TFW: Annotated Thermal Faces in the Wild Dataset

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

13-11-2021 / 03-03-2022

CITATION

Kuzdeuov, Askat; Aubakirova, Dana; Koishigarina, Darina; Varol, Hüseyin Atakan (2021): TFW: Annotated Thermal Faces in the Wild Dataset. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.17004538.v2>

DOI

[10.36227/techrxiv.17004538.v2](https://doi.org/10.36227/techrxiv.17004538.v2)

TFW: Annotated Thermal Faces in the Wild Dataset

Askat Kuzdeuov, *Member, IEEE*, Dana Aubakirova, Darina Koishigarina,
and Huseyin Atakan Varol, *Senior Member, IEEE*

Abstract—Face detection and subsequent localization of facial landmarks are the primary steps in many face applications. Numerous algorithms and benchmark datasets have been introduced to develop robust models for the visible domain. However, varying conditions of illumination still pose challenging problems. In this regard, thermal cameras are employed to address this problem because they operate on longer wavelengths. However, thermal face and facial landmark detection in the wild is an open research problem because most of the existing thermal datasets were collected in controlled environments. In addition, many of them were not annotated with face bounding boxes and facial landmarks. In this work, we present a thermal face dataset with manually labeled bounding boxes and facial landmarks to address these problems. The dataset contains 9,982 images of 147 subjects, collected under controlled and uncontrolled conditions. As a baseline, we trained YOLOv5 [1] object detection model and its adaptation for face detection, YOLO5Face [2], on our dataset. In addition to our test set, we evaluated the models on the external RWTH-Aachen [3] thermal face dataset to show the efficacy of our dataset. We have made the dataset, source code, and pre-trained models publicly available at <https://github.com/IS2AI/TFW> to bolster research in thermal face analysis.

Index Terms—Labeled thermal faces in the wild, thermal face detection, thermal facial landmark detection, YOLOv5.

I. INTRODUCTION

FACE detection is a fundamental step in many computer vision tasks, such as face recognition, age and gender classification, and facial expression analysis. Early face detection algorithms relied on classifiers built on top of feature extractors, such as Haar cascades [4] and Histogram of Oriented Gradients (HOG) [5]. These algorithms are computationally efficient and capable of accurately detecting faces under constrained conditions. However, because of the simplicity of the extracted features, their performance degrades in unconstrained environments. Later, the use of convolutional neural networks (CNNs) for image classification spurred rapid development in face detection [6]. Many annotated face databases and deep learning models have been introduced, significantly improving face detection in the wild [7]. However, variations in head pose, facial expression, occlusion, and especially illumination still pose challenges. The illumination problem is related to the traditional cameras that operate in the visible spectrum. To address this problem, thermal cameras are utilized because they are invariant to the illumination and can be used in complete darkness (see Fig. 1).

A. Kuzdeuov, D. Aubakirova, D. Koishigarina, and H. A. Varol are with the Institute of Smart Systems and Artificial Intelligence, Nazarbayev University, Nur-Sultan City, Kazakhstan. Email: {askat.kuzdeuov, d.aubakirova, darina.koishigarina, ahvarol}@nu.edu.kz

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

Fig. 1: Examples of visual and thermal images taken in different conditions: a) normal illumination, b) low illumination, and c) darkness.

Nowadays, thermal cameras are employed for a wide range of applications. For instance, in biometrics, thermal cameras are used to extract facial physiological patterns for face recognition [8]. The physiological features are also utilized for monitoring vital signs to detect stress [9], deception [10], and affective states [11]. In addition, thermal cameras have gained further popularity after the COVID-19 outbreak for remote fever detection [12]. These tasks require robust face detection, especially in uncontrolled environments such as airports, railway stations, and malls. However, face detection models trained on visual images are not accurate in detecting thermal faces because of the modality gap between the domains.

In this regard, many works on thermal face detection have been presented in recent years. For instance, a face detection method based on traditional image processing was introduced in [24]. The authors collected 300 images of 30 participants in a controlled environment to test the algorithm. The potential of the popular machine learning-based face detection models developed for the visual spectrum was explored in [25]. The researchers trained and evaluated the models on 2,935 thermal images using a cross-validation method. The results showed that the machine learning-based methods outperform specialized approaches when sufficient training data are available. Deep learning models for thermal face detection have been proposed in several works. In particular, a thermal face detection model based on the R-CNN architecture was trained using transfer learning to examine its performance under various conditions [26], where the original dataset (without augmentation) consisted of 287 images. Transfer learning was applied to train the YOLOv3 object detection model

TABLE I: Summary of thermal face datasets.

Dataset	Images	Participants	Resolution	Camera	Environment	Bounding box	Landmark
VIS-TH [13]	4,200	50	160×120	FLIR Duo R	controlled	–	–
IRIS [14]	4,228	30	320×240	Raytheon PalmIR Pro	controlled	–	–
IRIS-M3 [15]	2,624	82	640×480	Raytheon PalmIR Pro	controlled	–	–
NVIE [16]	N/A	215	320×240	SAT-HY6850	controlled	–	–
KTFE [17]	N/A	26	320×240	NEC R300	controlled	–	–
Carl [18]	2,460	41	160×120	TESTO 880-3	controlled	–	–
UL-FMTV [19]	N/A	238	640×512	FLIR Phoenix Indigo IR	controlled	–	–
SpeakingFaces [20]	4,581,595	142	464×348	FLIR T540	controlled	–	–
RWTH-Aachen [3]	2,935	90	1024×768	Infratec HD820	controlled	–	68
ARL-VTF [21]	549,712	395	640×512	FLIR Boson	controlled	✓	6
OSU Color-Thermal [22]	17,089	N/A	320×240	Raytheon PalmIR 250D	wild	–	–
TIV [23]	63,782	N/A	1024×1024	FLIR SC8000	wild	–	–
Our TFW dataset	9,982	147	464×348	FLIR T540	controlled, wild	✓	5

in [27]. The model was pre-trained on visual images and then fine-tuned using 2,859 thermal images. A comparison of the classical Viola-Jones algorithm with the YOLOv2 and Faster R-CNN object detection models for thermal face detection was presented in [28]. The models were trained on 1,148 thermal images, where each image contained one subject. According to the results, the Faster R-CNN model achieved the highest accuracy and computationally the most expensive. The YOLOv2 model provided a high detection rate along with faster performance.

