A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

# My Eyes Are Up Here: Promoting Focus on Uncovered Regions in Masked Face Recognition

Pedro C. Neto[1,2], Fadi Boutros[3,4], João Ribeiro Pinto[1,2], Mohsen Saffari[1,2],
Naser Damer[3,4] , Ana F. Sequeira [1], Jaime S. Cardoso[1,2]

**Abstract:** The recent Covid-19 pandemic and the fact that wearing masks in public is now mandatory in several countries, created challenges in the use of face recognition systems (FRS). In this work, we address the challenge of masked face recognition (MFR) and focus on evaluating the verification performance in FRS when verifying masked vs unmasked faces compared to verifying only unmasked faces. We propose a methodology that combines the traditional triplet loss and the mean squared error (MSE) intending to improve the robustness of an MFR system in the masked-unmasked comparison mode. The results obtained by our proposed method show improvements in a detailed step-wise ablation study. The conducted study showed significant performance gains induced by our proposed training paradigm and modified triplet loss on two evaluation databases.

**Keywords:** Face recognition, masked face recognition, Covid-19, triplet loss, vggface2.

## 1 Introduction

Computer vision tools have been successfully applied to face recognition (FR) in the past [SKP15]. New challenging conditions, such as the face occlusion caused by the use of face masks in public, mandatory during the SarsCov2 pandemic, raised limitations for well-performing and established FR methods. The pandemic has also stressed the importance of hygienic and contactless biometrics [Go21], such as FR. Recently, the National Institute of Standards and Technology (NIST), in the scope of the ongoing Face Recognition Vendor Test (FRVT), published a study on the effect of face masks on the performance of vendor's FR systems (FRVT -Part 6A). The NIST study concluded that the algorithm accuracy with masked faces declined substantially. The Department of Homeland Security has conducted an evaluation with similar goals, however on more realistic data[5]. They also observed the significant negative effect of wearing masks on the accuracy of automatic FR methods.

The lack of robustness of current systems to perform masked face recognition (MFR) fostered an interest in the research community to address this challenge [Di20; Ge20; Ho20; Li21]. Damer *et al*. [Da20] evaluated the verification performance drop in three

[1] INESC TEC, Porto, Portugal, pedro.d.carneiro@inesctec.pt
[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
[3] Fraunhofer Institute for Computer Graphics Research IGD, Germany
[4] TU Darmstadt, Germany
[5] https://mdtf.org/ Rally2020/Results2020

face biometric systems when verifying masked vs not-masked faces compared to verifying not-masked faces to each other. This study was extended [Da21a] to both synthetic and real masks, pointing out the questionable use of simulated masks to represent the real mask effect on face recognition. Furthermore, the performance of human inspectors in face verification has been shown to have a drop, consistent with automatic FR systems, when faces are masked [Da21b]. The effect of facial masks extends to other components of biometric systems as it has been shown to largely change the behaviour of face presentation attack detection [Fa21]. Recently, Boutros *et al*. [Bo21a] proposed a template unmasking approach that can be adapted on top of any face recognition network aiming at creating unmasked-like templates from masked faces by the proposed self-restrained triplet loss. Other initiatives were also incited by the lack of systems capable of handling this task, such as the "Competition on Masked Face Recognition" (IJCB-MFR-2021 [Bo21b]) and "The International Workshop on Face and Gesture Analysis for COVID-19" (FG4COVID19)[6].

The work proposed in this paper comprises the construction of a synthetic masked face dataset based on the VGGFace2 [Ca18] and proposes a solution to address the challenge of MFR. This solution is based on a proposed loss that combines the traditional triplet loss and the mean squared error (MSE) intending to improve robustness in the comparison between masked and unmasked samples.

The contributions of this work are: 1) A cascaded training paradigm that leverages the benefits of both a conventional identity classification learning in the first stage and the subsequent embedding optimization fine-tuning stage. 2) A specifically modified triplet loss function (for the embedding optimization) that incorporates a mean square error measurement to control the training process in a weighted manner. 3) A thorough ablation study on multiple databases (including our own created database), showing, in a step-wise manner, the benefit of our training paradigm and the specifically designed loss.

