https://doi.org/10.2352/EI.2022.34.4.MWSF-209

Robust Face Recognition: How Much Face Is Needed?

Niklas Bunzel
Fraunhofer SIT / ATHENE, Darmstadt, Germany
Email: niklas.bunzel@sit.fraunhofer.de

Abstract

Face recognition systems are used in high security applications for identification, authentication and authorization. Being robust, is essential, not only towards Adversarial Examples, but also towards occluding accessories, such as facial masks, which become particularly relevant through the COVID19 pandemic. We have identified three inconspicuous facial areas to wear adversarial examples to attack face recognition. These are the mouth-nose section, the forehead and the eye area. In this paper, we will address the question of how much of a face needs to be present for successful identification and whether removing the identified critical regions is a viable countermeasure against adversarial examples.

Introduction

Face recognition is used for identification and authentication, for example when unlocking the smartphone [1, 2, 3], for payment [4] or at immigration controls at airports [5].

The whole identification process is typically split in two parts: 1) face detection 2) face recognition. In the face detection phase, the presence of a face in the image is detected and key facial features are localized. The second phase, the face recognition, uses these facial features to identify people. Many commercial face recognition systems developed prior to the COVID-19 pandemic have difficulty capturing faces with masks. This was reflected in high error rates for masked faces. Systems developed or adapted after 2020 perform comparably on masked faces as the 2017 systems on unmasked faces[6].

Biometric facial recognition is used to evaluate characteristic features of the face and enables assignment to a person. A digital image is used to compare the characteristics with biometric references using analysis software. Here, mainly characteristics are analyzed that cannot be changed by facial expressions, such as side parts of the mouth or the distance between the eyes [7]. The biometric characteristics are translated into digital patterns (templates) and compared with each other. The simplest form of facial recognition is verification in a one-to-one comparison of two templates. In addition, identification can be performed by matching facial images from databases in a one-to-many comparison [8]. To improve the robustness of face recognition with glasses Guo et al. proposed the usage of synthetic images in [9]. Several previous works proposed automatic eyeglass removal. For example Wu et al. [10] located the eyeglasses based on MCMC and synthesised the face image without the eyeglasses. Hu et al. [11] presented a GAN to remove and synthesize faces without eyeglasses. Face recognition systems can handle single common occlusions as indicated by the above mentioned papers. Since adversarial patterns can get printed on objects, even common objects like a hat, eyeglasses and a face mask can be used to inconspicuously wear adversarial examples [12, 13]. Adversarial examples are small perturbations added to an input, with the aim of provoking a misclassification [14].

This paper explores the question of how much of a face must still be accessible to a face recognition system in order to allow for successful identification. Which parts of the face are important? To which extend can more training data compensate for the loss of facial features? As a defensive measure, can the parts of the face where adversarial examples can be worn inconspicuously be cut away?

The main findings/contributions of this paper are:

- Removal of areas by whitening is a feasible solution against attacks on facial recognition via adversarial examples on the *hat*, *mask*, *glasses* and even combinations of these areas.
- We can remove over 50% of the faces maintaining over 90% accuracy.
- More training data can compensate for the loss of facial feature - at least to some extend.

Adversarial Examples

Adversarial examples [14, 15] are specially crafted images, which are intended to provoke a misclassification in the target system, while being correctly classified by humans. An adversarial example can be crafted as a targeted or as an untargeted attack. A targeted attack leads to a misclassification into the target class and an untargeted attack results in any misclassification. Note that not only classifiers are vulnerable to adversarial examples, there are also attacks on object detection [16] and segmentation [17]. Adversarial examples exist not only in the digital space, but can also be applied in the "real world". Therefore an adversarial patch can be utilized. An adversarial patch is a region bounded adversarial example [18]. For example, Eykholt et al. [19] mislead the classification of road signs by selectively applying stickers. This type of attack can be used in the physical world and adopted to face detection and face recognition tasks.

Adversarial patches against face recognition

Face recognition is the matching of a face to a person. An attack on face recognition, as with adversarial examples in general, can be targeted and untargeted. A targeted attack means *impersonation*, i.e. that the person is recognized as a particular other person. An untargeted attack is *dodging*, which means that the person is recognized as someone else but not him/herself. To achieve such an attack Sharfi et al. generate an eyeglass frame with a adversarial pattern. This attack has been successfully applied to various neural networks for face recognition [12]. Such glasses could be used, e.g., to bypass automated passport controls such as those found at airports. Success rates of over 80% were achieved in the evaluation when this attack was used as an dodging attack.

For impersonation attacks, the chances of success vary greatly depending on the wearer and the target. Another attack on face recognition is *AdvHat* by Komkov et al. in which they propose to put an adversarial sticker on a hat [13]. During the crafting of the AdvHat, the authors use a parabolic and an affine transformation to simulate the bending and the rotation of the sticker on the hat. From their evaluation results we can assume, that their attack is transferrable to other Face ID models and robust against different viewing angles. Zolfi et al. print their attack on face masks [20], taking advantage of the fact that during the COVID-19 pandemic wearing face masks became common.