The bottleneck of training thermal face detection models is the lack of large-scale annotated thermal face datasets collected in the wild. For the visual spectrum, there are many benchmark datasets available, such as Face Detection Data Set and Benchmark (FDDB) [29], Multi-Attribute Labelled Faces (MALF) [30], and Wider Face [31] databases. The FDDB dataset contains 2,845 images with 5,171 labeled faces. The MALF dataset consists of 5,250 images and 11,931 labeled faces. The Wider Face dataset has 32,203 images and 393,703 labeled faces with a high degree of variability in scale, occlusion, expression, pose, and illumination, making it the most diverse dataset.

In this work, we present the Thermal Faces in the Wild (TFW) dataset with manually labeled bounding boxes and five-point facial landmarks (eye centers, nose tip, and mouth corners). In face detection, facial landmarks serves as an additional supervision signal and improve the detection of challenging faces [32]. In addition, facial landmarks are used for face alignment in face recognition. A few works have addressed the training of thermal facial landmark detection models [33]–[35]. In these works, the models were trained on a limited number of thermal images collected in controlled conditions. The TFW dataset contains 9,982 images and 16,509 labeled faces, of which 5,112 images (5,112 faces) were captured in a controlled environment, while 4,870 images (11,397 faces) in the wild. The dataset covers challenging scenarios, such as occlusion, head pose variations, different scales, face masks, different places and weather conditions. To demonstrate the effectiveness of our dataset, we trained the state-of-the-art YOLOv5 [1] object detection model and its version for face detection YOLO5Face [2] on our dataset. The models achieved highly accurate results on our test set and also on the external RWTH-Aachen thermal face dataset [3].

We have made the dataset, annotations, source code, and pre-trained models publicly available to facilitate research in thermal face analysis. To the best of our knowledge, our dataset is the first thermal face dataset developed in both controlled and uncontrolled environments and includes facial bounding boxes and landmarks.

The rest of the paper is organized as follows: In Section II, we review the existing thermal face datasets. Section III introduces our dataset including the data collection and annotation steps. In Section IV, we explain the rationale for choosing the YOLOv5 and YOLO5Face models in this work. Section V provides the details of the training the YOLOv5 and YOLO5Face models and the experimental results. We discuss and interpret the experimental results in Section VI. We conclude our work in Section VII.

II. RELATED WORKS

In this section, we review the existing thermal facial datasets. The characteristics of the datasets are provided in Table I. Most of the datasets have been collected in controlled environments. The VIS-TH dataset [13] contains 50 participants and 4,200 aligned thermal-visual image pairs with a resolution of 160×120 . The dataset was acquired using a FLIR Duo R thermal camera in a controlled environment under various illumination conditions, facial expressions, head poses, and occlusions. The dataset does not provide bounding boxes and facial landmarks. However, it is possible to automatically annotate the thermal images by applying the state-of-the-art face detection and facial landmark prediction models to their visual pairs. Nonetheless, the automatically detected bounding boxes and facial landmarks might require additional manual calibration [35].

The IRIS [14] and IRIS-M3 [15] datasets were collected in controlled environments using a Raytheon Palm-IR-Pro thermal camera. The IRIS dataset contains 4,228 thermal-visual image pairs (320×240 pixels) of 31 participants. The dataset was developed under different lighting conditions, head poses, and facial expressions. The IRIS-M3 dataset has 2,624 thermal-visual image pairs (640×480 pixels) of 82 individuals of different ethnicities. The distinguishing feature of the IRIS-M3 dataset is the presence of natural outdoor illumination. The IRIS-M3 dataset does not contain variations in head pose and facial expression. Also, annotations for the bounding boxes

and facial landmarks are not available in both datasets. In addition, the visual and thermal images are not aligned, which makes automatic annotation impossible.

The NVIE dataset [16] provides images of 215 individuals with posed and spontaneous facial expressions, such as disgust, fear, surprise, sadness, and anger. The dataset was acquired in a controlled environment under various illumination conditions using a SAT-HY6850 infrared camera. Similarly, the KTFE dataset [17] was designed for emotion recognition. It contains video clips of 26 individuals with seven spontaneous emotions that were recorded in a controlled setting, using a NEC R300 infrared camera. The thermal images in the NVIE and KTFE datasets have a same resolution of 320×240 pixels. The Carl dataset [18] contains 2,460 thermal, visual, and near-infrared images of 41 individuals. The thermal images with a resolution of 160×120 pixels were collected using a TESTO 880-3 camera in a controlled setting. The UL-MFTV [19] database has images of 238 individuals in the visible, near-infrared, short-wavelength infrared, middle-wavelength infrared, and long-wavelength infrared spectra. Moreover, the sessions were held with a time lapse of several years. In addition, the dataset contains images of individuals before and after exposure to cold temperatures or physical activity. The SpeakingFaces [20] is the largest database in terms of the number of frames. The dataset provides 4,581,595 thermal-visual image pairs of 142 individuals. The thermal images were acquired using a FLIR T540 thermal camera under various head poses in a controlled environment.

To the best of our knowledge, there are only a few datasets with labeled bounding boxes and facial landmarks. For instance, the RWTH-Aachen dataset [3] has 2,935 images of 90 individuals. The dataset was manually annotated with a 68-point landmark set. The images were captured in a controlled environment using an Infratec HD820 camera with a resolution of 1024×768 . The dataset contains various head poses and facial expressions. The bounding boxes and visual pairs are not available. However, the bounding boxes can be obtained using the coordinates of facial landmarks. The ARL-VTF dataset [21] contains more than 500,000 thermal-visual image pairs of 395 individuals. The dataset provides bounding boxes and six point landmarks. The images with a resolution of 640×512 were acquired in a controlled environment.

