceptionV4) achieved the best performance, with an F1 score of 96.54% and mIoU of 93.14%. This demonstrates the advantages of InceptionV4's hybrid architecture, which is specifically designed to capture features through multi-scale processing. Its ability to process multiple scale features simultaneously is especially useful for thermal segmentation, which requires accurate identification of diverse spatial patterns.

Unlike InceptionV4, other backbones such as ResNet and SE-ResNet employ sequential convolutional layers with fixed kernel sizes, limiting their ability to efficiently capture multi-scale details. As a result, they demonstrated lower performance, with ResNet achieving an F1 score of 95.58% and

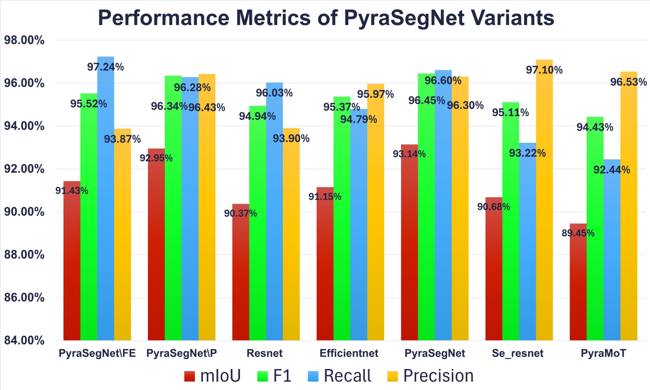


Fig. 5: Analysis of Pyramid Structure and Feature Extractor Effectiveness Across Backbones. PyraSegNet\P : PyraSegNet without Pyramid, PyraSegNet\FE : PyraSegNet without Feature Extractor

SE-ResNet scoring 93.96%. These backbones struggled with recall, particularly in regions requiring fine-grained feature extraction.

EfficientNet, while competitive, achieved an F1 score of 95.02% and an mIoU of 91.15%. Its scalable architecture optimized the parameter space well, however, it lacked the flexibility of InceptionV4's parallel convolutional operations, leading to decreased segmentation accuracy. Finally, PyraMoT [28], which utilized MobileNetV2 as the backbone, achieved the lowest scores (F1: 93.22%, mIoU: 89.45%) in this segmentation task. While PyraMoT has demonstrated strong performance in precise, pixel-level tasks such as facial landmark detection, its limitations become apparent in dense segmentation tasks. Thermal facial segmentation requires not only high precision in localized areas but also robust contextual understanding across the entire facial region. The representational depth of MobileNetV2, optimized for lightweight and efficient processing, is less suited for capturing the dense, high-dimensional feature maps required for segmentation tasks with complex boundaries.

2) Model Performance Without Pyramid: To evaluate the contribution of the pyramid structure, we removed it from PyraSegNet (PyraSegNet\P). While the model performed reasonably well, achieving an F1 score of 96.30% and an mIoU of 92.95%, there was a decline in recall (96.28% vs. 96.60%) and precision (96.43% vs. 96.30%) compared to the full model. This drop highlights the pyramid structure's role in effectively aggregating multi-scale features.

The pyramid structure fuses features from different resolutions, enabling the model to capture both global and local patterns in thermal images. Excluding it can lead to under-segmented or overly generalized predictions, showing that the pyramid is essential for enhancing PyraSegNet's robustness to subtle variations across facial regions.

3) Model Performance Without Feature Extractor: The feature extractor was removed to assess its impact on PyraSegNet (PyraSegNet\FE). This led to a significant performance drop, with the F1 score decreasing from 96.54% to 95.52% and the mIoU from 93.14% to 91.43%. Precision and recall were particularly affected, with precision falling to 93.87% from 96.30%, underscoring the feature extractor's importance in identifying fine-grained facial features.

The feature extractor plays a crucial role in isolating and amplifying relevant thermal patterns from raw input data. Its absence led to under-segmentation in critical areas, as the model failed to capture detailed spatial dependencies. Overall, the feature extractor is vital for enabling the model to process high-dimensional, thermal-specific features and produce reliable segmentation outputs.

VI. Conclusion

Our research represents a significant step forward in advancing the field of thermal facial region segmentation, a domain increasingly recognized for its importance in applications such as physiological monitoring, stress detection, and human-computer interaction. A central contribution of this work is the introduction of the TFR dataset—a comprehensive collection of 5050 annotated thermal images from 163 diverse subjects. This dataset addresses a critical gap in the field, providing a much-needed resource to support robust training and evaluation of segmentation models. Its breadth and quality offer a benchmark for future studies, paving the way for improvements in accuracy and reliability.

