N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb, A. Sequeira,
M. Todisco and A. Uhl (Eds.): BIOSIG 2023,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2023

Face verification explainability heatmap generation using a vision transformer[1]

Ricardo Correia, Fernando Pereira and Paulo L. Correia[2]

Abstract: Explainable Face Recognition (XFR) is a critical technology to support the large deployment of learning-based face recognition solutions. This paper aims at contributing to the more transparent usage of Vision Transformers (ViTs) for face verification (FV) tasks, by proposing a novel approach for generating FV explainability heatmaps, for both positive and negative decisions. The proposed solution leverages on the attention maps generated by a ViT and employs masking techniques to create masks based on the highlighted regions in the attention maps. These masks are applied to the pair of faces, and the masking technique with most impact on the decision is selected to be used to generate heatmaps for the probe-gallery pair of faces. These heatmaps offer valuable insights into the decision-making process, shedding light on the most important face regions for the verification outcome. The key novelty of this paper lies in the proposed approach for generating explainability heatmaps tailored for verification pairs in the context of ViT models, which combines the ViT attention maps regions of the probe-gallery pair to create masks that allow evaluating those region´s impact on the verification decision for both positive and negative decisions.

Keywords: Explainable face recognition, vision transformer, face verification heatmaps

1 Introduction

The increasing adoption of artificial intelligence (AI) tools, and deep learning (DL) models in particular, for multiple computer vision tasks, impacts users and their lives, thus accentuating the need for explainable artificial intelligence (XAI). In the context of face recognition (FR), this type of technology is known as explainable face recognition (XFR), acknowledging concerns about the lack of transparency of many FR models. In this context, understanding how a model works, especially when it fails, is crucial for improving and developing more effective FR solutions, and increase its societal acceptance. For FV, it is vital to know why impostors are wrongly validated or legitimate users are denied access.

A post-hoc XFR tool is applied after the FR model has made its decision, not changing

[1] This work has been partially supported by the European CHIST-ERA program via the *French National Research Agency (ANR) within the XAIface project (grant agreement CHIST-ERA-19-XAI-011)* and by FCT/MEC under the project UID/50008/2020.

[2] Instituto de Telecomunicações; Instituto Superior Técnico – Universidade de Lisboa; Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal, ricardo.r.nobre.correia@tecnico.ulisboa.pt, fp@lx.it.pt and plc@lx.it.pt

the model, while aiming at providing insights on how the model arrived at its decision. Various post-hoc XFR tools have been developed to enhance the explainability of FV decisions [Me22], [PDS18], [MM22], which can be categorized based on how they extract information from the FR model into propagation-based or perturbation-based tools. Propagation-based XFR tools leverage specific properties of the model being explained by considering the internal structure of the model. Perturbation-based tools make changes to the input face, e.g., masking, or altering specific input features, to assess their impact on the FV model's decision, not considering the inner working of the model.

ViTs [Va17] have recently emerged as a promising tool also for FR purposes [ZD21a]. The ViT self-attention mechanism can contribute to enhance FR explainability since propagation-based tools can explore it to provide insights about the decision made, and generate attention maps expressing the importance of different input image patches for the ViT created embeddings. Even if a ViT is often trained with a classification logic, it can still be used for FV tasks to compute embeddings for both the probe and gallery images and compare those embeddings to determine their similarity. While the ViT attention maps provide insights about the face salient regions with more influence in obtaining the desired output class, they are not necessarily appropriate to explain a FV decision, where the similarity of two faces is compared. As such, these attention maps provide important information, but they cannot be directly taken as FV explainability heatmaps. On the contrary, perturbation based XFR tools, such as Average Removal/Aggregation (AVG) [MM22] and MinPlus [Me22], do not consider the model internal structure but rather focus on the task decision, and therefore they directly output FV explainability heatmaps.

In this context, this paper proposes a novel XFR post-hoc tool for creating FV explainability heatmaps, leveraging the advantages of ViT propagation-based tools and their attention maps. Positive and negative FV decisions are treated differently since for the first case the goal is to highlight those facial regions contributing most to the similarity between probe and gallery images, while for the latter case the regions contributing to differentiate individuals should be highlighted. The key novelty of the proposed solution lies on the way attention maps regions for the probe-gallery pair are combined to create masks that allow evaluating those region's impact on the ViT FV decision, and then to generate effective FV explainability heatmaps.