Most thermal face datasets have been collected in controlled environments, where an individual sits at a predefined distance from the camera against a neutral background. Illumination, room temperature, expressions, and poses are predefined. However, in real-world scenarios, the number of individuals per image, background environment, temperature, illumination, scale, and pose vary greatly. One of the few datasets collected under realistic conditions is the OSU Color-Thermal Database [22]. The dataset contains 17,089 thermal-visual images of pedestrians captured by a Raytheon PalmIR 250D thermal camera mounted on the side of a building. This dataset is not suitable for face detection because of the tiny sizes of faces. In addition, bounding boxes and facial landmarks are not available. The Thermal Infrared Video (TIV) dataset [23] contains images with higher quality and resolution, thereby allowing the database to be used for face detection. However, of

TABLE II: Summary of the TFW dataset.

Dataset	# subjects	# images	# labeled faces
C-Indoor	142	5,112	5,112
S-Indoor	9	780	1,748
U-Outdoor	15	4,090	9,649
Combined	147	9,982	16,509

the published 63,782 frames, only a subset contains images of humans. Annotations of facial bounding boxes and landmarks are not available either.

In summary, many thermal datasets have been designed for cross-spectral face recognition, anti-spoofing, and emotion recognition. Most of the datasets have been collected in controlled environments. Moreover, only a few of them provide facial bounding boxes and landmarks. The existing datasets collected in wild conditions are not suitable for training face detection models. Therefore, the area of thermal face detection, especially in real-world environments, remains overlooked. Our dataset fills these gaps by providing thermal images taken in real-world settings, as well as face bounding boxes and facial landmarks.

III. TFW DATASET

A. Data Collection

The TFW dataset contains thermal images acquired in indoor and outdoor environments. The indoor dataset consists of images collected in controlled and semi-controlled environments. The controlled indoor (C-Indoor) dataset was constructed using our previously published SpeakingFaces dataset [20]. The semi-controlled indoor (S-Indoor) and uncontrolled outdoor (U-Outdoor) datasets were collected using the same

Fig. 2: Examples of cropped thermal facial images from the SpeakingFaces dataset showing nine participants from nine positions.

Fig. 3: Samples from a semi-controlled indoor environment: a) predefined motion (i.e. turn to the left), b) face masks, c) free motion, and samples from the uncontrolled outdoor environments: d) sunny weather, e) cloudy weather, and f) face masks.

FLIR T540 thermal camera with a resolution of 464×348 pixels, a wave-band of $7.5\text{--}14\ \mu\text{m}$, the field of view 24° , and an *iron* color palette. The characteristics of the TFW dataset are summarized in Table II. The dataset was collected with the explicit approval of the Institutional Research Ethics Committee of Nazarbayev University. All subjects participated voluntarily after signing a consent. Also, they were informed that we share the images as a dataset.

1) *C-Indoor dataset*: The SpeakingFaces dataset was collected in a controlled indoor environment. The dataset provides thermal-visual image pairs of 142 individuals, with each participating in two sessions conducted on different days. In each session, the subjects participated in two experiments. In the first experiment, they sat still and silent while the cameras recorded them from nine different positions (see Fig. 2). In the second experiment, the participants read a series of commands displayed on a screen, and the cameras recorded from the nine positions. In this work, we extracted one frame per position to avoid similarities. As a result, we obtained 5,112 thermal images ($142\text{ participants} \times 2\text{ sessions} \times 2\text{ experiments} \times 9\text{ positions}$) as shown in Table II. Considering that each image contains one participant, the total number of labeled faces is also 5,112.

2) *S-Indoor dataset*: In this dataset, subjects participated in four different sessions. In the first and second sessions, subjects walked under normal illumination with and without a face mask, respectively (see Fig. 3). In the third and fourth sessions, they walked under a low illumination condition with and without a face mask, respectively. We call this dataset semi-controlled because subjects were asked to do predefined commands such as turn head to the left/right and up/down at the beginning of each session, as shown in Fig. 3a. Then, they

were allowed to walk in random directions and perform any gestures and mimics (see Fig. 3c). In this dataset, we also recorded the visual pairs of thermal frames. We collected this dataset to compare our thermal face detection model with the original YOLO5Face model (trained on the WiderFace dataset) under different conditions. We tested our best model on the thermal images while the YOLO5Face model was evaluated on the visual images. In total, the dataset contains 780 thermal-visual pairs of 9 subjects, where 7 of them had participated in the collection of the SpeakingFaces dataset. The number of labeled faces is 1,748 in each domain.

3) *U-Outdoor dataset*: We used outdoor locations on Nazarbayev University campus to collect data in unconstrained environments. The dataset was collected on different days during the summer. Our main goal was to collect images of a randomly walking crowd in open spaces. Therefore, participants moved in front of the installed thermal camera. They were allowed to walk in any direction within the camera’s field of view and make any gestures. The only requirement was to be inside the camera’s field of view. We also aimed to collect data in different weather conditions, as the temperature directly affects thermal images. Thus, our dataset contains thermal images taken in sunny and cloudy weather conditions, as shown in Fig. 3. In addition, in some experiments, we asked participants to wear face masks as shown in Fig. 3f. In total, 15 subjects participated in this dataset, where 12 had participated in the SpeakingFaces dataset. In total, we conducted seven sessions at different locations. We extracted frames from the recorded videos and downsampled them to avoid similarities between the extracted frames. We then manually checked the extracted frames to remove similar images. As a result, we obtained 4,090 thermal images.

Fig. 4: a) RGB thermal image, b) the grayscale of the thermal image, c) the negative of the grayscale image, d) the enhanced negative image, and e) the enhanced image with bounding boxes and facial landmarks.

