This is a pre-print of the original paper accepted for presentation at the International Joint Conference on Biometrics (IJCB) 2024.

Watchlist Challenge: 3rd Open-set Face Detection and Identification

F. Kasım[a], T. E. Boult[b], R. Mora[c], B. Biesseck[d], R. Ribeiro[e], J. Schlueter[f], T. Repák[g], R. Vareto[c], D. Menotti[d], W. R. Schwartz[c], and M. Günther[a]

[a]*University of Zurich,* [b]*University of Colorado Colorado Springs,* [c]*Federal University of Minas Gerais,* [d]*Federal University of Paraná,* [e]*Federal Police of Brazil,* [f]*DERMALOG Identification Systems GmbH*

Abstract

In the current landscape of biometrics and surveillance, the ability to accurately recognize faces in uncontrolled settings is paramount. The Watchlist Challenge addresses this critical need by focusing on face detection and open-set identification in real-world surveillance scenarios. This paper presents a comprehensive evaluation of participating algorithms, using the enhanced UnConstrained College Students (UCCS) dataset with new evaluation protocols. In total, four participants submitted four face detection and nine open-set face recognition systems. The evaluation demonstrates that while detection capabilities are generally robust, closed-set identification performance varies significantly, with models pre-trained on large-scale datasets showing superior performance. However, open-set scenarios require further improvement, especially at higher true positive identification rates, i.e.*, lower thresholds.*

1. Introduction

In the evolving field of biometrics, face recognition technology stands as a cornerstone for security and surveillance systems worldwide. Surveillance systems, particularly in uncontrolled spaces, frequently encounter significant challenges such as blurry, partially occluded, or poorly illuminated facial images. In real-world use, the effectiveness of face recognition systems hinges on their ability to perform well under these tough conditions, which starkly differ from the controlled settings that most research focuses on and that often feature high-quality, cooperative subjects, typically containing only celebrities. Furthermore, the unpredictable nature of real-world surveillance necessitates a departure from traditional closed-set environments, where the gallery of subjects is identical to those employed for probing. Instead, the focus shifts to the watchlist problem, where the objective is to identify a few individuals listed on a watchlist while ignoring others (unknowns) [13]. Such watchlists can be used for identifying missing people, prohibition of unauthorized entry, or capturing criminals. Especially the latter scenario raises significant risks, including the wrongful targeting of innocent people [18], increased operational costs, and potential liabilities for security staff. These challenges underscore the critical need for solutions capable of reliable operation in open-set environments.

Sets	Images	Faces	Known	Unknown
Watchlist	—	10000	10000	—
Validation	7584	17689	9396	8293
Test	20534	57368	31512	25856

Table 1. LABEL DISTRIBUTION. *Distribution of images, faces, known and unknown labels are provided for all sets. The watchlist comprises cropped face regions corresponding to 1000 distinct identities, as shown in Fig. 1(b).*

In our challenge involving face detection and open-set identification, we exploit the UnConstrained College Students (UCCS) dataset introduced by Sapkota and Boult in 2013 [30], and substantially increased in size [14, 15] and label quality (see Supplemental Material). Specifically, we responded to previous criticism[1] by re-encoding the images to remove EXIF file information revealing details on the dataset collection, which itself was covered by IRB approval. This dataset is particularly well-suited to reflecting the surveillance system challenges discussed earlier, making it an ideal tool for addressing the critical issues encountered in real-world settings for the watchlist problem. In this dataset, individuals are typically unaware that their images are being recorded, which mirrors the non-cooperative behavior found in live surveillance scenarios and adds to the complexity of identifying watchlist subjects. One unique property of this dataset is that images are captured across different weather conditions, including sun with strong cast shadows, but also rainy and snowy conditions that severely influence imaging conditions. Additionally, the challenge

[1] https://exposing.ai/uccs

(a) Example Images

(b) Watchlist

Figure 1. EXAMPLE IMAGES AND WATCHLIST. *(a) shows two images with their annotations from the new version of UCCS dataset, including occlusions, different angles, and instances of significant blur. Faces marked with the same color indicate the same identity, whereas white boxes denote unknown subjects. (b) displays cropped faces in the watchlist, including 5 facial landmarks.*

incorporates face detection tasks where detectors may erroneously select regions of the background as faces. Along with a large pool of faces of unknown/innocent subjects that should not induce false identifications, these background detections are crucial to address, as they must also be treated as unknown by face recognition algorithms to provide a complete benchmark for evaluation. By focusing on these aspects, our challenge aims to foster research and development in face detection and recognition technologies, which are increasingly vital in today's surveillance applications.

This challenge has been conducted twice in previous research [14, 15]. Both instances yielded satisfactory outcomes in evaluating the detection capabilities and closed-set recognition of algorithms. However, open-set face recognition, which entails recognizing unknown faces and dealing with misdetections, remains a significant unresolved challenge. This complexity is particularly evident when probe faces are captured on different days than the gallery. Additionally, previous evaluation protocols displayed inherent biases as both enrollment and probe data were randomly selected from the same set, leading to unrealistic *same condition* matches. Past competitions also revealed numerous inaccuracies in the labels of some identities. To address these issues, the evaluation protocol has been revised to minimize the temporal overlap between gallery and probe data, especially in the test set. Simultaneously, the dataset underwent a significant cleanup, employing a combination of semi-automated and manual methods.