Defending against adversarial examples

Researchers are working on measures to improve the robustness of neural networks. For example, in adversarial training [21] adversarial examples are integrated into the training process. This makes neural networks robust to attacks in the l_p norm they have been trained with, but offers poor performance against attacks crafted with another l_p norm [22]. Preprocessing techniques such as JPEG compression, bit-depth reduction, cropping & rescaling [23] attempt to remove the attack character of the image before classification. So far for each preprocessing measure a successful attack could be mounted [24, 25]. Other defenses modify the structure of the neural network to increase its robustness, e.g. by changing its activation function [26] or try to detect adversarial examples [27, 28]. Most of theses defenses were also shown to be prone to adversarial attacks [29]. Most adversarial patch defenses for image classification and object-detection utilize multiple runs of occlusion and reclassification [30, 31, 32, 33]. Most of these defenses have not been evaluated for physical attacks on face recognition systems. Furthermore, the authors are not aware of any defensive measure that has been specifically developed to prevent physical attacks on face recognition.

Face Model

As Sinha et al. mention in [34], the most relevant parts of a face for the recognition task are the eye and eyebrow region as well as the mouth area. We have created a model of a general face that highlights the most important parts from an information-theoretical point of view. We used the entropy with a radius of 5 pixels to calculate the relevant parts of each face image and averaged it to get our final model[1]. This model supports the statement of Sinha et al. by emphasising the eyes, eyebrows and mouth area the most. As we can see in figure 1, the nose, the chin line and the area between the eyes and nose also provide important information.

Evaluation

In the evaluation we seek to answer the following question: How much of a face is needed for a successful recognition? We empirically evaluate to which extent an increase in training samples is able to compensate for a lesser amount of facial features.

Figure 1: The created face model based on the averaged entropy of all training images.

Data

In order to address these questions, we created a subset of the VGG Face 2 dataset [35]. We used the dlib [36] frontal_face_detector to detect and align the faces in the images. We have removed all pictures where another person is visible or where parts of the face are missing or unrecognisable, e.g. if the person is wearing sunglasses or has turned their face too far away. Furthermore, we removed photos with a large age difference, e.g., baby photos. The subset consists only of identities which have 500 or more images, which yields 175 identities (classes) and 94,304 images. An 80/20 split was performed on this data so that the train set consists of at least 400 images per identity and the test set of at least 100 images per identity, i.e. per class. In total, the training set contains 75,545 and the test set 18,759 images.

Facial surface reductions

We identified three ways to inconspicously mask or alter the face or wear adversarial face accessories. These are: 1) wearing a hat or a cap 2) wearing glasses 3) wearing a mask[2]. Since white has a lower effect on face recognition than most other colors [6], we reduced the facial surface by whitening the respective areas. We created versions of our dataset with all combinations of the three regions. We implemented the surface reduction by detecting the face and calcuating the 68 face landmarks with dlib.

To remove a hat or cap that can be pulled down over the forehead we removed every pixel slightly above the eyebrows as in equation 1 in the appendix. The face model obtained by this reduction is shown in figure 2a. We will call the resulting dataset *hat*. To remove the mask-area we remove a polygon that approximates a worn mask, as can be seen in 2b. The landmarks used to create the polygon can be viewed in the appendix. The resulting dataset is called *mask*. To remove the glasses we evaluated two approaches, one is using a binary mask of a pair of glasses as in [12], the other is to whiten the whole eye area covered by a rectangle. The equation for the rectangle can be found in the appendix. The face models of the two approaches can be seen in figures 2c and 2d and the corresponding datasets are named *glasses* and *eye_area*. The resulting face models for the individual combinations of accessories can be viewed in the appendix. To evaluate how much of a face is needed for recognition, we calculated the average removed area by each approach. We used semantic segmentation with a BiSeNet [37] trained on CelebAMask-HQ [38][3] to obtain all face pixels per image and calculate the portion of the removed

[1] To calculate the entropy we transformed the images into grayscale and computed the entropy with `https://scikit-image.org/docs/dev/api/skimage.filters.rank.html#skimage.filters.rank.entropy`

[2] In most countries this is only inconspiciusly due to COVID19 pandemic.

[3] We used the BiSeNet implementation from `https://github.com/shaoanlu/face_toolbox_keras`

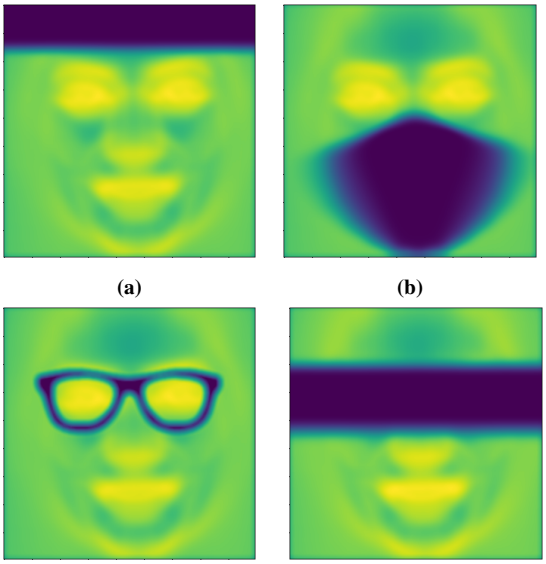


Figure 2: The entropy face model without a) the *hat* area, b) the *mask* area, c) the *glasses*, d) the complete *eye area*.

pixels. The percentage of removed pixels - both face pixels as well as image pixels in total - can be found in table 1.

