Demonstrating the vulnerability of RGB-D based face recognition to GAN-generated depth-map injection.

Valeria Chiesa, Chiara Galdi, and Jean-Luc Dugelay
Department of Digital Security, EURECOM, 450 Route des Chappes, 06410 Biot, France
valeria.chiesa2@gmail.com, {chiara.galdi, jean-luc}@eurecom.fr

Keywords: IST-EURECOM LFFD, IST LLFFSD, RGB-D, GAN, face recognition

Abstract: RGB-D cameras are devices able to collect additional information, compared to classical RGB devices, about the observed scene: its depth (D). This has made RGB-D very suitable for many image processing tasks, including presentation attack detection (PAD) in face recognition systems. This work aims at demonstrating that thanks to novel techniques developed in recent years, such as generative adversarial networks (GANs), face PAD systems based on RGB-D are now vulnerable to logical access attack. In this work, a GAN is trained to generate a depth map from an input 2D RGB face image. The attacker can then fool the system by injecting a photo of the authorized user along with the generated depth map. Among all RGB-D devices, this work focuses on light-field cameras but the proposed framework can be easily adapted for other RGB-D devices. The GAN is trained on the IST-EURECOM light-field face database (LFFD). The attack is simulated thanks to the IST lenslet light field face spoofing database (LLFFSD). A third dataset is used to show that the proposed approach generalizes well on a different face database.

1 INTRODUCTION

RGB-D cameras are a specific type of depth sensing devices that work in association with a RGB camera, that are able to augment the conventional image with depth information (related with the distance to the sensor) in a per-pixel basis (Nunes et al., 2012). In the past years, novel camera systems like the Microsoft Kinect or the Asus Xtion sensor that provide both 2D-texture and dense-depth images became readily available. Although these devices were born for motion-sensing applications and games, researchers have investigated their use for a number of other tasks, including face recognition. In more recent years, another RGB-D device has appeared on the market, namely the *plenoptic* or *light-field* camera. The plenoptic function is the 7-dimensional function representing the intensity of the light observed from every position and direction in 3-dimensional space (Adelson et al., 1991). Thanks to the plenoptic function, it is thus possible to define the direction of every ray in the light-field vector function. As described in (Ng et al., 2005), a light-field camera is similar to a traditional 2D RGB camera but has an additional micro-lens array placed between the main lens and the imaging sensor. This component allows to acquire multiple representations of the scene, each one captured from a slightly shifted point of view. Among other things, the information stored in a light-field image can be used for 3D scene reconstruction. As happened with the other RGB-D devices, also light-field imaging has been investigated for face analysis. For a recent overview on light fields for face analysis, the reader is referred to (Galdi et al., 2019).

Light fields have proven useful for face recognition (Sepas-Moghaddam et al., 2018c) (Raghavendra et al., 2013b) (Raghavendra et al., 2013c) (Raghavendra et al., 2013a) (Raja et al., 2015) (Sepas-Moghaddam et al., 2017b) (Chiesa and Dugelay, 2018b) (Sepas-Moghaddam et al., 2018a) (Sepas-Moghaddam et al., 2019). However, there is another field in which the depth information provided by RGB-D cameras has demonstrated to be very effective, namely face presentation attack detection (PAD) (Liu et al., 2019) (Kim et al., 2014) (Raghavendra et al., 2015) (Sepas-Moghaddam et al., 2018b) (Sepas-Moghaddam et al., 2018d) (Ji et al., 2016) (Chiesa and Dugelay, 2018a) (Zhu et al., 2020). A PAD method is any technique that is able to automatically distinguish between real biometric traits presented to the sensor and synthetically produced artefacts containing a biometric trait (Galbally et al., 2014). Thanks to the depth map, it is extremely easy to detect presentation attacks on flat surfaces, such as presenting to the system a face picture of an authorized user printed on paper or displayed on a screen.