This paper is structured as follows: Section 2 provides a brief overview of the state-of-the-art on explainable FV. Section 3 presents the proposed FV explainability XFR post-hoc tool, exploiting the ViT attention maps to derive ViT FV explainability heatmaps. Section 4 reports and discusses the results and findings obtained with the proposed XFR tool, comparing with state-of-the-art methods and highlighting the advantage of the proposed tool to also create explainability heatmaps for negative FV decisions. Section 5 presents final remarks and outlines future research directions.

2 Brief review on face verification explainability

Explainability tools play a crucial role to provide insights on the inner working of FV

models. While a few works propose XFR ante-hoc tools, i.e. intrinsically interpretable models that inherently provide transparency in the decision-making process [WBT22][JZ21], the main literature focus has been on XFR post-hoc tools. As discussed in the Introduction, post-hoc FV explainability tools (and FR tools) can be categorized as perturbation-based and propagation-based tools. Examples of the former include Local Interpretable Model-agnostic Explanations (LIME) [RSG16], Randomized Input Sampling for Explanation (RISE) [PDS18], AVG [MM22] and MinPlus [Me22]. These tools can be applied to any black-box model without changing the model architecture. LIME works by randomly selecting super-pixels and training a weighted model to determine their importance. RISE, MinPlus and AVG perturb different face regions to measure their effect on the model´s decision, thus creating a FV explainability heatmap.

Propagation-based post-hoc tools based on the ViT offer a more direct way to understand the internal functioning of a ViT model, by leveraging its attention mechanism [CGW20]. Examples of such tools applied to ViT include Rollout [AZ20], Gradient-weighted Class Activation Mapping (Grad-CAM) [Se17], Layer-wise Relevance Propagation (LRP) [Bi16], and a ViT-LRP tool [CGW20] which provide insights into the probe regions that are essential to its embedding representation. Rollout uses the attention matrices computed for the various attention layers to generate an attention map. These attention matrices represent the learned attention weights that capture the relationships between the different patches within the probe face image and are used to derive an attention map, visually representing the importance of each patch. Grad-CAM uses the attention matrices and combines them based on their gradients with respect to the output class to generate an attention map. ViT-LRP adapts LRP to the ViT architecture, propagating the output class relevance scores to the attention matrices of the various attention layers, enabling the identification of the most important probe patches for the obtained embedding representation. Unlike perturbation-based tools, propagation-based ViT tools do not directly provide FV explainability heatmaps, but rather attention maps that highlight important regions in the context of image classification.

3 Proposed face verification explainability tool

This section proposes a novel FV explainability heatmap generation tool that provides insights into the decision-making process by highlighting the regions that contribute the most to a positive or a negative FV decision. The architecture for the proposed FV explainability tool is depicted in Fig. 1, and its key modules to explain both types of FV decisions are detailed in the following.

3.1 Architecture and walkthrough

The proposed explainability tool complements a *Face Verification Pipeline* using a *ViT model* to perform feature extraction and produce embeddings describing the input probe and gallery images, see overall architecture in Fig. 1. The *Attention Map Generation*

module creates attention maps for the probe (P) and gallery (G) face images using a ViT propagation-based post-hoc tool. These classification-focused attention maps capture the regions that most influence the embedding representation of each input image, probe and gallery. Following the methodology outlined in [ZD21a], the *Face Verification Decision* is based on the distance, d, between the embeddings computed for the probe and gallery images using the square Frobenius norm, $d = \| e_P - e_G \|_F^2$, where $\{e_P, e_G\} \in \mathbb{R}^n$ are the probe and gallery embeddings, respectively. If this distance exceeds the decision threshold, a negative FV decision is taken. To find the optimal threshold, cross-validation is performed on the FV dataset, following the same methodology used by [ZD21a].