This paper is organised as follows. In this introductory section, we contextualise the challenge addressed within the related work and detail the contributions of the paper; and in the conclusion, we reflect on the findings and future work possibilities. In Section 2 we present the methodology used. Section 3 details the metrics, the datasets used (the creation of the synthetic face masks and the dataset with real masks), the implementation details, and finally, presents the results and its discussion.

## 2 Methodology

Performing the direct optimization of embedding predictions is often trickier than learning a model capable of performing classification. Hence, we propose a constrained triplet loss, specially crafted for masked face recognition (MFR), to be used after the classification optimization. Our approach redirects the focus of learned embeddings towards unmasked areas. In the embedding learning stage, the training focuses on comparing two images against a reference (anchor) and distinguishing between images of the same person (positives) and from a different person (negatives). Together, the anchor, the positive, and the

[6] https://fg4covid19.github.io/index.html

negative form a triplet. This will enable the model to capture the benefits of both strategies and therefore be more flexible to inter-class variations, as we will experimentally illustrate.

In this section, we start by describing the approach followed to train these models for classification. Afterwards, we expose our proposed change to the triplet loss that improves the representations learned by these models for MFR.

**Classification Training:** In our scenario, the number of classes and the universe of possible inputs is unknown once the model is deployed. Thus, supervised classification methods can only guide us up to a certain point. Nonetheless, they have been used as a technique to improve the convergence speed of the model for other tasks. Therefore, initially, the problem is approached as a closed-set recognition. Afterwards, the pre-trained model is fine-tuned towards learning meaningful embeddings.

The approach to train the classification model was based on minimizing the cross-entropy (CE). This loss function is frequently used for the classification of the input as a single class, which suits our use case since a picture can only belong to one subject. It attempts to minimize the confidence of the model on erroneous classes while maximizing its confidence in the correct class. The validation occurred after each epoch and it evaluated the accuracy of the model in the classification of masked pictures unseen during the training. And while these images were unseen, the network already knew the subject from past pictures. To separate validation and training sets, we followed an $80\%/20\%$ data split. This training process is similar to the one designed by Cao *et al*. [Ca18].

**Embedding Optimization:** Embedding optimization is a task that requires the network to learn the representations of the inputs instead of classifying them. This is no trivial task, and the hyperparameters search of this process has to be done carefully since this is an expensive task when compared to the use of classification losses [Yu20]. For this, the fully connected layer is removed, and another untrained embedding layer is added. Besides this last layer, all the weights of the network are frozen, and thus, further training does not update them. To train the embedding layer, which outputs an embedding vector with a size of 512, the triplet loss is used, based on Equation 1.

$$TripletLoss = \sum_{a,p,n} max(0, \alpha - ||x_a - x_n||_2^2 + ||x_a - x_p||_2^2) \tag{1}$$

$$x_i = W' \frac{\phi(l_i))}{||\phi(l_i)||_2} \tag{2}$$

Equation 2 describes the embedding layer added after the last convolutional layer. It receives as input the output of the convolutional layer (represented by $l$ on the equation) and normalizes it in the euclidean space.

The triplet loss has some aspects that serve our goals, for instance, it relies on three inputs, referred to as the "anchor", the "positive" and the "negative", which suits the structure of our evaluation method. Moreover, this loss verifies the distances between the anchor and

the positive and between the anchor and the negative. It penalizes the network if the last one is smaller. The formulation of the triplet loss is given by Equation 1, where it is possible to see that it penalizes the model if the distance between the negative and the anchor is shorter than the one of the positive and the anchor. Moreover, the equation includes a term $\alpha$, which is the margin. The margin, which in this case is set to 0.2, helps the model to define some separability between positives and negatives.