B. Data Annotation

We utilized a graphical image labeling tool [36] to annotate bounding boxes and facial landmarks. However, RGB thermal images are challenging to annotate because the contours of the eyes and lips are not entirely discernible because of the dissipating heat (see Fig. 4a). To address this problem, we processed thermal images before annotation. First, we converted the thermal images from RGB to grayscale as shown in Fig. 4b. Then, we inverted the grayscale images to obtain the negative ones as shown in Fig. 4c. Then, we applied gamma correction to the negative images to enhance the brightness of the faces. By doing so, we improved the visibility of facial features, especially the contours of the eyes and lips (see Fig. 4d). We then added bounding boxes and facial landmarks to the enhanced images as illustrated in Fig. 4e. We arranged the facial landmarks as follows: 1. the center of the right eye, 2. the center of the left eye, 3. the tip of the nose, 4. the right outer corner of the mouth, and 5. the left outer corner of the mouth. These landmarks are known as primary and play a crucial role in facial identity and face tracking [37]. The identical order of landmarks has been used to train many face detection models, such as MTCNN [38], RetinaFace [32], YOLO5Face [2], developed for visual images. The bounding box tightly contains the chin, forehead, and cheek. We annotated all the faces collected in the controlled environment. In the wild dataset, we ignored facial landmarks for highly occluded faces as in [32]. The dataset was labeled by four annotators and cross-checked by the main annotator. As a result, the TFW dataset contains 147 participants, 9,982 images, and 16,509 labeled faces.

The statistics for the annotated bounding boxes are given in Table III. We classified faces into three categories: small (< 25

pixels), medium (25–125 pixels), and large (≥ 125 pixels) based on their heights. All the faces captured in the controlled indoor environment are large because participants were close to the camera during data collection. Similarly, all the faces in the semi-controlled indoor environment are medium. The faces captured in the uncontrolled outdoor environment were almost evenly distributed between the small and medium categories. Overall, the TFW dataset consists of 29.84% small faces, 39.22% medium faces, and 30.94% large faces.

IV. FACE DETECTION MODELS

The state-of-the-art face detection methods leverage the achievements of deep learning-based object detection models. There are two major deep learning-based object detection architectures: one-stage (e.g., SSD [39], YOLO [40], and RetinaNet [41]) and two-stage (RCNN [42], Fast-RCNN [43], and Faster-RCNN [44]). The one-stage models directly predict classes and bounding boxes. On the other hand, the two-stage models generate the regions of proposals to classify and further refine the bounding boxes of the objects. Therefore, the one-stage detectors are usually faster than the two-stage detectors but are less accurate because the one-stage models tend to have a higher false-positive rate. However, they achieve a higher recall rate than the two-stage models.

In addition, research on face detection has been developing rapidly thanks to the extremely challenging Wider Face benchmark dataset [31]. A large number of face detection methods have been proposed to achieve accurate results on this benchmark dataset¹. The recent methods include RetinaFace [32], TinaFace [45], MogFace [46], and YOLO5Face [2]. Some methods are general object detection models adapted for face detection, while others explore unique facial features. For instance, the RetinaFace and YOLO5Face face detection architectures use five facial landmarks as an extra supervision signal, while the TinaFace is a general object detection model.

In this work, we employed the YOLOv5 [1] object detection model and its adapted version for face detection YOLO5Face [2]. We chose these models because they provide models of different sizes, such as nano, small, medium, and

TABLE III: Bounding box height statistics for the TFW dataset.

Dataset	< 25 pixels	25–125 pixels	≥ 125 pixels
C-Indoor	0%	0%	100%
S-Indoor	0%	100%	0%
U-Outdoor	51.12%	48.88%	0%
Combined	29.84%	39.22%	30.94%

¹http://shuoyang1213.me/WIDERFACE/WiderFace_Results.html

large, bringing more options for experiments and finding the best model. Also, an additional output block P6 can be attached to the output of each model to improve the detection of large faces. In addition, we investigate the benefits of using facial landmarks as a supervision signal by comparing the results of the YOLOv5 (without facial landmarks) and YOLO5Face (with facial landmarks) models.

V. EXPERIMENTS

A. Splitting the Dataset

We split the TFW dataset into training, validation, and test sets, as shown in Table IV. The controlled indoor dataset contains images of 142 individuals, and for each subject, we have 36 images. Therefore, we split the indoor dataset based on participants ids as [1-100] subjects (3,600 images, 3,600 faces) for training, [101-130] subjects (1,080 images, 1,080 faces) for testing, and [131-142] subjects (432 images, 432 faces) for validation. We split the uncontrolled outdoor dataset based on sessions so that the locations and weather conditions are different in the training, validation, and test sets. Therefore, we assigned the first 2,958 images (7,201 faces) to the training set, the next 332 images (649 faces) to the validation set, and the remaining 800 images (1,799 faces) to the test set. We use the semi-controlled indoor dataset to compare the performance of our best thermal face detection model with the original YOLO5Face model [1]. In total, the training set consists of 6,558 images (10,801 faces), the validation set contains 764 images (1,081 faces), and the test set consists of 2,260 images (4,508 faces). In other words, we used 65.7% of the data for training, 7.7% for validation, and 22.6% for testing. In addition to our test set, we evaluated the models on the RWTH-Aachen dataset, which contains 2,935 images (2,935 faces), to demonstrate the efficacy of our dataset. Considering that most of the existing thermal face datasets provide images in grayscale format, including the RWTH-Aachen dataset, we converted our images from the *iron* color palette to grayscale to train the models. Also, we mirrored the images and corresponding bounding boxes with facial landmarks to double the size of the dataset.

B. Training Settings

We used the training set to train the models while the validation set for tuning the hyperparameters. After finding the optimal parameters, we combined the training and validation sets to train the models on the combined set. We trained a series of YOLO5Face models of different sizes: nano (YOLOv5n-Face, YOLOv5n6-Face), small (YOLOv5s-Face,

YOLOv5s6-Face), medium (YOLOv5m-Face, YOLOv5m6-Face), and large-sized models (YOLOv5l-Face, YOLOv5l6-Face). As the name suggests, the main difference between these models is the size, – that is the depth (D) and width (W) multiples, and the presence of the P6 output block (see Table V). The YOLOv5n-Face and YOLOv5n6-Face models use ShuffleNetv2 as the backbone network, while the other models use CSPNet. The YOLOv5 models were trained on the COCO dataset [47], while the YOLO5Face models were trained on the Wider Face dataset. Both models provide pre-trained weights. Considering that our dataset is smaller than the COCO and Wider Face datasets, we applied transfer learning to train the pre-trained models further on our dataset.