The UCCS Watchlist Challenge[2] is structured into two distinct segments: (I) face detection and (II) open-set face recognition. Participants contributed by submitting their results for these specialized tasks. In Part I, the challenge is to detect all faces within the UCCS images, irrespective of the identity labels, ensuring comprehensive coverage of face detection capabilities. Part II requires participants to enroll a set of gallery identities and then compute the similarities between each detected face (including various false positive detections) from Part I and each identity in the watchlist. These similarity scores are critical, as they determine which faces match the watchlist identities, while effectively ignoring unknown/unimportant faces.

[2] https://www.ifi.uzh.ch/en/aiml/challenge.html

2. Dataset and Protocol

The UCCS dataset was collected over several months on the private premises of the University of Colorado Colorado Springs. Two exemplary images from the UCCS dataset are presented in Fig. 1(a), illustrating the diversity in facial orientations, occlusions, and blur. In response to previously identified issues with incorrectly labeled identities [14, 15], a systematic data cleaning process was implemented, straightening detection, missing labels, as well as intra-class and inter-class label issues. Further details on the original UCCS dataset and these cleaning procedures can be found in the Supplemental Material, as well as a comparison to related surveillance datasets SCface [12], PaSC [1], IJB-S [20], DroneSurf [21] and BRIAR [3].

2.1. Defining new Evaluation Protocols

In contrast to prior invocations of this competition [14, 15], this year's challenge introduces a separate watchlist, which comprises cropped expanded face regions from the dataset, eliminating the need for a traditional training set. Ten faces per watchlist identity were extracted from the images, and subsequently excluded from part II of our evaluation. Similarly to previous encounters of this competition, we did not include all labeled identities into our watchlist, but left some of them to be unknown. The selection of the watchlist prioritized identities that contain high-quality faces and appear in multiple sequences, while 10 gallery

faces were extracted from the same sequence only. These annotations of the selected faces include information on the positions of five automatically detected facial landmarks, including the eyes, nose, and mouth. Following these constraints, a selection process resulted in choosing precisely 1000 different identities that are treated as known subjects in our watchlist. About 40% of these identities appear over two or more days. Fig. 1(b) illustrates different gallery faces and their annotations originating from two identities.

The final version of the UCCS dataset consists of more than 85'000 faces in total. The distribution of labeled faces in our dataset can be found in Tab. 1. Following the completion of the gallery, we split up the images into validation and test sets. The validation set includes annotated images with lists of bounding boxes, each labeled either with an integral gallery identity label or with the unknown label -1. In contrast, the test set consists solely of raw images with anonymized filenames and no annotations, challenging participants to detect faces and provide similarity scores for each bounding box against all watchlist subjects. To enhance realism and reduce the potential for biased *same day* matches, the test set includes images from different sequences of the watchlist identities, if available, whereas the validation set predominantly uses faces from the same sequences. Ultimately, 996 known identities appear in the test set, compared to 932 in the validation set. In both sets, about half of the faces are categorized as *unknown* to emulate real-world scenarios, including the remaining subjects that are not enrolled in the gallery, and faces that are left with the unknown label in our dataset.

3. Challenge Participants

Participants were invited to contribute summaries of their algorithms. Together with baseline algorithms, they are presented in the order of submission and tagged with their respective institutions ($^a - ^g$, *cf*. list of authors).

3.1. Face Detection

MTCNN-Baseline: The baseline face detector simply uses the pre-trained MTCNN [34], with its Pytorch implementation.[3] Since the detector is not optimized for blurry, occluded, or full-profile faces, we had to lower the three detection thresholds to (0.2, 0.2, 0.2), which ended up in detecting most faces, but provides many false positive detections. Our implementation can be downloaded from PyPI.[4]

RetinaFace^{c,d,e}: The multi-task face detector [5] is designed to identify face bounding boxes and five key facial landmarks. This detector employs a feature pyramid and anchor boxes in its pipeline. It is trained on the WIDER FACE dataset [33] using a multi-task loss function. The model was used at multiple scales with image factor sizes of $(0.2, 0.5, 1.0)$, a confidence threshold of 0.3, and a Non-Maximum Suppression (NMS) IoU of 0.075.

F3Y640S/F3Y640L^f: The DERMALOG Face SDK[5] implements the entire proprietary facial recognition system, including face detection, landmark detection, and feature extraction. The F3Y640S and F3Y640L face detection models are based on the YOLOX architecture [11]. This single-shot architecture leverages spatial pyramid pooling, a feature pyramid, and decoupled heads. The F3Y640S model is less complex, featuring reduced network depth and fewer parameters compared to the F3Y640L. Face detections by these models are filtered based on the confidence scores they generate. Training for these systems has been performed using both publicly available and commercially usable datasets, along with additional internal data.

3.2. Face Identification

MagFace-Baseline: The baseline feature extractor[6] utilizes the MagFace model [27], which is pre-trained on the MS1MV2 dataset [16, 6] using an iResNet-100 backbone [9]. MagFace aims to increase the inter-class distance while maintaining a cone-like structure within each class to ensure ambiguous samples are pushed toward the origin and away from class centers. Faces detected by the baseline detector are aligned based on facial landmarks [27]. Each face is represented by a 512-dimensional embedding. The enrollment averages embeddings from 10 faces per subject. Probe faces are compared to the gallery via cosine similarity.