Building on this foundation, we developed PyraSegNet, a novel deep learning architecture tailored specifically for thermal imaging challenges. By integrating advanced multi-scale feature aggregation through its pyramid structure and leveraging a custom feature extractor, PyraSegNet has demonstrated exceptional performance. It surpasses established state-of-the-art models such as DeepLabV3+ and U-Net in accurately segmenting key facial regions, achieving an impressive mean Intersection over Union (mIoU) of 93.14% and an F1 score of 96.45% when utilizing landmarks. These results underscore PyraSegNet's capacity to handle the nuanced and often complex characteristics of thermal data.

Our analysis further highlights the critical role of methodological choices in achieving such advancements. Careful consideration of loss functions revealed that DiceLoss and TverskyLoss are particularly effective for thermal facial region segmentation tasks, especially when combined with facial landmark information. This approach consistently improved model accuracy across all tested architectures, emphasizing the interaction between data, architecture, and optimization strategies. Additionally, the ablation study reaffirmed the importance of PyraSegNet's core components—the pyramid structure for robust multi-scale feature representation and the feature extractor for detailed, fine-grained segmentation.

The implications of these findings extend beyond technical metrics. They demonstrate the necessity of domain-specific, high-quality datasets, such as TFR to address the unique challenges posed by thermal imaging. Furthermore, they validate the importance of designing specialized architectures, such as PyraSegNet that are explicitly tailored to such challenges.

VII. Ethical Impact Statement

This research presents PyraSegNet, a deep learning framework for thermal facial region segmentation, and introduces the TFR dataset, a large annotated thermal facial dataset. While the work has valuable applications in healthcare, security, and human-computer interaction, it also raises ethical concerns related to data privacy, potential misuse, and fairness in biometric technologies.

The primary ethical consideration involves the use of biometric thermal facial data, which, while less intrusive than visible-spectrum facial imaging, still presents potential privacy risks. Even though thermal imaging does not capture traditional facial features like skin tone, it remains possible to derive personally identifiable patterns. Additionally, biases in facial analysis models have historically led to disparities in performance across demographic groups, and it is crucial to consider these risks when developing segmentation models for human-centered applications.

To mitigate these risks, this research adheres to ethical data collection practices, including institutional review board (IRB) approval and informed consent from participants. The dataset is anonymized by excluding personally identifiable metadata, and the study does not involve facial recognition or identity verification. Moreover, the dataset includes a diverse range of subjects, ensuring variability in age, gender, and occlusions, such as glasses, to improve fairness and generalizability in segmentation performance.

The potential benefits of this work outweigh its risks, particularly in advancing non-contact physiological monitoring, stress detection, and human behavior analysis in healthcare and human-computer interaction settings. Unlike visible-spectrum imaging, thermal imaging provides a privacy-preserving alternative that reduces bias related to skin tone and lighting conditions. By promoting transparency in our methodology and acknowledging ethical considerations, this research aims to contribute responsibly to the field while encouraging future studies to incorporate fairness assessments and interpretability techniques for ethical AI deployment.

References

[1] A. S. Ahuja. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, 7:e7702, 2019.

[2] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H. A. Badreldin, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, 23(1):689, 2023.

[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] R. Azad, M. Heidary, K. Yilmaz, M. Hüttemann, S. Karimijafarbigloo, Y. Wu, A. Schmeink, and D. Merhof. Loss functions in the era of semantic segmentation: A survey and outlook. *arXiv preprint arXiv:2312.05391*, 2023.

[5] P. Buddharaju, I. T. Pavlidis, P. Tsiamyrtzis, and M. Bazakos. Physiology-based face recognition in the thermal infrared spectrum. *IEEE transactions on pattern analysis and machine intelligence*, 29(4):613–626, 2007.

[6] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier. Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *2017 Seventh international conference on affective computing and intelligent interaction (acii)*, pages 456–463. IEEE, 2017.

[7] C. B. Cross, J. A. Skipper, and D. T. Petkie. Thermal imaging to detect physiological indicators of stress in humans. In *Thermosense: thermal infrared applications XXXV*, volume 8705, pages 141–155. SPIE, 2013.