We used a single V100 graphics processing unit (GPU) on an NVIDIA DGX2 server to train the nano and small model series. We used the default parameters from the original work [2] that achieved state-of-the-art results on the Wider Face dataset, where the batch size and number of epochs were 64 and 250, respectively. To train the medium and large model series, we used the Distributed Data-Parallel mode in PyTorch to parallelize training on two V100 GPUs, setting the batch size and number of epochs to 32 and 250, respectively. Like in the original paper, we used the SGD optimizer, where the initial and final learning rates were 0.01 and 0.00001, respectively. The weight decay was 0.0005. The first three epochs were warm-up epochs with a momentum of 0.8. Afterward, the momentum was changed to 0.937.

The YOLOv5 model provides nano (YOLOv5n, YOLOv5n6), small (YOLOv5s, YOLOv5s6), medium (YOLOv5m, YOLOv5m6), and large (YOLOv5l, YOLOv5l6) models with and without the P6 output block. All models use the CSPNet as the backbone network. The depth and height multipliers are identical to those of the YOLOv5Face models. We trained the YOLOv5 models as a general object detection model. It predicts a single class, the face. We used parameters used in the YOLO5Face models for the YOLOv5 models to ensure a fair comparison.

C. Results

We tested the models on the test set of the C-Indoor and U-Outdoor datasets. In this way, we analyze the capability of models to detect large faces captured in the controlled condition and small/medium faces captured in the wild. The models require the width and height of the input images to be divisible by 64. Therefore, we resized the images to 512×384 by keeping the aspect ratio. We used the average precision (AP) metric to evaluate the accuracy of predicted bounding boxes. To cut out weak predictions, we set a confidence threshold of 0.5 to the output class probability of models. To measure the overlap ratio between the predicted and ground-truth bounding boxes, we used the intersection over union (IoU) metric with a threshold of 0.5.

The experimental results for YOLOv5 and YOLO5Face models are in Table V. It shows that the controlled indoor dataset is much easier to generalize because both models achieved the maximum AP score. However, the accuracy of models is lower on the uncontrolled outdoor and

TABLE IV: Training, validation, and test set specifications.

Dataset	Training		Validation		Test	
	Images	Faces	Images	Faces	Images	Faces
C-Indoor	3,600	3,600	432	432	1,080	1,080
S-Indoor	-	-	-	-	780	1,748
U-Outdoor	2,958	7,201	332	649	800	1,799
Combined	6,558	10,801	764	1,081	2,660	4,508

TABLE V: Results on the test set of the C-Indoor and U-Outdoor dataset, and on the RWTH-Aachen dataset [3]

Model	Backbone	AP ₅₀			NME			Speed	Params (M)	Flops (G)
		TFW		Aachen	TFW		Aachen	V100		
		C-Indoor	U-Outdoor		C-Indoor	U-Outdoor		(ms)		
YOLOv5n	CSPNet	100	97.29	99.15	–	–	–	6.16	1.76	0.99
YOLOv5n6	CSPNet	100	95.79	99.25	–	–	–	8.18	3.09	1.02
YOLOv5s	CSPNet	100	96.82	99.22	–	–	–	7.20	7.05	3.91
YOLOv5s6	CSPNet	100	96.83	99.35	–	–	–	9.05	12.31	3.88
YOLOv5m	CSPNet	100	97.16	99.45	–	–	–	9.59	21.04	12.07
YOLOv5m6	CSPNet	100	97.10	99.49	–	–	–	12.11	35.25	11.76
YOLOv5l	CSPNet	100	96.68	99.25	–	–	–	12.39	46.60	27.38
YOLOv5l6	CSPNet	100	96.29	99.45	–	–	–	15.73	76.16	110.2
YOLOv5n-Face	ShuffleNetv2	100	95.93	99.32	0.039	0.434	0.059	10.12	1.72	1.36
YOLOv5n6-Face	ShuffleNetv2	100	95.59	99.52	0.047	0.414	0.068	13.30	2.54	1.38
YOLOv5s-Face	CSPNet	100	96.73	99.25	0.039	0.391	0.055	8.29	7.06	3.67
YOLOv5s6-Face	CSPNet	100	96.36	99.63	0.041	0.418	0.062	10.86	12.37	3.75
YOLOv5m-Face	CSPNet	100	95.32	99.56	0.036	0.404	0.054	11.01	21.04	11.58
YOLOv5m6-Face	CSPNet	100	96.32	99.66	0.038	0.411	0.059	13.97	35.45	11.84
YOLOv5l-Face	CSPNet	100	96.18	99.45	0.035	0.405	0.051	13.57	46.59	25.59
YOLOv5l6-Face	CSPNet	100	95.76	99.63	0.038	0.401	0.059	17.29	76.67	113.2
YOLO5Face [1]	CSPNet	99.58	92.04	97.79	0.070	0.552	0.080	8.29	7.06	3.67

Aachen datasets. For instance, the YOLOv5n model showed the accuracy of 97.29% and 99.15% on the outdoor and Aachen datasets, respectively. The YOLOv5n6 model has an accuracy of 95.79% and 99.25% for the outdoor and Aachen datasets, respectively. Similar trends were observed for the small (YOLOv5s), medium (YOLOv5m), and large (YOLOv5l) models, where the additional P6 output block slightly increases accuracy on the Aachen dataset.