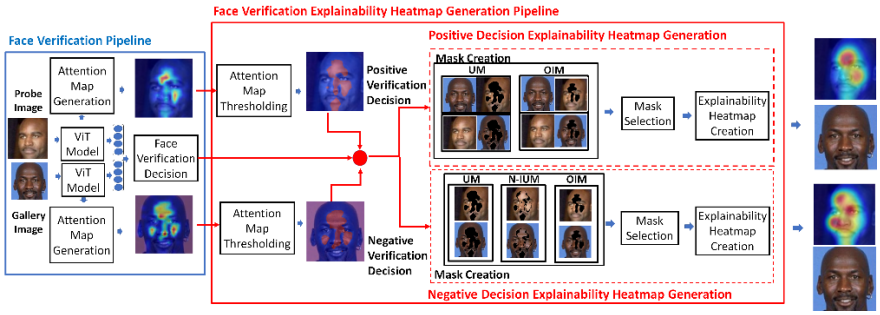

Fig. 1 Proposed face verification explainability architecture

The proposed *Face Verification Heatmap Generation Pipeline* starts with an *Attention Map Thresholding* module, applying Otsu thresholding [Ot79] to obtain a binary thresholded version of the ViT attention map, highlighting the most salient areas. Then, based on the FV decision, a different strategy for *Mask Creation* is adopted, as detailed in Section 3.2. The goal is to determine which input face areas are more relevant for the FV decision. While several masking techniques are discussed, the *Mask Selection* module selects the mask considered as most impactful for the FV decision. This involves feeding new pairs of images, obtained by applying the created masks to the probe and gallery face images, to the *Face Verification Pipeline*, and checking how the distance between the original and new embeddings changes. A selection metric is proposed to assess the impact of each candidate masking technique and guide the selection of the best mask creation technique. Finally, the *Explainability Heatmap Creation* module creates the FV explainability heatmap by considering the selected masking technique and the original probe and gallery attention maps. To get smoother heatmaps, they are filtered, applying an 8×8 dilation followed by a Gaussian filter of size 56×56 and 6.6 variance. Finally, for visualization purposes, the heatmap (with the warmer colours representing the more important regions) is overlaid with the probe face image luminance.

3.2 Positive decision explainability heatmap generation

The proposed FV explainability heatmap generation tool for positive verification decisions takes as input the probe and gallery face images and performs four main steps: (i) attention map thresholding; (ii) mask creation, where several alternative solutions are considered;

(iii) mask selection; and (iv) explainability heatmap creation. The *Attention Map Thresholding* module converts the attention maps created by the ViT propagation-based tool into a binary mask as described above. *Mask creation* considers alternative masking techniques for creating two image pairs, PP – positive probe, and PG – positive gallery, to be submitted to the FV pipeline as part of the *Mask Selection* module. The PP image pair includes the probe face image and a masked version of the gallery face image while the PG pair includes the gallery face image and a masked version of the probe face image, see Fig. 1. By excluding the regions identified as important in the corresponding attention maps, it is expected that the similarity of the new PP and PG pairs will decrease regarding the original probe and gallery face images. Two masking techniques are proposed for the positive FV case:

- **Only Intersection Masking (OIM)** – The thresholded attention maps computed in the *Attention Map Thresholding* module for both input images are combined to create a single mask, corresponding to the intersection of both the probe and gallery thresholded attention masks, thus including only the regions highlighted as important in both the probe and gallery thresholded attention maps.
- **Unified Masking (UM)** – The thresholded attention maps computed within the *Attention Map Thresholding* module for both input images are combined to create a single mask, corresponding to the union of both the probe and gallery thresholded attention masks, thus including the important regions from both face images.

