Engineering Applications of Artificial Intelligence 132 (2024) 107941

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Deep Face Decoder: Towards understanding the embedding space of convolutional networks through visual reconstruction of deep face templates

Janez Križaj [a,*,1], Richard O. Plesh [b,1], Mahesh Banavar [b], Stephanie Schuckers [b], Vitomir Štruc [a]

[a] University of Ljubljana, Faculty of Electrical Engineering, Tržaška cesta 25, Ljubljana, 1000, Slovenia
[b] Clarkson University, 8 Clarkson Ave, Potsdam, 13699, NY, United States

ARTICLE INFO

Keywords:
Deep face templates
Template inversion
Face recognition
Deep learning
Explainability
Interpretability

ABSTRACT

Advances in deep learning and convolutional neural networks (ConvNets) have driven remarkable face recognition (FR) progress recently. However, the black-box nature of modern ConvNet-based face recognition models makes it challenging to interpret their decision-making process, to understand the reasoning behind specific success and failure cases, or to predict their responses to unseen data characteristics. It is, therefore, critical to design mechanisms that explain the inner workings of contemporary FR models and offer insight into their behavior. To address this challenge, we present in this paper a novel *template-inversion approach* capable of reconstructing high-fidelity face images from the embeddings (templates, feature-space representations) produced by modern FR techniques. Our approach is based on a novel Deep Face Decoder (DFD) trained in a regression setting to visualize the information encoded in the embedding space with the goal of fostering explainability. We utilize the developed DFD model in comprehensive experiments on multiple unconstrained face datasets, namely Visual Geometry Group Face dataset 2 (VGGFace2), Labeled Faces in the Wild (LFW), and Celebrity Faces Attributes Dataset High Quality (CelebA-HQ). Our analysis focuses on the embedding spaces of two distinct face recognition models with backbones based on the Visual Geometry Group 16-layer model (VGG-16) and the 50-layer Residual Network (ResNet-50). The results reveal how information is encoded in the two considered models and how perturbations in image appearance due to rotations, translations, scaling, occlusion, or adversarial attacks, are propagated into the embedding space. Our study offers researchers a deeper comprehension of the underlying mechanisms of ConvNet-based FR models, ultimately promoting advancements in model design and explainability.

1. Introduction

Face recognition (FR) models are widely used in various applications such as video surveillance, access control, social media apps, and smart technologies, providing security and convenience benefits (Wang et al., 2023). This widespread deployment of FR technology can largely be attributed to advances in deep learning and particularly convolution neural networks (or ConvNets for short) that led to unprecedented success on various benchmarks as well as real-world recognition tasks (Wang and Deng, 2021). However, deep learning models are still often described as "*black boxes*", since they produce recognition results without revealing how they arrived at their decisions. Interpreting and understanding the underlying mechanisms behind the models' decisions, therefore, remains a challenge due to the abstract nature of the generated feature spaces and complex hierarchy of mappings applied to the input data (Li et al., 2022).

State-of-the-art ConvNet-based FR models typically accept a facial image as input and produce a fixed-size feature representation, commonly referred to as an embedding (or face template). Ideally, these embeddings are conditioned only on identity information and are invariant to changes in pose, illumination, expression, and other nuisance factors that are known to vary from image to image. With these characteristics, the embeddings of different images can be easily compared to determine if they belong to the same identity or not. However, since the embedding comparisons occur in an abstract high-dimensional feature-space, it is difficult to associate semantic meaning to the face templates or, in other words, to interpret the encoded embeddings w.r.t. the characteristics of the input face images.

A potential solution to these issues lies in *template inversion* methods. These methods aim to recover the information encoded in the face templates and generate reconstructions resembling input images. While

[*] Corresponding author.
E-mail address: janez.krizaj@fe.uni-lj.si (J. Križaj).
[1] J. Križaj and R. Plesh are first authors with equal contributions.

https://doi.org/10.1016/j.engappai.2024.107941
Received 6 September 2023; Received in revised form 1 January 2024; Accepted 17 January 2024
Available online 31 January 2024

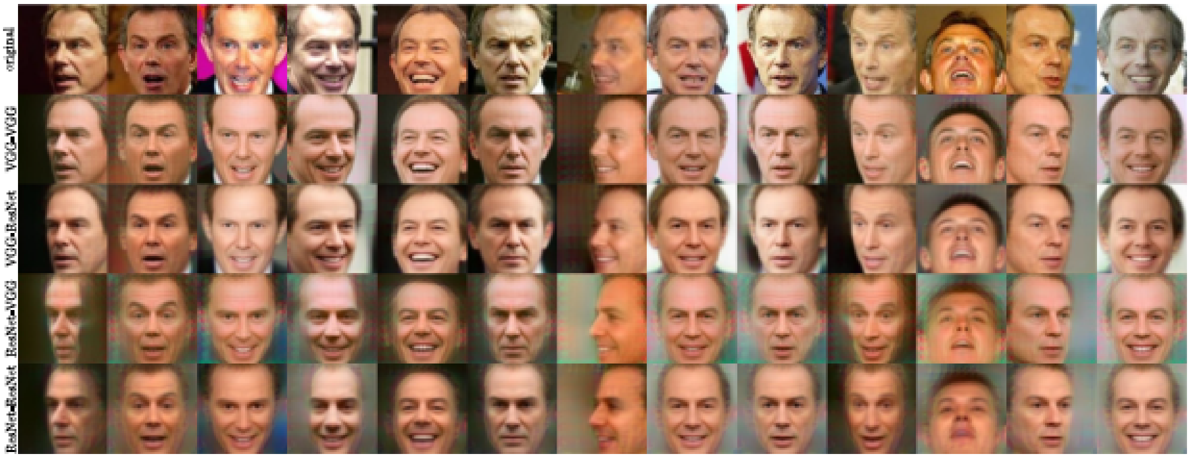


Fig. 1. We introduce a template inversion technique, named *Deep Face Decoder* (DFD), with the goal of analyzing, understanding and explaining the embedding space of ConveNet-based face recognition (FR) models. Above, we show inversion results (i.e., reconstructions, recovered images) for the embeddings, produced by two different FR models (with VGG and ResNet backbones) and two DFD variants. By comparing the original images (top row) and the generated reconstructions (rows 2–4), we are able to get insight into the characteristics of the embedding space of the FR models.

this task is challenging and requires inverting the feature extraction process of contemporary ConvNets to recover an approximation of the original face image from its embedding, it also has important implications for the understanding of the information encoded in the face templates. This capability can enhance the transparency of modern ConvNet-based facial recognition models, aid in interpreting the underlying decision-making procedures, and help discern the rationale behind model success and failure. Such transparency not only fosters a deeper grasp of contemporary deep learning-driven facial recognition technology but also aligns with the mandates of privacy laws and regulations, such as the General Data Protection Regulation (GDPR) (GDPR, 2023; Meden et al., 2021).

Although considerable progress has been made in template-inversion techniques (Dong et al., 2023; Akasaka et al., 2022; Dong et al., 2021; Mai et al., 2019), the majority of existing work focuses on the security threat of stolen biometric templates. Specifically, they aim to attack FR systems by inverting an acquired template and injecting the reconstructed image into the matching pipeline, in a so-called *template-inversion attacks*. As attack-motivated inversion techniques only care about the template distance between the reconstructed and original images, they do not necessarily maximize the visual correspondence with the original input image. As such, these techniques offer limited potential for the interpretation of the embedding space generated by modern ConvNet-based FR models. In this paper, we address this gap and develop a novel decoder model, named **Deep Face Decoder** (DFD), capable of adeptly inverting face templates and producing accurate image reconstructions that offer insights into the embedding space of contemporary FR models. We train the DFD model within a regression framework, employing specialized learning objectives. These objectives guide the model towards generating reconstructions closely resembling the original input faces, while avoiding model hallucinations often associated with inversion techniques based on Generative Adversarial Networks (Goodfellow et al., 2020).

Using the DFD model, we delve into the attributes of the embedding space of two contemporary FR models (VGG-16 and ResNet-based Simonyan and Zisserman, 2015; He et al., 2016), aiming to address key research inquiries. For instance, we investigate whether distinct ConvNet backbones in FR models encode facial details differently within their embedding spaces. We also analyze the effects of geometric face perturbations (i.e., rotations, translations, and scaling) on the generated face templates. Moreover, we examine the embedding space's response to adversarial noise. Finally, we analyze the influence of different embedding aggregation strategies on the encoded information and explore the repercussions of face template modifications. To explore these and related research inquiries, we conduct extensive experiments across three varied face datasets: VGGFace2 (Cao et al., 2018), Labeled Face in the Wild (LFW) (Huang et al., 2008), and CelebA-HQ (Lee et al., 2020). Our findings, previously unreported in the literature, shed light on these matters comprehensively.

The main contributions of this paper can be summarized into the following three points:

- We introduce the Deep Face Decoder (DFD), a state-of-the-art (SOTA) template inversion technique, designed to recover high-fidelity face images from the embeddings/templates of ConvNet-based FR models with the goal of visualizing the information encoded in the FR-model's embedding space, as also illustrated in Fig. 1.
- We utilize the proposed DFD model to gain insights into the characteristics of two (architecturally) distinct FR models and study the impact of geometric transformations, adversarial noise, occlusions, and different template computation strategies on the information encoded in the embedding space of the considered FR models.
- We make important observations about the properties of ConvNet-based FR models. For example: (*i*) We find that modern ResNet-based FR models abstract away multiple sources of image variability (e.g., pose, scale, position) when mapping input images into embeddings and do this more effectively than the earlier VGG-based FR models; (*ii*) We observe strong empirical evidence on the **equivariance** of the embedding space of the considered FR models w.r.t. geometric transformations, pointing towards the possibility of designing misalignment-correction schemes directly in the embedding space; (*iii*) We show that aggregating embeddings from multiple face images during template construction acts as a normalization process in the embedding space and produces templates that correspond to well aligned, frontal, neutral and well illuminated facial images.

2. Related work

In this section, we discuss relevant prior research with the goal of providing context for our work. We start the section with a brief review of modern face recognition techniques, continue with the current literature on the inversion of biometric templates, and finally discuss the latest research for understanding the embedding space of FR models.

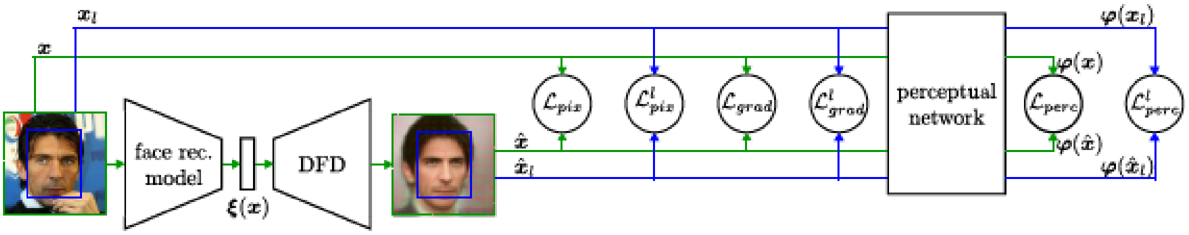


Fig. 2. High-level overview of the Deep Face Decoder (DFD) and illustration of the individual learning objectives. Given a (fixed and pretrained) face recognition model ξ and an input face image $\boldsymbol{x}$, the goal of DFD is to generate a reconstruction of the input image $\hat{\boldsymbol{x}}$ from the generated embedding $\xi(\boldsymbol{x})$ that can be used to explore the characteristics of the embedding space of ξ. DFD is trained using a multi-term loss function at both local and global scales, as further defined in Section 3.2.