In (Ratha et al., 2001), Ratha *et al.* identified several possible vulnerable points of a biometric recognition system. Figure 1 illustrates a biometric system pipeline and the possible attacks: (1) an artefact reproducing a biometric trait such as an artificial finger may be presented at the sensor, (2) illegally intercepted data may be resubmitted to the system, (3) the feature extractor may be replaced by a Trojan horse program that produces pre-determined feature sets, (4) legitimate feature sets may be replaced with synthetic feature sets, (5) the matcher may be replaced by a Trojan horse program that always outputs high scores thereby defying system security, (6) the templates stored in the database may be modified or removed, or new templates may be introduced in the database, (7) the data in the communication channel between various modules of the system may be altered, and (8) the final decision output by the biometric system may be overridden (Jain et al., 2005). In this paper, we focus on (2), in which the attacker has the possibility to directly inject a compromised biometric sample into the system, the latter is also referred to as "logical access attack". Injection attacks allow an attacker to supply untrusted input to a system, which gets processed as part of the query.

The aim of this work is to show that thanks to novel techniques developed in recent years, such as generative adversarial networks (GANs), face PAD systems based on RGB-D are now vulnerable to logical-access attack. A GAN is trained to generate depth maps from 2D RGB face images. The simulated attack is thus the following: an attacker only needs to retrieve a face picture of an authorized user (e.g. from the web) and generate the corresponding depth map using the GAN. Both the texture information (photo) and the generated depth map are injected in the system.

The GAN is trained on the IST-EURECOM light-field face database (LFFD). The attack is simulated thanks to the IST lenslet light field face spoofing database (LLFFSD). A third dataset is used to show that the proposed approach generalizes well on a different face database. According to the ISO/IEC standards, performance is reported in terms of (i) Attack Presentation Classification Error Rate (APCER): defined as the proportion of presentation attacks incorrectly classified as Bona Fide presentations (ii) Bona Fide Presentation Classification Error Rate (BPCER): defined as the proportion of Bona Fide presentations incorrectly classified as presentation attacks.

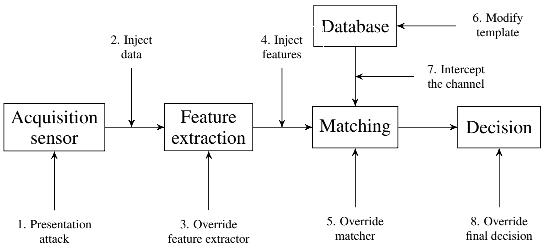

Figure 1: Possible attacks in a biometric recognition system (adapted from (Jain et al., 2005))

2 Proposed method

In order to demonstrate the vulnerability of RGB-D based face PAD systems, the following elements are required: an RGB-D based presentation attack detector; a GAN to generate depth maps from 2D-RGB face images; a dataset to train the GAN; a dataset to test the attack.

2.1 Presentation attack detector

The RGB-D face PAD system tested in this work was especially designed to detect presentation attacks in light-field based face recognition systems and is presented in (Chiesa and Dugelay, 2018a). This method can be used independently of the adopted face recognition algorithm. For example, it can be used in combination with a "classical" algorithm for face recognition based on the 2D face picture only, or with an algorithm exploiting the depth information provided by the RGB-D sensors. For a more detailed description of the latter, the reader is referred to the references provided in the introduction.

From a light-field face image it is possible to obtain a pair of RGB image and depth map of the acquired face, where each pixel in the RGB image has a corresponding depth value in the depth map. As described in (Chiesa and Dugelay, 2018a), the analysis of the depth values at specific face feature points leads to the detection, with high accuracy, of most of the tested presentation attacks, decreasing the vulnerability of the recognition system. Not only the algorithm works very well (100% correct detection) on flat presentation attacks such as printed or screen attack, but also detects with high accuracy ($99,26\% - 99,54\%$) the presentation attack consisting in wrapping a face picture around a cylindrical surface, outperforming other state-of-the-art algorithms.

The raw light-field images are processed with the proprietary software Lytro Desktop and for each image a pair of RGB and depth map is extracted. The software provides perfectly aligned RGB images and depth maps so that a depth value is associated with each pixel of the RGB picture. Landmark detection is performed with the DLIB library (DLIB: http://dlib.net/), which implements a method for face detection based on Histogram of Oriented Gradients (HOG) and a face pose estimator, which also provides 68 face landmarks, based on the algorithm described in (Kazemi and Sullivan, 2014). In the DLIB implementation, the pose estimator model is trained on the database used for the 300 faces In-the-wild challenge as described in (Sagonas et al., 2016). The algorithm identifies 68 landmarks for each face. Then, for each landmark, the corresponding depth value is considered. In order to smooth eventual noise, the depth map is convoluted with a 7×7 pixels average filter. Landmark Depth Features (LDF) are defined as the set of depth values corresponding to the 68 landmarks of the face.

