MULTIDISCIPLINARY SCIENCE JOURNAL

RESEARCH ARTICLE
Published Online: July 31, 2024
https://doi.org/10.31893/multiscience.2025065

Fully automated real-time approach for human temperature prediction and COVID-19 detection-based thermal skin face extraction using deep semantic segmentation

Adil Al-Azzawi[a] ✉ | **Mohammed Khaleel Hussein**[b]

[a]University of Diyala, College of Science, Computer Science Depratmenet, Diyala, Iraq.
[b]Ministry of Higher Education and Scientific Research, Private Higher Education Directorate, Baghdad, Iraq.

Abstract The COVID-19 pandemic caused by the coronavirus has affected every country. The World Health Organization has declared it a global health emergency by 2020. Early intervention provides a more effective cure and relieves patients of unnecessary intensive care from hospitals and other healthcare facilities. However, that requires an accurate method for early disease detection. COVID-19 illness may eventually infect people, making it a severe condition. Research shows that COVID-19 pneumonia shares many clinical features with other forms of pneumonia, suggesting that COVID-19 can be a serious condition for those affected. The main symptom of the disease is high body temperature. Noncontact thermal imaging is one of the most important methods for measuring body temperature, although it has some performance limitations. Many of these thermal cameras are privately owned, not widely available, and are often owned by for-profit organizations. In this paper, we propose a real-time method for human temperature prediction and tracking based on facial skin temperature extraction for accurate COVID-19 case detection. The main idea of this paper is to use an automated deep-learning model for thermal skin temperature extraction. The optimization is based on using the deep semantic segmentation approach for fully automated thermal skin binary mask prediction using the thermal video only. The predicted mask is projected onto the original thermal video to extract the thermal skin mask, which is used to calculate the average thermal skin face temperature. The proposed deep semantic segmentation model is trained on the thermal skin binary mask dataset that is generated by using our first model a fully automated unsupervised learning approach for camera calibration and thermal skin binary mask extraction. The experimental results show that the proposed approach can automatically predict better binary skin-mak than our first model which leads to a very high-efficiency human temperature tracking using thermal videos only compared with the Ground Truth using the Speaking Faces thermal datasets.

Keywords: supervised learning approach, deep learning, semantic segmentation, COVID-19, human temperature tracking, machine learning

1. Introduction

The World Health Organization (Leong et al., 2018) declared the disease affecting every country in the world an epidemic on March 11, 2020. Hospitals and facilities offer the possibility of survival without the need for intensive care when discharged early in disease progression, as this makes it more responsive to medical interventions. According to research, COVID-19 pneumonia shares many clinical signs and symptoms with other forms of pneumonia, but COVID-19 patients experience a greater loss of liver function than non- COVID-19 patients. This indicates that COVID-19 is a disease with great potential to spread to humans (Toniato et al., 2020). One of the symptoms of this disease is a high body temperature. Thus, one of the best methods for patient diagnosis is thermal examination (Makino et al., 2023; Abdrakhmanova et al., 2021). Non-contact thermal imaging of humans is one of the most important methods for obtaining human temperature, though there is a dearth of thermal video datasets. And if they exist, most of them are private and not available for use, and some belong to for-profit institutions. A thermal detector, as shown in Figure 1, needed a cooling method such as a Stirling refrigerator, argon gas, or nitrogen (Krišto et al., 2020). The perfect arrangement for a thermal imaging detector is shown in Figure 1 (a) (Vollmer, 2020). The first attempt to determine the quantity of infrared radiation from human skin was carried out 78 years after the discovery of "dark heat," even though infrared radiation was first identified in 1800. The first skin temperatures based on radiometry were computed in 1921 and physically measured in 1923. The development of infrared imaging as a military instrument followed, and it wasn't until the late 1950s that it was declassified for civilian use (Yeganeh et al., 2012).

 Received: May 23, 2024 | Accepted: July 11, 2024

There are two primary models for determining body temperature (Rottmann et al., 2023). An infrared thermometer, which is used to measure temperature over a short distance, and a thermography camera, is used to remotely measure an object's temperature. The inspector must be near pedestrians, even though the infrared short-distance body temperature meter is simple to use and inexpensive. When the pedestrian is infected with the virus, it is unsafe since it is most likely appreas to separates the virous. Thermal cameras that can monitor body temperature from a distance have been created to enable a secure assessment of body temperature (Su et al., 2020).

(a) (b) (c)

Figure 1 Block diagram demonstrates the proper thermal imaging room setup [7], (a) sample form the thermal video, (b) sample form the RGB video.

Recently, Machine Learning (ML) and Deep Learning (DL) have transformed computer vision, offering powerful solutions to countless complex problems. These technologies in medical imaging accurate the diagnosis of conditions such as cancer, heart disease, and neurological disorders through accurate image analysis. A real-time fully automated supervised learning approach for human temperature extraction and tracking is proposed in this paper. The proposed system is based on the first model, which is a fully automated unsupervised learning approach for human temperature extraction based on thermal skin face extraction (Al-Azzawi, A.,2024). The main purpose of this paper is to design a deep semantic segmentation model for fully automated thermal skin face temperature extraction. It intends to be used for early detection of influenza, the SRV virus, and developing coronavirus disease (COVID-19) and the mutant Omicron. The temperature of the human body, as determined by the data gathered from Speaking Faces (Abdrakhmanova et al., 2021), may be used by this model to identify and reliably categorize whether the patient has certain illnesses or not.

2. Related Works

Recently, many reseracher have proposed different approaches for human temperature extraction, tracking and prediction using the machine learning. A summary of these recent reseraches are discussed below:

Rottmann et al. (Rottmann et al., 2021) presented an approach to measure the task of detecting naming errors by dropping labels from a Cityscapes dataset, as well as from a dataset extracted from the CARLA driving simulator, where in the latter case they have the labels under control. Their experiments showed that their proposed approach is able to detect the vast majority of naming errors while controlling the number of misnaming errors detected. Moreover, their method was applied to semantic segmentation datasets frequently used by the computer vision community, and they provided a set of naming errors along with sample statistics. Diakogiannis et al. (Diakogiannis et al., 2020) proposed a reliable performance outcome framework for the semantic segmentation task of high-resolution single-time aerial photographs. Their framework consisted of a new deep learning architecture, ResUNet-a, and a new loss function based on dice loss. ResUNet-a backbone UNet was used for encoder/decoder as well as residual connections, atrial gyrus clustering, hierarchical scene analysis, and multitask inference. ResUNet-a sequentially inferred the boundaries of the objects, the distance for the segmentation mask, and the segmentation mask, and reconstructed the color input. Each of the tasks is conditioned by the inference of the previous tasks, and thus a conditional relationship is established between the different tasks. They analyzed the performance of several flavors of generalized dice loss for semantic segmentation and introduced a new variable loss function for semantic segmentation for objects that have excellent affinity properties and behave well even in the presence of highly unbalanced classes. They evaluated the performance of their modeling framework on the ISPRS 2D Potsdam dataset. The results showed them a top performer with an average F1 score of 92.9% across all classes for their best model. Shenyu et al. (Ji et al., 2023) proposed a deep learning-based soft edge optimized network to extract the semantic labels of each sculpted element from multichannel images projected from 3D point clouds of the engravings. They were able to clearly extract the soft edges in the engravings using a new blurring-based edge rendering method. By mapping the extracted semantic markers into 3D points of the relief data, the proposed method provides comprehensive 3D semantic segmentation results for Borobudur inscriptions. Zhou et al. (Zhou et al., 2023) proposed a segmentation model and transfer learning protocol where the segmentation model adopted a point-reduction structure.