For each masking technique, the generated mask is applied to both the probe and gallery images and the percentage of removed area is denoted as *RA*. The *Mask selection* module takes the PP and PG pairs generated with each mask creation technique and feeds them to the *Face Verification Pipeline*. The distances between the corresponding embeddings, d_{PG} and d_{PP} , against the original FV embeddings distance, d, are then computed; since the areas identified as important in the attention maps were excluded, the distance between embeddings is expected to increase. The selection metric used to evaluate the effectiveness of the mask creation techniques considers: (i) the variation of the distance between the embeddings prior and after the masking, which is related to the relevance of the removed areas; and (ii) the size of the removed areas, to ensure that techniques masking out larger image areas do not receive an unfair advantage, as larger masked areas tend to result in a larger increase in the embeddings distance. The proposed selection metric for positive FV decision cases (SM+) is computed as:

$$\text{SM+} = ((d_{PG} - d)/ d) / 2RA + ((d_{PP} - d)/ d) / 2RA , \quad (1)$$

where $d_{PG} = \| e_G - e_{MP} \|_F^2$ and $d_{PP} = \| e_P - e_{MG} \|_F^2$ correspond to the distance between embeddings for the PG and PP image pairs, and e_{MP} and e_{MG} correspond to the embeddings resulting from the *ViT model* module after masking the probe (*P*) and gallery (*G*) images, respectively. The best mask creation strategy is the one leading to a larger SM+ value, as this metric captures the contribution of the masked areas to the FV decision. Finally, the *Explainability Heatmap Creation* module considers the areas of the selected mask as those contributing the most to explain the FV decision. The FV explainability

heatmap values are obtained from the original attention maps for the selected area. The OIM heatmap considers only the regions common to both attention maps and, as such, it is generated by computing the average values of the attention maps for these shared regions. The UM heatmap follows the same approach, except when only one attention map contributes to a part of the UM mask, in which case its value is directly used. Finally, the filtering process described in Section 3.1 is applied.

3.3 Negative decision explainability heatmap generation

The proposed explainability heatmap generation tool for negative FV decisions differs from the positive case on the mask creation process and the selection metric. *Mask creation* aims to create one image pair, NPG – negative probe and gallery, to be fed to the *Face Verification Pipeline* as part of the *Mask Selection* module, consisting of masked versions of both the probe and gallery face images. The masking techniques aim at removing the regions that contribute to the differentiation between individuals and, by doing so, it is expected that the similarity of the new NPG pairs will increase regarding the original probe-gallery similarity. In addition to the masking techniques proposed in Section 3.2 (OIM and UM), the *Non-Intersecting Unified Masking (N-IUM)* technique combines the thresholded attention maps for the probe and gallery face images to form a single mask, including the union minus the intersection of both masks, to keep regions that are considered important in only one of the attention maps.

The *Mask Selection* module feeds the NPG image pair to the *Face Verification Pipeline* and computes the d_{NGP} distance between the new embeddings, comparing it against the original FV embeddings distance, d. By masking out areas identified as important in the attention maps, a smaller distance between embeddings is expected. For the negative FV decisions, the proposed mask technique selection metric (*SM-*) is:

$$\text{SM-} = ((\, d - d_{NGP} \,)/\, d \,)/RA, \tag{2}$$

where $d_{NGP} = \| e_{MG} - e_{MP} \|_F^2$ is the squared Frobenius norm distance between e_{MP} and e_{MG}, which correspond to the embeddings resulting from the *ViT model* module after masking the probe (P) and gallery (G) images, respectively. The best mask creation technique produces the larger *SM-* value, corresponding to a larger impact on the FV decision with a larger reduction of the d_{NGP} value weighted by RA. The *Explainability Heatmap Creation* process is completed with the same filtering described in Section 3.1.

4 Results and discussion

The ViT code available from [CGW20] was used in this paper for training and testing the ViT model as specified in [ZD21a]. Training used the large-scale database MS-Celeb-1M [Gu16], which contains 5.3 million images of 93,431 celebrities, with the CosFace loss function [Wa18]. For evaluation purposes, several FR datasets were used, notably LFW

[Hu07], Similar-looking LFW (SLLFW) [De16], Cross-Age LFW (CALFW) [ZDH17], Cross-Pose LFW (CPLFW) [ZD18] and Transferable Adversarial LFW (TALFW) [ZD21b]. The LFW dataset is split into 10 subsets of image pairs, each with 300 positive and 300 negative pairs; the LFW variants allow testing in more challenging scenarios.

