ENHANCING GENDER PRIVACY WITH PHOTO-REALISTIC FUSION OF DISENTANGLED SPATIAL SEGMENTS

Peter Rot, Janez Križaj, Peter Peer, Vitomir Štruc

University of Ljubljana
Kongresni trg 12, SI-1000 Ljubljana, Slovenia

ABSTRACT

Soft-biometric privacy enhancing techniques (SB-PETs) transform facial images to preserve identity while preventing the automatic extraction of soft-biometrics by confusing machines through noise injections or attribute obfuscation. However, existing SB-PETs often sacrifice image quality for privacy enhancement, limiting practical usage, especially in applications that allow for human inspection. To address these issues, we introduce a novel SB-PET that (i) generates photo-realistic images with obscured gender information, which makes attribute extraction challenging for machine-learning models, but also human observers, and (ii) preserves identity to a significant extent. The proposed approach, abbreviated PriDSS, operates in the latent space of the StyleGANv2 model and aims to (i) preserve the appearance of facial parts from the input image carrying identity information, and (ii) incorporate global context from images of the opposite gender, thus, obscuring the original gender information. PriDSS shows promising results when compared to state-of-the-art techniques from the literature, and leads to competitive gender-privacy and face-verification performance, while ensuring superior photo-realism.

Index Terms— soft–biometrics, privacy, verification

1. INTRODUCTION

When people provide facial biometrics for identification or verification purposes, soft-biometric attributes like gender, age, and ethnicity can inadvertently be extracted (without consent), posing risks to one's personally privacy [1, 2]. Soft-biometric privacy-enhancing techniques (SB-PETs) [3, 4] have emerged to protect privacy by confusing automated machine learning models attempting to extract personal information. However, current SB-PETs often trade privacy enhancement for image quality as they focus on confusing machines rather than humans [1]. This hampers practical use, especially in applications where human inspection is possible and the added noise and image artifacts that impact automatic models have little effect on inferring soft-biometric attributes, as illustrated in the middle column of Fig. 1.

A considerable amount of work has been done on soft-biometric privacy over the years [2]. State-of-the-art (SoTa) techniques typically aim to modify the facial appearance in a way that still allows legitimate use cases, such as face verification, but obscures other types of potentially sensitive information, such as age or gender. This is achieved through various mechanisms, such as the addition of adversarial noise [5], image perturbations [3] or image-to-image translation techniques [4] among others. While existing techniques achieve a competitive level of privacy enhancement, they typically produce visually compromised (or low-quality) images that still allow for attribute inference by humans. Our goal, illustrated in Fig. 1, is, therefore, to achieve state-of-the-art soft-biometric privacy enhancement, while maintaining photo-realism and producing artifact-free results.

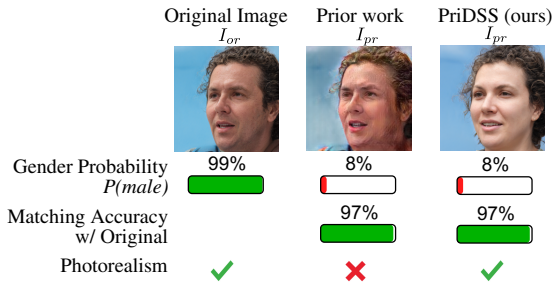

Fig. 1: **Illustration of the main idea behind PriDSS**. Existing SB-PETs aim to transform the original image I_{or} into a privacy-enhanced image I_{pr}, preserving identity and preventing soft-biometric attribute extraction. However, this often degrades photo-realism, which we address through PriDSS.

Specifically, we propose a novel SB-PET that focuses on safeguarding the gender attribute while preserving visual quality. The proposed approach leverages the insight that certain facial parts (e.g., eyes, nose, mouth) contain more identity information, while broader global context can predict soft-biometric attributes (e.g., skin smoothness indicating gender). Our technique, called PriDSS (**Pri**vacy through fusion of **D**isentangled **S**patial **S**egments) combines identity-related facial parts with contextual information from the opposite gender through image fusion in the latent space of the pre-trained StyleGANv2 model [6]. This approach confounds machine classifiers, minimizes modification traces, and enhances gender attribute privacy for human observers. We evaluate PriDSS in comparison to its closest competitor, PrivacyNet [4], and report highly encouraging results.

Supported by ARRS through P2-0250(B), P2-0214 and J2-2501 (A).