3. Proposed System

This paper proposes a real-time fully automated supervised learning approach for human temperature tracking based-deep semantic segmentation model. The main flowchart of the proposed system is shown in the following Figure 2.

Figure 2 General flowchart of the proposed system.

As shown in Figure 2, the proposed system is designed based on two stages. The first stage is the video calibration and thermal skin binary masks extraction. In this stage, our first model (Al-Azzawi, A., 2024) which is a fully automated unsupervised learning approach for thermal camera calibration and face skin temperature extraction has been used to automated thermal and RGB videos calibration, and automated thermal skin binary mask generation. In this model, an optimized unsupervised learning approach has been developed and implemented.

The first stage of the proposed system has been developed and implemented based on five steps. The first step is the video preprocessing step, where each video is converted into a sequence of frames. The second step is the first videos calipration step, where each frame (from the RGB and thermal videos) is calibrated using image scaling and padding. After each step, both frames are measured using diferent similarity metrics such as correlation and SSIM. The fourth step is the second videos calipration step, in which different image registration algorthims have been used for an accurate calibration results. The final step is the automated thermal skin face binary maskextraction. In this step, an automated unsupervised learning approach using k-mean clustering is implemented for an automated RGB face skin mask extraction. In this case, each thermal video frame has associate binary mask that comes for the optimized k-means algorthim. The second stage is an automated thermal skin face mask prediction using a Deep Semantic Segmentation model. After the thermal binary mask have been generated using the sophisticated unsupervised learning approach-based image registration and optimized k-means clustering. These skin binart face masks are used to train our deep semantic segmentation model. Then, the trained deep semantic segmentaion model is used to automatically predict the thermal face skin mask directly from the thermal videos without the calibration

step. Finally, the predicted binary masks of the thermal skin face are used to extract the average of the human temperature and track individual cases.

3.1. Stage 1: Fully automated unsupervised learning approach for video calibration and face skin binary mask generation

The first stage of the proposed system is based on our first model (Al-Azzawi, A., 2024). In this stage, face skin binary masks are automatically generated for each frame in the thermal video of the dataset. During the first stage, a sophisticated unsupervised learning approach-based image registration and optimized k-means clustering is designed to calibrate the RGB and thermal video first and extract the face skin masks from the calibrated RGB video second.

3.1.1. First step of video calibration

Speaking Faces thermal video dataset (Zhou et al., 2020) is used in this paper, which was captured manually by operating the camera to cover the human faces based on nine positions. Both RGB and thermal videos have different dimensions and views (camera pose and location). For this reason, video calibration is proposed in this stage to align both frames (RGB and Thermal) in both cameras. In multi-camera systems, calibration also ensures that the camera positions and orientations are described in the same coordinate system. In this stage, each RGB video frame is resized by (0.77). The number has been tunned from the gange of [0.1-0.99]. Then, the size of the thermal frames is zero padded to ensure the best matching results. Two metrics are used to measure the similarity. The first metric is the Structure Similarity Information Matching (SSIM) (Setiadi, 2021) see Equation (1).

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (1)$$

Where μ_x is the pixel sample mean of x; μ_y the pixel sample mean of y; σ_x^2 the variance of x; σ_y^2 the variance of y; $\sigma_{xy}$the covariance of x and y; $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$ two variables to stabilize the division with weak denominator; L the dynamic range of the pixel-values (typically this is $2^{\#\, bits\ per\ pixel}$-1); k_1= 0.01 and k_2= 0.03 [8].

The second measure is a correlation. It allows us to understand how thermal and RGB frames relate to one another. The population correlation coefficient ($\rho x, \gamma$) of two random variables (X, Y) with expected values (μ_x, μ_y) and standard deviations (μ_x, μ_y) is given in Equations (2) (Yuan et al., 2021).

$$corr(x,\gamma) = \frac{E[(X-\mu_x)(Y-\mu_y)]}{(\sigma_x\sigma_y)}, \quad \text{if } \sigma_x\sigma_y > 0 \quad (2)$$

Where E denotes as the expected error, μ_x is the mean of the first image, and μ_y is the mean of the second image in which the pure signal for both the first and second images are calculated. σ_x, and σ_y are donated as the standrated deviation in which the amount of noise in both images are calculated respectively.

Figure 3 shows an example of the video calibration-based image scaling step. Figure 3 (a) illustrates the overlapping calibrated RGB and thermal image before the calibration step. Figure 3 (b) shows the thermal image after the calibration. Figure 3 (c) illustrates the overlapping of the calibration results.

(a) (b) (c)

Figure 3 An example of the first step of the video calibration results. (a) original thermal and RGB image projection before the first step of the image calibration, (b) Calibrated thermal image, (c) Calibrated thermal and RGB image projection.

3.1.2. Second step of video calibration

Image registration is used when multiple images need to be combined or overlaid to align the two images accurately. Image registration is a fundamental process in image processing and computer vision that involves aligning two or more images of the same scene or object taken from different perspectives, times, or with different sensors. Different image registration approaches have been used in our first model. The original RGB face frame is isolated to the R-channel, G-channel, and B-channel frames. Then, each of the RGB channels is registered using four different registration methods such as translation,

affine, rigid, and similarity transformation. Affine image registration includes different types of transformation such as translation, rotation, scaling, and shearing, as is given in Equation (3) (Fu et al., 2020; Haskins et al., 2020).

$$\begin{vmatrix} x_2 \\ y_2 \end{vmatrix} = A \times \begin{vmatrix} x_1 \\ y_1 \end{vmatrix} + B \tag{3}$$

Image translation employs either 2D (x, y) translation or 3D (x, y, z) translation as is given in Equation (4) [18].

$$X' = X + t_x \tag{4}$$

The pair (t_x, t_y) is referred to as the shift or translation vector. The column vectors may also be used to illustrate the Equations (5)-(7) (Pang et al., 2021).

$$P = [X][Y] \tag{5}$$

$$P' = [X'][Y'] \tag{6}$$

$$T = [t_x][t_y] \tag{7}$$

These equations can write it as:

$$P' = P + T \tag{8}$$

Translation and rotation make up the third approach, which is referred to as static (Rigid) transformation as it is given in the following Equation (9) (Pang et al., 2021).

$$T(v) = R\,v + t \tag{9}$$

Where R is an orthogonal transformation which can be written as the following Equation (10) and t is a vector indicating the translation of the origin (Pramanik et al., 2020).

$$R^T = R^{-1} \tag{10}$$

The fourth technique, which combines translation, rotation, and scaling, is known as (Similarity) transforming non-reflexive similarity as is given in Equation (11) (Ding et al., 2020).

$$B = P^{-1}AP \tag{11}$$

Figure 4 shows an example of the second step of the video registration stage based on different image registration approaches. Figure 4 (a) is the corresponding thermal frame to the RGB (h). Figure 4 (b), (c), and (d) are the R, G, and B channels respectively. Figure 4 (e) is the registration results using different transformations based on R-channel, Figure 4 (f)-(q) is the registration results using different transformations based on G-channel, and Figure 4 (g) is registration results using different transformations based on B-channel. Finally, Figure 4 (h) represents the reconstructed calibrated RGB frame.