AdaFace^{c,d,e}: ResNet-100 [17] is trained on the MS1MV3 [16] dataset with the AdaFace [22] loss function, which introduces an angular margin loss that utilizes image quality to scale the gradient during training and adjusts margins for different classes based on their recognition difficulty. Image quality is assessed using the norm of the 512-dimensinoal embedding. The process for enrollment and probing aligns with the baseline method.

MEL^{c,d,e}: This method [32] integrates the principles of Maximal Entropy and Objectosphere Loss [8] to enhance face recognition capabilities. Maximal Entropy increases the entropy in feature representations, effectively reducing certainty about any unknown category to distinguish more accurately between known and unknown faces. Concurrently, Objectosphere loss improves the separation of known and unknown classes within the feature space. To train this MEL model, it receives the AdaFace embeddings as inputs, projects them to a new space where the unknown class is more compacted and separated from known person classes. This training includes the known gallery embeddings and an unknown class that is composed of 1600 unknown subjects extracted from the UCCS validation set.

[3] https://pypi.org/project/facenet-pytorch
[4] https://pypi.org/project/challenge.uccs
[5] https://www.dermalog.com/products/software/face-recognition
[6] https://github.com/IrvingMeng/MagFace

Enrollment and scoring follow the baseline protocol.

F3Y640S/F3Y640L[f]: Before feature extraction, a novel facial landmark detection[5] assesses face quality by evaluating the occlusion of each landmark, setting thresholds to exclude faces that could yield poor matches. Features are then extracted using the same face recognition model[5] for faces detected separately by the F3Y640S and F3Y640L detectors. All enrollment faces for each identity are encoded into templates and used during the matching process. The query template (probe face) is compared against the templates of all enrolled identities. Only the highest matching score achieved between the query template and each enrolled identity's templates is reported in the score file.

DaliFace[a,b]: DaliFace [28] uses ResNet-100 [17] as its backbone, trained on the WebFace4M [36] dataset with the AdaFace [22] loss function. It incorporates distortion augmentations during its training to simulate real-world distortions like motion blur and atmospheric turbulence, maintaining feature-level invariance against severe image quality degradations. Additionally, an adaptive weighting schedule adjusts the intensity of these distortions progressively, allowing the model to adapt gradually without overwhelming initial learning stages. DaliFace processes the detection results from JointFaceDetectID, which provides only bounding boxes; therefore, the faces are resized for inference. Enrollment and scoring follows the baseline.

EnsembET/EnsembETMN/ET[g]: EnsembET is an ensemble model composed of three identical EVA-02-Ti transformer-based visual representation models [10], each utilizing a different loss function or training strategy. All models are trained on facial images resized to 224×224, sourced from the LFW [19], CelebA [25], and the UCCS validation set. The first model in the ensemble is trained from scratch with triplet loss [31] using Euclidean distance, the second with cosine distance. The third starts with pre-trained weights and is fine-tuned using triplet loss with cosine distance. Similarly, the EnsembETMN model is an ensemble setup that changes only the first model with MobileNetV2 [29], while keeping the same training strategies and loss functions as EnsembET. Finally, the ET model is a single instance of EVA-02-Ti that uses pre-trained weights and is fine-tuned with triplet loss using cosine distance. The enrollment for the watchlist is similar to the baseline. For ensemble models, the final scores are derived by averaging scores from individual components of the ensemble.

3.3. Joint Face Detection and Identification

JointFaceDetectID[a,b]: The JointFaceDetectID model represents a novel approach that integrates face detection and recognition tasks into a single model. This unified approach uses the same feature space for both tasks, potentially improving accuracy by leveraging their interdependencies. Designed as a single-stage, anchor-based detector, the model's architecture includes iResNet-50 [9] backbone, a feature pyramid network [23] neck, and heads that consist of three branches. Two detection branches handle anchor classification and provide bounding boxes, employing focal loss [24] and distance-IoU loss [35]. The last branch generates 256-dimensional embeddings to represent faces, utilizing ArcFace loss [6] to enhance the discriminative capabilities of the embeddings. Furthermore, Entropic Open-Set loss [7] is used to manage the uncertainty associated with unknown faces and high-confidence false postive detections. The model was initially trained on the IJB-C [26] and PaSC [1] datasets, which collectively contain 3966 identities, before being fine-tuned on the UCCS gallery and validation sets, using faces labeled as -1 as negatives.

During inference, the model generates multiple anchors for each face. To finalize the embedding, a confidence threshold of 0.5 filters out low-confidence boxes, and NMS with a 0.4 IoU threshold removes highly overlapping boxes. This process ensures that the remaining bounding box accurately represents the face. The enrollment and scoring processes follow the same methodology as the baseline.

4. Evaluation

For evaluating face detection participants submitted bounding boxes for detected faces, each accompanied by a confidence score. For face recognition, participants also provide a similarity score for each watchlist subject associated with the detected faces. Since the faces on the watchlist are cropped from the original images, those are omitted from the evaluation process.