[8] A. Duarte, L. Carrão, M. Espanha, T. Viana, D. Freitas, P. Bártolo, P. Faria, and H. Almeida. Segmentation algorithms for thermal images. *Procedia Technology*, 16:1560–1569, 2014.

[9] P. Ekman and W. V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*, volume 10. Ishk, 2003.

[10] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[11] H. Ito, K. Oiwa, and A. Nozawa. Image segmentation-based face tracking on thermal images for automatic estimation of psychophysiological states using facial skin temperature distribution. *Journal of Bioinformatics and Neuroscience (JBINS)*, 4(1):142–146, 2018.

[12] D. Jaramillo-Quintanar, J. K. Gomez-Reyes, L. A. Morales-Hernandez, B. Dominguez-Trejo, D. A. Rodriguez-Medina, and I. A. Cruz-Albarran. Automatic segmentation of facial regions of interest and stress detection using machine learning. *Sensors*, 24(1):152, 2023.

[13] J. Jeoung, S. Jung, T. Hong, M. Lee, and C. Koo. Thermal comfort prediction based on automated extraction of skin temperature of face component on thermal image. *Energy and Buildings*, 298:113495, 2023.

[14] J. Joshi, N. Bianchi-Berthouze, and Y. Cho. Self-adversarial multi-scale contrastive learning for semantic segmentation of thermal facial images. *arXiv preprint arXiv:2209.10700*, 2022.

[15] E. Khvedchenya. Pytorch toolbelt. `https://github.com/BloodAxe/pytorch-toolbelt`, 2019.

[16] Z. Kütük and G. Algan. Semantic segmentation for thermal images: A comparative survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 286–295, 2022.

[17] 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.

[18] 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.

[19] J. Kwon, O. Kwon, K. T. Oh, J. Kim, and S. K. Yoo. Breathing-associated facial region segmentation for thermal camera-based indirect breathing monitoring. *IEEE Journal of Translational Engineering in Health and Medicine*, 2023.

[20] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2020.

[21] 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.

[22] P.-E. Martin, G. Kachel, N. Wieg, J. Eckert, and D. B. Haun. Apeti: A thermal image dataset for face and nose segmentation with apes. *Signals*, 5(1), 2024.

[23] E. Mostafa, R. Hammoud, A. Ali, and A. Farag. Face recognition in low resolution thermal images. *Computer Vision and Image Understanding*, 117(12):1689–1694, 2013.

[24] D. Müller, A. Ehlen, and B. Valeske. Convolutional neural networks for semantic segmentation as a tool for multiclass face analysis in thermal infrared. *Journal of nondestructive evaluation*, 40(1):9, 2021.

[25] T. Olugbade, Y. Cho, Z. Morgan, M. Abd El Ghani, and N. Bianchi-Berthouze. Toward intelligent car comfort sensing: New dataset and analysis of annotated physiological metrics. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

[26] K. Panetta, K. S. Kamath, S. Rajeev, and S. S. Agaian. Ftnet: Feature transverse network for thermal image semantic segmentation. *IEEE Access*, 9:145212–145227, 2021.

[27] 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.

[28] K. Riani, S. Sharak, and M. Abouelenien. Pyramot: A novel framework for enhanced facial thermal landmarks detection. In *2024 IEEE 18th International Conference on Automatic Face and Gesture*

Recognition (FG), pages 01–09. IEEE, 2024.
[29] 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.
[30] F. J. Rodriguez-Lozano, F. León-García, M. Ruiz de Adana, J. M. Palomares, and J. Olivares. Non-invasive forehead segmentation in thermographic imaging. *Sensors*, 19(19):4096, 2019.
[31] S. S. M. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.
[32] V. Selvaraju, N. Spicher, J. Wang, N. Ganapathy, J. M. Warnecke, S. Leonhardt, R. Swaminathan, and T. M. Deserno. Continuous monitoring of vital signs using cameras: A systematic review. *Sensors*, 22(11):4097, 2022.
[33] Supervisely. Supervisely computer vision platform. `https://supervisely.com`, jul 2023. visited on 2023-07-20.
[34] A. A. Taha and A. Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15:1–28, 2015.
[35] V. A. Weidlich. Thermal infrared face recognition. *Cureus*, 13(3), 2021.
[36] H. Xiong, W. Cai, and Q. Liu. Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene. *Infrared Physics & Technology*, 113:103628, 2021.