Figure 4 RGB frame registration based thermal frame (frame are taken as an example form video frame no. (13), which was calibrated using four methods registration. (a) the thermal frames, (b), (c) and (d) the Red, Geen, and blue channels respectively, (e), (f), and (g) registration results using different transformations based on R-channel, (e)-(q) registration results using different transformations based on G-channel, (f)-(r) registration results using different transformations based on B-channel, (h) reconstructed calibrated RGB frame.

3.1.3. Thermal skin face mask extraction

In this stage, the thermal skin is extracted based on the RGB skin face frame in the GBR videos. K-means unsupervised image clustering (Sinaga et al., 2020) is used after some sophisticated modification is made to the original algorithm. The main contribution of the modification is the automated desired cluster selection. An automated approach for the desired cluster is designed and added to the original k-means algorithm to make the cluster that has the skin color from the calibrated RGB selection fully automated.

The k-means clustering algorithm is a popular unsupervised learning method for classifying data into k distinct groups based on similarity. For image clusters, this works by reducing the cumulative classes score (WCSS), which measures the differences in each cluster. The algorithm repeatedly assigns each data point to the nearest cluster center and updates the cluster centers. The basic similarities of k-means clustering is based assigning each data point x_i to the nearest cluster center μ_i. The assignment is based on the Euclidean distance as is given in the following Equation (12) (Sinaga et al., 2020).

$$c_i = \underset{j}{\operatorname{argmin}} \left\| x_i - \mu_j \right\|^2 \tag{12}$$

where c_i is the cluster assigned to data point x_i. Then, update the cluster centers μ_j to be the mean of all data points assigned to that cluster as is given in the following Equation (13)[21].

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \tag{13}$$

where C_j is the set of all data points assigned to cluster j and $|C_j|$ is the number of data points in cluster j. he objective is to minimize the within-cluster sum of squares (WCSS) is given in the following Equation (14) (Sinaga et al., 2020).

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|^2 \tag{14}$$

Figure 5 shows the three clusters for the calibrated RGB video frame using the k-means unsupervised learning algorithm. The best cluster from the k-means has been automatically selected based on using our modification for the k-means algorithm.

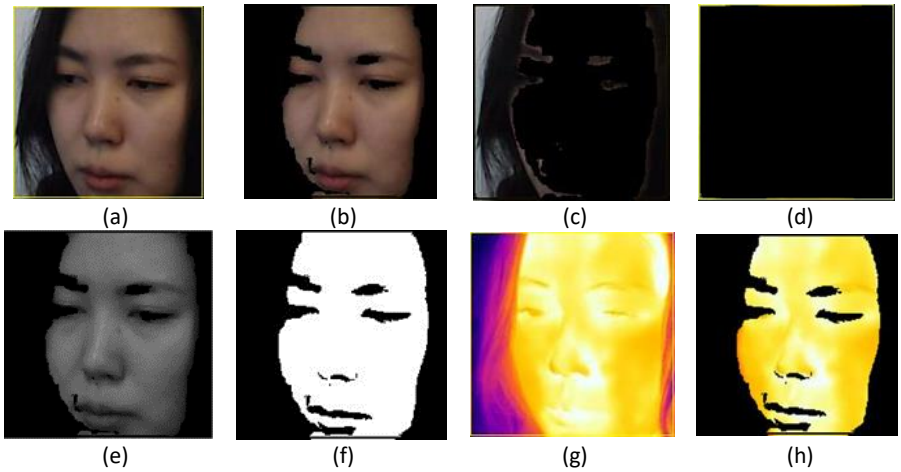


Figure 5 RGB and thermal face skin mask generation-based K-means image clustering using the calibrated RGB video frames-based, (a) calibrated RGB video frame, (b)-(d) cluster 1, cluster 2, and cluster 3 of the k-means respectively, (e) the grayscale version of the RGB face skin image, (f) cleaned binary mask of (e) after some morphological image operations. (g) corresponding thermal frame of (a), (h) unsupervised thermal face skin image.

The extracted RGB skin face video frame is converted to a grayscale image. Then, the gtayscale image is converted to a binary. Moreover, some postprocessing steps such as morphological image operations (image opening and closing) are used to a clean cluster (skin binary mask) as is shown in Figure 5. Figure 5 (a) calibrated RGB video frame, Figure 5 (b)-(d) is the k-means image clustering results respectively, Figure 5 (e) is the grayscale version of the RGB face skin image, Figure 5 (f) is the cleaned binary mask of Figure 5 (e) after some morphological image operations. Figure 5 (g) is the corresponding thermal frame of (a), finally, Figure 5 (h) is the final unsupervised thermal face skin image.

3.2. Stage 2: Fully automated supervised learning approach for thermal skin face mask prediction using deep semantic segmentation

A fully automated supervised learning approach for thermal skin face mask prediction using deep semantic segmentation uses the power of deep learning to accurately detect and classify thermal skin areas in face images. This approach uses deep semantic segmentation model for studying the shapes and complexities associated with hotspots of the trained binary skin mask. The model can create two accurate masks describing these regions. This fully automated process includes several steps such as data preprocessing, model training, and mask prediction. Standard metrics such as accuracy, precision, recall, and F1 scores are used to evaluate the efficiency of the model. This method is particularly valuable for applications that require real-time human temperature monitoring and tracking, as it provides a robust and flexible solution for the removal of hot facial skin masks, and for temperature measurement accuracy and reliability increase significantly in a range of practical situations. Various semantic segmentation methods mostly employ CNN as its architecture (Yuan et al., 2021). The architecture of a structure may be altered by the addition of new levels and features or by completely altering the architectural plan (Groschner et al., 2021). The second stage of the proposed system is based on a fully automated deep semantic segmentation for an accurate thermal skin face temperature extraction. The general diagram of the second stage of our proposal system is illustrated in the following Figure 6.