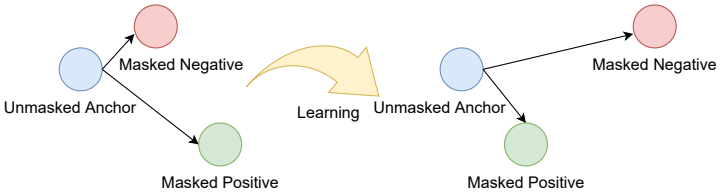



Fig. 1: Triplet loss effect in the euclidean space

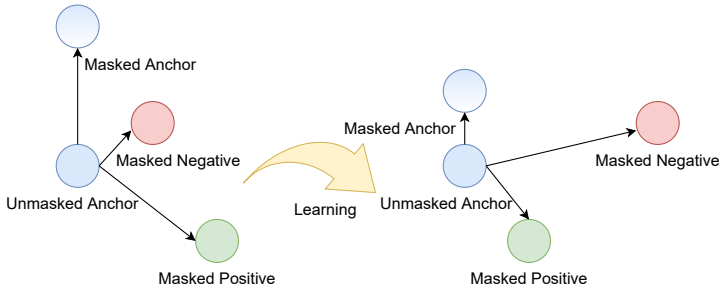



Fig. 2: Proposed Triplet loss effect in the euclidean space

It is possible to see on Figure 1 the effect that optimizing with triplet loss has on the embeddings. The anchor is a randomly selected image (without mask) to be used as a comparison point. The positive image, is a masked image from the same identity as the anchor image, whereas the negative is from a different identity. Our proposed approach, $TripletLoss_{Prop}$, constraints the loss of the original triplets, through the minimization of the distance between the masked and unmasked anchor embeddings. Our loss is formulated in Equation 3 and the effects of its optimization are visible on Figure 2.

$$TripletLoss_{Prop} = \sum_{a,am,p,n} TripletLoss(a,p,n) + MSE(am,a) \tag{3}$$

Since the mask occlusions are always on the same facial area, it is known that the network should focus its attention on areas that will not have occlusions. And thus, as seen on Equation 3, a mean squared error term is added to the loss. This way, we introduce more information to the model so that embedding optimization can be done more effectively.

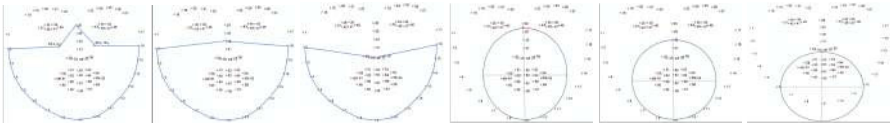
## 3 Experimental Setup and Results

**Evaluation Metrics:** To report the results we present the *false non-match rate (FNMR)*; the *FMR100* and *FMR10* which are the lowest *FNMR* for a *false match rate (FMR)* $< 1.0\%$

and $< 10.0\%$, respectively; the *equal error rate (EER)*; and the *area under the receiver operating characteristic curve (AUC)*. We also report the genuine mean (GMean) and impostors mean (IMean), which represent the mean distances between the embeddings of the same individual and from different people respectively.

**Face Data**: The development of face recognition methods requires large and diverse datasets. When the Covid-19 pandemic started and it became evident that the use of face masks had a negative impact on FR systems there was no ready-to-use data for research. The creation of synthetic data allows leveraging from existing data so that it fits the problem. Still, using real data is crucial as the ultimate test to the models and the community started to also collect face samples of individuals using face masks. The creation of these two types of data used in our method is described as follows.

*Synthetic masked face data (SMFD)*: Here we describe the synthetic masked face data (SMFD) creation process. Adding a facial mask requires information regarding facial landmarks. Moreover, the position and inclination of the face affect the positioning of the mask. Due to the lack of large-scale pairs of masked and unmasked identities, in this work, we synthetically generate different types of masks and adjust them on the unmasked samples of the VGGFace2 dataset [Ca18]. The dataset includes 3,310,000 face images from 8,631 train identities and 500 disjoint identities for the test, and includes a diverse set of samples regarding the various poses, ages, and ethnicity. Mask generation is carried out using the proposed algorithm by NIST [NGH20]. The algorithm exploits the Dlib C++ toolkit to obtain 68 facial landmarks for each image; afterward, using the extracted facial landmarks and interpolation between the points, various synthetic masks are generated. The details of the landmark extraction and mask generation are described in [Ki09; NGH20]. Due to variability regarding shape (Wide vs. Round) and face coverage (High, Medium, Low) we obtained six possible combinations. Figure 3 shows the result of applying the mask generation algorithm on a randomly selected sample from VGGFace2. Both the shape of the mask and its colour are randomly selected, for each image, while generating the masks.