3. Deep Face Decoder

In this section, we present the proposed Deep Face Decoder (DFD) that allows us to visualize (and, consequently, interpret) the information contained in the face templates, produced by contemporary, ConvNet-based FR models.

3.1. Overview

In order to analyze the information encoded in the embedding space of modern face recognition models, we design DFD as a decoder model capable of mapping the computed face templates back into the visual domain, as also illustrated in Fig. 2. Formally, given a face recognition model ξ that produces a face template (or embedding), i.e., $\boldsymbol{e} = \xi(\boldsymbol{x})$, from the provided input face image $\boldsymbol{x}$, the DFD decoder D aims to generate a reconstructed image $\hat{\boldsymbol{x}}$ that is as close (and as similar) to the input image $\boldsymbol{x}$ as possible. The parameters of D, θ_D, are optimized using a reconstruction-oriented loss function $\mathcal{L}_r$, i.e.:

$$\theta_D^* = \arg\min_{\theta_D} \mathbb{E}_x \left\{ \mathcal{L}_r(D(\xi(\boldsymbol{x}); \theta_D), \boldsymbol{x}) \right\}. \tag{1}$$

The loss function aims to recover the information contained in the template $\boldsymbol{e}$, so it becomes interpretable for humans. The optimal decoder parameters θ_D^* are typically learned over a dataset of N suitably preprocessed face images $\boldsymbol{x}$.

3.2. Loss definition

The overall objective function for training the DFD model consists of multiple losses, designed to reconstruct as much of the initial visual information from the given face template $\boldsymbol{e}$ as possible. Because face recognition models ξ are expected to produce image representations that are invariant to various nuisance factors, including pose, age, expression and others, a considerable amount of information is typically abstracted away during the template extraction step and a perfect reconstruction is, in general, not possible. We, therefore, design the learning objective as a combination of low-level per-pixel losses ($\mathcal{L}_{pix}$) and higher-level gradient-domain ($\mathcal{L}_{grad}$) and perceptual losses ($\mathcal{L}_{perc}$) of the following form with the goal of recovering an approximate face image $\hat{\boldsymbol{x}}$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = {} & \mathcal{L}_{pix} + \lambda_{grad}\mathcal{L}_{grad} + \lambda_{perc}\mathcal{L}_{perc} \\ & + \lambda_{pix}^l \mathcal{L}_{pix}^l + \lambda_{grad}^l \mathcal{L}_{grad}^l + \lambda_{perc}^l \mathcal{L}_{perc}^l, \end{aligned} \tag{2}$$

where $\lambda_{grad}, \lambda_{perc}, \lambda_{pix}^l, \lambda_{grad}^l$ and λ_{perc}^l are balancing weights. The individual losses are applied separately over the narrow/local facial area $\boldsymbol{x}_l$ (marked with the superscript ^{l} above), but also the complete input image $\boldsymbol{x}$ that contains a larger degree of contextual information. Such an approach allows for the tuning of the overall objective towards the most expressive regions of the face, while still taking the encoded contextual information into account (Hu and Ramanan, 2017). Details on the individual loss terms (defined over $\boldsymbol{x}$) are given below:

- The **pixel loss** ($\mathcal{L}_{pix}$) encourages the DFD model to reconstruct images $\hat{\boldsymbol{x}}$ that are as close as possible to the original images $\boldsymbol{x}$ in terms of low-level pixel intensities. In other words, the loss aims to recover the exact visual appearance of the input image from the face template $\boldsymbol{e}$ and is defined by a squared L_2 error norm:

$$\mathcal{L}_{pix} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \|\mathbf{x} - D(\xi(\mathbf{x}); \theta_D)\|^2. \tag{3}$$

The L_2 loss considers each pixel individually and neglects correlations between neighboring pixels. To address this issue, we incorporate gradient and perceptual losses into our overall learning objective, as defined in the following sections.

- The **gradient loss** ($\mathcal{L}_{grad}$) serves a complementary role to the pixel loss defined above. When using only pixel-level losses, the reconstructed images $\boldsymbol{x}$ tend to be overly smooth and without sharp edges that are typically key for perceiving the structural content of an image (Ma et al., 2020). To this end, we define the gradient loss as a squared L_2 error norm in the gradient domain of the image. Formally, this can be written as:

$$\mathcal{L}_{grad} = \|\nabla\mathbf{x} - \nabla\hat{\mathbf{x}}\|^2 = \|\nabla\mathbf{x} - \nabla D(\xi(\mathbf{x}); \theta_D)\|^2. \tag{4}$$

- The **perceptual loss** ($\mathcal{L}_{perc}$) is the final component of the optimization objective and helps to penalize differences in higher-level semantics between the original and reconstructed images. This loss is paramount for the capabilities of the DFD model, as it allows to recover images $\hat{\boldsymbol{x}}$ that contain similar semantic content to the inputs, while not requiring perfect per-pixel correspondences. This aspect is particularly important for DFD since some of the information initially contained in the face image $\boldsymbol{x}$ may have been discarded or abstracted away during the template-computation process, i.e., $\xi(\mathbf{x})$. Inspired by the work in Johnson et al. (2016), we define the perceptual loss in the form of a squared L_2 error norm in the feature space of a perceptual network φ, i.e.,

$$\mathcal{L}_{perc} = \|\varphi(\mathbf{x}) - \varphi(\hat{\mathbf{x}})\|^2 = \|\varphi(\mathbf{x}) - \varphi(D(\xi(\mathbf{x}); \theta_D))\|^2. \tag{5}$$

Note that the above losses are defined only over the input images $\boldsymbol{x}$, but the same definitions also apply for the local facial regions $\boldsymbol{x}_l$ and the corresponding losses $\mathcal{L}_{pix}^l$, $\mathcal{L}_{grad}^l$ and $\mathcal{L}_{perc}^l$. The use of local and global losses allows us to put additional emphasis on the central region, while also reconstructing the context in a meaningful manner. As we demonstrate empirically in the experimental section, this leads to minor reconstruction improvements.

It is also worth emphasizing that we intentionally avoid adversarial losses (in a GAN framework) when learning the DFD decoder. While such losses help with the photo-realism of the reconstructions, they are known to lead to visual distortions that impact the interpretation of the recovered visual information, as emphasized in Korkmaz et al. (2022) and Blau and Michaeli (2018).

3.3. Model architecture and training

We use an inverted VGG architecture for the implementation of the DFD decoder D (Yasrab, 2018). Such a decoder architecture has

[illegible]

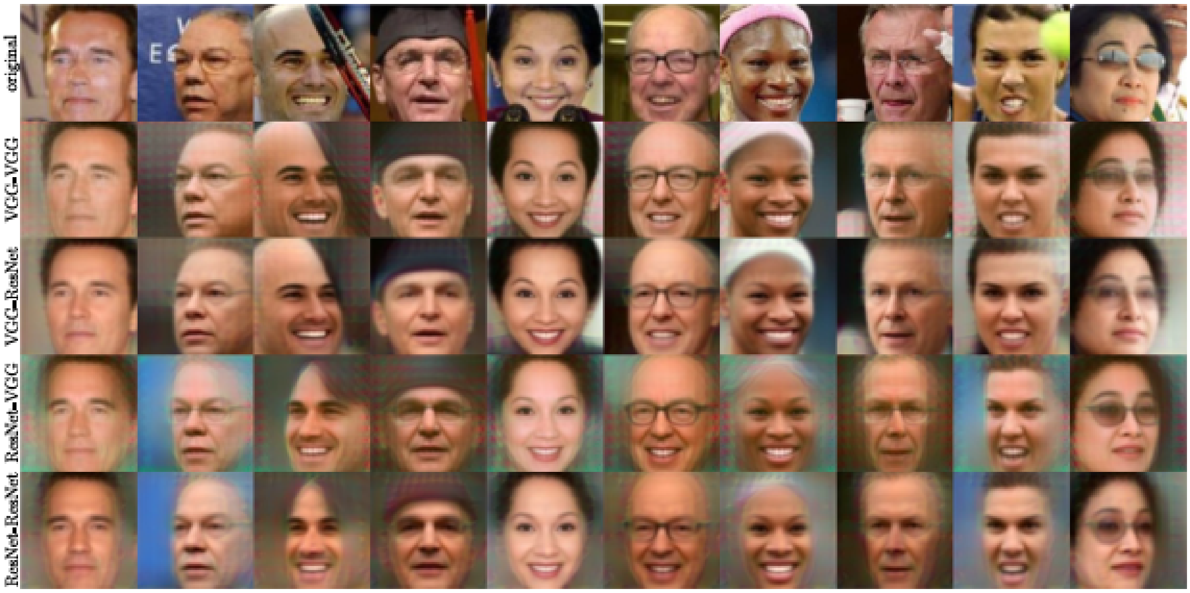


Fig. 3. Example DFD reconstructions of selected sample images from the LFW dataset: original images (1st row), reconstructions of the VGG–VGG DFD model (2nd row), reconstructions of the VGG–ResNet DFD model (3rd row), reconstructions of the ResNet–VGG DFD model (4th row), reconstructions of the ResNet–ResNet DFD model (5th row).

the perceptual network during training utilizes the ResNet model (referred to as ResNet–ResNet). It is important to highlight that the perceptual features are derived not from the embedding space of the FR model, but rather from specific internal convolutional layers, so $\xi \neq \varphi$. This distinction holds true despite the fact that the same ConvNet model is utilized for the implementation of ξ and φ.

- **Black-box experiments**: In this configuration, we operate under the premise of having access solely to the calculated face templates, without direct access to the actual FR model. Consequently, a distinct network from the targeted FR model (to be analyzed) serves as the origin of perceptual features. This configuration corresponds to a more realistic setting, where the DFD decoder has to be learned from a collection of face embeddings and corresponding enrollment images. The black-box experiments involve training the DFD decoder for the VGG-16 model using perceptual features provided by the ResNet model (referred to as VGG–ResNet), as well as optimizing the ResNet DFD decoder with perceptual features from the VGG-16 model (referred to as ResNet–VGG).

4.2. DFD validation

In the first series of experiments, we investigate the reconstruction capabilities of the proposed DFD decoder and study some initial characteristics of the VGG and ResNet FR models through qualitative experiments. To validate the suitability of the DFD models as a visualization tool that can be used to investigate the characteristics of the embedding space of ConvNet-based FR models, we also analyze the model in comparison to competing solutions from the literature and explore the impact of the individual loss terms on performance within an ablation study.