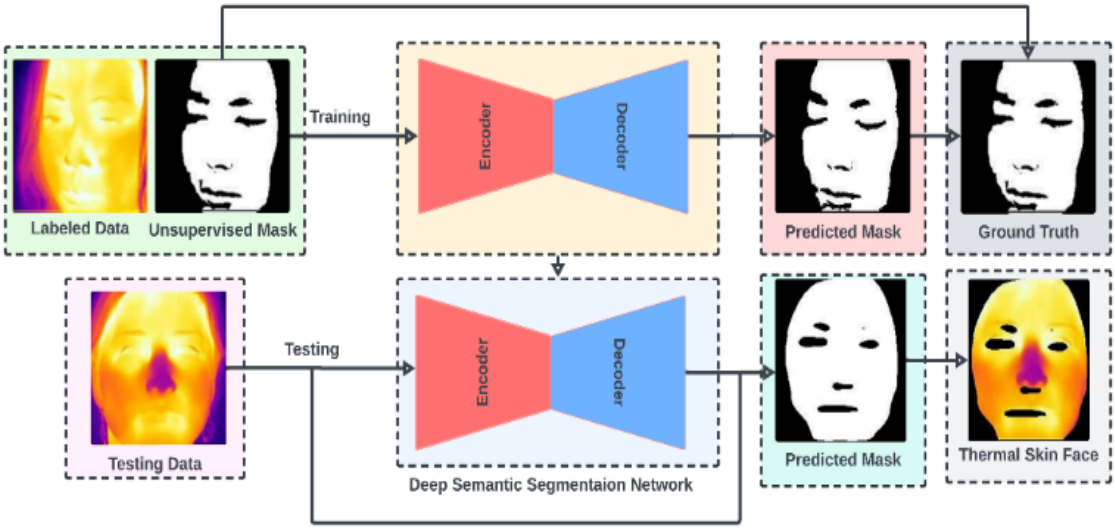


Figure 6 Fully automated supervised learning approach for thermal skin face binary mask Prediction using deep semantic segmentation network-based U-net network structure.

In this stage, the selected deep semantic segmentation network is trained on the generated binary masks that have been automatically generated using the first stage of our model (Azzawi, A., 2021). The binary face skin masks are constructed and predicted using a loss function, and the accuracy of the prediction is assessed using the dice coefficient to measure how similar the two faces are to one another. Based on the initial binary mask of the skin face, the second model that is being offered allows for the identification of the real human face. The following illustrates the main structure of the deep semantic segmentation model that is designed based on the U-Net architecture (Groschner et al., 2021). The U-Net framework is a widely used deep learning model specifically designed for semantic classification tasks. It was introduced in 2015 by Olaf Ronneberger, Philip Fischer and Thomas Brox, primarily for biomedical image segmentation. The architecture was nicknamed the "U-Net" because of its unique U-shaped structure, which consists of two main parts: a narrow channel (encoder) and a wide channel (decoder).

3.2.1. Encoder (Contract method)

The encoder part of the U-Net system, also known as the contract method, was responsible for sequentially downsampling and identifying the input image, extracting features at multiple levels of abstraction This method follows a typical convolutional neural network (CNN); structure afterwards. The encoder part in the U-Net architecture has the following parts.

Convolutional Layers: Each section of the contraction path consists of two 3x3 convolutional layers, each tracked by a rectified linear unit (ReLU) activation function as it is given in the following Equation (15) (Zhou et al., 2018).

$$f^{(l)} = \mathrm{ReLU}\left(W^{(l)} \times f^{(l-1)} + b^{(l)}\right) \quad (15)$$

where $f^{(l)}$ is the feature map at layer l, $W^{(l)}$ is the weight matrix, $b^{(l)}$ is the bias term, and $\times$ denotes the convolution operation.

Maximum Pooling Layer: A 2x2 maximum pooling level with stride 2 is applied after each section, in order to reduce the spatial resolution of the feature maps by half by using the following Equation (16) (Zhou et al., 2018).

$$f^{(l)} = \mathrm{MaxPool}\left(f^{(l-1)}\right) \quad (16)$$

Feature Channel: As spatial dimensions decrease, the number of feature channels doubles, enabling the network to recognize more complex features.

3.2.2. Decoder (Expansive Path)

The decoder part of the U-Net architecture, also known as the detailed path, consists of upsampling the feature maps and combines them with the corresponding feature maps from the encoder to achieve accurate localization and detailed segmentation .

Transposed Convolutional Layers: Each section in general begins with a transposed (or up-) convolutional layer that elevates the feature map, effectively increasing its spatial dimension by two times using the following Equation (17) (Zhou et al., 2018).

$$f^{(l)} = \mathrm{ReLU}\left(W_T^{(l)} \times f^{(l-1)} + b^{(l)}\right) \quad (17)$$

Concatenation: The unsampled feature map is merged with the corresponding feature map from the contract path. This step provides high-resolution features in the network that were lost during downsampling using the following Equation (18) (Zhou et al., 2018).

$$f^{(l)} = \mathrm{Concate}\left(f^{(l-1)}, f^{(l-1)'}\right) \quad (18)$$

where $f^{(l-1)'}$ is the corresponding feature map from the contracting path.

Convolutional Layers: Each concatenated feature map passes through two 3x3 convolutional layers, each of which follows a ReLU activation task, resolving the segmentation (Zhou et al., 2018).

Fully Connected Layer: The last part of the U-Net structure consists of a 1x1 convolutional layer that maps each feature vector to an arbitrary number of classes, generating a partition map (Zhou et al., 2018).

Overall, the U-Net architecture efficiently combines the strengths of both the encoder-decoder architecture and pass connections to attain modern overall performance in various photograph segmentation duties, inclusive of biomedical picture analysis, satellite picture segmentation, and more. Its versatility, simplicity, and effectiveness make it a popular desire for researchers and practitioners in the imaginative and prescient.

4. Experimental Results

In this paper, as the first experimental results, 900 video frames were extracted from both RGB and thermal videos Speaking faces thermal video dataset (Abdrakhmanova et al., 2021). The main description the whole dataset is illusrtaed below.

4.1. Speaking faces thermal video database

Speaking faces thermal video dataset (Abdrakhmanova et al., 2021) consists of a FLIR T540 thermal camera (resolution 464×348, wave band 7.5–14 μm, and 24◦ field of view) with an attached visual spectrum camera, a Logitech C920 Pro HD web-camera (resolution 1920×1080 and field of view 78◦), which has a built-in dual stereo microphone (44.1 kHz). A sample of the dataset is shown in Figure 7. The web camera was attached on the top of the thermal camera to facilitate the subsequent alignment of the image pairs. The duration of data collection for each position was set to 900 frames. Given the data collection rate of 28 fps for both cameras, this is equivalent to approximately 32 s of video, yielding on average 4.5 min of total video per subject. In total 142 persons make up the conversational interfaces, with 19 videos for different angles for each person, which are gender and ethnically balanced (in total 2,698 videos). The dataset takes over 13,000 occurrences of spoken instructions and over 45 hours of video (in total 3.7 million video frame pairs) are produced by having each subject be closely filmed while speaking approximately 100 English urgent phrases or orders.

