Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cranio-Diff: Diffusion-based Cross-domain Craniofacial Reconstruction with 2D X-ray Skull

Ravi Shankar Prasad
d23033@students.iitmandi.ac.in
Indian Institute of Technology Mandi
Naresh Gurjar
Indian Institute of Technology Mandi
Shashank Baghel
Indian Institute of Technology Mandi
Chirag Chirag
Indian Institute of Technology Mandi
Dinesh Singh
Indian Institute of Technology Mandi

Research Article

Keywords: Craniofacial Identification, Generative AI, Forensics, skull-to-face reconstruction, 2D-to-2D mapping

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11402689/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cranio-Diff: Diffusion-based Cross-domain Craniofacial Reconstruction with 2D X-ray Skull

Ravi Shankar Prasad[1*], Naresh Gurjar[1], Shashank Baghel[1], Chirag[1], Dinesh Singh[1]

[1*]Visual Intelligence and Machine Learning Group, School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Kamand, Mandi, 175075, Himachal Pradesh, India.

*Corresponding author(s). E-mail(s): d23033@students.iitmandi.ac.in, dineshsingh@iitmandi.ac.in;
Contributing authors: ng4955607@gmail.com; s25038@students.iitmandi.ac.in; t25089@students.iitmandi.ac.in;

Abstract

The identification of unknown human remains is a persistent humanitarian and forensic challenge. In forensic investigations, it becomes even more challenging when common biometric DNA identifications, such as soft-tissue or hair strands are not present, as they degrade with time. To narrow down the process of identification, few attempts have been made to identify an individual by reconstructing a face from the given skeletal remains (i.e., skull) using generative AI models, such as CycleGAN, Pix2Pix, and diffusion models. However, they fail to effectively capture semantic information in craniofacial reconstruction, due to a mismatch in the alignment of structural information across modalities. To address this issue, we propose *Cranio-Diff*, a diffusion-based framework for craniofacial reconstruction from 2D X-ray skull images. The proposed approach integrates skull-conditioned structural guidance through ControlNet with biometric text prompts to generate a face. The proposed *Cranio-Diff* method is evaluated on skull-face dataset obtained from X-ray scans of 120 subjects in lateral and frontal views. To enable controlled evaluation, each face image is artificially synthesized across three age groups (25, 45, 65) and three BMI variations **±10%**, yielding 4320 paired samples. This dataset is also curated across different facial modalities (RGB and sketch) with different appearance conditions (hair and without hair). To the best of our knowledge, this is the only X-ray-face dataset with this magnitude. Extensive experiments showed that the proposed method achieves improved performance over the recent existing approaches in both generated

image quality and retrieval task. Finally, to evaluate the performance of our proposed method, we have evaluated the quality of the generated image using FID, SSIM, LPIPS score. Additionally, retrieval performance is evaluated using recall@k, mAP@k and MRR@k. Obtained experimental results demonstrate that the proposed method can be used as an alternate tool in providing aid in forensic investigations.

Keywords: Craniofacial Identification, Generative AI, Forensics, skull-to-face reconstruction, 2D-to-2D mapping

1 Introduction

Identifying a deceased individual from skeletal remains is a challenging task, particularly in the absence of common biometric information such as soft tissue, hair strands, etc. Forensic Craniofacial Reconstruction (FCR) is one of the techniques that is widely used in forensic and anthropological applications. Traditionally, FCR is done by manually overlapping clay onto the skull with the help of soft-tissue depth information. But this manual process is time-taking, requires anatomical expertise and is highly subjective. To bypass this manual FCR, a few studies were conducted on computerized methods [1–3]. However, due to complex structure of craniofacial features, these methods were not effective in mapping skull-to-face features.