4.2.1. Visualizing face templates with DFD

Exploring backbones. In Fig. 3, we present the reconstructions of a diverse set of images from the LFW dataset in the white- and black-box decoding scenarios. Several interesting observations can be made from the presented examples: (*i*) The white and black-box scenarios both lead to visually similar results when decoding the embeddings of a specific FR model. This suggests that the source of perceptual features is less important when learning the DFD model than the characteristics of the FR embedding space, where the visual information is encoded. More importantly, this finding implies that even without access to the targeted FR model, it is possible to reconstruct a similar amount of information, as in the case when the targeted FR model is available and completely transparent. This also suggests that certain types of model obfuscations are more effective than others at preventing template inversion attacks. Namely, that concealment of model architecture is insufficient at preventing template inversion attacks if the attacker has a means of obtaining image-feature pairs to train a template decoder. (*ii*) The reconstructions from the VGG embeddings exhibit higher correspondence with the original input samples than the reconstructions, produced from the ResNet templates. With the VGG embeddings, many variable image attributes, such as pose, accessories, and background information, still appear to be present in the recovered images, whereas the same attributes are largely removed through the ResNet embedding. This observation points towards better FR robustness of the ResNet model and more suitable encoding in the embedding space. (*iii*) Partial face occlusions are treated differently by the two FR models. While the VGG embeddings seem to encode the occlusions as semantically meaningful image attributes (e.g., as part of the background — see 3rd and 4th column, as part of the body/hair — see 5th and 7th column, face color changes — see 8th and 9th column of Fig. 3), the ResNet embedding only retain information from the informative part of the facial region and discard the rest. This point to fundamental differences in the behavior of both models and provides insight into the performance of both models observed in the literature (Grm et al., 2018).

Exploring loss functions. Fig. 4 shows reconstructions obtained using two distinct loss functions for training the embedding network: SoftMax and ArcFace, both implemented using the ResNet50 backbone. ArcFace, built upon an angular-margin softmax loss, is specifically designed to enhance inter-class separation in facial recognition and robustness to identify variation. Despite this, our results demonstrate that ArcFace embeddings still encode common identity variations in pose, illumination, and expression. Furthermore, we observe that the embeddings also encompass details of facial accessories, like eyeglasses. Since this information does not inherently pertain to identity, its presence indicates that there is still room for refinement in achieving more compact and discriminative facial representations.

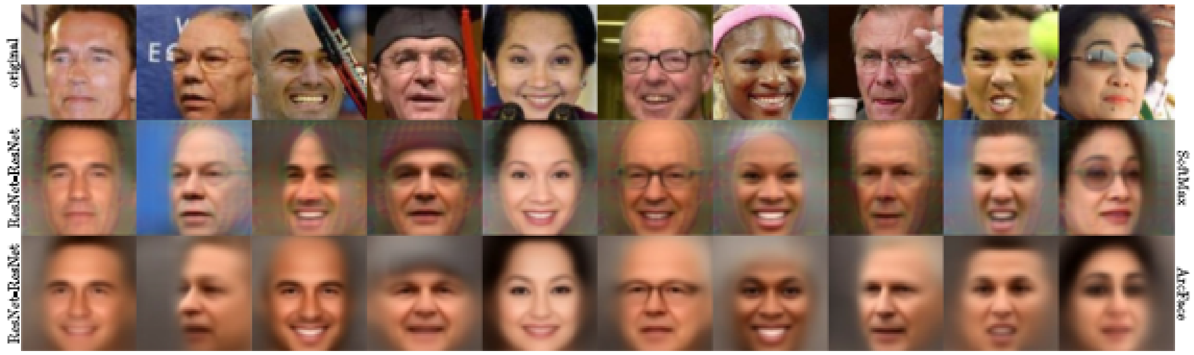


Fig. 4. Reconstructions derived using SoftMax and ArcFace loss functions with the ResNet50 backbone. Although ArcFace is designed to enhance inter-class separation in facial recognition, both embeddings capture variations in pose, illumination, and expression. ArcFace reconstructions differ from SoftMax in the nuanced encoding of facial accessories, such as eyeglasses, and the delineation of facial borders.

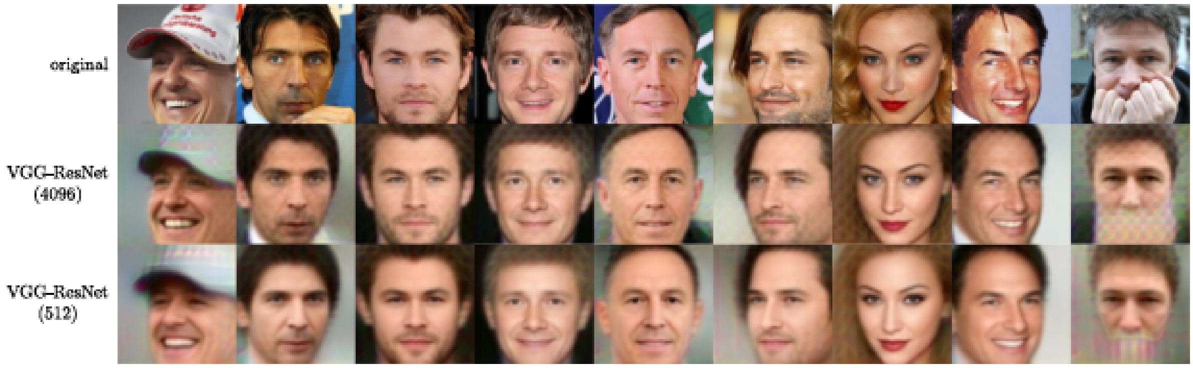


Fig. 5. Reconstructions of embeddings of two different sizes with VGG-16 backbone. The 4096-dimensional embeddings are computed from the last fully connected layer, with the classification layer being discarded. The 512-dimensional embeddings are derived by averaging the output of the last convolutional layer.

Exploring embedding dimensionality. Fig. 5 depicts the reconstructions based on embeddings of two distinct dimensions, both utilizing the VGG-16 architecture as their backbone. Embeddings of 4096 dimensions are derived from the final fully-connected layer, excluding the classification layer. Conversely, the 512-dimensional embeddings are obtained by averaging the outputs from the terminal convolutional layer. Both these layers are frequently employed in the literature for embedding extraction. Observations suggest that the number of dimensions in the embeddings exerts minimal influence on the quality of the reconstructions. However, the reconstructions from the lower-dimensional embeddings exhibit fewer artifacts. This distinction might be attributed more to the inherent characteristics of the embedding layers (convolutional versus fully-connected) rather than the dimensionality itself.

4.2.2. *Comparison with competing inversion techniques*

To put the decoding results produced by our DFD model into perspective, we conduct a visual comparison between the DFD reconstructions and the reconstructions generated by two contemporary state-of-the-art (SOTA) template inversion techniques, proposed by Dong et al. (2021) and Mai et al. (2019). We note, however, that the competing techniques were developed to study template inversion attacks, where the goal is to produce a sample image from the given template that successfully matches with a subject enrolled in a face recognition system. Thus, the generated images are allowed to look differently from the input image, as long as the FR model recognizes the subject in the two images as being the same.

For a fair comparison, we extracted the visual results directly from the relevant original papers (i.e., Dong et al. (2021) and Mai et al. (2019)) and applied the DFD model to the same test images. In Figs. 6 and 7 we present a visual comparison of the generated results together with Peak-Signal-to-Noise-Ratio (PSNR) scores that measure the quality of the reconstructions in comparison to the original input images. As can be observed, the DFD model consistently yields higher PSNR values than the two competitors, while ensuring competitive visual quality of the reconstructions both in the white-box as well as in the black-box setting. Compared to the results of the competing techniques, DFD generates reconstructions with higher correspondence to the input samples, competitive identity-recovery/visualization capabilities, and visual characteristics that to a large extent depend on the FR model utilized to produce the initial face templates. Thus, the presented results reaffirm the suitability of the DFD model as a tool for studying and interpreting the embedding space of ConvNet-based FR models.

4.2.3. *Ablation study*

To further validate the DFD model, we present in Table 1 an ablation study that explores the impact of the individual loss terms from Eq. (2). As the different DFD configurations are affected by the loss terms similarly, we present results for the VGG–ResNet DFD variant exclusively to maintain table brevity. Two important observations can be made from these results: (*i*) All of the considered losses contribute to the perceptual quality of the recovered images and improve the (average) Peak Signal To Noise Ratio (PSNR), Structural Similarity (SSIM) (Horé and Ziou, 2010), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) scores of the reconstructions that measure the correspondence with the original input images. As can be seen, incorporating local losses significantly enhances the reconstruction of the most expressive facial regions such as the mouth, nose, and eyes while leading to minor quantitative improvement. The use of perceptual loss adds minor low-level image artifacts and when used exclusively causes global color errors, causing a washed-out appearance in the reconstructed image. Similarly, exclusively utilizing either the pixel loss or gradient loss also yields subpar facial reconstructions when compared to the output ensured by the complete loss function from (2).

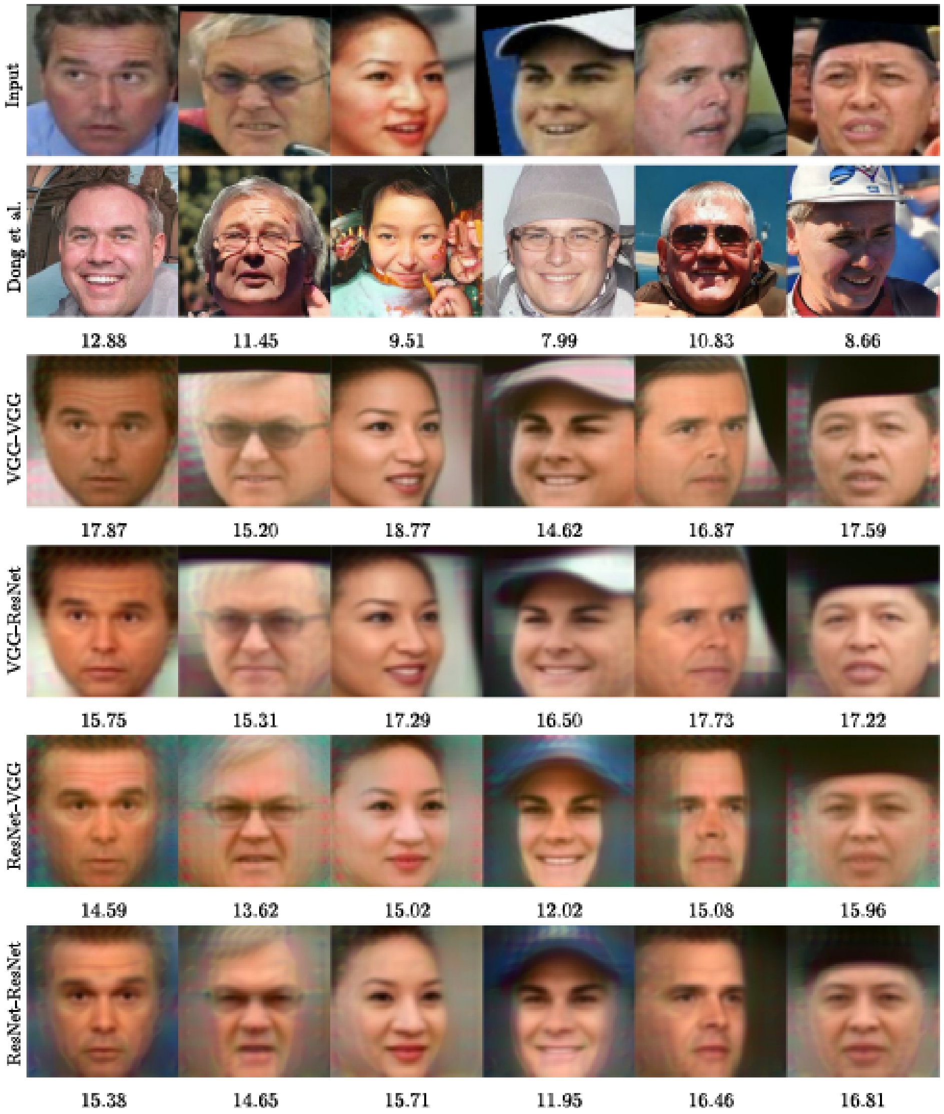


Fig. 6. Visual comparison with SOTA: original images from LFW (1st row), reconstructions from Dong et al. (2021) (2nd row), DFD reconstructions (from 3rd row onward). Below the images are the PSNR values obtained in comparison to the originals from the first row.