4.2. Fully automated rgb skin face mask training dataset generation based unsupervised learning model

In terms of automated human temperature tracking based on thermal skin face extraction, the proposed deep semantic segmentation needs a labeled binary mask for each thermal frame in the thermal videos dataset. In this case, our previous model has been used to automatically generate the deep semantic segmentation dataset as is shown in the following Figure 8 which illustrates several examples of the human face RGB video frames (see Fig.8 (a), (d), (g), (j), (m), and (p)), the corresponding thermal video frames (see Figure 8 (b),(e),(h),(k),(n), and (q)), the face skin binary masks (see Fig.9 (c),(f),(i),(l),(o), and (r)).

Figure 7 Pairs of visual and thermal facial images of nine subjects taken from the predefined nine position, (1-3) RGB video samples, (1-9) Thermal video samples (Abdrakhmanova et al., 2021).

(a) (b) (c) (d) (e) (f)

Figure 8 Automated deep semantic training dataset generation (face skin binary mask) using fully automated unsupervised learning approach or thermal camera calibration and face skin temperature extraction. (a), (d), original RGB video frames, (b), (e) corresponding thermal video frames, (c), (f) the fully automated thermal skin face binary masks extraction.

Preparing the training dataset, which is a binary mask for each human face that is created by the first model, is the initial stage in this model. In the second model, the whole dataset is split into distinct halves, such as 20% of the data was used for testing and 80% were used for training. Table 1 below presents the training, validation, and testing datasets for the proposed second model. This form is used to extract a clearer face mask. As a result, this data is separated as test data, which is a negligible percentage, and then the 20% percentile is replaced with some data from the training part.

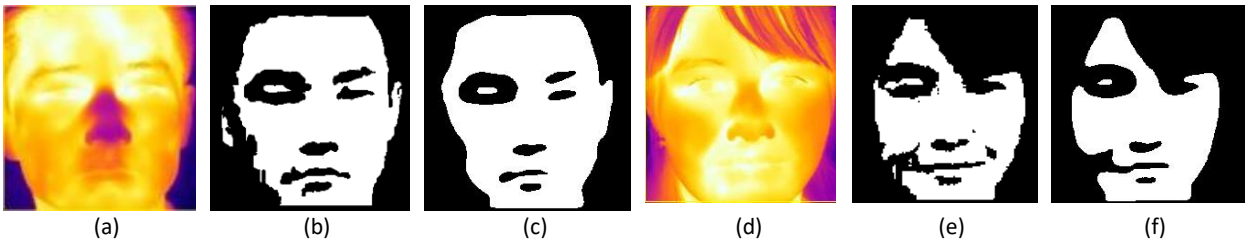
Table 1 Thermal video and binary face mask dataset training and testing.

Dataset Partition	Number Frames
Training 80%	5,373 video frames
Testing 20%	1,343 video frames
Total	1,679 video frames

4.3. Fully automated thermal face skin mask prediction based supervised deep semantic semination model

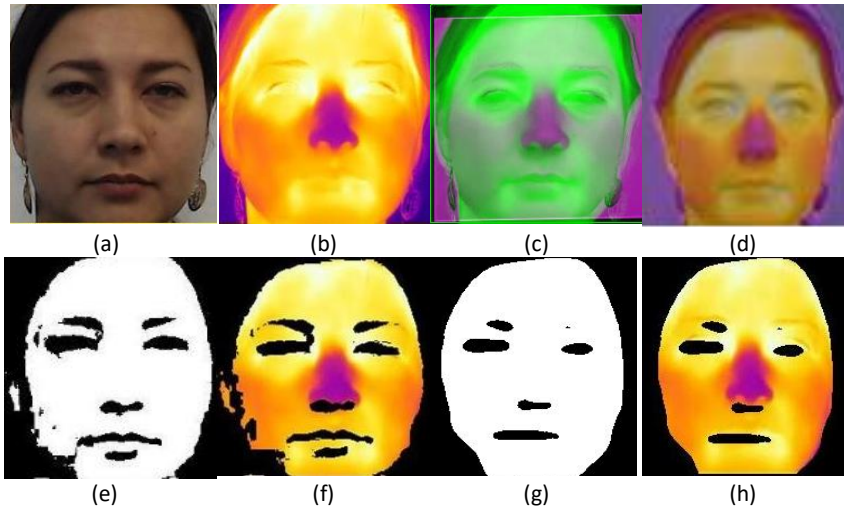
The first stage of the proposed system which is “Fully Automated Thermal Face Skin Extraction” has been tested using Deep Semantic Segmentation based on previous training set. Performance metrics were assessed over several time periods, including 5, 10, 15, 20, 25, and 30 time points. Table 2 below shows the results of the training and validation datasets using 30 epochs. The maximum IOU score was 87% in the training data set and 86% in the validation data set. Moreover, the minimum loss value is 6.675% in the training data set, while it was 7.492% in the validation data set. This indicates that the training level of the model has improved compared to the previous iterations, indicating that it is trainable, accurate, and efficient Figure 9 below shows experimentally results from training and validation datasets using 30 epochs, and further confirm that max IOU score for training and validation datasets are 87% and 86%, respectively Furthermore, the minimum loss value is 6.675% for the training data set and 7.492% for the validation data set, confirming the improved training rate, accuracy and efficiency of the model.

Figure 10 below shows the experimental results of the thermal skin binary prediction-based deep semantic segmentation model. Figure 10 (a) shows the original RGB video frame, Figure 10 (b) shows the corresponding thermal video frame, Figure 10 (c) shows the overlap between the calibrated RGB video frame and thermal video frame, Figure 10 (d) shows the calibrated and registered RGB video frame, Figure 10 (e) shows the face skin binary mask based unsupervised learning (k-means clustering) while Figure 10 (f) shows the thermal skin face extraction based unsupervised learning approach. Figure 10 (g) shows the predicted thermal skin face binary mask based on deep semantic segmentation and Fig.12 (h) shows the face skin thermal prediction from which the temperature is extracted for the proposed second model.



(a) (b) (c) (d) (e) (f)

Figure 9 Fully automated thermal face skin binary face prediction using deep semantic segmentation using (100_1_1_1), (60_1_1_1), (35_1_1_1) thermal video. (a), and (b) original thermal video frame, (b) and (e) face binary mask based unsupervised learning, (c) and (f) predicted face binary mask.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 10 Thermal face skin mask prediction based deep using 103_1_1_1 thermal video, (a) original RGB video frame, (b) original thermal video frame, (c) overlap calibrated RGB video frame and thermal video frame, (d) calibrated and registered RGB video frame, (e) face binary mask based unsupervised learning (k-means clustering) using (d), (f) face skin thermal extraction based unsupervised learning approach using (d) and (e), (g) predicted face binary mask based deep semantic segmentation using (b) and (e), (h) face skin thermal prediction using (b) and (g).

4.4. Fully automated human temperature tracking based thermal skin face extraction

The experimental results of the fully automated human temperature tracking based deep semantic segmentation approach are illustrated in the following Figure 11 using a sample for the Speaking Face Dataset (Abdrakhmanova et al., 2021) based on video case no. 103_1_1_1. Figures 11 (a) and (g) show the Ground Truth face skin binary mask that is extracted using an unsupervised learning approach (Abdrakhmanova et al., 2021) while (b) and (h) show the extracted face skin thermal frame while Figure 11 (c) and (i) show the extracted thermal face temperature. Figures 11 (d) and (j) show the predicted face binary mask using the proposed deep semantic segmentation while Figures 11 (e) and (k) show the extracted thermal face skin image and (f) and (l) show the predicted human temperature.