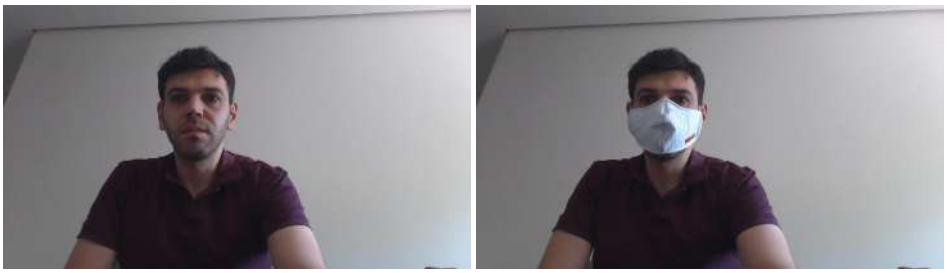


(a) Wide, high (b) Wide, medium (c) Wide, low (d) Round, high (e) Round, medium (f) Round, low

Fig. 3: Examples of face landmarks obtained from one image with different types of masks added. These masks vary in shape and face coverage.

*Real masked face dataset*: We evaluated our proposed solution using masked face recognition competition dataset (MFRC-21) [Bo21b]. MFRC-21 dataset was collected on 3 different days from 47 subjects. The first day is considered as a reference session, while the second and third sessions are considered probe sessions. In each session, 3 (two with mask and one without) videos are recorded using a webcam, while the subjects are requested to look directly into their camera. An overlapping database, and the same capture and frame selection procedure is describe in [Da20; Da21a]. In total, the references contain 470 unmasked images and 940 masked images. The probes contain 940 unmasked images and

1880 masked images. We evaluate our proposed solution under two evaluation scenarios. The first is between unmasked references and masked probes (U-M) and the second is between masked references and masked probes (M-M).



(a) Example from the test set of a reference image without a mask

(b) Example from the test set of a probe image with a mask

Fig. 4: Examples of images from the real masked faces dataset: images of the same individual with (a) and without a mask (b) (from [Da20]).

**Implementation Details**: To implement the mask FR model, we exploit ResNet50 architecture [He16] to extract the features from masked/unmasked facial samples. We trained the model by making use of cross-entropy (CE). The training of this model required around 150 thousand iterations. It trained with an initial learning rate of 0.1. It was decreased by a factor of 10 whenever the validation accuracy decreased. Stochastic gradient descent was used, with a batch size of 400, momentum of 0.9, and 0.0005 weight decay. We did not use any face alignment of the VGGFace2 dataset, giving as input of the model 224x224 images. After achieving convergence, it was fine-tuned with triplet loss and the combination of triplet loss with the mean squared error. Triplets were randomly generated at training time, and thus, it was unlikely to have them seen by the network more than once. Furthermore, we did not use any triplet mining. The models trained for 65 thousand iterations with a batch size of 200. And thus, 13 million triplets were created and used for the weight updates. The margin hyper-parameter $\alpha$ in Triplet loss is empirically determined as 0.2.

**Results and discussion**: We evaluated our method on two distinct datasets. One evaluation used synthetic masked face data (SMFD), with all the identities used for testing being disjoint from the training identities. We also evaluated the model with real masked face data (RMFD). Evaluating on these datasets allows us to infer the generalization capabilities of the model for unknown identities and images with different characteristics from the gallery images (e.g.real masks). The results are provided using the already mentioned metrics: GMean, IMean, AUC, EER, FMR100, and FMR10.