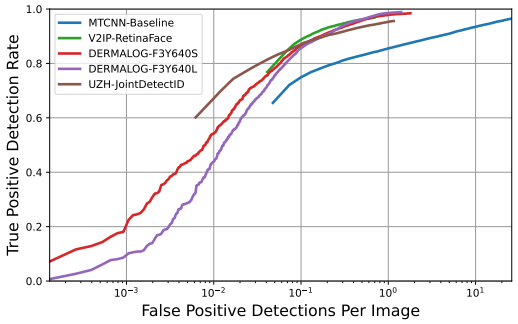
Participants were provided with the challenge's evaluation scripts and ground-truths[4] to facilitate the evaluation on the validation set. Here, we use the exact same evaluation framework on the test set.

4.1. Face Detection

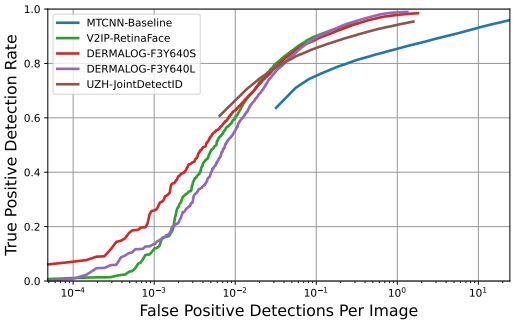
To assess the accuracy of each bounding box, the standard Jaccard index, also known as the Intersection Over Union (IoU), is used to compare the detected bounding box with the ground truth. In our evaluation, we accept all bounding boxes with an IoU threshold, $IoU \geq 0.2$. This threshold is selected to compensate for potential inaccuracies[7] in the ground truth boxes, allowing for the inclusion of loosely matched detections.

Face detection evaluation is conducted using an adaptation of the Free-response Receiver Operating Characteristic (FROC) curve [2]. Specifically, confidence scores c are classified as C^+ when $IoU \geq 0.2$ and as C^- when $IoU < 0.2$, based on the overlap between the ground-truth and detected bounding boxes. The True Positive Detection Rate (TPDR) is calculated using labeled faces **M** from

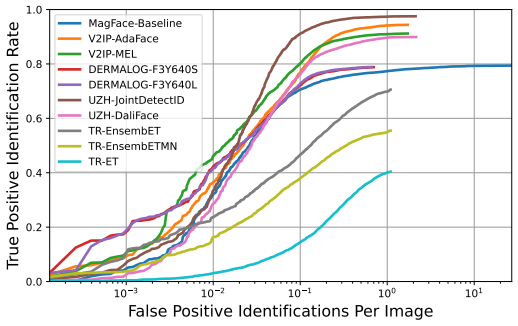
[7] There were a few faces for which the landmark detector failed in the annotation process – for these we kept the original bounding boxes [14].



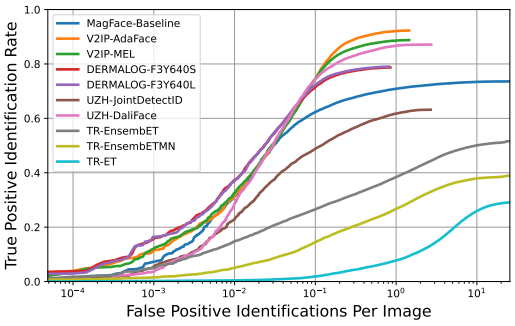
(a) Detection on Validation Set



(b) Detection on Test Set



(c) Identification on Validation Set



(d) Identification on Test Set

Figure 2. FACE DETECTION AND RECOGNITION EVALUATION. *A Free-response Receiver Operating Characteristic (FROC) curve is shown for the (a) validation and (b) test set. The horizontal axis includes the number of false positive detections normalized by the number of images, while the vertical axis outlines the relative number of true positive detections of faces. Open-set ROC curve at rank 1 is shown for (c) validation and (d) test set. The horizontal axis includes the number of false positive identifications normalized by the number of images, while the vertical axis outlines the relative number of correctly identified faces.*

probe samples if the confidence score exceeds the operating threshold θ:

$$\text{TPDR}(\theta) = \frac{1}{|\mathbf{M}|}\left|\{c \mid c \in C^{+} \wedge c \geq \theta\}\right| \qquad (1)$$

Instead of calculating False Positive Detection Rates on the horizontal axis, this method measures the number of False Positive Detections Per Image (FPDPI) [15], calculated by dividing total false positive detections by *the number of probe images* $\mathbf{I}$. This approach accounts for the varying number of false positive detections that a face detector might produce across different images. A false positive detection occurs when the confidence score of a misdetection is larger than threshold θ:

$$\text{FPDPI}(\theta) = \frac{1}{|\mathbf{I}|}\left|\{c \mid c \in C^{-} \wedge c \geq \theta\}\right| \qquad (2)$$

By varying the threshold θ, the FROC plots TPDR over FPDPI. Finally, a single score to compare across algorithms is computed by obtaining the thresholds θ_i that result in FPDPI $= 10^i$:

$$\sum \text{TPDR} = \sum_{i \in \{-3,-2,-1,0\}} \text{TPDR}(\theta_i) \quad \theta_i = \text{FPDPI}^{-1}(10^i) \qquad (3)$$

We add 0 when a low FPDPI is reached by no threshold θ.