The training curve for all epochs (30) which represents the validation accuracy of predicting the binary face mask using the thermal video frames only during the training phase (see the curve in Figure 12 (a)). In contrast, the validation curve represents data testing (prediction the correct thermal face skin mask) during the testing/validation phase is represented by the blue curve in Figure 12 (b)). It has been noticed that the training loss has been stabilized in the period between epochs 28-30 by reach the loss value (0.063). It is also important to note that the training curve for IOU starts with accuracy value (78%) and ends with (89%) as is shown in Figure 12 (a) and (b) respectively. Also, Figure 12 (a) shows Intersection over Union (IOU) values for training and validation datasets at 30 epochs for deep semantic segmentation model The blue line representing the training IOU shows a significant increase in initial epochs, stable from 0.86 to 0.87, with validation IOU contradicts The red dashed line fluctuates from 0.76 to 0.78, with small variations across seasons. Despite the variation, there is a slight increase towards the end, indicating that although the model performs consistently on the training data, it experiences variability when applied to the validation data, which may prove necessary that retuning is performed to increase overfitting or generalization. In contrast, Figure 12 (b) shows the loss values at 30 training and verification periods for the deep learning model. The blue line representing the training loss shows a sharp decrease initially, stabilizing from 0.06 to 0.07 as the number of times increases This trend indicates the optimization of the model and reduces the error of the training dataset on. In contrast, the red lines representing the validation loss fluctuate from 0.13 to 0.15 across the epochs, including small peaks and troughs, indicating

variation in performance applied to validation data Regardless about these variables, the validation loss is generally stable, indicating that although the model overfits to some extent it remains a challenging task.

On summary, the experimental results of the training data set and the validation dataset using 30 epochs shows that the IOU score obtained in the training data set was (86%), where it is observed that the percentage has increased from the previous one. Also, the highest score is obtained in the validation dataset by achiveing (77%). In the training data set, the lowest performance loss value was (7.37%), while it was in the validation data set (12.68%). It is noted that the IOU value of the training data increased by one percent, while its value for the validation data remained constant. As for the value of the loss function for the training data, it decreased by a large percentage, while its value for the validation data increased by a few percent. And note that the model gave satisfactory results.

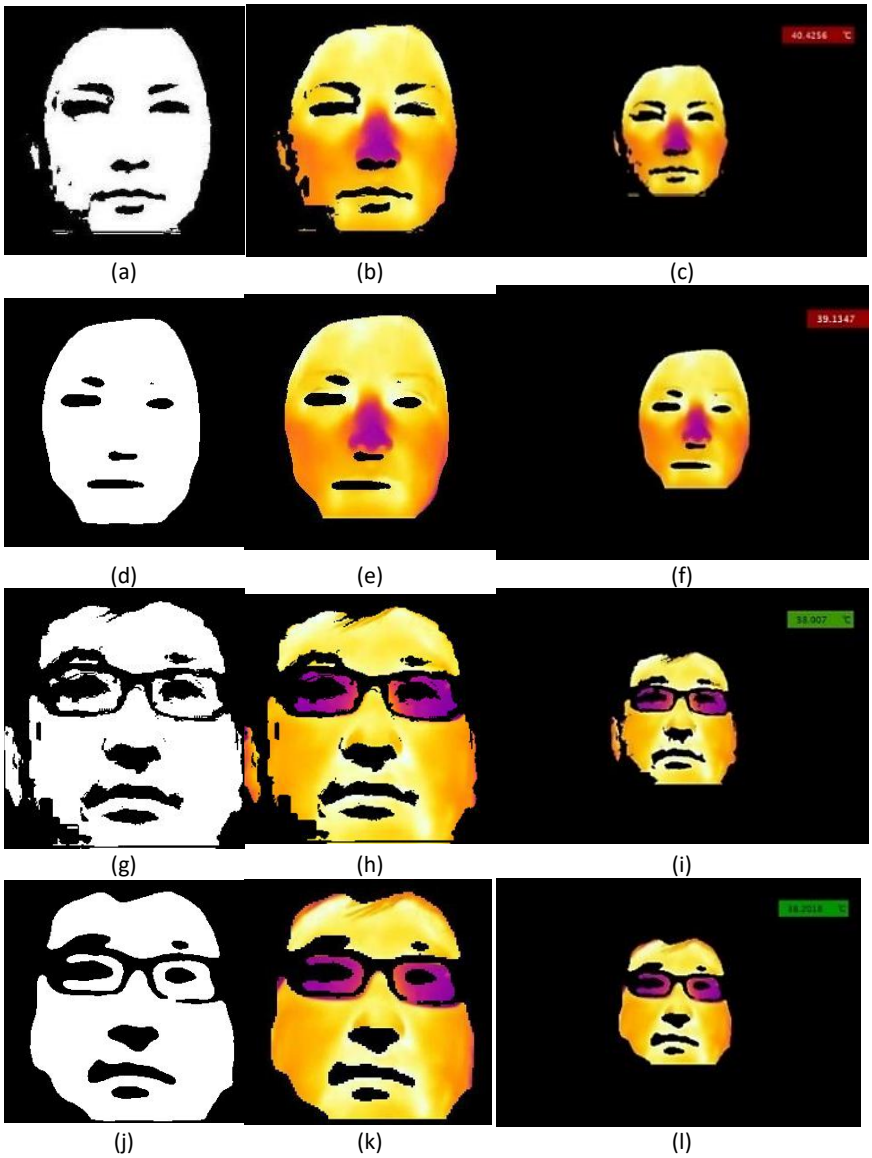
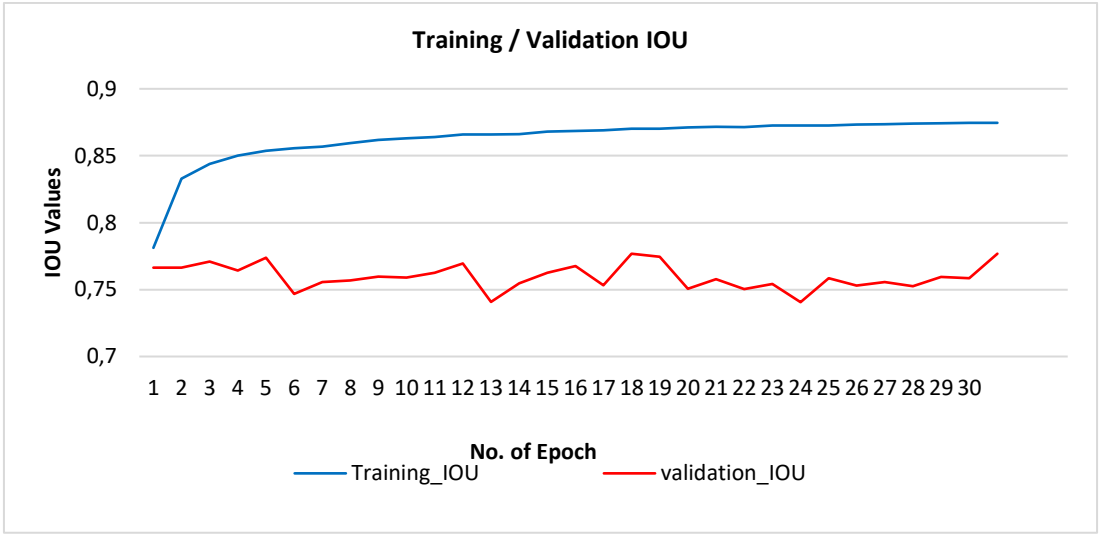