(*ii*) While the different losses impact the visual appearance and perceptual quality of the reconstructions, they do not affect how the encoded facial information is visualized. For example, image attributes, such as pose, hats, hair, and partial occlusions are interpreted similarly, visual background information is still comparable in all images, and overall, all aspects important for the interpretation of the embedding space remain stable when using different loss combinations, which is important for the applicability of the DFD model.

In the lower part of Table 1, we show visual results for all four considered DFD variants when using the complete loss function from Eq. (2). Note that the reconstructions from the ResNet embeddings generally lead to somewhat lower correspondence scores (PSNR, SSIM, LPIPS) than their VGG-based counterpart. However, this can be ascribed to the properties of the ResNet embeddings, which appear to abstract away a considerable amount of pose and background information, (which is highly desired to ensure the robustness of FR models) and consequently lead to lower correspondence with the initial input samples.

4.3. *Understanding appearance variations*

In the next series of experiments, we apply the DFD model to explore how various appearance perturbations impact the information encoded in the face templates. We study three different types of perturbations, i.e.: (*i*) geometric perturbations, specifically, face rotation, translation and scaling, (*ii*) partial occlusions of salient facial regions, and (*iii*) additions of adversarial noise.

4.3.1. *Geometric perturbations*

When investigating the impact of geometric perturbations on face embeddings, we are particularly interested in the **equivariance** properties of the considered FR models. In other words, we are interested in whether the geometric transformations of the input images, such as rotations or translations, can also be modeled in the embedding space of ConvNet-based FR techniques. Such properties have important implications for the design of FR systems in practice, because of their potential for designing transformation-invariant face representations and for enhancing the robustness of existing recognition models towards off-center and off-angle faces through a template augmentation process.

To explore the equivariance of ConvNet-based FR models, we learn the geometric transformations directly in the face embedding space using a subset of training images from the VGGFace2 dataset. Here, we first apply specific geometric transformations to the training images and then calculate the least squares estimate (Golub and van Loan,

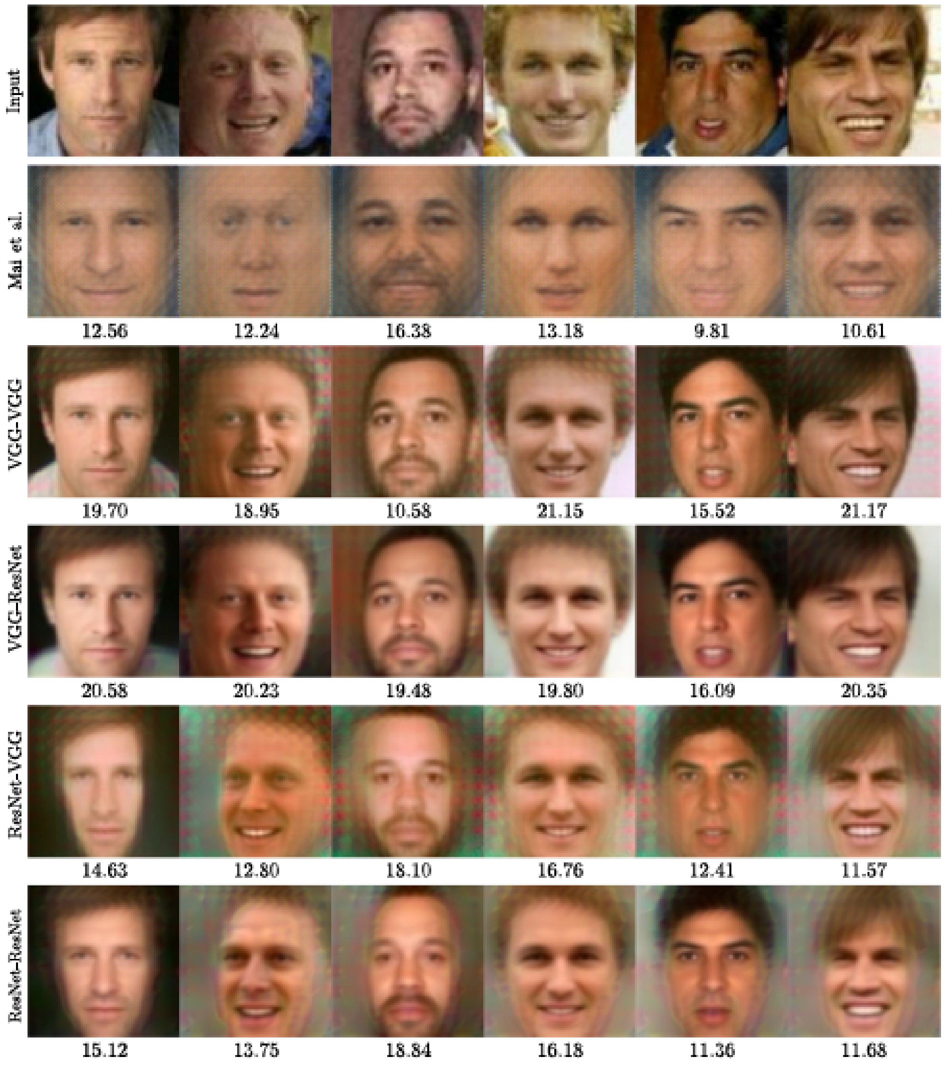


Fig. 7. Visual comparison with SOTA: original images from LFW (first row), reconstructions from Mai et al. (2019) (second row), DFD reconstructions (third row). Below the images are the PSNR values obtained in comparison to the originals from the first row.

2013) of the mapping parameters between the original image embeddings and the embeddings of the corresponding transformed images. The relationship is defined by

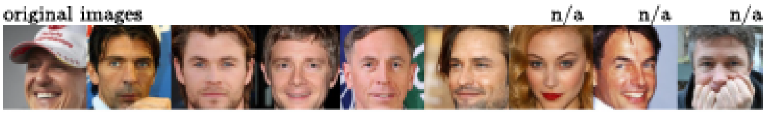
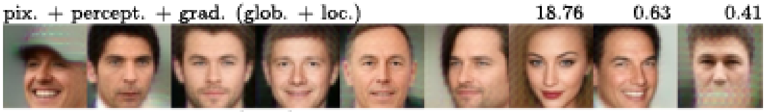
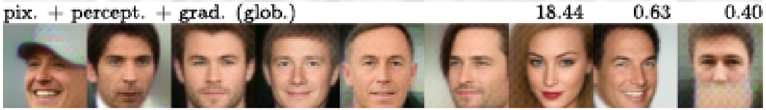
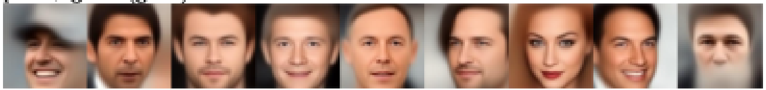
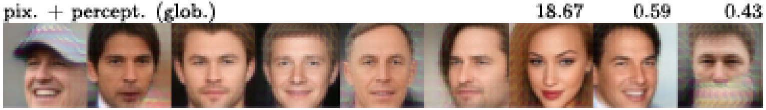
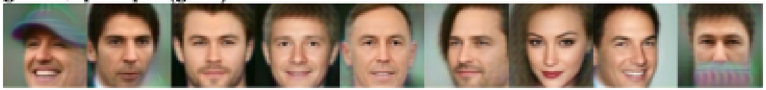
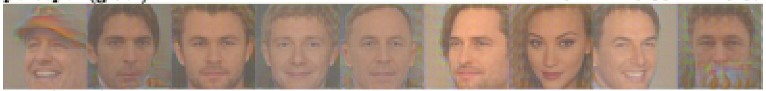
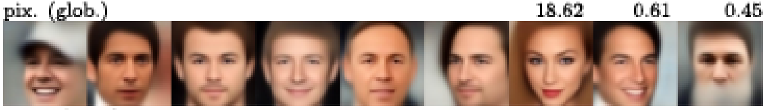
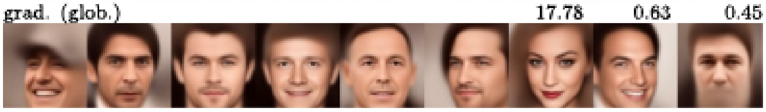
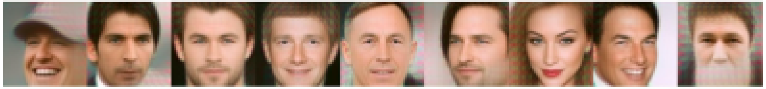
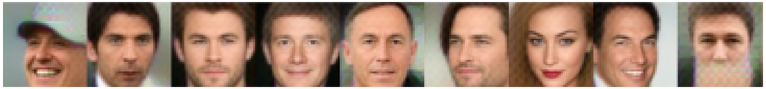
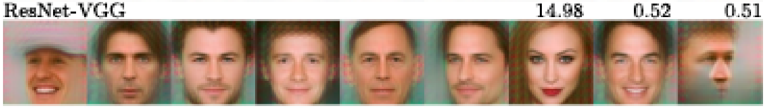
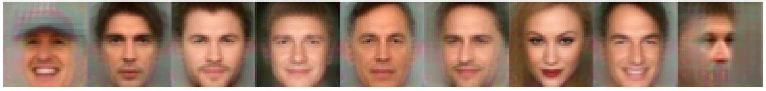
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \|\boldsymbol{e}_i - \boldsymbol{M}_\theta(\boldsymbol{e}'_i)\|^2, \tag{6}$$

where $\hat{\theta}$ represents the estimated mapping parameters, while $\boldsymbol{e}_i$ and $\boldsymbol{e}'_i$ denote the embeddings of the original and transformed images, respectively, and N denotes the number of images used to compute the mapping. The (linear) mapping function $\boldsymbol{M}_\theta$ is parameterized by θ and applied to the transformed image embeddings. In case the considered FR models are in fact equivariant with respect to the geometric transformation, the learned mapping for each transformation type should allow us to modify the embeddings of a given transformed face image such that the reconstruction of the modified embedding appears in its initial form, i.e., with the transformation undone.