Classification is based on One-Class Support Vector Machine, a particular classifier trained with samples belonging to a single class only. The classifier is trained with the LDFs from the bona fide (genuine face images) class.

While most of the PAD algorithms presented in literature, including (Chiesa and Dugelay, 2018a), are able to detect presentation attacks performed at sensor level, they may be vulnerable to logical access attacks. However, for an attacker it would be still difficult to retrieve the data to be injected in the system, that is: a pair RGB image - depth map of the face of an authorized user. While for systems based on RGB acquisition only a simple picture retrieved from the web would be enough, in this case the attacker would need to retrieve or generate the face depth map as well.

In this work we demonstrate that thanks to novel techniques developed in recent years, such as generative adversarial networks (GANs), it is now possible to use the RGB face picture to synthesize a realistic depth map to be injected in the system and able to fool the PAD module.

2.2 Conditional GANs

Given a training set, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) learn to generate new data with the same statistics as the training set. For example, a GAN trained on photos can generate new photos that look, at least superficially, authentic to human observers. GANs learn a loss that tries to classify if the generated output image is real or fake, while simultaneously training a generative model to minimize this loss. Because GANs learn a loss that adapts to the data, they can be applied to a multitude of tasks that traditionally would require very different kinds of loss functions (Isola et al., 2017). In (Isola et al., 2017), Isola et al. present an image-to-image translation algorithm based on conditional adversarial networks. They made the code publicly available (Pix2pix: https://phillipi.github.io/pix2pix/), namely *pix2pix*, and thus their method has been adopted in many works.

GANs are generative models that learn a mapping from a random-noise vector z to an output image y, $G : z \rightarrow y$ (Goodfellow et al., 2014). Conditional GANs learn a mapping from an observed image $\mathbf{x}$ and a random-noise vector z to an output image y, $G : \{\mathbf{x}, z\} \rightarrow y$. The generator G is trained to produce outputs that cannot be distinguished from "real" images by an adversarially trained discriminator, D, which is trained to detect the generator's "fakes". A key feature of image-to-image translation is that the input and output differ in surface appearance but both are renderings of the same underlying structure. Therefore, structure in the input is roughly aligned with structure in the output (Isola et al., 2017).

The *pix2pix* translator is thus used in this work to convert RGB face images in the corresponding depth maps. An example of such translation is given in Figure 2.

Figure 2: Image-to-image translation example: on the left the RGB face image; in the centre the ground truth depth map obtained with Lytro's software; and on the right the generated depth map using pix2pix. Images from the LFFD. (Sepas-Moghaddam et al., 2017a).

2.3 Datasets

As mentioned before, at least two datasets are needed for this work: one for training the conditional GAN (i.e. pix2pix) for generating the face depth maps and one for testing the logical access attack.

Pix2pix is trained thanks to the IST-EURECOM Light Field Face Database (LFFD) (Sepas-Moghaddam et al., 2017a). The LFFD is a dataset made publicly available (IST: http://www.img.lx.it.pt/LFFD/, EURECOM: http://lffd.eurecom.fr/) to serve as basis for the design, testing and validation of novel light-field imaging based face recognition systems. The database consists of 100 subjects, with images captured by a Lytro ILLUM light-field camera. Two separate sessions were performed for each subject with a temporal separation between 1 and 6 months. The database includes 20 image shots per person in each of the two sessions, with several facial variations including expressions, actions, poses, illuminations, and occlusions.

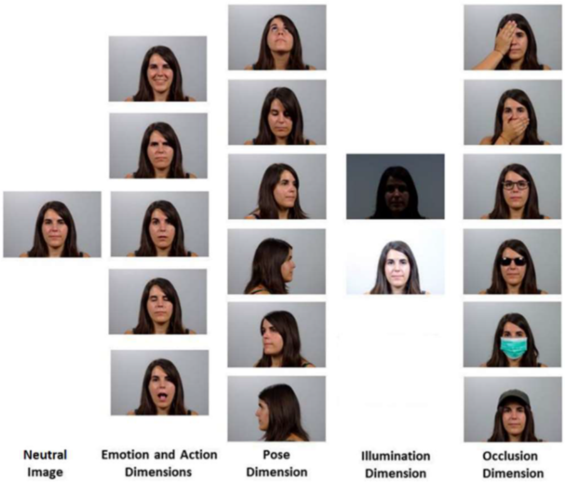

Figure 3: An example of pictures captured for each subject in the LFFD (Sepas-Moghaddam et al., 2017a).