Removed area	% facial pixels	% pixels
Hat	16%	17%
Glasses	22%	20%
eye area	32%	33%
Mask	37%	30%
Hat & glasses	38%	35%
Hat & eye area	45%	48%
Mask & hat	53%	47%
Mask & glasses	54%	48%
Mask & eye area	66%	60%
Mask & hat & glasses	72%	65%
Mask & hat & eye area	80%	77%

Table 1: The percentage of the average removed facial pixels and the percentage of the average removed pixels of the whole image by the different face reduction methods.

Models

For the face recognition we chose to use a ResNet [39] architecture with ArcFace loss[4] before the softmax activation as proposed in [40]. On top of the ResNet we put a BatchNormalization layer, then a Dropout layer with a rate of 0.5, then a Flatten layer and a FullyConnected layer with an L_2 regularization of $5e^{-4}$, finally a BatchNormalization and then the ArcFace loss with the softmax activation as described in [40]. For the ArcFace loss we tried different values for the hyperparameter s ($s \in [10, 11, 15, 20, 25, 30]$), the radius of the hypersphere all identities are distributed on. We set the margin penalty $m = 0.5$ as in [40]. We evaluated ResNet50 and ResNet101[5] with pretrained weights on imagenet. We used stochastic gradient descent as optimizer with a learning rate of 0.01, a momentum of 0.9 and a decay of 2×10^{-4}. We trained for 50 epochs with early stopping and a patience of 4.
The ResNet50 with $s = 11$ performed best on the validation-split with an average accuracy of 97%, therefore we take this model as a base line for our further experiments. On the full face test set it performed with an average accuracy of 96.9%. We trained models for every combination of the face reductions as described in the previous section. These are named corresponding to the dataset variation they are trained on e.g. *mask model* for the model trained with the whitened/removed masked area. We refer to the particular augmentation with which the models were trained as "in distribution", while all other variants, including the non-augmented full-face images, are considered "out of distribution".

[4]We use the ArcFace implementation from `https://github.com/4uiiurz1/keras-arcface/`

[5]We use the implementation of the ResNet models from Keras `https://keras.io/api/applications/`.

Results

Each model was trained 10 times on the different portions of the training set and the results are averaged. Face recognition models trained on complete faces are able to accurately recognize faces with removed parts (>90%), given about 120 (30%) frontal face images. For the out of distribution data the accuracy rates differ strongly depending on which area of the face is removed. The sheer amount of information removed is not decisive. For example, for the *masked* test data the model reaches up to 72% and for the glasses test data the model reaches only 52% accuracy, despite the fact that for the mask 37% of the facial information was removed and for the glasses only 22%. Since the drop in accuracy is the highest when the eye area is removed, we can deduce that the eye area is the most important feature, which can also be seen from figure 3. Adding training data increases the model accuracy even for out of distribution data. This is espacially true for regions that contain little or unimportant information like the forehead (*hat*).

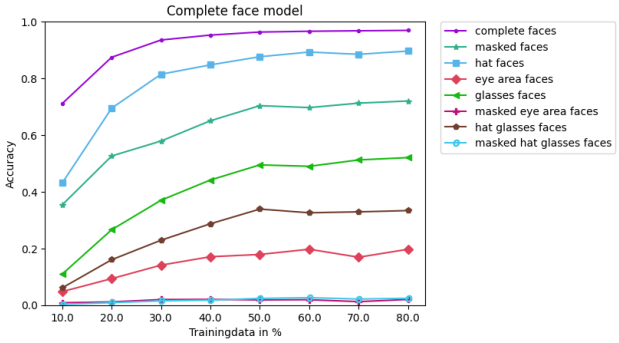


Figure 3: The model trained with complete faces evaluated on test sets with complete faces and different face reductions: *mask*, *hat*, *glasses*, *eye area* and combinations of face reductions: *mask, eye area*; *hat, glasses*; *mask, hat, glasses*.

Most models reach over 90% accuracy on their in distribution data with 30 - 50% of the training images. Training models with single areas removed like *mask*, *hat*, *glasses* yields accurate classifiers with over 90% accuracy for their respective in distribution data and complete face images as can be seen in 4. The only exception is the *eye area* model with 80% accuracy on full face images as can be seen in the appendix in figure 16c. The evaluations of the other models can also be viewed in the appendix.
Models trained on combinations of facial reductions result in an

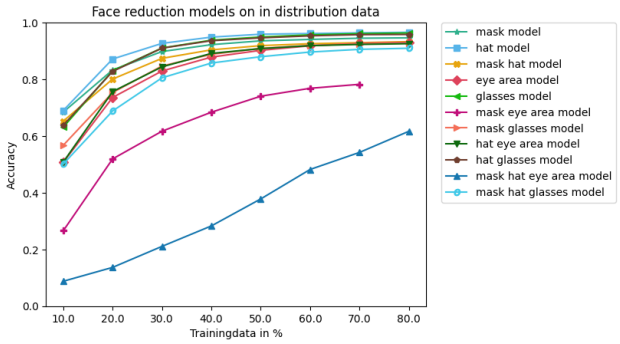


Figure 4: The face reduction models evaluated their in distribution data.

accuracy of over 90% on their specific facial reduction test sets - and are therefore better suited for application to facial reductions than a model trained on complete faces. Exceptions are models trained on images where the *mask* and *eye area* have been removed, as can be seen in figure 6. In these cases, over 65% of the facial information was removed. The models trained on combinations of face reductions even achieve accuracies of 80% and more on the full face images, up to a removal of over 54% of the facial information as depicted in 5.