Figure 11 Fully automated real-time human temperature prediction and tracking using 103_1_1_1 thermal video, (a) and (g) face skin binary mask based unsupervised learning (k-means clustering), (b) and (h) face skin thermal extraction based unsupervised learning approach, (c) and (j) temperature prediction and tracking, (d) and (j) predicted face binary mask based deep semantic segmentation, (g) and (j) thermal skin face mask prediction, (h) and (k) human temperature prediction and tracking.

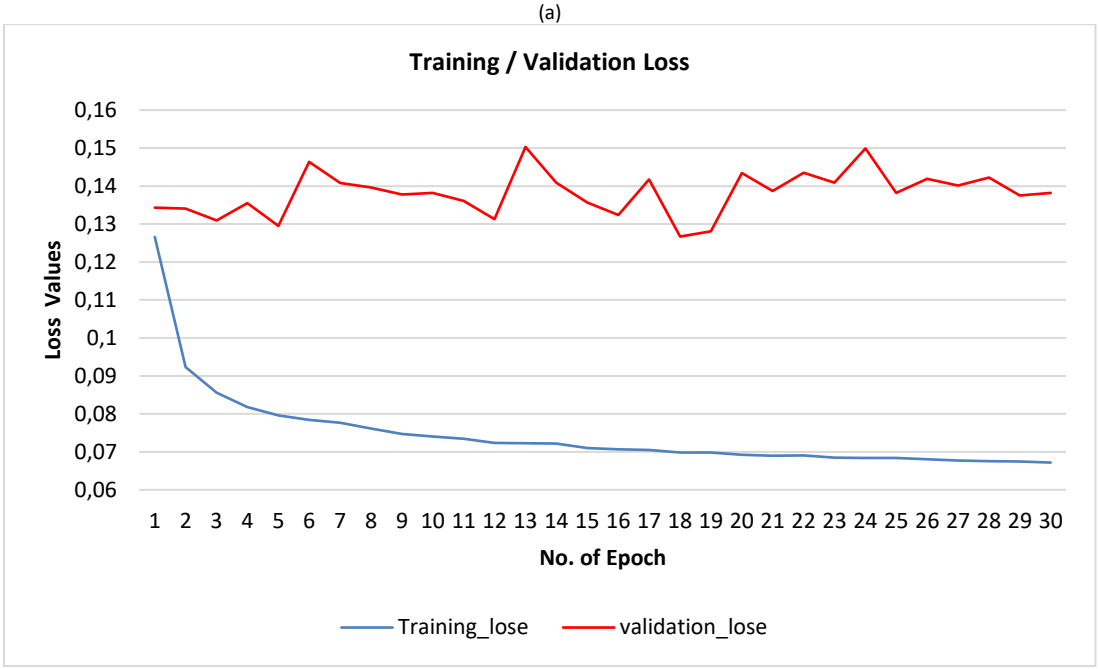
4.5. Experimental Results of Testing Approach

The experimental results of the testing dataset is illusrtaed in the following Table 2. Table 2 presents the test results of the deep semantic segmentation model using the test dataset. The model achieved an accuracy of 95.72%, indicating that most of the predictions were correct. The precision that measured the accuracy of good predictions was 86.14%, while the F1 measure that measures the ability of the model to detect all contexts was slightly higher at 87.64%, and the harmonic mean of accuracy

and recall was 86.89 %. Together, these metrics demonstrate the robust performance and reliability of the model in classifying accuracy and identify relevant features in the test dataset.



(a)



(b)

Figure 12 Experimental results of the training and validation performed of the deep semantic segmentation model using 30 epochs.

Table 2 Experimental results of the deep semantic segmentaion using the Testing Dataset.

Criteria	Average values
Accuracy	95.72%
Precession	86.14%
Recall	87.64%
F1-measure	86.89%

5. Discussion

The experimental results of deep semantic classification using the test dataset show high performance, with 95.72% accuracy, 86.14% accuracy, 87.64% recall, and 86.89% F1-measurement. These metrics show that there is a strong distribution effect, which is necessary for accurate human temperature filtering in timely ways. High precision ensures reliable identification of relevant features, while a balance of accuracy and recall emphasizes the effectiveness of the system in correctly identifying and classifying the target fields without missing no significant omissions or false positives. The F1 measure further emphasizes the balanced performance of the model, combining precision and recall as a single metric. These results suggest that the

combination of such classification methods based on deep learning can improve the accuracy and reliability of human heat removal systems, in particular when infrared thermology or similar technologies are used. A comparison of using the proposed deep semantic segmentation results with state-of-the-art approaches that used for the human temperature extraction using thermal cameras is illustrated in the following Table 2.

Table 3 Comparison of Deep Semantic Segmentation Results with State-of-the-Art Approaches for Human Temperature Extraction Using Thermal Cameras.

Criteria	Deep Semantic Segmentaion	State-of-the-Art (Traditional IRT)	State-of-the-Art (Machine Learning)	State-of-the-Art (Deep Learning)
Accuracy	95.72%	80-90%	90-93%	Up to 95%
Precession	86.14%	Variable (lower realiablity)	85-88%	Approximitily 85%
Recall	87.64%	80-85%	85-88%	Approximitily 85%
F1-measure	86.89%	Genrally lower due to balanced issues	Typically 86-88%	Approximitily 86-88%

Translational deep classification results show a high accuracy (95.72%) compared to traditional infrared thermodynamics (IRT), typically ranging from 80% to 90% (Usamentiaga, R., et al., 2014), 90 to 93% using machine learning (Kazemi N., et al., 2021), and up to 95% using deep learning (Choi, Y., et al., 2016). This shows the potential of deep learning methods to provide accuracy has improved in heat release testing. The deep classification method has slightly better accuracy (86.14%) and recall (87.64%) than machine learning based methods, traditional IRT, and Deep Learning (85-88%) (Naik, K., et al., 2021; Mambou, S. J., et al., 2018; Bagavathiappan, S., et al., 2018; Pagani, S., et al., 2020; Manssor, S. A., et al., 2021),and shows more reliability as the target areas will be identified and classified. It is also highly consistent with 88% demonstrating a balanced performance and confirming its effectiveness in assessing the accuracy and reliability of human heat extraction such as F1-meausement (Peng, J., R., et al., 2023; Chen, Y. Y., et al., 2020; Ahalya, R. K., et al., 2023).