Our method is evaluated through a detailed step-wise ablation study that allowed us to understand the impact of the proposed modification of the triplet loss. Hence, we evaluate the model, in both datasets, with different training frameworks. This allows us to capture information regarding the impact of individual components of the model, such as, training with triplet loss, or not optimizing the embeddings. Besides, we also included the results of the method referred to in the tables as "VGG Face" [Ca18; SKP15] consisting of an Inception-ResNet pre-trained on the original VGGFace2 datasets.

In Table 1, it can be observed that good performance is achieved with just cross-entropy training (CE). Nevertheless, optimizing the produced embeddings with triplet loss (CE+TL) led to significant improvements in performance, lower distances for impostors, and higher distances for genuines. Moreover, our proposed adapted triplet loss resulting from the addition of the MSE constraint (CE+TL+MSE) lead to even more significant improvements, for example, approaching the AUC to 0.99, besides improvements in all the other metrics.

Tab. 1: Results obtained for the synthetic masked face data (EER, FMR100, FMR10 in %).

| Method | GMean | IMean | AUC | EER | FMR100 | FMR10 |
|---|---|---|---|---|---|---|
| VGG Face[Ca18; SKP15] | 0.505 | 0.325 | 0.951 | 11.8 | 38.2 | 13.5 |
| CE Loss | 0.528 | 0.426 | 0.941 | 13.2 | 38.5 | 21.5 |
| CE + TL | 0.601 | 0.320 | 0.977 | 7.8 | 28.9 | 11.9 |
| CE + TL + MSE **(Ours)** | 0.596 | 0.319 | **0.985** | **6.2** | **18.5** | **4.1** |

In Table 2, can be observed that, the model's performance degrades when compared to the previous table. Regardless of that, for the targeted comparison mode - U versus M - the results show the superior performance of our proposed loss. The distance of impostors increases as the distance of genuines increases too. Furthermore, it is possible to conclude that the model is competent in the task despite being trained only on synthetic data.

Tab. 2: Results obtained for the real masked face data (in the column "Mode": U and M stands for unmasked and masked data, respectively; EER, FMR100, FMR10 in %).

| Method | Mode | GMean | IMean | AUC | EER | FMR100 | FMR10 |
|---|---|---|---|---|---|---|---|
| VGG Face[Ca18; SKP15] | U-M | 0.523 | 0.426 | 0.769 | 29.419 | 90.587 | 58.959 |
| | M-M | 0.616 | 0.461 | 0.847 | 23.552 | 68.979 | 38.159 |
| CE Loss | U-M | 0.610 | 0.475 | 0.931 | 11.687 | 32.041 | 12.852 |
| | M-M | 0.702 | 0.503 | 0.936 | **9.002** | **16.628** | **8.791** |
| CE + TL | U-M | 0.647 | 0.396 | 0.943 | 11.213 | 34.744 | 11.874 |
| | M-M | 0.699 | 0.414 | 0.945 | 10.806 | 26.457 | 11.249 |
| CE + TL + MSE **(Ours)** | U-M | 0.649 | 0.383 | **0.957** | **9.799** | **28.252** | **9.678** |
| | M-M | 0.699 | 0.390 | **0.959** | 9.292 | 23,507 | 9.035 |

It should be noted that, the M-M mode experimental results do not keep up with the U-M one, in order words while our method improves the U-M, it offers no improvements for M-M recognition. This can be due to the fact that the embedding optimization process is made in a way that the model is trained to minimise the distance between masked and unmasked (U-M) genuine pairs, thus aiming at making it greater than the distance between imposter pairs. However, this was performed because the main application scenario commonly would contain unmasked references, such as automatic border control with an unmasked passport-stored reference and a possible masked live probe.