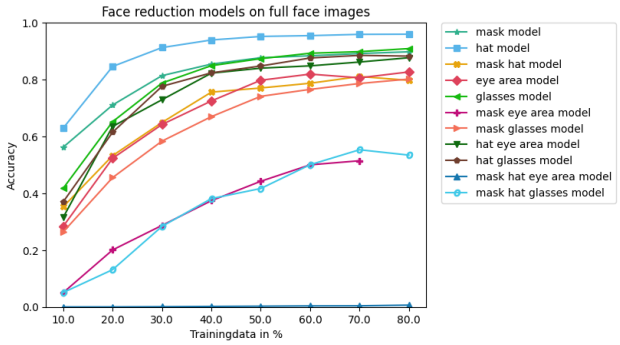


Figure 5: The face reduction models evaluated on full face images.

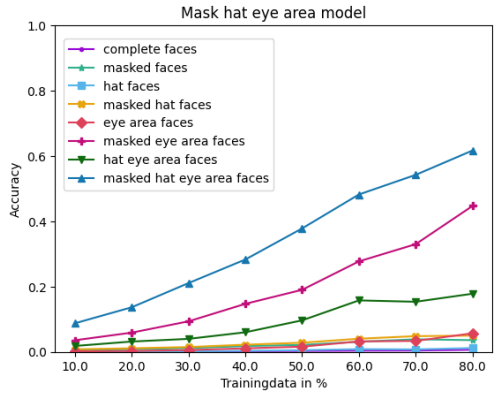


Figure 6: The model without the *mask, hat* and *eye area*.

We trained a model on all data which reaches over 90% accuracy for most face reductions with 70% of the training data. With 80% of the training data, the accuracy drops drastically. A closer inspection of the grad cam heatmaps generated by models trained on 70% of the training samples, and 80% respectively (see figure 8), reveals that the model trained with only 70% of the data has learned the upper left corner of the image as a relevant feature. At 80% of the training data, this image corner is no longer considered relevant. So we can assume that there was a spurious correlation in the data here. The performance seen at 80% of the training data is probably a correct approximation as would occur with a larger data set.

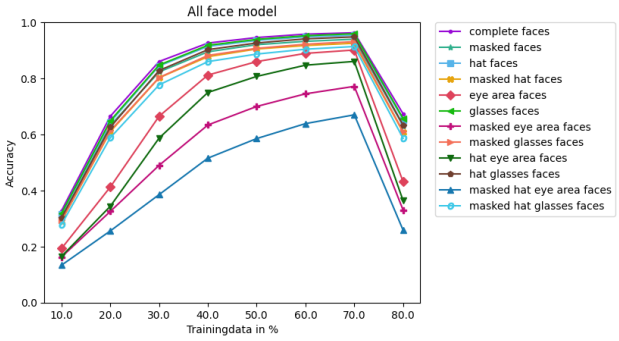
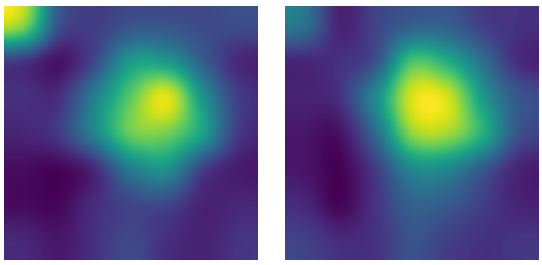


Figure 7: The face model trained with the full face images and all variants of face reductions. Therefore 10% of the training data means about 12 times as much as for the other models.



(a) **(b)**

Figure 8: The grad cam heatmaps a) with 70%, b) with 80% of the training data of various face reductions.

Based on the heatmaps 9, we can assume that the cnn models learned or paid attention to the features as we represented them in our face models. Removing the hat, the *mask* or the *glasses* results in robust and accurate face recognition models, even combinations of these face reductions yield competitive classifiers. When we remove too much information we end up with a classifier robust against adversarial patches, but depending on fragile features like the hair, noise or background elements as we can see e.g. in the heatmap of the model trained without the *mask, hat* and *eye area* in figure 10.

Conclusion & Futurework

For security authorities, biometric facial recognition offers enormous support potential. But for the use in high-risk applications, facial recognition must be robust against wearing accessories and, in particular, against adversarial pertubated accessories. To ensure this, we have investigated to what extent removing the facial regions typically used for wearing accessories is a viable solution. Thereby the trade-off between accuracy and robustness has to be considered. We also investigated how additional training images

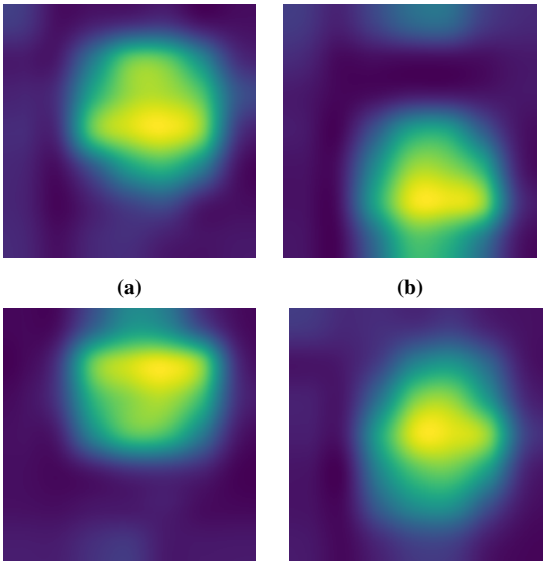


Figure 9: GradCam heatmaps of the different models. a) The complete faces model, b) the model without the *hat* and the complete *eye area*, c) without the *mask* d) without *glasses*.