We visually examine the impact of the mapping for the following geometric perturbations:

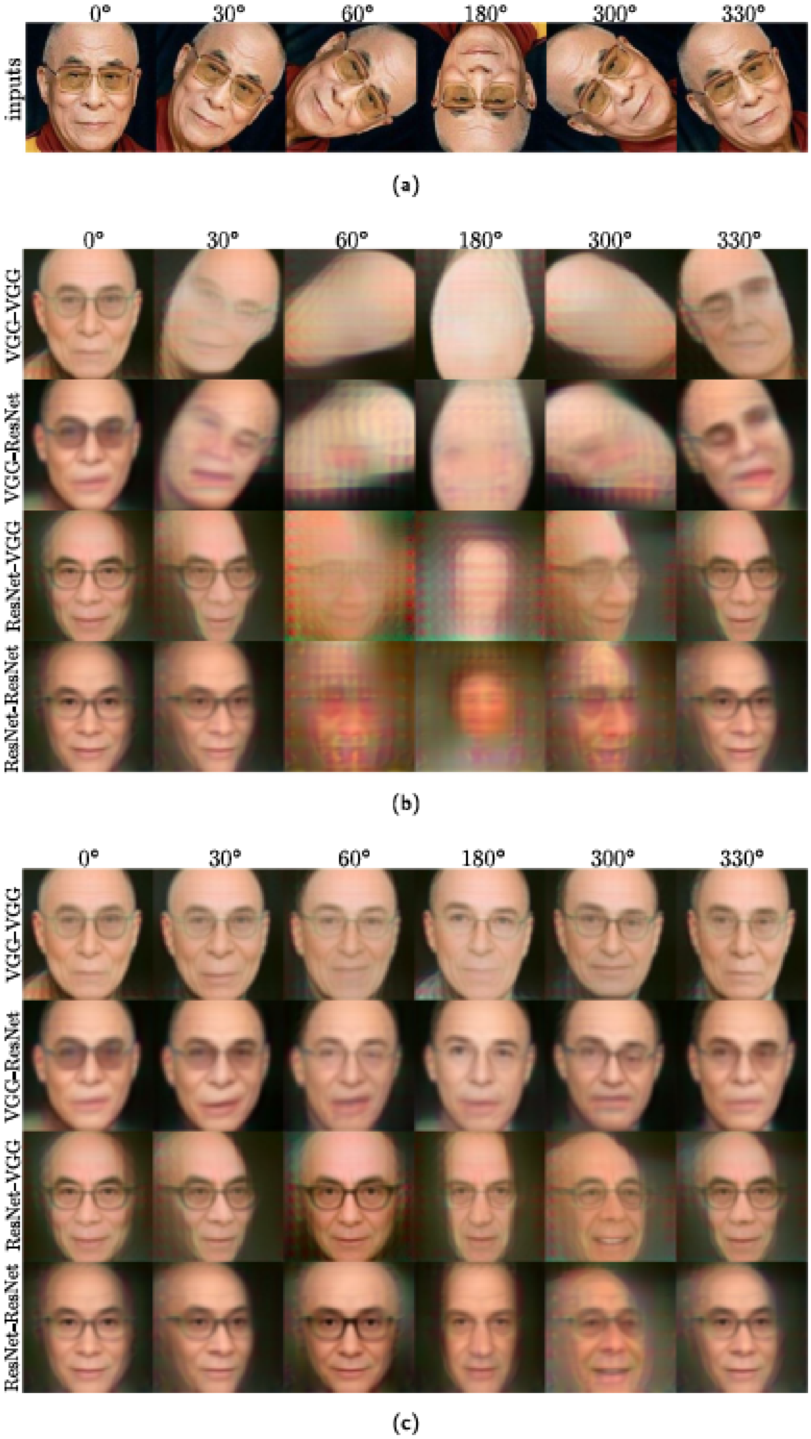
- **Rotations.** To estimate the mapping that models rotations in the embedding space, we analyze the relationship between a set of embeddings for the original face images (in an upright orientation) and the embeddings of the input images rotated in 30° increments, as shown in Fig. 8(a). Through this process, we learn the mapping for each 30° rotation step, which allows us to transform the embedding of any given rotated face image, such that the reconstructed image appears upright. A comparison of the reconstruction results without (Fig. 8(b)) and with (Fig. 8(c)) the learned transformations demonstrates that the considered FR models indeed exhibit a certain level equivariance with respect to rotations. The recovered faces are virtually unrecognizable for both FR models with rotation angles greater than 30° (in either direction) when decoded from the original unaltered embeddings (Fig. 8(b)). Conversely, the faces become properly discernible when the mapping is applied in the embedding space. In this case, the rotation is not only compensated for, the reconstructed faces also correspond reasonably well in terms of appearance to the reconstructions of the unrotated images, as seen from Fig. 8(c).
- **Translations.** To model translations within the embedding space, we estimate the mapping $\boldsymbol{M}_\theta$ on training images shifted in four distinct directions. Through this process, we acquire embedding transformations for each translation direction, enabling us to modify the embedding of a given shifted face image such that its reconstruction appears centered. When looking at the reconstruction results without (Fig. 9(a)) and with (Fig. 9(b)) the learned transformations, we observe that: (*i*) the VGG embeddings are impacted severely from translations of the facial images and poorly encode identity information if the faces are not well aligned.

Table 1
Ablation study exploring the impact of different loss functions and model variants.

		PSNR	SSIM	LPIPS
	original images	n/a	n/a	n/a
				
loss	pix. + percept. + grad. (glob. + loc.)	18.76	0.63	0.41
				
	pix. + percept. + grad. (glob.)	18.44	0.63	0.40
				
	pix. + grad. (glob.)	18.41	0.63	0.43
				
	pix. + percept. (glob.)	18.67	0.59	0.43
				
	grad. + percept. (glob.)	17.84	0.63	0.43
				
	percept. (glob.)	14.02	0.38	0.69
				
	pix. (glob.)	18.62	0.61	0.45
				
	grad. (glob.)	17.78	0.63	0.45
				
architecture	VGG-VGG	17.71	0.62	0.40
				
	VGG-ResNet	18.44	0.63	0.40
				
	ResNet-VGG	14.98	0.52	0.51
				
	ResNet-ResNet	15.28	0.53	0.49
				

While one could argue that this is a property of the DFD model that is trained on aligned face images, the results for the ResNet embeddings suggest the opposite, since reasonable reconstructions are seen for the recovered images in Fig. 9(a) despite the translated inputs. This result again speaks of the robustness of ResNet embeddings, which appear to exhibit better invariance w.r.t. to input translations than the VGG FR counterparts. *(ii)* After applying the embedding transforms, both VGG and ResNet models lead to more consistent reconstructions, suggesting that it is possible to devise misalignment-compensation techniques directly in the embedding space of ConvNet-based FR models through simple linear transforms, which is particularly strong finding, not reported earlier in the open literature, to the best of our knowledge.

- **Scaling.** To model scaling in the embedding space, we estimate the mapping $\boldsymbol{M}_\theta$ on a set of training images scaled by 0.5 and 1.5, as illustrated in Fig. 10. From this process, we first learn the embedding transformations for each scale change and then transform the embedding of a given scaled face image, so that the reconstruction appears with a neutral scaling. From the results in Fig. 10(a), we again see that the VGG FR model is sensitive to the scale of the input face images. with the generated embeddings



(a)

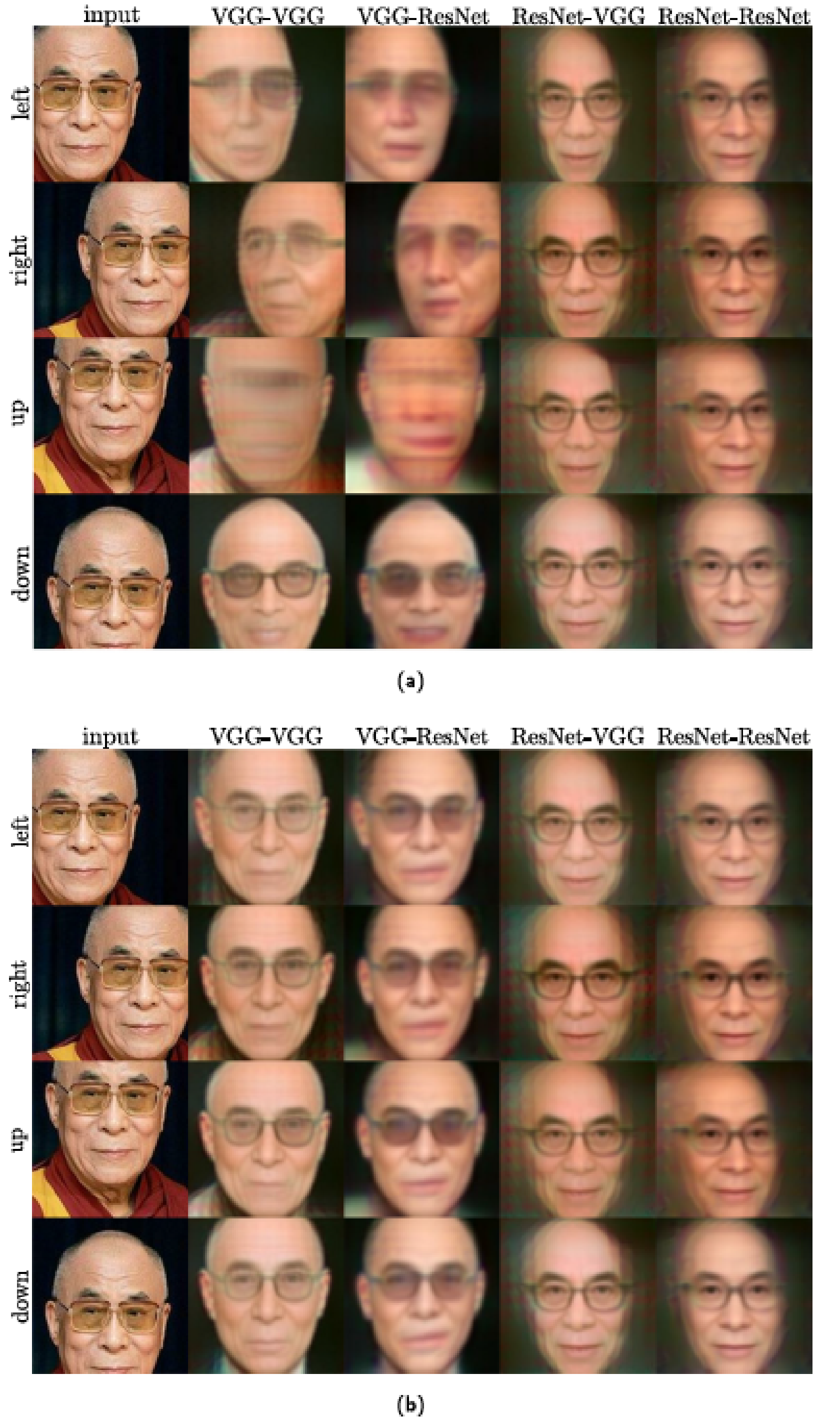
(b)

(c)

Fig. 8. Rotations in the embedding space: (a) input images rotated by different degrees; (b) reconstructions from the non-transformed embeddings of the input images; (c) reconstructions from the transformed embeddings of the input images.

leading to image reconstructions with poorly visible features and limited identity correspondence. The ResNet embeddings, on the other hand, still encode identity information to a certain extent and naturally account for the input scale changes. When the learned mapping is applied in the embedding space, the transformed templates (VGG and ResNet) better compensate for the scale changes and lead to consistent and scale-normalized reconstructions.

We demonstrate the importance (and some of the implications) of the observations made above through face-verification experiments on the LFW dataset. Specifically, we compare the verification performance of the original LFW images according to the standard protocol and the performance, when one of the images in each pair is geometrically perturbed (i.e., either rotated by 30°, horizontally translated by 20% of the width of the detection window, scaled to 0.5 of the original size). We consider both scenarios, with and without the mapping procedure in the embedding space. As can be seen from the results in Table 2, for each type of geometric perturbation, the mapping procedure leads to improved verification performance. Notably, this improvement is more pronounced in the case of the VGG embeddings, which appear to be less



(a)

(b)

Fig. 9. Translations in the embedding space: (a) reconstructions from the non-transformed embeddings of the input images; (b) reconstructions from the transformed embeddings of the input images.