Besides the quantitative evaluation of the proposed method, a more qualitative approach was also studied. In order to infer the importance of each pixel for the overall embedding produced we used the Smooth Grad-CAM++method. The output of this gradient-based method was computed for each of the 512 features. Afterwards, all the map outputs were summed and divided by 512, thus generating a final map with the average importance of

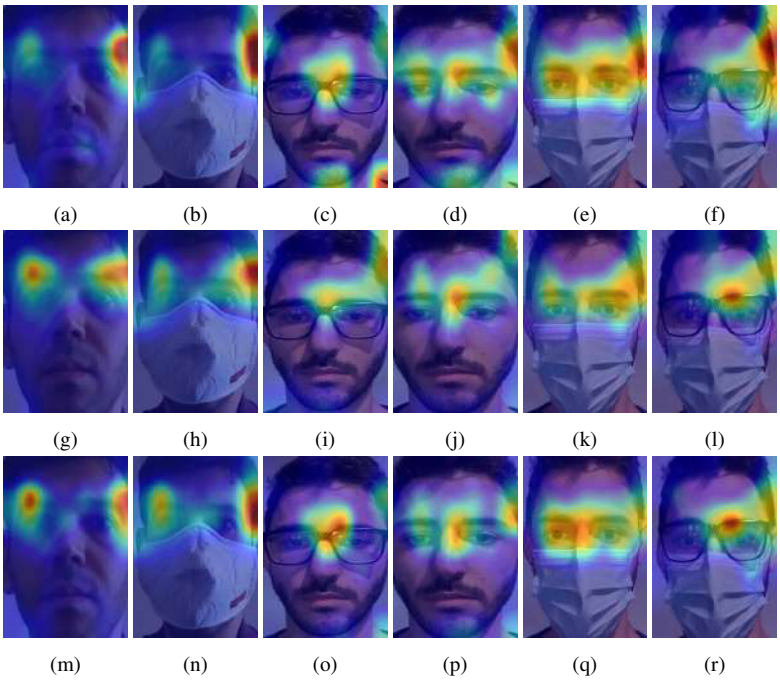



Fig. 5: Output of the Smooth Grad-CAM++ computed for each of the 512 features of the feature vector and normalized. Subfigures (a)-(f) computed from the cross-entropy model; (g)-(l) computed from the triplet loss model; and (m)-(r) computed from the mean squared error and triplet loss model.

the pixels to the overall embedding. Figure 5 displays the outputs for three of the studied methods, from the top to the bottom we have the CE, the TL and the TL+MSE methods. While the first is already capable of ignoring the masks, it still constructs the embedding of the unmasked images based on the chin area. Between the other two models, the main difference seems to be that the model with the MSE uses a wider area of pixels to construct the embedding, thus, capturing more information.

## 4 Conclusion and Future Work

In this work we addressed the challenge of masked face recognition motivated by the recent Covid-19 pandemic causing that wearing masks is now essential to prevent the spread of contagious diseases and has been currently forced in public places in many countries. However, recent research has shown that the performance, and thus the trust in contactless identity verification through face recognition, can be impacted by the presence of a mask. The scenario addressed is the evaluation of the verification performance in face recognition systems when verifying masked vs not-masked faces compared to verifying not-masked faces to each other. It was already noted in the literature, that the effect of masks was stronger on genuine pairs decisions in comparison to imposter pairs decisions. In this work, we proposed a methodology that targeted that observation and aimed at improving the

performance of MFR systems in the comparison of unmasked versus masked faces. The results obtained by our proposed method showed consistent improvements in a detailed step-wise ablation study. The ablation studies performed showed that our proposed triplet loss modification improved the performance of the models in the addressed scenario.

## Acknowledgements

This work was financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020, and within the PhD grants "2021.06872.BD" and "SFRH/BD/137720/2018". This research work has been also funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

## References

[Bo21a] Boutros, F.; Damer, N.; Kirchbuchner, F.; Kuijper, A.: Unmasking Face Embeddings by Self-restrained Triplet Loss for Accurate Masked Face Recognition. arXiv preprint arXiv:2103.01716, 2021.