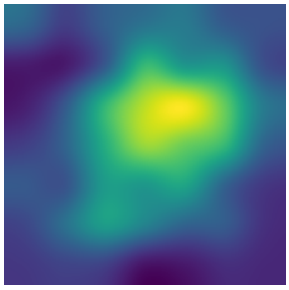


Figure 10: Grad Cam heatmap of the *mask hat eye area* model trained with 80% of the training data.

can compensate for the loss in accuracy.

Removal of areas by whitening is a feasible solution against attacks on facial recognition via adversarial examples on the areas often covered by hats, masks or glasses and even combinations of these three regions. With up to 54% of the facial pixels removed, we were nevertheless able to maintain over 90% accuracy on the models in distribution data. Most of the face reduction models even achieve over 80% accuracy on full face images. We have also validated that more training data is - at least to some extend - able to compensate for the loss of facial features.

In the future we want to evaluate the performance of Vision Transformers instead of CNNs as they can handle different image ablations better than CNNs [31]. We also want to investigate the performance if Generative Adversarial Networks (GANs) as proposed in [41] inpaint the removed areas. The performances of each model on its in distribution test data are superior to the one of a model trained on a combined dataset of all the herein described augmentations. Therefor the performance of an ensemble of models trained on augmented face data will also be investigated.

Acknowledgement

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] A. Inc., "Face id auf dem iphone oder ipad pro verwenden," https://support.apple.com/en-en/HT208109, 2021, visited: 20.01.2022.

[2] G. LLC, "Unlock your pixel phone with your face," https://support.google.com/pixelphone/answer/9517039?hl=en, 2021, visited: 20.01.2022.

[3] Samsung, "Use facial recognition security on your galaxy phone," https://www.samsung.com/us/support/answer/ANS00062630/, 2021, visited: 20.01.2022.

[4] F. Liu, "Making cutting-edge technology approachable: A case study of facial-recognition payment in china," https://www.nngroup.com/articles/face-recognition-pay/, 2020, visited: 20.01.2022.

[5] S. McCartney, "Are you ready for facial recognition at the airport?" https://www.wsj.com/articles/are-you-ready-for-facial-recognition-at-the-airport-11565775008, 2019, visited: 20.01.2022.

[6] M. Ngan, P. Grother, and K. Hanaoka, "Ongoing face recognition vendor test (frvt) part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms," 2020-11-30 2020.

[7] Bundesamt für Sicherheit in der Informationstechnologie, "Gesichtserkennung," https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Biometrie/Gesichtserkennung_pdf.pdf?__blob=publicationFile&v=1, 2008, stand: 20.01.2022.

[8] C. Rath, "Was gesichtserkennung kann," *Deutsche Richterzeitung*, vol. 2017, no. 1, pp. 8–10, 2017.

[9] J. Guo, X. Zhu, Z. Lei, and S. Z. Li, "Face synthesis for eyeglass-robust face recognition," 2021.

[10] C. Wu, C. Liu, H.-Y. Shum, Y.-Q. Xy, and Z. Zhang, "Automatic eyeglasses removal from face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 322–336, 2004.

[11] B. Hu, Z. Zheng, P. Liu, W. Yang, and M. Ren, "Unsupervised eyeglasses removal in the wild," *IEEE Transactions on Cybernetics*, pp. 1–13, 2020.

[12] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Trans. Priv. Secur.*, vol. 22, no. 3, Jun. 2019.

[13] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 819–826.

[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013.

[15] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases*, H. Blockeel, K. Ker-

sting, S. Nijssen, and F. Železný, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 387–402.

[16] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector," in *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2019, pp. 52–68.

[17] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[18] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2018.

[19] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] A. Zolfi, S. Avidan, Y. Elovici, and A. Shabtai, "Adversarial mask: Real-world adversarial attack against face recognition models," 2021.

[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[22] A. Araujo, L. Meunier, R. Pinot, and B. Negrevergne, "Advocating for multiple defense strategies against adversarial examples," 2020.

[23] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=SyJ7ClWCb

[24] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 274–283.

[25] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[26] C. Xiao, P. Zhong, and C. Zheng, "Enhancing adversarial defense by k-winners-take-all," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=Skgvy64tvr

[27] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *Proceedings 2018 Network and Distributed System Security Symposium*, 2018. [Online]. Available: http://dx.doi.org/10.14722/ndss.2018.23198

[28] K. Roth, Y. Kilcher, and T. Hofmann, "The odds are odd: A statistical test for detecting adversarial examples," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 5498–5507.