The YOLOv5n-Face model with the ShuffleNetv2 backbone network achieved an AP score of 95.93% on the outdoor dataset and an AP score of 99.32% on the Aachen dataset. The YOLOv5n6-Face with the P6 output block improved the AP score on the Aachen dataset up to 99.52% but was slightly less accurate (95.59%) on the outdoor dataset. As with the YOLOv5 models, using the P6 output block yielded better results on the Aachen dataset. For instance, the YOLOv5s-Face small model showed AP scores of 96.73% and 99.25% on the outdoor and Aachen datasets, respectively. Then, the YOLOv5s6-Face small model with the P6 block improved the accuracy to 99.63% for the Aachen dataset. In addition, the YOLOv5m6-Face model improved the AP scores of the YOLOv5m-Face model from 95.32% and 99.56% (outdoor) to 96.32% and 99.66% (Aachen). Similarly, the YOLOv5l6-Face large model with the P6 block showed better results than the YOLOv5l-Face model on the Aachen dataset.

To measure the performance of the YOLO5Face models in detecting facial landmarks, we used the normalized mean error (NME) metric

$$NME = \frac{1}{N} \sum_{i=1}^N \frac{\|\hat{l}_i - l_i\|}{K \times D_i} \quad (1)$$

where $\|\hat{l}_i - l_i\|$ is L2 norm between the coordinates of the ground-truth $\hat{l}_i$ and predicted l_i facial landmarks, N is the number of detected faces, K is the number of facial landmarks, and D_i is the distance between eye centers. The results in Table V reveal interesting insights. For instance, the P6 output block slightly reduces performance in terms of NME, despite providing higher results in terms of AP. The reason might be

that the P6 block reduces the feature map size twice, leaving fewer features for regressing the facial landmarks. Moreover, the NMEs on the outdoor dataset are noticeably larger than on the indoor and Aachen datasets. The reason is that the outdoor dataset is much more challenging because of the large variations in scale, pose, and temperature of the environment. Also, subjects are far from the camera in the outdoor dataset. Therefore, facial features are not poorly observable in thermal images. The NME on the Aachen dataset is slightly greater than on the indoor dataset. The main reason is that the Aachen dataset is more diverse in terms of facial expressions. Also, we did not use images from the Aachen dataset to train the models.

We used the YOLOv5s-Face model for further experiments because it provided accurate results and was faster than larger models. We compared its performance on thermal face detection with the original YOLO5Face (it's YOLOv5s version) model [1]. The results of the YOLO5Face on the test set of our dataset are in Table V. It achieved 99.58% on the indoor dataset, 92.04% on the outdoor dataset, and 97.79% on the Aachen dataset. The NMEs for the indoor, outdoor, and Aachen datasets are 0.07, 0.552, and 0.08, respectively. The same model but trained on our dataset, YOLOv5s-Face, reduced the error for the indoor dataset to zero, for the outdoor dataset 2.4 times, and for the Aachen dataset 6.5 times. The NMEs were reduced 1.8 times for the indoor dataset, 1.4 times for the outdoor dataset, and 1.5 times for the Aachen dataset. These results show that the visual face detection model requires additional training on thermal images despite it being trained on a large-scale visual face dataset.

We used the semi-controlled indoor dataset to compare the performance of our YOLOv5s-Face thermal face detection model and the YOLO5Face visual face detection model under different conditions. The experimental results are in Table VI. In the first case, the illumination is normal, and subjects are without face masks. The YOLO5Face model successfully detected all visual faces in the dataset, while our YOLOv5s-Face model showed 99.42% accuracy on the thermal images.

TABLE VI: Comparison of thermal and visual face detection models.

Illumination	Face mask	YOLO5Face [1]		YOLOv5s-Face	
		Visual images	Thermal images	Visual images	Thermal images
		AP50	NME	AP50	NME
Normal	No	100.0	0.074	99.42	0.118
Normal	Yes	99.65	0.157	84.07	0.250
Low	No	79.84	0.137	97.20	0.135
Low	Yes	65.55	0.159	85.89	0.184

The NME of our model is noticeably larger than the error of the YOLO5Face model. In the second case, the illumination is normal, and subjects wear face masks. In this case, the performance of the YOLO5Face model dropped slightly (99.65%), while our YOLOv5s-Face model's performance dropped significantly (84.07%). The NME errors increased considerably for both models. In the third case, the illumination is low, and subjects are without face masks. In this case, the accuracy of the YOLO5Face model dropped dramatically to 79.84%, while our model showed high accuracy (97.2%) as in the normal illumination case. The NME error of the YOLO5Face model increased substantially compared to the normal illumination case, while for the YOLOv5-Face, the error was almost the same. In the fourth case, the illumination is low, and subjects wear face masks. In this case, the accuracy of the YOLO5Face model decreased further to 65.55%. The YOLOv5-Face showed the accuracy (85.89%) was similar to the normal illumination case (84.07%). The same trend was observed for the NMEs.

VI. DISCUSSION

The experimental results show that the YOLOv5 and YOLO5Face models provide accurate AP scores. However, using facial landmarks as an additional supervision signal improves the accuracy of detecting large faces. The results also revealed that our thermal face detection model is more robust to illumination changes than the visual face detection model. However, our model is less accurate in detecting faces with masks. It can be explained by the limited amount of thermal faces with masks in our dataset.

The trade-off between the inference speed and accuracy on the outdoor dataset at 0.5 IOU metric is visualized in

Fig. 5: Trade-off between the speed and accuracy on the AP of the outdoor dataset at 0.5 IOU metric.

TABLE VII: Comparing the results in [25] with the results of our YOLOv5s-Face model on the Aachen dataset. Precision and recall scores are given for an IoU of 0.5.

Model	VJ	LBP	HOG	DPM	PICO	YOLOv5s-Face
Precision	0.93	0.97	1.00	1.00	0.90	1.00
Recall	0.98	0.92	0.93	0.98	1.00	0.99

Fig. 5. The YOLOv5n model with 1.76M parameters is the fastest, processing one image with a resolution of 512×384 within 6.16 ms. At the same time, this model is the most accurate on the uncontrolled outdoor dataset. The YOLOv5l6-Face model with 113.2M parameters is the slowest, processing one image within 17.29 ms. However, it is less accurate than many smaller models. The reason might be that the capacity of our dataset is enough for training smaller models but not for the larger ones.