2. BACKGROUND AND RELATED WORK

SB–PETs. Given an original face image I_{or} and an attribute classifier ξ_a, soft-biometric privacy-enhancement (ψ) aims to produce privacy-enhanced images (I_{pr}) that prevent confident prediction of attribute labels by ξ_a [1,2,4]. The goal of ψ is to obscure attribute information while maintaining (visual) similarity to the original image (I_{or}) in terms of retained identity information. Two strategies are commonly used to enhance the privacy of a specific facial attribute in an image [7]: (1) modifying the image to confuse the classifier into making incorrect predictions (confusing *male* for *female* or vice versa when targeting gender), or (2) modifying the image to yield near-random performance from the classifier for the given examples. A technique following the first strategy based on adversarial noise was presented in [5]. Examples from the second group, on the other hand, include [8–10] for methods trying to enhance privacy by modifying face templates, and [3,4] for techniques modifying visual information for privacy enhancement. Our work also focuses on this latter strategy and aims to ensure near-random gender recognition performance with the privacy-enhanced image.

Face editing using StyleGAN. In recent years, face-image editing techniques have made significant progress, covering deep fakes for identity alteration, beauty filters and emotion manipulation [6, 11, 12]. While privacy-enahncing methods like face swaps for deidentification have garnered attention, face editing for soft-biometric privacy is still underexplored.

A considerable body of recent face-editing work leverages the capabilities of StyleGAN2 [6], a deep generative model capable of generating highly realistic images. Image editing with StyleGAN2 involves manipulating the latent space, a low-dimensional representation of an image capturing its essential features [13]. Among the different works in this area, a particularly interesting approach, called StyleFusion was presented in [14]. StyleFusion utilizes the so-called $\mathcal{S}$ latent space for semantic manipulation, and allows combining diverse semantic components (e.g., eyes, nose) from a set of original images $I_1, I_2, \ldots, I_n$ into an artificially generated output image. This fusion process is executed through FusionNet modules, each dedicated to a specific semantic unit (e.g., eye swapping, mouth swapping). These modules align images semantically and produce a fused latent space, allowing control over multiple semantic factors within a single image. The proposed PriDSS approach, described in the next section, extends the outlined ideas into privacy-enhancing technique for gender obfuscation.

3. METHODOLOGY

In this section, we propose a novel SB-PET, named PriDSS, that enhances **Pri**vacy in facial images by photo-realistically fusing **D**isentangled **S**patial **S**egments from the original input face and an artificially generated image of the opposite gender, as shown in Fig. 2. This process involves two steps: (i) the integration of various spatial segments from both images, and (ii) the selection of the image of the opposite gender.

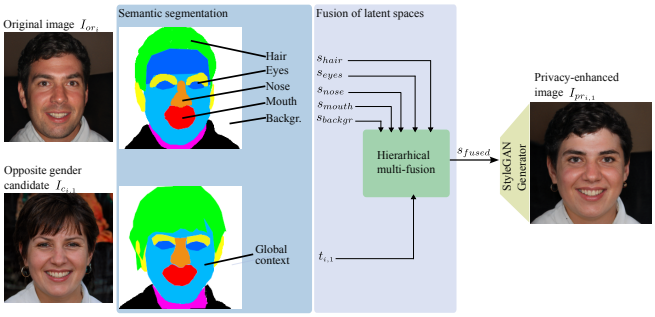

Fig. 2: **High-level overview of PriDSS.** PriDSS modifies the original image I_{or_i} by fusing it's StyleGAN latent code, s_i, with the latent code t of an image I_c of the opposite gender. The fusion process preserves identity, while incorporating gender-related features from I_c, such as roughness/smoothness of the skin and presence of beard.

Fusing spatial segments (Step 1). Let $\mathcal{I} = \{I_{or_i}\}_{i=1}^{N}$ be a set of N original face images, with binary gender labels $a_i \in \{male, female\}$. The goal of PriDSS is to produce privacy-enhanced images I_{pr_i} using a dedicated fusion technique, where a_i is confounded for both human observers and machine classifiers, while maintaining photo-realism.

For each I_{or_i} there exists a corresponding StyleGAN2 latent space representation denoted as s_i. The privacy-enhancing fusion that forms the basis for PirDSS, involves modifying s_i by incorporating relevant attributes of the opposite gender. To achieve this, for each I_{or_i}, we define a set of M images representing individuals with the opposite gender to a_i, denoted as $I_{c_{i,j}}$, where the subscript c stands for 'candidates', i is the index of the original image, and j enumerates the candidates. Each image in this set serves as a candidate image, ensuring diversity in the fusion processm which allows us to select the best-fused candidate based on a high similarity score with the original and low gender predictability. Each $I_{c_{i,j}}$ also has a corresponding StyleGAN2 latent vector $t_{i,j}$.