[29] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1633–1645. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf

[30] M. McCoyd, W. Park, S. Chen, N. Shah, R. Roggenkemper, M. Hwang, J. X. Liu, and D. A. Wagner, "Minority reports defense: Defending against adversarial patches," in *ACNS Workshops*, 2020.

[31] H. Salman, S. Jain, E. Wong, and A. Madry, "Certified patch robustness via smoothed vision transformers," 2022. [Online]. Available: https://openreview.net/forum?id=t2Mzgc9JEjZ

[32] B. Liang, J. Li, and J. Huang, "We can always catch you: Detecting adversarial patched objects with or without signature," 2021.

[33] A. Levine and S. Feizi, "Robustness certificates for sparse adversarial attacks by randomized ablation," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 4585–4593.

[34] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.

[35] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," 2018.

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

[37] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[38] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[40] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[41] X. Zhang, X. Wang, C. Shi, Z. Yan, X. Li, B. Kong, S. Lyu, B. Zhu, J. Lv, Y. Yin, Q. Song, X. Wu, and I. Mumtaz, "Degan: Domain embedded gan for high quality face image inpainting," *Pattern Recognition*, p. 108415, 2021.

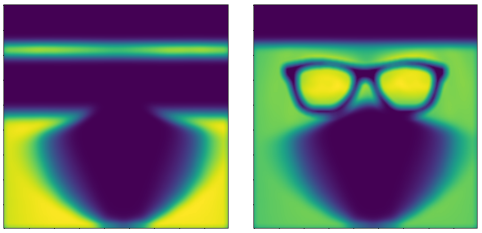
Author Biography

Niklas Bunzel received his B.Sc. and M.Sc. degrees in computer science and IT security from Technical of University Darmstadt 2015 and 2020, respectively. He is currently a research scientist

at Fraunhofer Institute for Secure Information Technology (SIT) and a PhD student at the Techincal University of Darmstadt. His major research interests include artificial intelligence, adversarial machine learning, IT security and steganography.

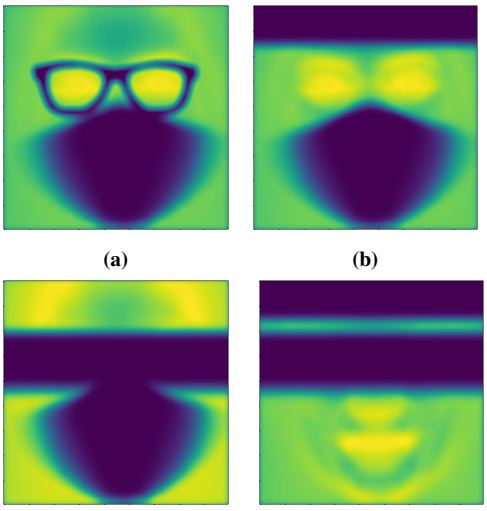
Appendix

Face models



(a) **(b)**

Figure 11: The face model without a) the mask, hat and complete eye area, b) the mask, hat and glasses area.



(a) **(b)**

(c) **(d)**

(e)

Figure 12: The face model without a) the mask and glasses, b) the mask and hat area, c) the mask and the complete eye area, d) the hat and the complete eye area, e) the hat and glasses.

Facial surface reductions

Equation to remove the hat area We take the minimal y, which is the y-coordinate of the pixel at the different landmarks $0, 16, 17, ..., 27$. Then we whiten the whole rectangle which spans between $(0,0)$ and (width, y).

$$y = min_y(0, 16, 17, ..., 27) - \varepsilon, \text{ with } \varepsilon = 10 \quad (1)$$

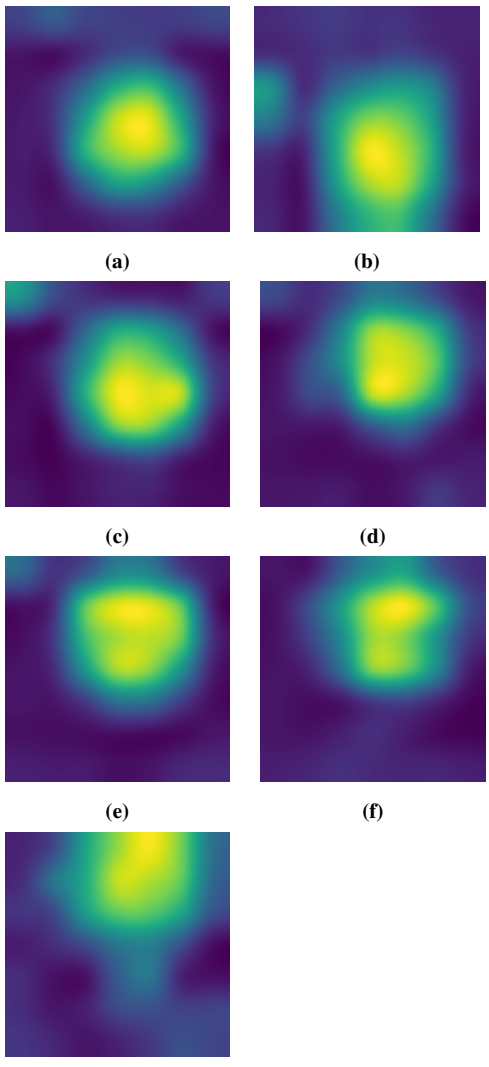
Mask removal We whiten the polygon with the landmarks $2, 3, ..., 14, 28$.