Fig. 2(a) and Fig. 2(b) display the participants' detection results on both sets, while Tab. 2(a) provides detailed information about the rankings on the test set, evaluated via (3). All models demonstrate consistent behavior across both the validation and test sets, except for V2IP-RetinaFace. Notably, V2IP-RetinaFace does not register any TPDR at the lower FDPIs on the validation set, yet it does show performance on the test set. This discrepancy could be due to V2IP-RetinaFace being particularly sensitive to the specific characteristics or quality of the data in the test set. DERMALOG-F3Y640S is the best-performing detector with a $\sum$TPDR score of 2.752. It shows the highest by far TPDR at the FPDPI 10^{-3}, indicating its capability in environments where minimizing false positive detections is crucial. Additionally, it maintains consistently high performance across other FPDPI thresholds. DERMALOG-F3Y640L and V2IP-RetinaFace also perform well, particu-

Table 2. RANKING FOR DETECTION AND RECOGNITION TASKS. *(a) shows the detection ranking, whereas the identification ranking is indicated in (b).*

(a) Detection

Method	@FPDPI				∑ TPDR
	10^{-3}	10^{-2}	10^{-1}	10^{0}	
DERMALOG-F3Y640S	**0.2585**	*0.6271*	0.8882	*0.9782*	**2.752**
DERMALOG-F3Y640L	*0.1350*	0.5530	*0.8958*	**0.9875**	*2.5713*
V2IP-RetinaFace	0.1142	0.5993	**0.9011**	0.9469	2.5615
UZH-JointDetectID	-	**0.6583**	0.8556	0.9409	2.4548
MTCNN-Baseline	-	-	0.7424	0.8519	1.5943

(b) Identification

Method	@FPIPI				∑ TPIR
	10^{-3}	10^{-2}	10^{-1}	10^{0}	
V2IP-AdaFace	0.1106	0.3157	0.7434	**0.9208**	**2.0905**
V2IP-MEL	0.1196	0.3261	**0.7457**	*0.8859*	*2.0773*
DERMALOG-F3Y640L	**0.1640**	*0.3695*	0.7226	0.7901	2.0462
DERMALOG-F3Y640S	*0.1543*	**0.3739**	0.7142	0.7871	2.0295
UZH-DaliFace	0.0361	0.2857	*0.7435*	0.8668	1.9321
MagFace-Baseline	0.0721	0.3135	0.6227	0.7086	1.7169
UZH-JointDetectID	0.0522	0.2275	0.4880	0.6175	1.3852
TR-EnsembET	0.0490	0.1479	0.2651	0.3840	0.8460
TR-EnsembETMN	0.0152	0.0503	0.1449	0.2671	0.4775
TR-ET	0.0019	0.0043	0.0186	0.0774	0.1022

larly at higher FPDPI thresholds (more lenient conditions), which suggests these models maintain a good balance between detecting existing faces and controlling false positives under less restrictive conditions. UZH-JointDetectID does not report data for the strictest threshold (10^{-3}), which might suggest a limitation in its ability to handle extremely low false positive detections. However, it performs exceptionally well at the 10^{-2} threshold, achieving the best TPDR among all methods, indicating its effectiveness in slightly less stringent conditions. To reach a similar TPDR as other methods, the MTCNN baseline records a notably higher number of false positive detections likely due to its overly permissive thresholds.

4.2. Face Identification

For identification, we rely on comparing a gallery template T_g and a probe face F_p via scoring function s. In this open-set context, a face recognition algorithm aims to achieve three objectives: First, for a probe face of a known identity, the corresponding gallery template T_{g^*} should show the highest similarity among all templates. Second, if the probe face belongs to an unknown identity, the similarities to all gallery templates should be low. Finally, any false positive *detections* should be treated as unknown. In our evaluation, we rely on our adaptation [15, 13] of the Open Receiver Operating Characteristic (O-ROC) curve. We split the probe faces into a set of *known* faces $\mathbf{K}$, as well as a set of unknown faces and false positive detections $\mathbf{U}$. We plot the True Positive Identification Rate (TPIR) over the False Positive Identifications Per Image (FPIPI):

$$\mathrm{TPIR}(\tau) = \frac{1}{|\mathbf{K}|}\left|\left\{F_p \in \mathbf{K} \mid \arg\max_{g\in G} s(T_g, F_p) = g^* \wedge s(T_{g^*}, F_p) \geq \tau\right\}\right| \quad (4)$$

$$\mathrm{FPIPI}(\tau) = \frac{1}{|\mathbf{I}|}\left|\left\{F_p \in \mathbf{U} \mid \max_{g\in G} s(T_g, F_p) \geq \tau\right\}\right| \quad (5)$$

Similarly to the face detection evaluation, we also define a single number for defining a ranking across algorithms:

$$\sum \mathrm{TPIR} = \sum_{i\in\{-3,-2,-1,0\}} \mathrm{TPIR}(\tau_i) \quad \tau_i = \mathrm{FPIPI}^{-1}(10^i) \quad (6)$$