As illustrated in Fig. 2, we obtain privacy-enhanced images $I_{pr_{i,j}}$ by fusing the latent code s_i that captures identity information with the latent code of the opposite gender $t_{i,j}$ that corresponds to gender-specific attributes in the synthesized image. The hierarchical multi-fusion module, as described in [14], smoothly fuses relevant parts of the latent codes generated through a face parser, preserving identity in eyes, nose, and mouth, while incorporating gender-related features from the global context of $I_{c_{i,j}}$, such as roughness/smoothness of the skin and presence of a beard. Additionally, the fusion considers hair and background from I_{or_i} to enhance the overall similarity between I_{or_i} and $I_{pr_{i,j}}$. The fused latent space s_{fused} is then passed to the StyleGAN generator to produce $I_{pr_{i,j}}$. To obtain the final privacy-enhanced image I_{pr_i}, we utilize a **best-candidate selection procedure** that, in an on-line fashion, evaluates all output images $I_{pr_{i,j}}$ with respect to the matching score with the original and the

Fig. 3: **Sample images generated using StyleGANv2.** Female samples are shown on top, male at the bottom.

gender prediction, produced by a pretrained gender classifier.

Best-candidate selection (Step 2). For each original image I_{or}, we first generate M privacy enhanced images I_{pr_j}, for $j \in \{1, \ldots, M\}$ using M sampled candidate images of the opposite gender $I_{c_{i,j}}$. Next, we select the best privacy-enhanced image from the set of M generated candidates I_{pr_j} using the following steps: (i) we subject the candidates to a gender classifier and sort the candidates by their gender scores, giving priority to those with a target gender probability of $P(gender) = 0.5$, (ii) we compute the (cosine) similarity scores with the original input face images in the embedding space of selected face recognition model (denoted as SIM_{score}), and (iii) finally, select the candidates by the lowest privacy–gain identity–loss coefficient (PIC) [10]. PIC is in our case defined as $PIC = |2P(male) - 1| - SIM_{score}$.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

Investigations. We conduct comprehensive experiments to explore various aspects of PriDSS, including:

- **Gender classification performance:** One of the main goals of PriDSS is to obscure (soft-biometric) gender information and make it challenging for machine learning models to extract such information automatically. To evaluate this aspect, we conduct gender recognition experiments on the original images and their privacy–enhanced versions, and utilize the state-of-the-art DeepFace gender classifier for this task [15]. We treat gender recognition as a binary classification problem, and, following standard evaluation methodology, use the Area Under the Curve (AUC) corresponding to the generated Receiver Operating Characteristics (ROC) curves to quantify performance [4, 8, 9].
- **Identity verification performance.** To evaluate the identity-preservation capabilities of PriDSS, we adopt the state-of-the-art CosFace [16] face recognition model. We extract face embeddings from the original images and their privacy–enhanced versions, and then perform verification experiments with 1000 mated pairs and 1000 non-mated embedding pairs. For the mated comparisons, we use the original input images and the corresponding privacy–enhanced counterparts, similarly to [4].
- **Evaluation of photo–realism.** The Fréchet Inception Distance (FID) is a common metric regularly used in the literature to assess image synthesis quality by comparing feature embeddings of real and synthetically generated images produced by pre-trained neural networks [17]. A lower FID score indicates better correspondence between the distributions of the real and synthesized images, and is therefore often used as a measure of (photo) realism. In our case, we utilize FID scores to assess the quality of the privacy-enhanced images by comparing them to the original ones using the standard InceptionV3 features.
- **Robustness against recovery attempts:** Given the known vulnerability of SB-PETs to reconstruction attacks, where privacy-enhanced regions are reconstructed to recover soft biometric information, we conduct an evaluation of their susceptibility to such attacks using the PrivacyProber framework [7]. This framework has been purposefully developed for assessing the robustness of SB-PETs against reconstruction attempts. We conducted tests using a set of 3 recovery strategies and report results in terms of the AUC of the gender classifier after the recovery phase.

Performance indicator	**Original**	**PrivacyNet**	**PriDSS (ours)**
Gender (AUC)	0.981	0.5400	0.5900
Verification (EER) w/ Original	n/a	0.0070	0.0680
Verification (FNMR@FMR10^{-1}) w/ Original	n/a	0.0005	0.0010
Photo-realism (FID) w/ Original	n/a	57.499	25.386

Table 1: **Performance evaluation and SoTa comparison.** The table shows results for gender-classification (AUC), verification (EER), and photo-realism evaluation (FID).