The LFFD provides a large set of RGB face images and corresponding depth maps necessary to train pix2pix to generate depth maps from RGB face images. Eight face variations are considered, discarding occlusions and pose variations.

To test the logical access attack, the IST Lenslet Light Field Face Spoofing Database (LLFFSD)(Sepas-Moghaddam et al., 2017c) (Sepas-Moghaddam et al., 2018e) is used (LLFFSD: http://www.img.lx.it.pt/LLFFSD/). The LLFFSD is generated from the IST LFFD, by using the RGB face images from the LFFD to reproduce presentation attack items (PAIs), such as printed paper, wrapped printed paper, laptop, tablet, and two different mobile phones, for a total of 600 presentation attack images. The PAIs are then acquired with the same camera used for LFFD, that is the Lytro ILLUM plenoptic camera.

A third dataset is used to demonstrate that the proposed approach generalises well on different datasets. The Visible and Thermal Pair Face Database (VIS-TH), presented in (Mallat and Dugelay, 2018), is composed by 4200 images of 50 subjects collected with a dual camera collecting pairs of images in the visible and the thermal spectra at once. The VIS-TH database includes 21 face variations, including different illumination, expressions, and occlusions. From this dataset only the RGB face images are selected and used to generate the PAI pairs to attack the RGB-D based PAD.

3 Experimental evaluation

3.1 Depth-map generation

The pix2pix network is trained with 700 RGB-D pairs of face images from the LFFD database. The 700 images are randomly sampled from the set of images composed by the following face variation sets: ”Smile”, ”Angry”, ”Surprise”, ”Closed eyes”, ”Open month”, ”High illumination”, and ”Low illumination”. The images in the ”Neutral” face variation category are used for testing. Regarding the parameter setting of pix2pix, the maximum number of epochs is set to 100. This number is empirically selected by considering the trade-off between the generator’s and discriminator’s loss functions.

Two approaches are adopted to evaluate how similar to the original the generated depth maps are: (i) by computing the disparity map between the original and the generated depth maps; (ii) by attacking the PAD system with the generated depth maps and considering the success rate of the attacks.

Approach (ii) is discussed in next Section. Regarding (i), the RGB images from the LFFD belonging to ”Neutral” variation set (not used in training) are processed with the pix2pix network in order to obtain the synthesized depth maps. The disparity map in Figure 4 shows the comparison between ground-truth and synthesized depth maps. The disparity map is obtained by computing the pixel difference over all original and generated depth maps. Lighter areas indicates a larger difference between the original and the generated depth map. As can be observed from the figure, the depth value prediction is less accurate around the nose and the eyes (lighter color).

3.2 PAD vulnerability assessment

The PAD method (Chiesa and Dugelay, 2018a), is trained using the bona fide samples belonging to ”Smile”, ”Angry”, ”Surprise”, ”Closed eyes”, ”Open month”, ”High illumination” and ”Low illumination” sets not used to train the pix2pix network.

Three testing sets are considered: LFFD, LLFFSD and VIS-TH. For each of them, the RGB-D pairs are created using the pix2pix network on the RGB face images. The pairs are then submitted to the PAD method. Performance is reported in terms of Attack Presentation Classification Error Rate (APCER): defined as the proportion of presentation attacks incorrectly classified as Bona Fide presentations. The results are discussed in detail in the following for each of the three testing sets:

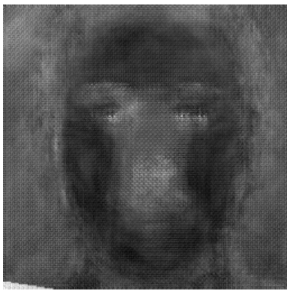

Figure 4: Pixel difference between original and synthesized depth maps: the lighter pixels around the nose and eyes areas show that it is more challenging to predict the correct depth value in those areas.