Fig. 2(c) and Fig. 2(d) display the participants' recognition results on both sets, while Tab. 2(b) provides detailed information about the rankings on the test set, evaluated at four different FPIPI levels via (6). V2IP-AdaFace stands out as the top performer with the highest overall TPIR, showcasing robust capabilities across varying thresholds of false positive identification. In particular, this model performs exceptionally well (92%) at the most lenient threshold, demonstrating its adaptability in environments where higher rates of false identifications are permissible. DERMALOG-F3Y640L and DERMALOG-F3Y640S, on the other hand, show superior performance at the strictest thresholds, making them ideal for applications demanding high accuracy with minimal false positive identifications. Both V2IP-MEL and UZH-DaliFace exhibit commendable performances at higher tolerance levels for false identifications, successfully recognizing 86-88% (@FPIPI $= 1$) of watchlist subjects. Almost all models display consistent performance between the validation and test sets. One exception is UZH-JointDetectID, which excels on the validation set but shows a marked decrease in effectiveness on the test set, underlining potential overfitting issues. Furthermore, TR-EnsembET, TR-EnsembETMN, and TR-ET lag significantly behind the other models at all assessed thresholds, struggling to identify true positives accurately, which is likely caused by small number of identities in their training datasets. However, it is noteworthy that among these, the ensemble models (TR-EnsembET and TR-EnsembETMN) demonstrate better performance compared to TR-ET, indicating that the ensemble approach does offer advantages even among the lower-performing models.

5. Discussion

Additionally to the main results, we explore key aspects of the facial recognition challenge, including closed-set performance, threshold effects, and handling of unknowns. A detailed analysis outlining specific detection and identification failure cases, along with limitations and potential improvements, is available in the Supplemental Material.

Table 3. CLOSED-SET PERFORMANCE. *Closed-set performance of methods is shown for both tasks.*

Method	TPDR(%)	TPIR(%)	FPIPI
MagFace-Baseline	**98.81**	73.58	25.63
V2IP-AdaFace	98.45	**92.27**	1.449
V2IP-MEL		*88.77*	
DERMALOG-F3Y640S	83.30	78.71	0.848
DERMALOG-F3Y640L	83.62	79.01	0.813
UZH-JointDetectID	*98.80*	63.17	2.684
UZH-DaliFace		87.12	2.724
TR-EnsembET	**98.81**	51.59	25.63
TR-EnsembETMN		38.90	
TR-ET		29.07	

5.1. Closed-Set Performance

While not the main purpose of this challenge, it is worth looking into closed-set performances of the algorithms, which correspond to the right-most points in Fig. 2(b) and Fig. 2(d). These numbers are provided in Tab. 3. For the identification task, also the number of FPIPI corresponding to the highest TPIR is reported. While nearly all known faces are detected by most models' detectors at a rate close to 99% – with the exception of the DERMALOG detectors – TPIR varies significantly across algorithms. DERMALOG's models detect approximately 83% of known subjects, fewer than other models, due to their facial landmark detection step that assesses each landmark's quality, effectively eliminating low-quality faces to reduce false positives during the recognition phase. However, DERMALOG models still demonstrate commendable consistency, correctly identifying about 79% of the faces (83%), and they boast the lowest FPIPI of 0.84. Among the models, AdaFace stands out by correctly identifying 92% of the faces it detects (98%), showcasing the best performance at the second-lowest FPIPI. We can conclude that DERMALOG models are particularly suited for environments where minimizing false positives is important, whereas models like V2IP-AdaFace, V2IP-MEL, and UZH-DaliFace are better suited for settings where higher false identifications are acceptable due to their higher TPIR.

5.2. Analysis of Threshold

When evaluating detection and identification models via FROC and O-ROC, thresholds are estimated on the test set directly. This does not correspond to operation conditions where thresholds have to be determined before deployment [4]. To test this behavior, we determine detection thresholds θ_i in (3) and recognition thresholds τ_i via (6) on the validation set, and compute all metrics on the test set. In Fig. 3, we show the effects of the different ways of selecting the thresholds on the final evaluation. While for some methods, the thresholds from the validation set translate well to the results on the test set, i.e., the FPDPI and FPIPI do not change much, for other methods these numbers are less stable, resulting in a large performance difference. This highlights the need for more realistic evaluation metrics used in face detection and open-set face recognition tasks.

5.3. Analysis of the Unknown

Since our test data contains two different types of unknowns, *i.e.*, unknown faces $\mathbf{U}_{-1}$ and false positive detections $\mathbf{U}_{\text{FPD}}$, we investigate the behavior of the recognition systems on both types separately. Similar to [14, 15], we plot the Correct Unknown Rejection Rate (CURR) over the True Positive Identifications Per Image (TPIPI):

$$\text{TPIPI}(\tau) = \frac{1}{|\mathbf{I}|}\Big|\Big\{F_p \in \mathbf{K} \mid \arg\max_{g\in G} s(T_g, F_p) = g^* \wedge s(T_{g^*}, F_p) \geq \tau\Big\}\Big| \tag{7}$$