Experimental data. We harness the capabilities of the pre-trained StyleGANv2 model [6] to directly generate faces through latent-space sampling. This process allows us to bypass the embedding step needed to project real-world faces into the StyleGAN latent space [18, 19] and follows standard research methodology in face editing [20]. Additional, it enabled us to synthesize a balanced and representative test dataset in terms of gender distribution. The final test dataset used for the experiments, thus, consists of 1000 female and 1000 male facial images, as shown in Fig. 3, i.e., $N = 2000$.

4.2. Quantitative Results

Privacy-enhancement and verification. In the first series of experiments, we investigate the privacy-enhancement and identity-preservation capabilities of PriDSS as well as the photo-realism of the generated images. The result of the experiments are presented in Table 1, together with a comparison with the (conceptually) closest state-of-the-art (SoTa) competitor, PrivacyNet [4]. We set PrivacyNet to only enhance gender privacy, and evaluate it on the same set of generated images as PriDSS. As can be seen, the proposed PriDSS method achieves a promising level of soft-biometric privacy enhancement for the gender attribute, with an AUC score of 0.59 in gender recognition experiments, indicating (ideal) near-random performance. PrivacyNet achieves a comparable AUC value of 0.54, suggesting that both tech-

Input image		PrivacyNet	PriDSS (ours)
Privacy enhanced		0.540	0.590
Recovered with	PP–D	0.545	0.603
	PP-I	0.540	0.602
	PP–A	0.550	0.554

Table 2: **Robustness to image restoration attempts.** We use three different strategies from [7] to restore the obscured gender information using an autoencoder (PP-A), a deonising (PP-D) and an inpainting (PP-I) procedure. Shown are AUC scores generated in gender recognition experiments.

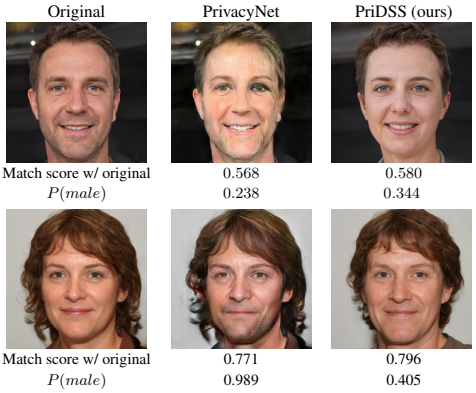

Fig. 4: **Visual examples of privacy-enhanced images.** Note that both methods produce faces that lead to high matching scores with the originals and effective gender obfucation. However, the resuls produced by PriDSS exhibit a higher level of photo-realism and are free of artifacts.

niques effectively conceal gender information. In terms of verification performance, PrivacyNet leads to an EER score of 0.007 when matching the privacy-enhanced images to the originals, while PriDSS results in an EER of 0.068, which again points to a considerable identity-preservation ability of both tested techniques. Similar results are also observed for other operating points.The **strong point of PriDSS**, however, **is the photo-realism**. Here, the proposed approach significantly outperforms PrivacyNet with an FID score of 25.386 - an improvement of more than $2\times$ over PrivacyNet.

Robustness against reconstruction attacks. Next, we explore how robust the proposed privacy-enhancement is w.r.t. attempts to recover the initial visual appearance of the facial images, or in other words, to attempts aiming to reverse the privacy enhancement. To this end, we implement 3 variants of the PrivacyProber (PP) recovery strategies, proposed specifically to probe the robustness of SB-PETs [7]. Here, the symbols A, D, and I represent PP variants that try to recover the obscured soft-biometric information using an auto–encoder (A), a denoising procedure (D) and an inpainting scheme (I). From the AUC scores, generated in gender recognition experiments with the privacy enhanced and PP restored images, in Table 2, we can see that both PrivacyNet and PriDSS are quite robust to recovery attempts. The AUC scores do not change much and the gender-recognition performance is still close to random despite the application of the restoration strategies.

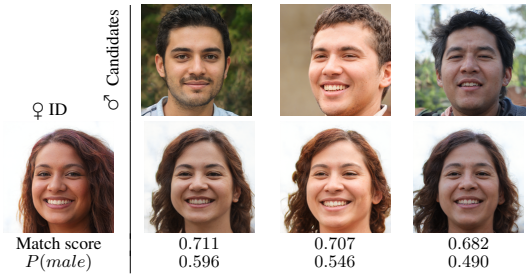

Fig. 5: **Ablation study results.** We explore the impact of the candidate-image selection procedure on the results.

Computational Complexity. On a Desktop PC with an RTX 3090 GPU, PriDSS takes approximately 1800 ms to (i) generate a candidate face, (ii) compute it's similarity score, and (iii) calculate the gender probability. Note that this procedure can also be parallelized using multiple GPUs.