It is important to note that the accurate results on the Aachen dataset illustrate the efficacy of our dataset. The reason is that the Aachen dataset was not used to train and validate the models. Therefore, the models 'saw' the Aachen dataset only during the testing step. In comparison, Kopaczka et al. [25] trained five different machine learning-based face detection algorithms on the Aachen dataset using the leave-one-subset-out cross-validation method. The algorithms considered are the Haar cascade classifier (Viola-Jones (VJ)) [4], Haar cascade classifier with local binary patterns (LBP) [48], HOG [5], deformable parts model (DPM) [49], and pixel intensity comparisons organized in decision trees (PICO) [50]. The performance of the tested algorithms in [25] and our best model YOLOv5m6-Face is given in Table VII. The best result in [25] was obtained using the DPM, achieving 1.00 precision and 0.98 recall scores for an IoU of 0.5. In contrast, the YOLOv5s-Face model achieves 1.00 precision and 0.99 recall scores on the Aachen dataset for an IoU of 0.5, using only the TFW dataset for training and validation. It shows that our dataset is large and diverse enough to train deep learning models to produce accurate results on an unseen dataset..

The examples of the predicted bounding boxes and facial landmarks on randomly selected images from the test set are shown in Fig. 6. As can be seen from the figure, the model can accurately detect faces of varying sizes, head poses, facial expressions, face masks, and glasses. Despite its many advantages, our dataset has some limitations. For instance, the number of unique subjects who participated in the unconstrained conditions is much smaller than in the controlled environment. In addition, the visual pairs are only available for the images collected in controlled and semi-controlled indoor environments.

VII. CONCLUSION

Thermal cameras have unique properties, such as insensitivity to illumination, the ability to measure the temperature without contact, and the capability to extract unique physiological patterns from the human face. Therefore, they are gaining popularity in many areas of computer vision, such as face recognition, vital sign recognition, and emotion recognition. Face detection and localization of facial landmarks are the

Fig. 6: Examples of the detected bounding boxes, facial landmarks, and corresponding confidence scores predicted by the YOLOv5s-Face model on randomly selected images from the test of a) C-Indoor, b) S-Indoor, c) U-Outdoor, and d) Aachen dataset.

fundamental steps in these tasks. There are many thermal face datasets to train deep learning models. However, most datasets have been collected in controlled environments and provided without facial bounding boxes and landmarks.

To address this problem, we presented the TFW dataset that was collected in both controlled and uncontrolled environments. The dataset contains 9,982 thermal images of 147 individuals. We thoroughly annotated the dataset with facial bounding boxes and five-point facial landmarks. To the best of our knowledge, it is the first annotated thermal face dataset that covers both controlled and uncontrolled environments. As

a baseline, we trained the YOLOv5 object detection model and its modified version of the YOLO5Face face detection model on our dataset.

The experimental results show that both models achieve high accuracy. However, the additional use of facial landmarks as a supervision signal provided more accurate results for large faces. In addition to our test set, we evaluated the models on the external Aachen thermal face dataset. The models achieved accurate results on this dataset, highlighting the efficacy of our dataset. However, the accuracy on the outdoor dataset is noticeably lower than on the indoor and

Aachen datasets, leaving room for improvement. To reinforce research in thermal face analysis, we have made the source code, dataset, annotations, and pre-trained models publicly available. In our future work, we will focus on improving the accuracy of our models in detecting faces with masks.