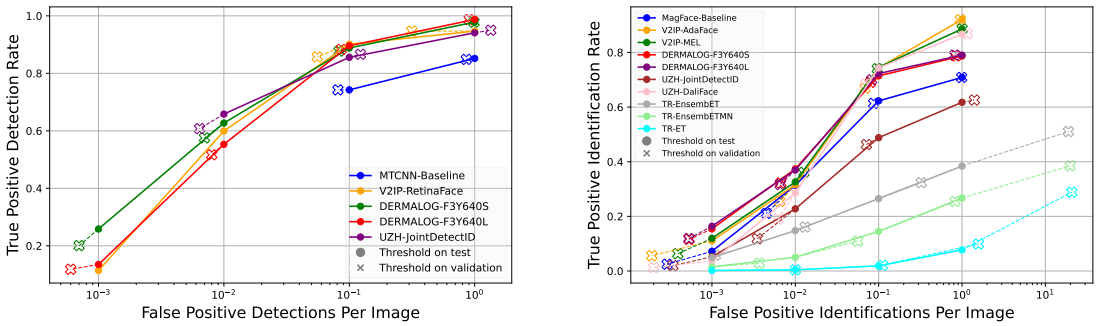
$$\text{CURR}(\tau) = \frac{1}{|\mathbf{U}_\bullet|}\Big|\Big\{F_p \in \mathbf{U}_\bullet \mid \max_{g\in G} s(T_g, F_p) < \tau\Big\}\Big| \tag{8}$$

with $\mathbf{U}_\bullet \in \{\mathbf{U}_{-1}, \mathbf{U}_{\text{FPD}}\}$. The CURR analysis for unknown subjects $\mathbf{U}_{-1}$ as shown in Fig. 4(a) reveals that most algorithms are prone to assign watchlist identity label to all unknown subjects at their low thresholds τ (high TPIPI). UZH-DaliFace, V2IP-AdaFace, and DERMALOG diverge from this trend, albeit still exhibiting very low CURR of about 2% at their highest TPIPI. This common issue might show potential problems with the labels of unknown subjects, which we analyze in the Supplemental Material. Among all models, UZH-DaliFace, V2IP-AdaFace, and DERMALOG consistently outperform others, maintaining better TPIPI across every CURR level. Similarly, they sustain a high CURR of 90% up to a TPIPI of 1.2 but experience a sharp decline in CURR like other all models as their rejection thresholds τ are lowered further.

In the analysis of rejection of false positive detections $\mathbf{U}_{\text{FPD}}$, UZH-DaliFace distinguishes itself by maintaining the highest CURR across all levels of TPIPI, as highlighted in Fig. 4(b). Impressively, it manages to avoid assigning identities to all background detections, holding a CURR of 36% even at the highest TPIPI. Similarly, MagFace-Baseline continues with high CURR, and then sharply starts declining the CURR as the rejection threshold decreases. V2IP-AdaFace and V2IP-MEL also perform commendably, with relatively stable CURR as TPIPI increases. Conversely, the TR series models – TR-EnsembET, TR-EnsembETMN, and TR-ET – display significantly lower performance across both unknown categories, indicating that enhancements in their algorithms could be necessary.

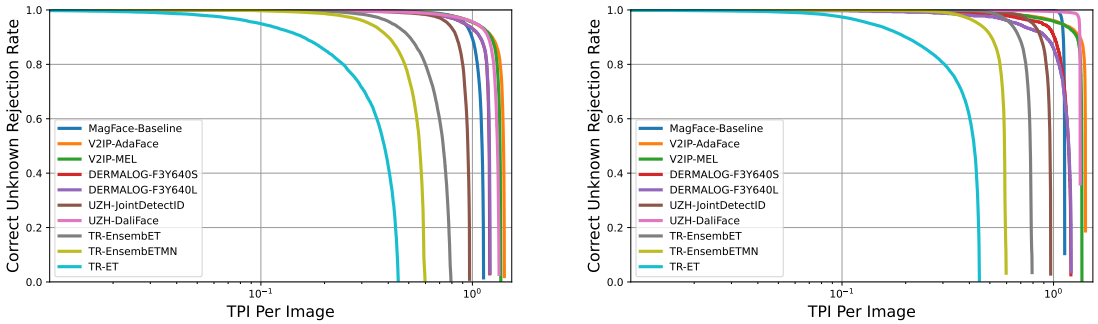
6. Conclusion

We present a comprehensive evaluation of the results from participants in the watchlist challenge, focusing on the



(a) Detection (b) Identification

Figure 3. THRESHOLD SELECTION. *We depict the effect of selecting the thresholds on the validation and test sets. In (a), we show differences in detection scores, while (b) highlights differences in identification performances.*



(a) Unknown Faces $\mathbf{U}_{-1}$ (b) False Positive Detections $\mathbf{U}_{\text{FPD}}$

Figure 4. REJECTION RATES BY TYPE. *Unknown samples are split into (a) unknown subjects and (b) false positive detections, illustrating the rate of correctly rejected samples (i.e., samples not identified as any known subject) across varying thresholds that are based on the number of correctly identified known subjects per image. X-axes are plotted in logarithmic scale.*

critical aspects of real-world surveillance scenarios where open-set face detection and recognition are pivotal. This challenge is designed to foster collaboration and establish an ideal benchmark for assessing the robustness and performance of facial recognition algorithms in surveillance settings, incorporating revised data and protocols to reflect more realistic conditions. Detection results generally meet expectations, even in challenging cases presented by the dataset. However, a handful of faces under extreme conditions are not detected by any model, highlighting areas for potential improvement in detection capabilities.