Equation to remove the eye area We whiten the rectangle which spans between $(0, y_1), (\text{width}, y_2)$, with

$$y_1 = min_y(19, ...24) \quad y_2 = max_y(0, 16, 29, 41, 47) \quad (2)$$

Where y is the y coordinate at the landmarks.

Grad Cam Heatmaps



(a) **(b)**

(c) **(d)**

(e) **(f)**

(g)

Figure 13: GradCam heatmaps of the face models without a) the hat and glasses, b) complete eye area, c) hat, d) mask, hat and glasses, e) mask, hat, f) mask, glasses g) mask and eye area.

Evaluations

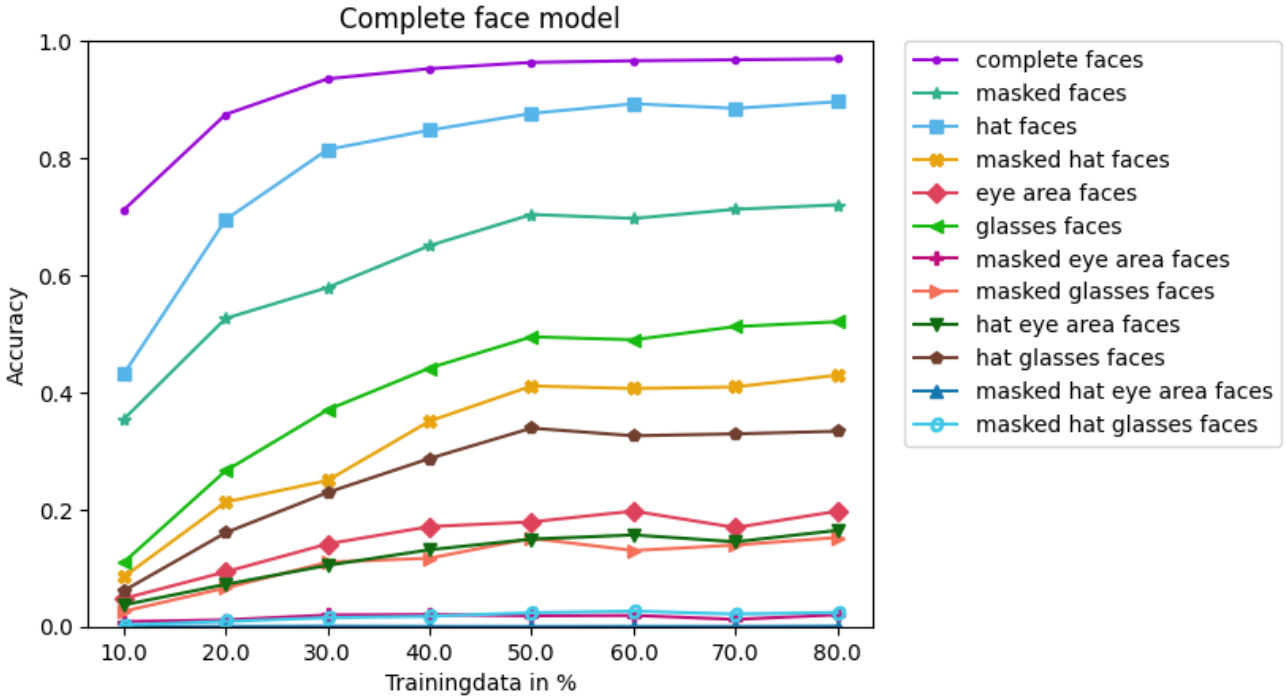


Figure 14: The face model trained with complete faces evaluated on all test sets.