REFERENCES

- [1] G. Jocher. (2020) Yolov5. Last accessed on 2021-08-25: "https://github.com/ultralytics/yolov5".
- [2] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why reinventing a face detector," *arXiv preprint arXiv:2105.12931*, 2021.
- [3] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof, "A thermal infrared face database with facial landmarks and emotion labels," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 5, pp. 1389–1401, 2019.
- [4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. 1–1.
- [5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>
- [7] S. Minaee, P. Luo, Z. Lin, and K. Bowyer, "Going deeper into face detection: A survey," 2021.
- [8] 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*, vol. 29, no. 4, pp. 613–626, 2007.
- [9] S. Kumar, A. S. M. Iftikhar, M. Goebel, T. Bullock *et al.*, "StressNet: Detecting stress in thermal videos," in *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 998–1008.
- [10] B. A. Rajoub and R. Zwiggelaar, "Thermal facial analysis for deception detection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 6, pp. 1015–1023, 2014.
- [11] B. R. Nhan and T. Chau, "Classifying affective states using thermal infrared imaging of the human face," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 979–987, 2010.
- [12] D. Perpetuini, C. Filippini, D. Cardone, and A. Merla, "An overview of thermal infrared imaging-based screenings during pandemic emergencies," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, 2021. [Online]. Available: <https://www.mdpi.com/1660-4601/18/6/3286>
- [13] K. Mallat and J.-L. Dugelay, "A benchmark database of visible and thermal paired face images across multiple variations," in *Proc. of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2018, pp. 1–5.
- [14] B. Abidi, "IRIS thermal/visible face database. IEEE OTCBVS WS series bench," <http://vcip-okstate.org/pbvs/bench/>, accessed: 2021-07-02.
- [15] H. Chang, H. Harishwaran, M. Yi, A. Koschan, B. Abidi, and M. Abidi, "An indoor and outdoor, multimodal, multispectral and multi-illuminant database for face recognition," in *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 54–54.
- [16] S. Wang, Z. Liu, S. Lv, Y. Lv *et al.*, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.
- [17] H. Nguyen, K. Kotani, F. Chen, and B. Le, "A thermal facial emotion database and its analysis," in *Image and Video Technology*, R. Klette, M. Rivera, and S. Satoh, Eds. Springer, Berlin Heidelberg, 2014, pp. 397–408.
- [18] V. Espinosa-Duró, M. Faundez-Zanuy, and J. Mekyska, "A new face database simultaneously acquired in visible, near-infrared and thermal spectrums," *Cognitive Computation*, vol. 5, pp. 119–135, 03 2013.
- [19] R. Ghiass, H. Bendada, and X. Maldague, "Université Laval face motion and time-lapse video database (UL-FMTV)," 01 2018.
- [20] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol, "SpeakingFaces: A large-scale multimodal dataset of voice commands with visual and thermal face streams," *Sensors*, vol. 21, no. 10, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/10/3465>
- [21] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman *et al.*, "A large-scale, time-synchronized visible and thermal face dataset," in *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1559–1568.
- [22] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 162–182, 2007. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1077314206001834>
- [23] Z. Wu, N. Fuller, D. Thériault, and M. Betke, "A thermal infrared video benchmark for visual analysis," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 201–208.
- [24] Y. K. Cheong, V. V. Yap, and H. Nisar, "A novel face detection algorithm using thermal imaging," in *Proc. of the IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, 2014, pp. 208–213.
- [25] M. Kopaczka, J. Nestler, and D. Merhof, "Face detection in thermal infrared images: A comparison of algorithm-and machine-learning-based approaches," in *Proc. of the International Conference on Advanced Concepts for Intelligent Vision Systems*, 2017, pp. 518–529.
- [26] T. Vuković, R. Petrović, M. Pavlović, and S. Stanković, "Thermal image degradation influence on R-CNN face detection performance," in *Proc. of the Telecommunications Forum (TELFOR)*, 2019, pp. 1–4.
- [27] D. Li, X. Zhu, X. Chen, D. Tian, X. Hu, and G. Qin, "Thermal imaging face detection based on transfer learning," in *Proc. of the International Conference on Smart Grid and Electrical Automation (ICSGEA)*, 2021, pp. 263–266.
- [28] M. Kowalski, A. Grudzień, and W. Ciurapiński, "Detection of human faces in thermal infrared images," *Metrology and Measurement Systems*, vol. 28, 2021.
- [29] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.
- [30] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Fine-grained evaluation on face detection in the wild," in *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition*, vol. 1, 2015, pp. 1–7.
- [31] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [32] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211.
- [33] M. Kopaczka, K. Acar, and D. Merhof, "Robust facial landmark detection and face tracking in thermal infrared images using active appearance models," in *Proc. of the International Conference on Computer Vision Theory and Applications (VISAPP)*, 2016, pp. 150–158.
- [34] W.-T. Chu and Y.-H. Liu, "Thermal facial landmark detection by deep multi-task learning," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–6.
- [35] J. Keong, X. Dong, Z. Jin, K. Mallat, and J.-L. Dugelay, "Multi-spectral facial landmark detection," in *Proc. of the IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, pp. 1–6.
- [36] K. Wada, "labelme: Image Polygonal Annotation with Python," <https://github.com/wkentaro/labelme>, 2016.
- [37] O. Celiktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP Journal on Image and Video Processing*, vol. 2013, pp. 1–27, 2013.
- [38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
- [39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.
- [40] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
- [41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

- [42] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587.
- [43] R. Girshick, "Fast R-CNN," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
- [44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [45] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "TinaFace: Strong but simple baseline for face detection," *ArXiv*, vol. abs/2011.13183, 2020.
- [46] Y. Liu, F. Wang, B. Sun, and H. Li, "MogFace: Rethinking scale augmentation on the face detector," *ArXiv*, vol. abs/2103.11139, 2021.
- [47] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick *et al.*, "Microsoft COCO: Common objects in context," 2015.
- [48] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning multi-scale block local binary patterns for face recognition," in *Advances in Biometrics*, S.-W. Lee and S. Z. Li, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 828–837.
- [49] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.
- [50] N. Markuš, M. Frljak, I. S. Pandžić, J. Ahlberg, and R. Forchheimer, "Object detection with pixel intensity comparisons organized in decision trees," 2014.

Askat Kuzdeuov (M'19) received his B.S. in automation and control engineering from Almaty University of Power Engineering and Telecommunication, Kazakhstan, in 2014. From 2014 to 2019, he was a software engineer in the oil & gas industry, participating in more than thirty local and international projects. In 2019, he received M.S. in robotics from Nazarbayev University, Nur-Sultan, Kazakhstan. He is currently a data scientist with the Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University. His research in-

terests include computer vision, machine learning, and robotics.

Dana Aubakirova is a senior student in mathematics with a focus on machine learning and statistics at Nazarbayev University. Currently, she is a research assistant at the Institute of Smart Systems and Artificial Intelligence (ISSAI) in Nazarbayev University focusing on thermal face and facial landmark detection. Her research interests include computer vision, deep learning, and human-robot interaction.

Darina Koishigarina is an undergraduate student in computer science at Nazarbayev University, Nur-Sultan, Kazakhstan. She is currently a research assistant at the Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University. Her research interests are computer vision, deep learning, and statistical learning.

Huseyin Atakan Varol (M'09, SM'16) received his B.S. degree in mechatronics engineering from Sabanci University, Istanbul, Turkey, in 2005, and M.S. and Ph.D. degrees in electrical engineering from Vanderbilt University, Nashville, TN, USA, in 2007 and 2009, respectively. He was a postdoctoral research associate and then a research assistant professor at Vanderbilt University from 2009 to 2011. In 2011, he joined the faculty of Nazarbayev University, Nur-Sultan, Kazakhstan, where he currently chairs the Department of Robotics and Mechatronics and directs the Institute of Smart Systems and Artificial Intelligence (ISSAI). His research interests include soft robotics, machine learning, computer vision, sensor fusion, and tensegrity. He has published over 80 technical papers on related topics in international journals and conferences.

Dr. Varol was a finalist for the KUKA Innovation Award in 2014. He was also the recipient of the IEEE International Conference on Rehabilitation Robotics Best Paper Award in 2009, and the IEEE Engineering in Medicine and Biology Society Outstanding Paper Award in 2013. He served as Technical and Associate Editor for prominent journals and conferences.