- **LFFD**: The testing set consists in face images from the ”Nautral” category, these images were not used for training the pix2pix network nor to train the PAD method. All the generated image pairs are recognized by the classifier as bona fide. APCER = 100%.

- **LLFFSD**: For this dataset, we must draw a distinction since the ground-truth images already represents an attack, that is a presentation attack. On the original pairs, the classifier recognizes 100% of the attacks for ”Paper”, ”Laptop”, ”Mobile 1”, and ”Mobile 2” PAIs and 99% of the ”Wrapped Paper” attacks. In order to test the logical access attack, the presentation attack RGB images are used to synthesize the corresponding depth maps using pix2pix. The classifier is not able anymore to detect the attacks and classifies all samples as bona fide. APCER = 100%.

- **VIS-TH**: The RGB images from VIS-TH are used to synthesize the depth maps. Also in this case the PAD method classifies all samples as bona fide. APCER = 100%. This test is done to demonstrate that the proposed approach generalises well on a different dataset. While the images of the LFFD and LLFFSD datasets are collected with the same camera model – the Lytro ILLUM plenoptic camera – the images in VIS-TH are collected with a different sensor, a dual visible and thermal camera, namely the FLIR Duo R.

Table 1 summarizes the results for the LFFD and VIS-TH databases. The results for the LLFFSD are reported in Table 2, in this case the results on the pairs with generated depth maps are compared with the PAD results on the original pairs representing presentation attacks. The results on the LLFFSD are very interesting as they show that depth maps generated from re-captured face photos are able to spoof the PAD system.

	APCER
LFFD	100%
VIS-TH	100%

Table 1: Attack presentation classification error rate on the LFFD and VIS-TH testing sets.

LLFFSD PAI	Original depth maps APCER	Generated depth maps APCER
Paper	0%	100%
Laptop	0%	100%
Mobile 1	0%	100%
Mobile 2	0%	100%
Wrapped paper	1%	100%

Table 2: Attack presentation classification error rate on the LLFFSD testing set.

4 Conclusions

In this work we demonstrated that recently developed AI techniques, such as GANs, represent a threat to RGB-D face recognition systems. A logical access attack is simulated in order to demonstrate the vulnerability of a PAD method that in normal conditions is able to detect up to 99% of the attack presentations. While presentation attacks are performed at the sensor level, the considered logical access attack operates between the acquisition and the feature extraction module. The pix2pix Conditional Adversarial Network is used to synthesize depth maps from 2D RGB face images. In this way, an attacker would only need to retrieve a face picture of an authorized user and generate the corresponding depth map to bypass the RGB-D based PAD module. The performances of the presentation attack detector are tested on the synthesize depth maps obtained from three datasets: the IST-EURECOM Light Field Face Database, the Lenslet Light Field Face Spoofing Database, and the Visible and Thermal Pair Face Database. The PAD method fails to classify the generated depth maps as presentation attacks for all the attempts. Thus, in order to detect such attack, a PAD module for an RGB-D face recognition system should incorporate a discriminator able to distinguish between a real depth map and a generated one. At the moment, it might be easy to develop such a discriminator. However, with the continuous improvement of techniques based on artificial intelligence, it is easy to foresee that this task may become increasingly difficult.

REFERENCES

Adelson, E. H., Bergen, J. R., et al. (1991). *The plenoptic function and the elements of early vision*, volume 2. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology.

Chiesa, V. and Dugelay, J.-L. (2018a). Advanced face presentation attack detection on light field images. In *17th International Conference of the Biometrics Special Interest Group, BIOSIG*, Darmstadt, Germany.

Chiesa, V. and Dugelay, J.-L. (2018b). On multi-view face recognition using lytro images. pages 2250–2254.

Galbally, J., Marcel, S., and Fierrez, J. (2014). Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552.

Galdi, C., Chiesa, V., Busch, C., Lobato Correia, P., Dugelay, J.-L., and Guillemot, C. (2019). Light fields for face analysis. *Sensors*, 19(12):2687.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Jain, A. K., Ross, A., and Uludag, U. (2005). Biometric template security: Challenges and solutions. In *2005 13th European signal processing conference*, pages 1–4. IEEE.