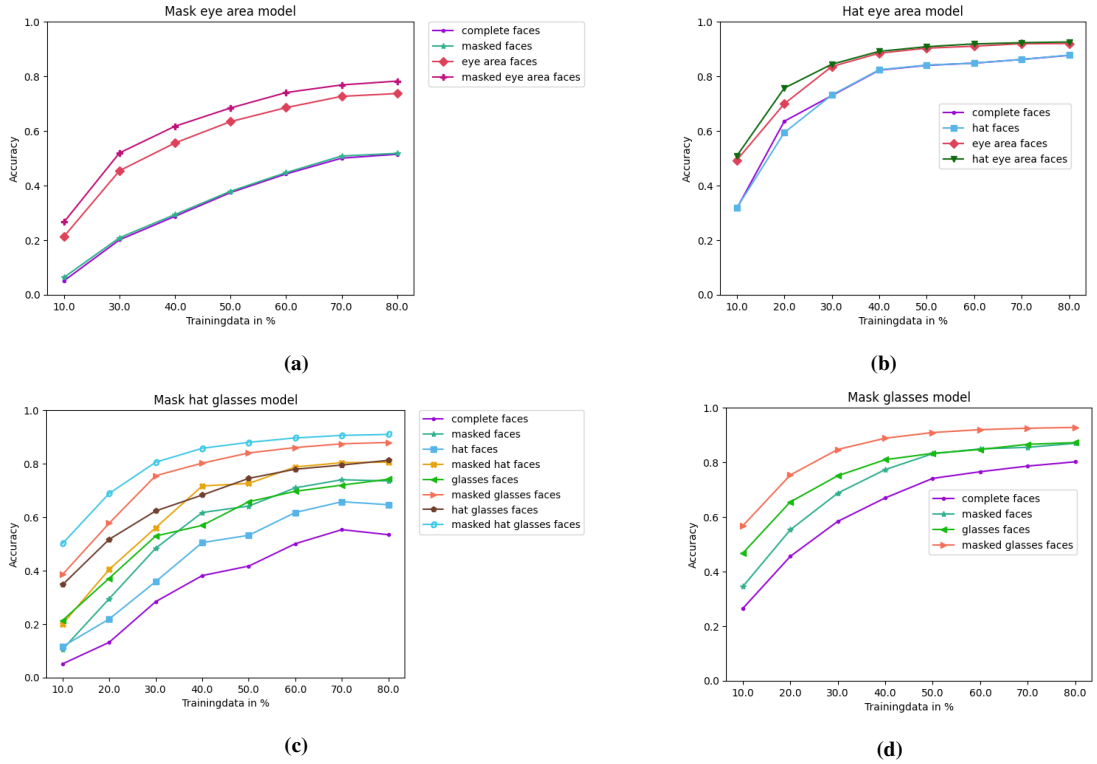


Figure 15: The face model without a) the mask and the complete eye area, b) the hat and the complete eye area, c) the mask, hat and glasses, d) the mask and glasses

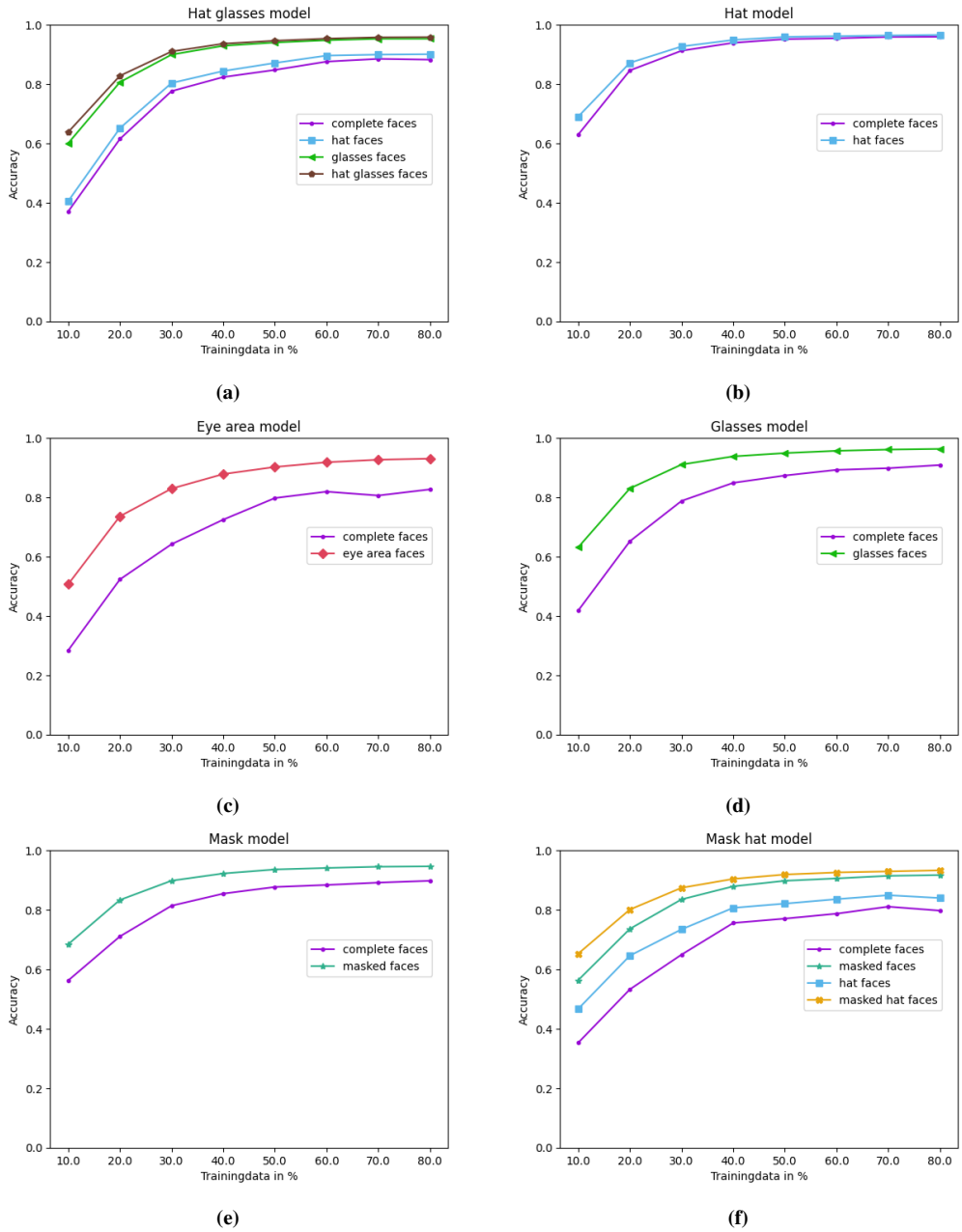


Figure 16: The face model without a) the hat and glasses, b) the hat, c) the complete eye area, d) the glasses, e) , the mask f) the mask and hat.