6. Conclusion

Many companies have developed systems that measure human temperatures and integrate them with cameras for economic benefit, these cameras cost $4,000 This high cost is especially burdensome for countries that need more such cameras. Dealing with this requires appropriate machine learning. This paper proposes a fully automated real-time supervised learning method for human temperature prediction and tracking based on thermal facial temperature extraction The proposed system uses deep semantic segmentation to predict thermal skin face masks automatically. The deep semantic segmentation model was trained on thermal video frames and their corresponding thermal skin-face binary masks, from which the first model was extracted Experimental results show that this method provides human temperature prediction and tracking accurate compared to ground truth.

Acknowledgment

Thanks for Ahmed Abdulstar for his assistance in this project.

Ethical considerations

Not applicable.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

Research did not receive any financial support.

References

Abdrakhmanova, M., Kuzdeuov, A., Jarju, S., Khassanov, Y., Lewis, M., & Varol, H. A. (2021). Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. *Sensors, 21*(10), 3465.

Al-Azzawi, A. (2024). Fully automated unsupervised learning approach for thermal camera calibration and an accurate COVID-19 human temperature tracking.

Bagavathiappan, S., Lahiri, B. B., Saravanan, T., Philip, J., & Jayakumar, T. (2013). Infrared thermography for condition monitoring–A review. *Infrared Physics & Technology, 60*, 35-55.

Choi, Y., Kim, N., Hwang, S., & Kweon, I. S. (2016, October). Thermal image enhancement using convolutional neural network. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 223-230). IEEE.

Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing, 162*, 94-114.

Ding, K., Ma, K., Wang, S., & Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(5), 2567-2581.

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2020). Deep learning in medical image registration: A review. *Physics in Medicine & Biology, 65*(20), 20TR01.

Groschner, C. K., Choi, C., & Scott, M. C. (2021). Machine learning pipeline for segmentation and defect identification from high-resolution transmission electron microscopy data. *Microscopy and Microanalysis, 27*(3), 549-556.

Haskins, G., Kruger, U., & Yan, P. (2020). Deep learning in medical image registration: A survey. *Machine Vision and Applications, 31*(1), 8.

Ippalapally, R., Mudumba, S. H., Adkay, M., & HR, N. V. (2020, December). Object detection using thermal imaging. In *2020 IEEE 17th India Council International Conference (INDICON)* (pp. 1-6). IEEE.

Ji, S., Pan, J., Li, L., Hasegawa, K., Yamaguchi, H., Thufail, F. I., ... & Tanaka, S. (2023). Semantic segmentation for digital archives of Borobudur reliefs based on soft-edge enhanced deep learning. *Remote Sensing, 15*(4), 956.

Kazemi, N., Abdolrazzaghi, M., & Musilek, P. (2021). Comparative analysis of machine learning techniques for temperature compensation in microwave sensors. *IEEE Transactions on Microwave Theory and Techniques, 69*(9), 4223-4236.

Krišto, M., Ivasic-Kos, M., & Pobar, M. (2020). Thermal object detection in difficult weather conditions using YOLO. *IEEE Access, 8*, 125459-125476.

Leong, D. P., Teo, K. K., Rangarajan, S., Lopez-Jaramillo, P., Avezum Jr, A., & Orlandini, A. (2018). World Population Prospects 2019. Department of Economic and Social Affairs Population Dynamics. New York (NY): United Nations; 2019 (https://population.un.org/wpp/Download/, accessed 20 September 2020). The decade of healthy ageing. *Geneva: World Health Organization. World, 73*(7), 362k2469.

Makino Antunes, A. C., Aldred, A., Tirado Moreno, G. P., de Souza Ribeiro, J. A., Brandão, P. E., Barone, G. T., ... & Gomes, G. (2023). Potential of using facial thermal imaging in patient triage of flu-like syndrome during the COVID-19 pandemic crisis. *PloS One, 18*(1), e0279930.

Mambou, S. J., Maresova, P., Krejcar, O., Selamat, A., & Kuca, K. (2018). Breast cancer detection using infrared thermal imaging and a deep learning model. *Sensors, 18*(9), 2799.

Manssor, S. A., Ren, Z., Huang, R., & Sun, S. (2021, October). Human activity recognition in thermal infrared imaging based on deep recurrent neural networks. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (pp. 1-7). IEEE.

Naik, K., Pandit, T., Naik, N., & Shah, P. (2021). Activity recognition in residential spaces with internet of things devices and thermal imaging. *Sensors, 21*(3), 988.

Pagani, S., Manoj, P. S., Jantsch, A., & Henkel, J. (2018). Machine learning for power, energy, and thermal management on multicore processors: A survey. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 39*(1), 101-116.

Pang, Y., Lin, J., Qin, T., & Chen, Z. (2021). Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia, 24*, 3859-3881.

Pramanik, S., Singh, R. P., & Ghosh, R. (2020). Application of bi-orthogonal wavelet transform and genetic algorithm in image steganography. *Multimedia Tools and Applications, 79*, 17463-17482.

Rottmann, M., & Reese, M. (2023). Automated detection of label errors in semantic segmentation datasets via deep learning and uncertainty quantification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3214-3223).

Setiadi, D. R. I. M. (2021). PSNR vs SSIM: Imperceptibility quality assessment for image steganography. *Multimedia Tools and Applications, 80*(6), 8423-8444.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access, 8*, 80716-80727.

Su, Y., Ma, C., Chen, J., Wu, H., Luo, W., Peng, Y., ... & Li, H. (2020). Printable, highly sensitive flexible temperature sensors for human body temperature monitoring: A review. *Nanoscale Research Letters, 15*, 1-34.

Toniato, E., Ross, R., & Kritas, S. K. (2020). How to reduce the likelihood of coronavirus-19 (CoV-19 or SARS-CoV-2) infection and lung inflammation mediated by IL-1. *J Biol Regul Homeost Agents, 34*(2), 11-16.

Usamentiaga, R., Venegas, P., Guerediaga, J., Vega, L., Molleda, J., & Bulnes, F. G. (2014). Infrared thermography for temperature measurement and non-destructive testing. *Sensors, 14*(7), 12305-12348.

Vollmer, M. (2020). Infrared thermal imaging. In *Computer Vision: A Reference Guide* (pp. 1-4). Cham: Springer International Publishing.

Yeganeh, H., & Wang, Z. (2012). Objective quality assessment of tone-mapped images. *IEEE Transactions on Image Processing, 22*(2), 657-667.

Yuan, X., Shi, J., & Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications, 169*, 114417.

Zhou, Y., Ji, A., Zhang, L., & Xue, X. (2023). Sampling-attention deep learning network with transfer learning for large-scale urban point cloud semantic segmentation. *Engineering Applications of Artificial Intelligence, 117*, 105554.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* (pp. 3-11). Springer International Publishing.