Masking technique selection

The performance of the various proposed masking techniques when integrated in the proposed FV explainability heatmap generation tool is evaluated by calculating the average value of *SM+* for each masking technique (OIM, UM) across the positive decision probe-gallery pairs for each dataset, and the average value of *SM-* for each masking technique (OIM, UM, N-IUM) for the negative decision probe-gallery pairs for each dataset. A summary of the results obtained is included in Tab. 1. ViT-LRP [CGW20] is used to generate the attention maps, and all FV pairs of each dataset are considered.

Tab. 1 Masking techniques evaluation results for positive and negative face verification decisions.

FV decision	Masking technique	LFW	TALFW	CALFW	SLLFW	CPLFW
Positive	OIM	**10.00**	**5.93**	**4.97**	**9.72**	**3.89**
	UM	4.80	2.82	2.59	5.1	1.49
Negative	OIM	**1.84**	**2.01**	**2.40**	**2.84**	**2.62**
	UM	1.56	1.52	1.69	1.95	1.39
	N-IUM	1.70	1.61	1.82	2.13	1.48

The results in Tab. 1 show that, the OIM masks perform the best in capturing the important features for explaining both positive and negative decisions. For the alternative ViT attention map generation tools, notably No-Grad ViT-LRP [CGW20] (a variant of ViT-LRP not integrating gradient information from the attention matrices) and Rollout [AZ20], they both provide consistent results with the ViT-LRP results reported in Table 1 in terms of the best mask creation techniques. In this context, the *Mask Selection* module in the pipeline may not be necessary, as the results clearly indicate that OIM is the most effective masking technique for both types of FV decision. Therefore, to provide an explanation for a FV decision, No-Grad ViT-LRP OIM, ViT-LRP OIM or Rollout OIM should be used.

Evaluation of face verification explainability heatmaps

To evaluate the FV explainability heatmaps generated by the proposed explainability tool against the state-of-the-art, a set of perturbation tests were conducted by progressively masking the probe pixels in descending relevance order, in line with the approach from [CGW20]. Tests were applied to 1000 LFW true positive pairs, as the state-of-the-art explainability tools only explain positive decisions. Using No-Grad ViT-LRP, ViT-LRP and Rollout in the *Attention Map Generation* module and the OIM masking technique, these are compared against the MinPlus, LIME and RISE benchmarks, using Recall as the evaluation metric - see Fig. 2 (left). The tool leading to a faster decrease in Recall is considered the most effective. For negative decisions, only the proposed explainability

tool is evaluated as MinPlus, LIME and RISE focus only on similar regions to explain positive decisions. Tests were performed for 1000 LFW true negative pairs, with both the probe and gallery images being progressively masked based on their respective FV explainability heatmaps; here, the True Negative Rate (TNR) metric was used for evaluation. By gradually masking regions that contribute to distinguishing individuals, the expectation is that TNR decreases and the tool leading to a faster decrease will be the most effective, see Fig. 2 (right).

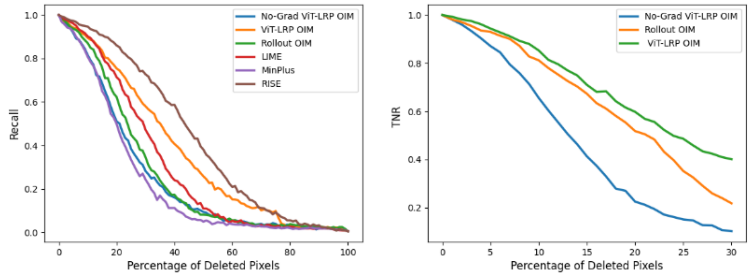

Fig. 2 Post-hoc tools evaluation: left) Recall evaluation; right) TNR evaluation

The Recall results in Fig. 2 show that the proposed No-Grad ViT-LRP OIM tool achieves the best explainability performance for the positive FV decisions along with MinPlus. For negative decisions, the TNR results show that the proposed No-Grad ViT-LRP OIM tool also performs the best. In summary, the proposed No-Grad ViT-LRP OIM tool allows generating effective explainability heatmaps for both positive and negative decisions, a major advantage over perturbation-based tools. Fig. 3 includes a few examples of FV explainability heatmaps for both types of FV decisions.