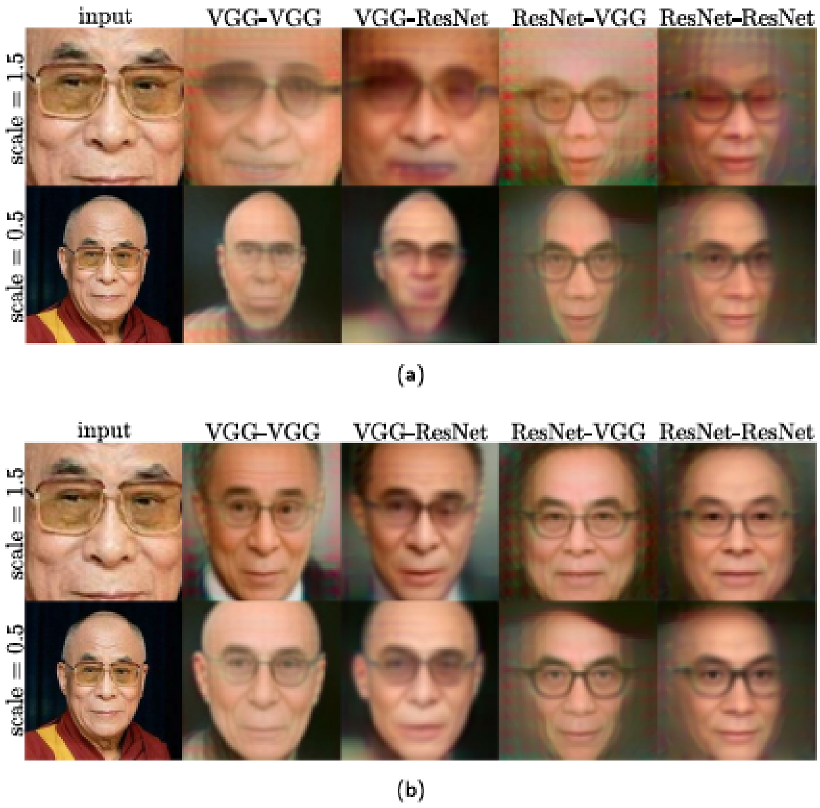
robust to image transformation than the ResNet embeddings. Nonetheless, the observed consistent performance improvements suggest that normalizing for misalignment in the embedding space is feasible and leads to performance gains even if a simple scheme, such as the one used in this paper, is utilized.

4.3.2. Facial occlusions

Next, we analyze the impact of partial occlusions of prominent facial areas on the information encoded in the face templates. To this end, we consider homogeneous block occlusions of two key face regions, i.e., the eyes and the mouth. The results in Fig. 11 present reconstructions derived from a couple of occluded face images for all DFD variants

Table 2
Verification performance (TARs (%) at 0.1% FAR) on the LFW using original and mapped embeddings.

Mapping	VGG/ResNet			
	Orig. vs. orig.	Orig. vs. rotated	Orig. vs. translated	Orig. vs. scaled
Without	79.2/99.3	62.0/98.9	76.4/99.1	31.4/88.9
With	n/a	67.8/99.0	78.8/99.2	48.3/94.0



(a)

(b)

Fig. 10. Scaling in the embedding space: (a) reconstructions from the non-transformed embeddings of the input images; (b) reconstructions from the transformed embeddings of the input images.

and explore how the two FR models interpret occlusions, which are generally considered to be problematic for contemporary FR models.

Interestingly, all models appear to consistently interpret the artificial eye occlusions as glasses. This implies a shared underlying mechanism for dealing with eye region occlusions and suggests that ConvNet-based FR models map images into embeddings that lie on a learned manifold that corresponds to semantically meaningful facial images, in our case, faces with sunglasses. Conversely, the occlusions of the mouth region are predominantly perceived as open and smiling mouths. This interpretation is likely again a consequence of the morphological similarity between the type of occlusion and the typical appearance of an open/smiling mouth, and the mapping onto the learned semantically-meaningful embedding manifold. It is also interesting to observe that the identity information is still largely discernible in the reconstructed images, that attribute information is retained (e.g., gender), and that the local occlusions remain comparably local in the recovered images.

4.3.3. *Adversarial attacks*

The last type of appearance perturbation we study in this section is adversarial noise. Adversarial noise is typically generated through an adversarial attack that aims to modify the input image in such a way that a ConvNet FR model produces incorrect (or ambiguous) recognition results. Due to the importance and implication of adversarial attacks for the security of biometric systems, it is critically important to understand their impact on the embedding space of modern FR models. For the experiments, we implement two distinct targeted adversarial-attack techniques: the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) and the method proposed by Carlini and Wagner (CW) (Carlini and Wagner, 2017), and consider the original targeted (tFGSM and tCW) as well as the iterative targeted (itFGSM and itCW) variant (Kurakin et al., 2017). The latter allows for adversarial attacks with lower noise levels. Both types of techniques rely on a softmax layer to attack facial images.

In Fig. 12, we present the results of our experiment with an input image of “A. Carr” and the target identity provided by the image of “T. Maze”. The second row of the figure shows the attacked “A. Carr” image distorted by different adversarial attacks, the remaining rows show reconstructions from the embeddings of these distorted images using different DFD variants. The identity probability above each image is determined by the ResNet-50 model (He et al., 2016), trained on 8631 identities from the VGGFace2 database (Cao et al., 2018). Notably, when the input image “A. Carr” is distorted by an adversarial attack to resemble “T. Maze”, the identity classifier tends to make a correct prediction more frequently if the embedding of the attacked image is decoded through our DFD model before classification. This is especially true for the VGG model embeddings, which we already observed earlier to lead to reconstructions with a high level of correspondence with the original input image. Nevertheless, we also see correct identity predictions (and altered to identities different than “T. Maze”) for the ResNet models and all investigated adversarial attacks. No significant differences are observed between white and black-box experiments. The presented observations have interesting implications: (*i*) The adversarial attacks appear to have a limited impact on the face embeddings and are mostly causing incorrect predictions at the softmax layer, suggesting that similarity-based matching schemes should be less affected by adversarial attacks than classifier based models (i.e., as far as the considered attacks are concerned). (*ii*) While primarily designed as a visualization tool, the DFD model offers a certain level of defense against adversarial attacks, as evidenced by the identity probabilities reported above the images.

To further support this last observation, we perform a number of verification experiments on the LFW dataset (Huang et al., 2008), where one of the images in each matching verification pair is first distorted by FGSM attack (red curve in Fig. 13) and later reconstructed by the proposed decoder (dotted red curve in Fig. 13). As a benchmark, we plot the verification rates of the original (unattacked) image pairs

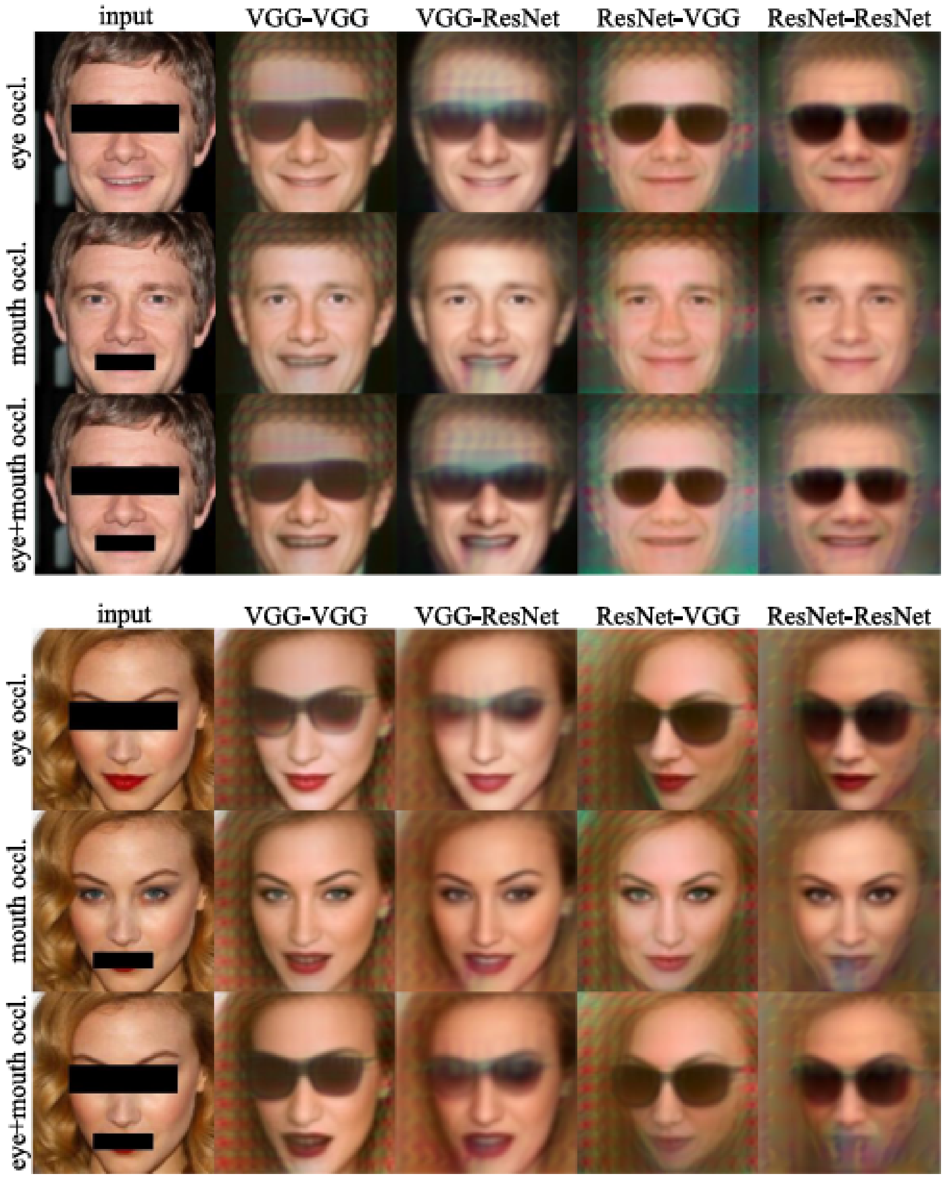


Fig. 11. Analysis of reconstructions from occluded images. The images illustrate how different models interpret eye and mouth occlusions. All models consistently interpret artificial eye occlusions as glasses, whereas mouth occlusions are predominantly perceived as open mouths. Notably, the models utilizing ResNet input embeddings (4th and 5th columns) demonstrate slight orientation changes in the reconstructions compared to those from VGG embeddings.