[Bo21b] Boutros, F.; Damer, N.; Kolf, J. N.; Raja, K.; Kirchbuchner, F.; Ramachandra, R.; Kuijper, A.; Fang, P.; Zhang, C.; Wang, F.; Montero, D.; Aginako, N.; Sierra, B.; Nieto, M.; Erakin, M. E.; Demir, U.; Ekenel, H. K.; Kataoka, A.; Ichikawa, K.; Kubo, S.; Zhang, J.; He, M.; Han, D.; Shan, S.; Grm, K.; Štruc, V.; Seneviratne, S.; Kasthuriarachchi, N.; Rasnayaka, S.; Neto, P. C.; Sequeira, A. F.; Pinto, J. R.; Saffari, M.; Cardoso, J. S.: MFR 2021: Masked Face Recognition Competition. In: 2021 IEEE International Joint Conference on Biometrics (IJCB). Pp. 1–10, 2021.

[Ca18] Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.; Zisserman, A.: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In. Pp. 67–74, 2018.

[Da20] Damer, N.; Grebe, J. H.; Chen, C.; Boutros, F.; Kirchbuchner, F.; Kuijper, A.: The Effect of Wearing a Mask on Face Recognition Performance: an Exploratory Study. In: BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, online, 16.-18. September 2020. Vol. P-306. LNI, Gesellschaft für Informatik e.V., pp. 1–10, 2020.

[Da21a] Damer, N.; Boutros, F.; Süßmilch, M.; Kirchbuchner, F.; Kuijper, A.: An Extended Evaluation of the Effect of Real and Simulated Masks n Face Recognition Performance. IET Biometrics, 2021.

[Da21b] Damer, N.; Boutros, F.; Süßmilch, M.; Fang, M.; Kirchbuchner, F.; Kuijper, A.: Masked Face Recognition: Human vs. Machine.
arXiv preprint arXiv:2103.01924, 2021.

[Di20] Ding, F.; Peng, P.; Huang, Y.; Geng, M.; Tian, Y.: Masked Face Recognition with Latent Part Detection. In: Proceedings of the 28th ACM International Conference on Multimedia. Pp. 2281–2289, 2020.

[Fa21] Fang, M.; Damer, N.; Kirchbuchner, F.; Kuijper, A.: Real Masks and Spoof Faces: On the Masked Face Presentation Attack Detection.
CoRR abs/2103.01546, 2021.

[Ge20] Geng, M.; Peng, P.; Huang, Y.; Tian, Y.: Masked Face Recognition with Generative Data Augmentation and Domain Constrained Ranking. In: Proceedings of the 28th ACM Int. Conference on Multimedia. Pp. 2246–2254, 2020.

[Go21] Gomez-Barrero, M.; Drozdowski, P.; Rathgeb, C.; Patino, J.; Todisco, M.; Nautsch, A.; Damer, N.; Priesnitz, J.; Evans, N. W. D.; Busch, C.: Biometrics in the Era of COVID-19: Challenges and Opportunities.
CoRR abs/2102.09258, 2021.

[He16] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Pp. 770–778, 2016.

[Ho20] Hong, Q.; Wang, Z.; He, Z.; Wang, N.; Tian, X.; Lu, T.: Masked Face Recognition with Identification Association. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp. 731–735, 2020.

[Ki09] King, D. E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research 10, pp. 1755–1758, 2009.

[Li21] Li, Y.; Guo, K.; Lu, Y.; Liu, L.: Cropping and attention based approach for masked face recognition. Applied Intelligence 515, pp. 3012–3025, 2021.

[NGH20] Ngan, M. L.; Grother, P. J.; Hanaoka, K. K.: Ongoing Face Recognition Vendor Test (FRVT) Part 6B: Face recognition accuracy with face masks using post-COVID-19 algorithms. 2020.

[SKP15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Pp. 815–823, 2015.

[Yu20] Yuan, Y.; Chen, W.; Yang, Y.; Wang, Z.: In Defense of the Triplet Loss Again: Learning Robust Person Re-Identification with Fast Approximated Triplet Loss and Label Distillation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Pp. 1454–1463, 2020.