Ji, Z., Zhu, H., and Wang, Q. (2016). Lfhog: A discriminative descriptor for live face detection from light field image. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1474–1478.

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874.

Kim, S., Ban, Y., and Lee, S. (2014). Face liveness detection using a light field camera. *Sensors (Basel, Switzerland)*, 14(12):22471–22499.

Liu, M., Fo, H., Wei, Y., Rehman, Y., Po, L., and Lo, W. (2019). Light field-based face liveness detection with convolutional neural networks. *SPIE, Electronic Imaging*, 28(1).

Mallat, K. and Dugelay, J.-L. (2018). Light field-based face presentation attack detection: reviewing, benchmarking and one step further. In *17th International Conference of the Biometrics Special Interest Group, BIOSIG*.

Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P., et al. (2005). Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR*, 2(11):1–11.

Nunes, J. F., Moreira, P. M., and Tavares, J. M. R. (2012). Human motion analysis and simulation tools. *XX Encontro Português de Computação Gráfica (EPCG)*.

Raghavendra, R., Raja, K., and Busch, C. (2015). Presentation attack detection for face recognition using light field camera. *IEEE Trans. on Image Processing*, 24(3):1060–1074.

Raghavendra, R., Raja, K. B., Yang, B., and Busch, C. (2013a). Comparative evaluation of super-resolution techniques for multi-face recognition using light-field camera. In *2013 18th International Conference on Digital Signal Processing (DSP)*, pages 1–6.

Raghavendra, R., Raja, K. B., Yang, B., and Busch, C. (2013b). Improved face recognition at a distance using light field camera and super resolution schemes. In *SIN*.

Raghavendra, R., Raja, K. B., Yang, B., and Busch, C. (2013c). A novel image fusion scheme for robust multiple face recognition with light-field camera. In *Proceedings of the 16th International Conference on Information Fusion*, pages 722–729. IEEE.

Raja, K. B., Raghavendra, R., Alaya Cheikh, F., and Busch, C. (2015). Evaluation of fusion approaches for face recognition using light field cameras.

Ratha, N. K., Connell, J. H., and Bolle, R. M. (2001). Enhancing security and privacy in biometrics-based authentication systems. *IBM systems Journal*, 40(3):614–634.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18.

Sepas-Moghaddam, A., Chiesa, V., Correia, P. L., Pereira, F., and Dugelay, J.-L. (2017a). The ist-eurecom light field face database. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE.

Sepas-Moghaddam, A., Correia, P. L., Nasrollahi, K., Moeslund, T. B., and Pereira, F. (2018a). Light field based face recognition via a fused deep representation. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Sepas-Moghaddam, A., Correia, P. L., and Pereira, F. (2017b). Light field local binary patterns description for face recognition. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3815–3819.

Sepas-Moghaddam, A., Haque, M. A., Correia, P. L., Nasrollahi, K., Moeslund, T. B., and Pereira, F. (2019). A double-deep spatio-angular learning framework for light field based face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1.

Sepas-Moghaddam, A., Malhadas, L., Correia, P., and Pereira, F. (2018b). Face spoofing detection using a light field imaging framework. *IET Biometrics*, 7(1):39–48.

Sepas-Moghaddam, A., Malhadas, L., Correia, P. L., and Pereira, F. (2017c). Face spoofing detection using a light field imaging framework. *IET Biometrics*, 7(1):39–48.

Sepas-Moghaddam, A., Pereira, F., and Correia, P. (2018c). Ear recognition in a light field imaging framework: A new perspective. *IET Biometrics*, 7(3):224–231.

Sepas-Moghaddam, A., Pereira, F., and Correia, P. (2018d). Light field based face presentation attack detection: Reviewing, benchmarking and one step further. *IEEE Trans. on Information Forensics and Security*, 13(7):1696–1709.

Sepas-Moghaddam, A., Pereira, F., and Correia, P. L. (2018e). Light field-based face presentation attack detection: reviewing, benchmarking and one step further. *IEEE Transactions on Information Forensics and Security*, 13(7):1696–1709.

Zhu, S., Lv, X., Feng, X., Lin, J., Jin, P., and Gao, L. (2020). Plenoptic face presentation attack detection. *IEEE Access*.