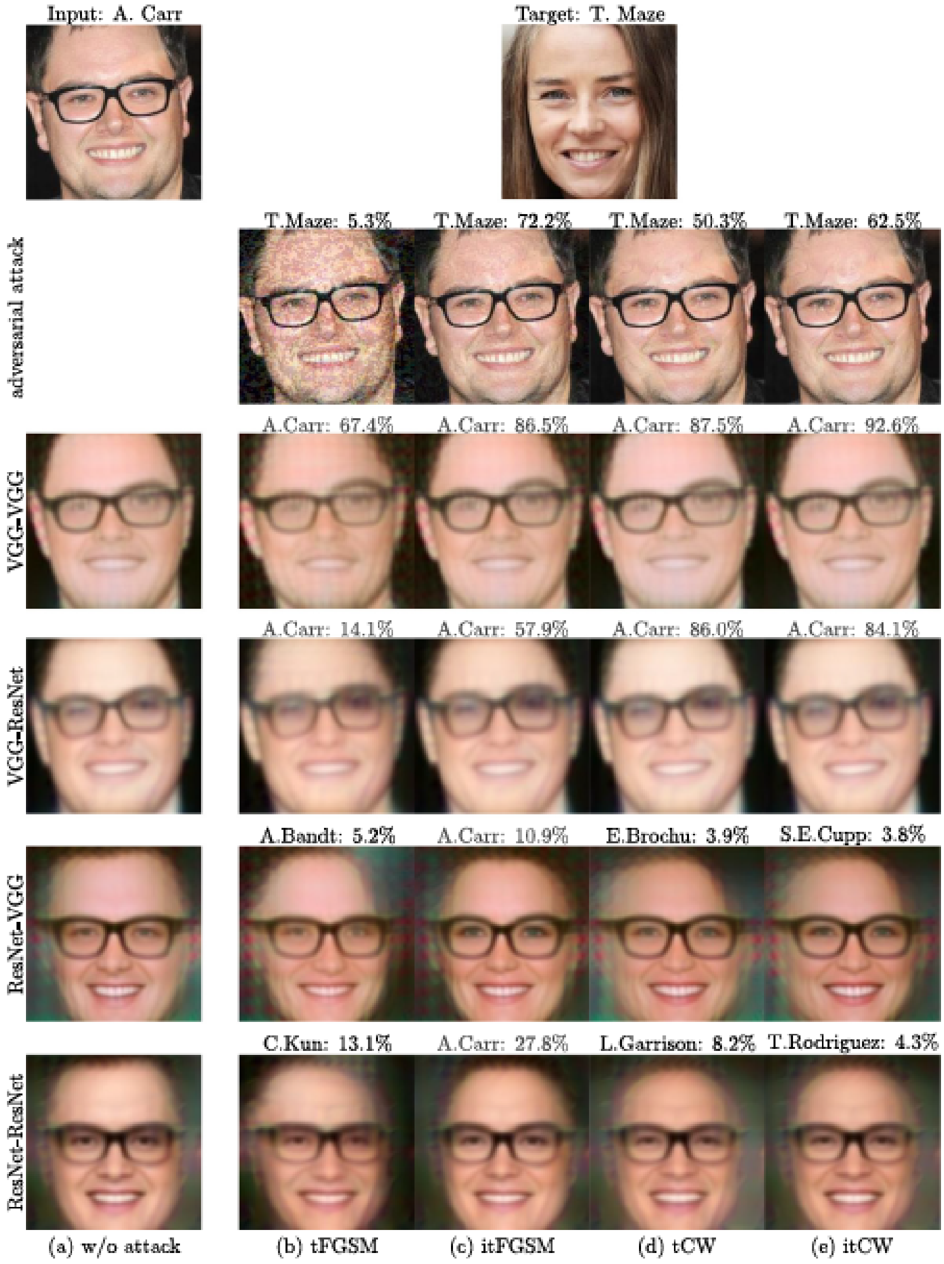
in green. As expected, we observe a significant decline in performance when comparing results produced by the original images (solid green line) and the attacked ones (solid red line). However, when the same experiment is conducted on images reconstructed through the DFD model, we observe a minor performance decrease in the non-attack scenario (solid green line versus dotted green line), but see a considerably less pronounced performance decline due to the adversarial attack (dotted green line versus dotted red line). This result demonstrates that the DFD model can effectively provide a degree of defense against the FGSM attack.

4.4. Understanding template-construction procedures

The performance of FR models strongly depends on the procedure utilized to construct face templates during the enrollment process. While academic recognition problems often assume a single input image for the construction of the reference face template (in a sort of single-shot learning setting), industry solutions often capture a larger set of images and derive a more elaborate face template from the captured enrollment data. Therefore, in the next series of experiments, we investigate how different template-construction strategies impact the information encoded in the templates. In the experiments, we consider two settings, where: (*i*) the face template is represented by multiple face embeddings, e.g., cluster centroids of the enrolled image embeddings, and (*ii*) the reference face template is represented by some aggregation (e.g., arithmetic mean) of all enrolled image embeddings. For the qualitative part of this series of experiments, we use 500 images from LFW, all representing the same identity.

4.4.1. Face templates from cluster centroids

In an operational setting, multiple face images of the same subject are commonly available to construct the face template to be stored in the system's database for later matching operations. One strategy on how to utilize the available images is to store all corresponding embeddings in the system and probe for the best match when a probe image arrives. Alternatively, to reduce redundancy, computational costs, and storage requirements, these embeddings are also clustered and only the



(a) w/o attack (b) tFGSM (c) itFGSM (d) tCW (e) itCW

Fig. 12. Adversarial attack results without (row 2) and with (rows 3–6) the use of DFD reconstructions. Predicted identity above each image. Adversarial attack designed to fool the system into classifying A. Carr as T. Maze. DFD reconstructions show resistance to adversarial attacks, especially when reconstructing from the VGG embeddings. ResNet-based reconstructions (bottom two rows) also exhibit some characteristics of the target gender. Each column is a different attack method: (b) targeted FGSM, (c) iterative targeted FGSM, (d) targeted CW, (e) iterative targeted CW.

embeddings of the cluster centroids (i.e., means) are stored for later comparison purposes.

To explore the effect of this latter approach on the information encoded in the centroids, we use agglomerative clustering over the selected 500 LFW images and then invert the centroids using the DFD model. As can be seen from Fig. 14, each cluster captures distinct forms of image variability, including factors such as facial expressions, poses, and the presence of accessories like hats and sunglasses, when the VGG model is used to produce the embeddings. With the ResNet model, the cluster centroids still differ from each other, but are closer in appearance, again suggesting that the ResNet-based FR model has a stronger tendency towards making the information encoded in the embeddings more robust to typical sources of image variability, such as pose or facial expression. Overall, the presented results suggest that ResNet-based models produce compacter data distributions in the embedding space (with comparably lower intra-class variability) compared to the VGG-based model, where much larger variability is observed among the reconstructed images.

4.4.2. *Face templates through embedding aggregation*

Another possibility to construct a face template from multiple face images is to aggregate the corresponding embeddings through, e.g., averaging. This procedure results in a single vector that is used for matching purposes in operational settings. Using the same set of 500 images from LFW as in the previous section, we present in Fig. 15 the effect of averaging different numbers of (randomly selected) face embeddings for both FR models and all DFD variants. As can be seen, increasing the number of embeddings to aggregate appears to help "normalize" the information encoded in the template, so it corresponds to better aligned, frontal, neutral, and homogeneously illuminated

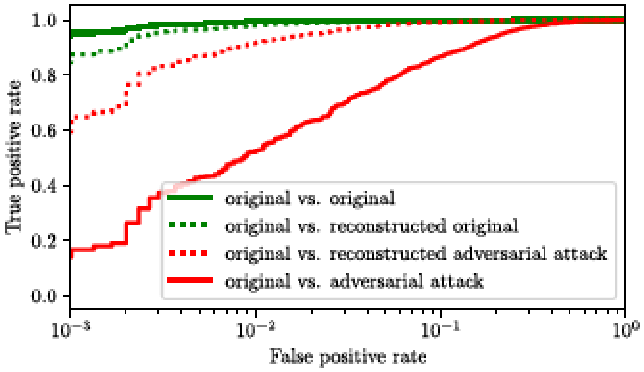


Fig. 13. Impact of the FGSM attack on verification performance, as illustrated by the ROC figure. The vertical axis denotes the true positive rate for four distinct scenarios: (*i*) both images in the verification pair are untouched originals (solid green line), (*ii*) one image in the verification pair remains original, while the other is reconstructed using the DFD (dotted green line), (*iii*) one image in the verification pair is the original, and the other is a reconstruction derived from an image distorted by the adversarial attack (dotted red line), (*iv*) one image in the verification pair is the original, and the other has been manipulated by the adversarial attack (solid red line).

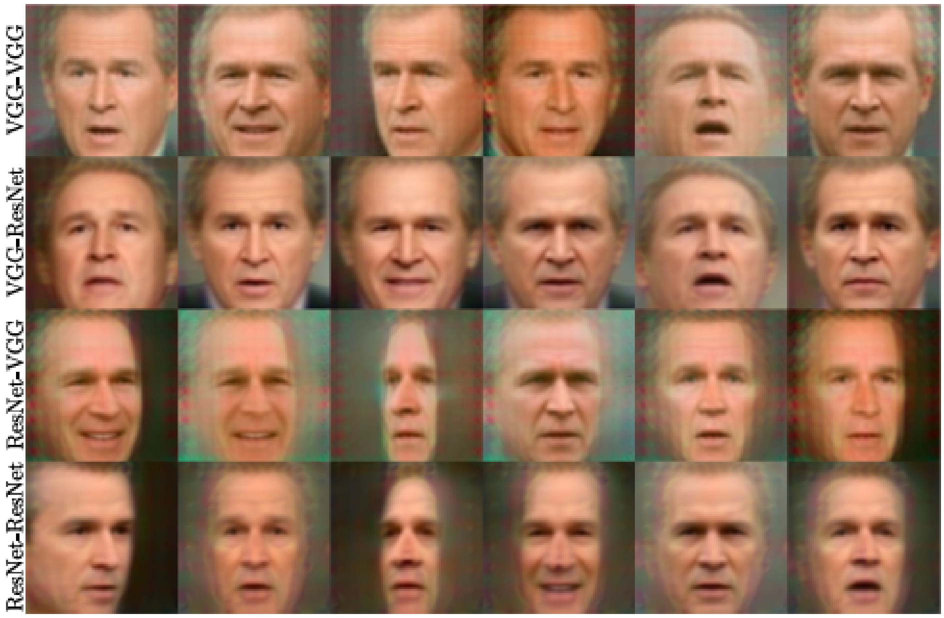


Fig. 14. Visualization of clustered embeddings: Each image corresponds to a reconstruction derived from the average of all templates within a specific cluster. This approach encapsulates the overall characteristics of the respective cluster, providing a representative snapshot of its inherent variability.

faces, which should make it easier to match the templates to potential probe samples and enhance the template's generalization power. This behavior is again more obvious for the VGG-based FR model since the ResNet model already reduces the amount of non-identity-related information in the templates and comparably benefits less from the aggregation process. Interestingly, we observe some differences in the reconstruction with the white- and black-box configurations, but in general, the reported observations still apply.

To quantitatively assess the impact of the embedding aggregation process, we conduct verification experiments on the LFW dataset with the VGG FR model, where the gallery templates are computed by aggregating the embeddings of different numbers of gallery images of each of the 50 most represented subjects from the LFW database. The results in Fig. 16 show that the verification performance progressively improves when the number of embeddings to aggregate increases. The performance stabilizes when 10 or more embeddings are used to compute the gallery template, which is also the point, where the only minute difference in the reconstructed images is observed in Fig. 15 - see top two rows.

4.5. Understanding template modification techniques

In the last series of experiments, we investigate techniques that modify the face templates produced by ConvNet-based FR models with some specific goal. Specifically, we are interested in a special type of Biometric Privacy-Enhancement Technique (B-PET) (Meden et al., 2021) that aims to remove information on soft biometric attributes (e.g., gender) from the face embeddings to ensure higher levels of privacy. To this end, we experiment with the PFRNet model, proposed by Bortolato et al. (2020), which excels in disentangling identity information from facial attributes. Consequently, this capability allows for the suppression of various soft biometrics in face templates. PFRNet is designed for the ResNet-50 FR model trained on VGGFace2, so we conduct experiments with the white-box ResNet–ResNet DFD variant. Gender and identity metrics are computed on the reconstructed images using the DeepFace gender classifier (Serengil and Ozpinar, 2021) and with embeddings produced by the VGG-16 face recognition network (Simonyan and Zisserman, 2015), respectively.