4.3. Qualitative Results

Visual evaluation. While both PrivacyNet and PriDSS lead to comparable verification performance in general, PriDSS ensures significantly higher-quality images after privacy-enhancement, as already demonstrated by the FID scores in Table 1. In Fig. 4, we now further capitalize on this characteristic with some visual examples. As can be seen from the presented results, PriDSS generates high-quality artifact-free images that well preserve identity information, while effectively obscuring gender information. To support these observations with empirical evidence, matching scores and classifier probabilities (for male) are also reported in Fig. 4.

Ablation study. Finally, we present an ablation study in Fig. 5 that demonstrates the impact of the candidate-image selection procedure. Here, we observe three instances of a female identity being merged with multiple male candidate images. It is noteworthy that the identity of the fused/combined result remains similar to that of the original image, corroborated by consistently high matching scores. On the other hand, the gender scores are close to 0.5 in all cases, making it difficult to robustly infer gender information. While, the final selection of the most suitable candidate involves a comprehensive assessment of both verification and gender probability scores, the presented visual examples offer insight into the intricacies of the proposed approach.

5. CONCLUSION

In this paper, we presented a novel soft-biometric privacy-enhancing technique, called PriDSS, capable of obscuring gender information in facial images, while preserving identity and ensuring highly photo-realistic results. Our experiments point to highly promising results, but still offer room for future research, where the same concept could be extended towards other soft-biometric attributes.

6. REFERENCES

[1] Vahid Mirjalili, Sebastian Raschka, and Arun Ross, "Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers," *IEEE Access*, vol. 7, pp. 99735–99745, 2019.

[2] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J. Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc, "Privacy–enhancing face biometrics: A comprehensive survey," *IEEE Transactions on Information Forensics and Security*, vol. 16, 2021.

[3] Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *International Conference on Biometrics (ICB)*, 2018, pp. 82–89.

[4] Vahid Mirjalili, Sebastian Raschka, and Arun Ross, "Privacynet: Semi-adversarial networks for multi-attribute face privacy," *IEEE Transactions on Image Processing*, vol. 29, pp. 9400–9412, 2020.

[5] Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta, "Anonymizing k-Facial Attributes via Adversarial Perturbations," *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of StyleGAN," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] Peter Rot, Klemen Grm, Peter Peer, and Vitomir Štruc, "PrivacyProber: Assessment and detection of soft–biometric privacy–enhancing techniques," in *IEEE Transactions on Dependable and Secure Computing*, 2023.

[8] Blaž Bortolato, Marija Ivanovska, Peter Rot, Janez Križaj, Philipp Terhörst, Naser Damer, Peter Peer, and Vitomir Štruc, "Learning privacy-enhancing face representations through feature disentanglement," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 495–502.

[9] Pietro Melzi, Hatef Otroshi Shahreza, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Sébastien Marcel, and Christoph Busch, "Multi-ive: Privacy enhancement of multiple soft-biometrics in face embeddings," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 323–331.

[10] Philipp Terhörst, Kevin Riehl, Naser Damer, Peter Rot, Blaž Bortolato, Florian Kirchbuchner, Vitomir Štruc, and Arjan Kuijper, "Pe-miu: A training-free privacy-enhancing face recognition approach based on minimum information units," *IEEE Access*, vol. 8, pp. 93635–93647, 2020.

[11] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, 2020.

[12] Blaž Meden, Manfred Gonzalez-Hernandez, Peter Peer, and Vitomir Štruc, "Face deidentification with controllable privacy protection," *Image and Vision Computing*, vol. 134, pp. 104678, 2023.

[13] Andrew Melnik, Maksim Miasayedzenkau, Dzianis Makarovets, Dzianis Pirshtuk, Eren Akbulut, Dennis Holzmann, Tarek Renusch, Gustav Reichert, and Helge Ritter, "Face generation and editing with stylegan: A survey," *arXiv preprint arXiv:2212.09102*, 2022.

[14] Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel Cohen-Or, "StyleFusion: Disentangling spatial segments in StyleGan-generated images," *ACM Transactions on Graphics*, vol. 41, no. 5, pp. 1–15, 2022.

[15] Sefik Ilkin Serengil and Alper Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4.

[16] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "CosFace: Large margin cosine loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5265–5274.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[18] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021.

[19] Rameen Abdal, Yipeng Qin, and Peter Wonka, "Image2stylegan++: How to edit the embedded images?," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8296–8305.

[20] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou, "Interfacegan: Interpreting the disentangled face representation learned by gans," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 4, pp. 2004–2018, 2022.