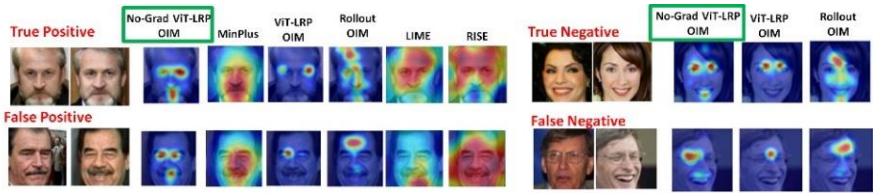

Fig. 3 Face verification explainability heatmap examples

5 Conclusions and future work

This paper proposes a novel ViT post-hoc FV explainability tool based on heatmaps, achieving comparable performance to perturbation-based tools for positive decisions. The best solution uses the original attention maps generated by No-Grad ViT-LRP [CGW20] and the OIM masking technique. A key advantage of the proposed propagation-based approach regarding the perturbation-based explainability tools is its ability to effectively explain both positive and negative FV decisions. As future work, the goal is to further leverage the ViT attention mechanisms to create a FV ante-hoc tool.

References

[AZ20] Abnar, S.; Zuidema W.: Quantifying Attention-flow in Transformers. Annual Meeting of the Association for Computational Linguistics, CoRR, 2020.

[Bi16] Binder, A. *et. al.*: Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. International Conference on Artificial Neural Networks, 63-71, 2016.

[CGW20] Chefer, H.; Gur S.; Wolf, L.: Transformer Interpretability Beyond Attention Visualization. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 782-791, 2021.

[De16] Deng W. *et. al.*: Fine-grained Face Verification: FGLFW Database, Baselines, and Human-DCMN Partnership. Pattern Recognition, 2016.

[Gu16] Guo, Y. *et. al.*: Ms-Celeb-1M: A Dataset and Benchmark for Large-scale Face Recognition. European Conference on Computer Vision, 2016.

[Hu07] Huang G. *et. al.*: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, University of Massachusetts, Amherst, 2007.

[JZ21] Jiang, H.; Zeng D.: Explainable Face Recognition Based on Accurate Facial Composition. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 1503-1512, 2021.

[Me22] Mery, D.:True Black-Box Explanation in Facial Analysis. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 1595-1604, 2022.

[MM22] Mery, D.; Morris, B.: On Black-Box Explanation for Face Verification. 2022 IEEE/CVF Winter Conference on Application of Computer Vision (WACV), Waikoloa, Hi, USA, 1194-1203, 2022.

[Ot79] Otsu, N.; A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics, 62–66, 1979.

[PDS18] Petsiuk, V.; Das A.; Saenko, K.: Randomized Input Sampling for Explanation of Black-box Models. CoRR, 2018.

[RSG16] Ribeiro M.; Singh S.; Guestrin C.: Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, San Diego, California, USA, 97-101, 2016.

[Se17] Selvaraju, R. *et. al.*: Grad-cam Visual Explanation from Deep Networks via Grandient-based Localization. 2017 IEEE International Conference on Computer Vision (ICCV)., Venice, Italy, 618-626, 2017.

[Va17] Vaswani, A. *et. al.*: Attention Is All You Need. In (I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett): Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017.

[Wa18] Wang H. *et. al.*: Cosface: Large Margin Cosine Loss for Deep Face Recognition. 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT,

USA, 5236-5274, 2018.

[WBT22] Winter, M.; Bailer W.; Thallinger G.: Demystifying Face-Recognition with Locally Interpretable Boosted Features (LIBF). 2022 10th European Workshop on Visual Information Processing (EUVIP), Lisbon, Portugal, 1-6, 2022.

[ZD21a] Zhong, Y.; Deng W.: Face Transformer for Recognition., CoRR, 2021.

[ZD21b] Zhong Y.; Deng W.: Towards Transferable Adversarial Attack Against Deep Face Recognition. IEEE Transactions of Information Forensics and Security, 1452-1466. 2021.

[ZDH17] Zheng T.; Deng W.; Hu J.: Cross-age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments. CoRR, 2017.

[ZD18] Zheng T.; Deng W.: Cross-Pose LFW: A Database for studying Cross-Pose Face Recognition in Unconstrained Environments, CoRR, 2018.