The main idea behind PFRNet is to disentangle the face embeddings into two distinct components, where the first encodes identity information and the second encodes gender information only. In Fig. 17,

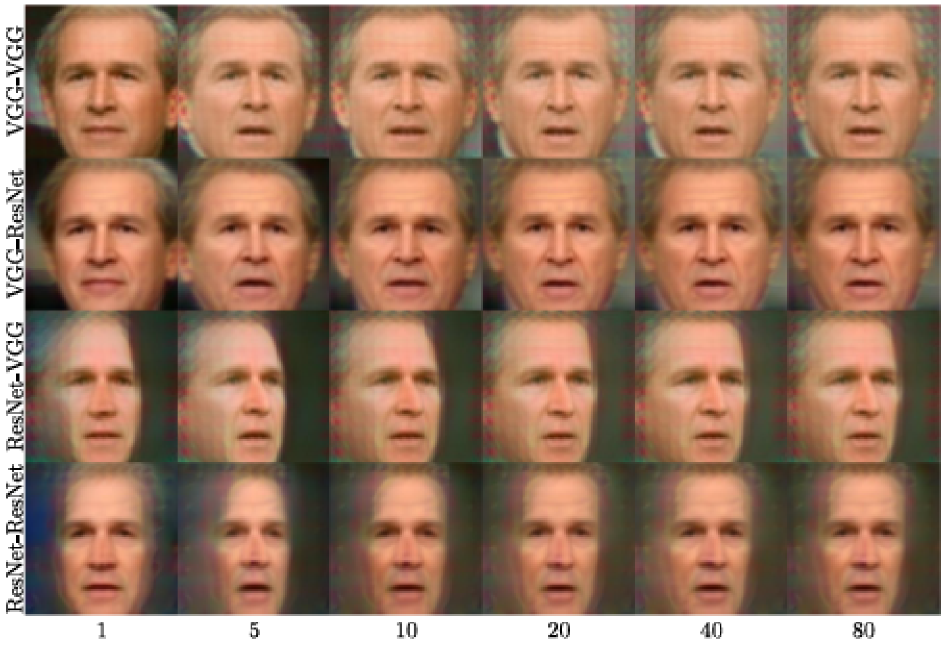


Fig. 15. Visual representation of template convergence. Each image is reconstructed from a face template that consists of aggregated embeddings. The four rows correspond to different DFD configurations. The columns represent the varying numbers of embeddings aggregated to form a template, as indicated below each column.

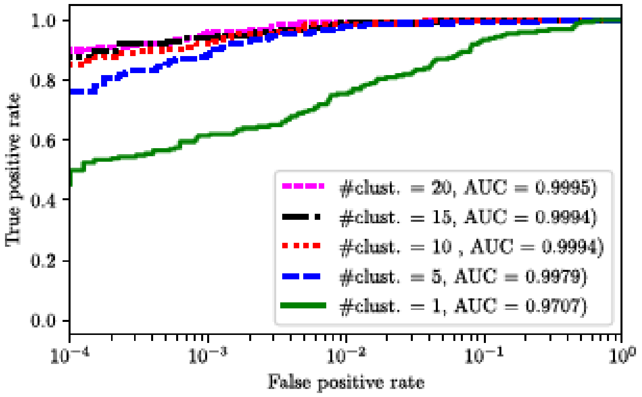


Fig. 16. Verification performance for the case where a gallery template is defined as an average embedding of a variable number of gallery images. It can be seen that verification performance improves as the number of embeddings aggregated for a gallery template increases and stabilizes when 10 or more embeddings are used to calculate the gallery template.

we show the impact of this disentanglement process on the visual appearance of a few reconstructed images from the CelebA dataset. For baseline comparisons, we first reconstruct the embeddings without modifying the gender or identity information and show results in Fig. 17b. To evaluate the information encoded in the gender component, we replace this component with the mean value for each gender class in the dataset: female (Fig. 17c), male (Fig. 17d), and combined (Fig. 17e). As can be seen, this manipulation has a strong apparent effect on the gender information in the reconstructed images, making them appear female, male, and androgynous, respectively while maintaining identity content to a certain degree. Additionally, evaluating the gender information contained in the identity component using a similar manipulation for female (Fig. 17f), male (Fig. 17g), and combined (Fig. 17h) vector averages, shows little effect in the corresponding reconstructions in terms of apparent gender. On the other hand, the identity information of the reconstructed image is substantially altered but shows little dependence on which gender label was used to compute the mean vector. This can be attributed to the fact that the identity-related part of the disentangled embedding is primarily charged with encoding identity information and contains little to no gender information.

The results of these experiments again point to the usefulness of the proposed DFD for interpreting template manipulation techniques and validation of their characteristics.

5. Conclusion

In this paper, we have presented a novel template inversion technique, the Deep Face Decoder (DFD), for examining the characteristics of face image embeddings of contemporary ConvNet-based face recognition (FR) models. Our experiments with two FR models (with different backbones) and multiple face datasets showed that the proposed DFD model is able to produce informative (high-fidelity) image reconstructions from the embeddings, both in a white-box as well as a black-box setting. Additionally, we demonstrated how DFD can be used to analyze and interpret the characteristics of the embedding space of ConvNet-based FR models and to explore the impact of appearance perturbations, occlusions, adversarial attacks, and various template modification procedures on the information encoded in the generated face templates.

Our analysis led to several interesting findings. The results related to geometric perturbations showed that such perturbations can directly

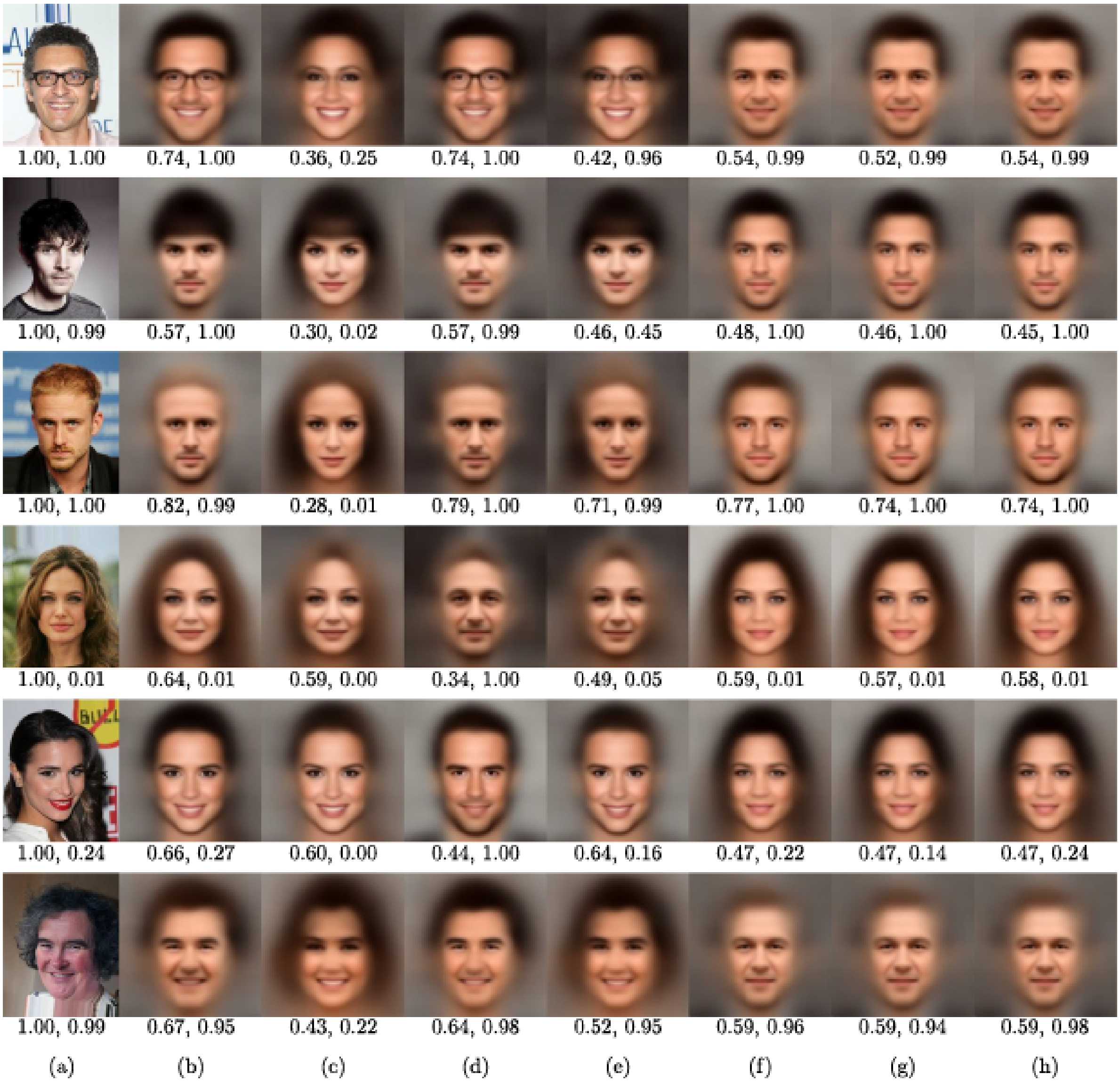


Fig. 17. Reconstructions from PFRNet latent space. (a) Original images. Reconstructions from disentangled latent space: (b) non-modified disentangled embeddings, (c) embeddings using the female average for the dependent section, (d) embeddings using the male average for the dependent section, (e) embeddings using the overall average for the dependent section, (f) embeddings using the female average for the independent section, (g) embeddings using the male average for the independent section, (h) embeddings using the overall average for the independent section. The first value below each image corresponds to the cosine similarity [−1,1] against the original image, while the second value corresponds to the gender classifier's probability of the face being male.

be modeled in the embedding space of FR models and that it is possible to learn simple linear mappings that normalize for misalignment at the template level. Additionally, we showed that occlusions of the facial area are often interpreted as semantically meaningful objects in the embedding space, and that adversarial noise infused through softmax classifiers has only a limited impact on the facial embeddings. When looking at different strategies for template construction from multiple face images, we managed to associate a semantic interpretation to the template-construction process that justifies the commonly observed performance improvement associated with aggregated templates. Finally, we showed that the DFD can also be employed as a highly useful tool for validating the performance of template modification procedures, e.g., soft-biometric privacy-enhancing techniques.

Taken together, our findings illuminate several significant characteristics of face image embeddings and their implications, offering valuable insights to the academic community and the industry. This understanding could pave the way for more sophisticated, reliable, and privacy-preserving facial recognition systems in the future. Future work could extend these findings by exploring more complex and diversified scenarios, as well as by addressing the challenges raised in this study.

CRediT authorship contribution statement

Janez Križaj: Investigation, Validation, Visualization, Writing – original draft, Data curation, Formal analysis, Software. **Richard O. Plesh:** Methodology, Writing – original draft. **Mahesh Banavar:** Funding acquisition, Resources. **Stephanie Schuckers:** Funding acquisition, Resources. **Vitomir Štruc:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.