In terms of identification, some models excel at strict thresholds, making them suitable for applications where minimizing false positive identifications is vital. Conversely, other models perform exceptionally well at softer thresholds, achieving the highest TPIR. The selection of models can therefore be tailored based on system preferences and the specific security requirements of the deployment environment. The open-set performance of the models, particularly in terms of the Correct Unknown Rejection Rate (CURR) at lower thresholds, needs improvement. The results highlight that all models struggle to maintain high rejection rates as the threshold for true positive identifications decreases, indicating a crucial area for future research.

The analysis indicates that models pre-trained on large-scale datasets typically surpass others, highlighting the significant impact of extensive training on model performance. Two models that were fine-tuned to the UCCS validation set show promising capabilities. However, due to a limited number of identities in the training data, these designs currently exhibit poorer identification performance compared to those trained on more extensive datasets. It is anticipated that with further training on datasets containing a larger array of identities, the performances of those models increase.

Acknowledgements

This research was supported by the Student Abroad Program of the Republic of Türkiye Ministry of National Education. We are thankful for their financial support and dedication to fostering academic and professional growth.

References

[1] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013. 2, 4

[2] D. Chakraborty. Statistical power in observer-performance studies: comparison of the receiver operating characteristic and free-response methods in tasks involving localization. *Academic radiology*, 9(2):147–156, 2002. 4

[3] D. Cornett, J. Brogan, N. Barber, D. Aykac, S. Baird, N. Burchfield, C. Dukes, A. Duncan, R. Ferrell, J. Goddard, G. Jager, M. Larson, B. Murphy, C. Johnson, I. Shelley, N. Srinivas, B. Stockwell, L. Thompson, M. Yohe, R. Zhang, S. Dolvin, H. J. Santos-Villalobos, and D. S. Bolme. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2023. 2

[4] T. de Freitas Pereira, D. Schmidli, Y. Linghu, X. Zhang, S. Marcel, and M. Günther. Eight years of face recognition research: Reproducibility, achievements and open issues. *arXiv*, 2022. 7

[5] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[7] A. R. Dhamija, M. Günther, and T. Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 4

[8] A. R. Dhamija, M. Günther, and T. E. Boult. Improving deep network robustness to unknown inputs with objectosphere. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 3

[9] I. C. Duta, L. Liu, F. Zhu, and L. Shao. Improved Residual Networks for Image and Video Recognition. In *International Conference on Pattern Recognition (ICPR)*, 2021. 3, 4

[10] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 4

[11] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3

[12] M. Grgic, K. Delac, and S. Grgic. SCface - surveillance cameras face database. *Multimedia Tools and Applications*, 51(3), 2011. 2

[13] M. Günther, A. R. Dhamija, and T. E. Boult. Watchlist adaptation: Protecting the innocent. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2020. 1, 6

[14] M. Günther, P. Hu, C. Herrmann, C.-H. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler, et al. Unconstrained face detection and open-set face recognition challenge. In *International Joint Conference on Biometrics (IJCB)*, 2017. 1, 2, 4, 7

[15] M. Günther, W. Scheirer, and T. E. Boult. Open-set recognition challenge. Poster presented at the IAL Workshop, ECCV, 2018. 1, 2, 5, 6, 7

[16] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[18] K. Hill. Wrongfully accused by an algorithm. *New York Times*, June 2020. https://www.nytimes.com. 1

[19] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, Univ. of Massachusetts, Amherst, 2007. 4

[20] N. D. Kalka, B. Maze, J. A. Duncan, K. O'Connor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. IJB–S: IARPA Janus surveillance video benchmark. In *International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018. 2

[21] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. B. Sujit. DroneSURF: Benchmark dataset for drone-based face recognition. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2019. 2

[22] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *International Conference on Computer Vision (ICCV)*, 2017. 4

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015. 4

[26] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. IARPA Janus benchmark-C: Face dataset and protocol. In *International Conference on Biometrics (ICB)*, 2018. 4

[27] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. MagFace: A universal representation for face recognition and quality assessment. In *Conference on Computer Vision and Patter Recognition (CVPR)*, 2021. 3

[28] W. Robbins, G. Bertocco, and T. E. Boult. DaliID: Distortion-adaptive learned invariance for identification – a robust technique for face recognition and person re-identification. *IEEE Access*, 2024. 4

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[30] A. Sapkota and T. E. Boult. Large scale unconstrained open set face database. In *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013. 1

[31] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4

[32] R. H. Vareto, Y. Linghu, T. E. Boult, W. R. Schwartz, and M. Günther. Open-set face recognition with maximal entropy and objectosphere loss. *Image and Vision Computing*, 141, 2024. 3

[33] S. Yang, P. Luo, C. C. Loy, and X. Tang. WIDER FACE: A face detection benchmark. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[34] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *Signal Processing Letters*, 23(10), 2016. 3

[35] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *AAAI Conference on Artificial Intelligence*, 2020. 4

[36] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4