In the direction of craniofacial identification, direct skull-to-face matching [4–6] can be seen as an alternate solution but this task is overburdened by model to map the representation of skull with the face, also it does not aid to the deceased individual's family and friend to visually identify them. The recent methods in FCR perform skull-to-face reconstruction [7, 8] using different generative AI models (i.e., GANs). Although these generative models have claimed good performance in terms of mapping skull structure to the facial appearance. But, due to very limited dataset and absence of any public benchmark dataset in this domain, these generative models often fail to generalize in preserving structural identity of individuals under different conditions such as age and gender variations. In the recent years, diffusion-based models [9, 10] have emerged as state-of-the-art methods in generating improved facial images from 3D CT skull images with improved training stability. Furthermore, the introduction of multimodal text-conditioned diffusion models [11] has enabled flexible and controllable image synthesis by incorporating semantic guidance through textual prompts. Also, these methods primarily focus on visual appearance and lack explicit mechanisms to enforce anatomical consistency. However, existing diffusion-based approaches for craniofacial reconstruction conducted on 3D CT scans, as collection of 3D CT scan has following limitations. First, high radiation exposure. Second, need high cost device and setup. Third, high compute and processing time.

Despite advances in these generative models, we have found three key limitations that remain in existing craniofacial reconstruction methods. **First**, there is a lack of multimodal conditioning that integrates both structural (skull/X-ray) and semantic (textual) information. **Second**, absence of explicit geometric constraints to enforce

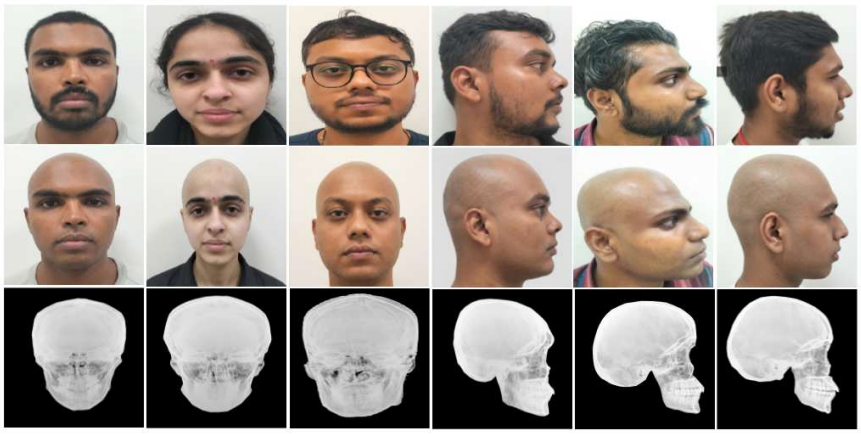


Fig. 1 Face-skull dataset with cleaned and ground truth face images.

anatomical consistency. **Third**, limited exploration of cross-modal reconstruction from 2D X-ray images.

Hence, to address these challenges, we propose *Cranio-Diff*, a diffusion-based framework for forensic skull-to-face reconstruction from skull images. The proposed framework integrates a trainable ControlNet-guided [11] Stable Diffusion v1.5 [12] backbone, which has been fine-tuned for photorealistic facial synthesis (Realistic Vision v5.1[1]). This enables to generate photorealistic facial reconstructions directly from skull anatomy. Also, few studies were conducted on 3D skull and 3D face due unavailability of public benchmark dataset and these 3D scans requires large amount of resources to collect and process the data. Hence, this study benefits from the use of 2D skull and 2D face image datasets, which significantly reduce redundant information compared to 3D and are also easier to collect than 3D. To the best of our knowledge, no comprehensive automated craniofacial reconstruction techniques have been proposed to directly recreate high quality 2D images of the face from the 2D X-ray scan data with variation of age and body mass index.

The major contributions of this research are as follows:

- We propose *Cranio-Diff*, a diffusion-based framework for craniofacial reconstruction from multi-view 2D X-ray images with variation of age and Body mass index.
- We introduce a multi-modal conditioning strategy that integrates structural information and textual prompts for anatomically consistent and controllable face generation.
- We have conducted extensive experiments on multimodal skull-to-face generation and evaluated their performance using different metrics such as FID, SSIM, LPIPS score, demonstrating the effectiveness of the proposed approach.
- We synthesize proposed dataset with controlled age (25, 45 and 65) and BMI (-10%, baseline, +10%) variations, and also curated the dataset across four facial

[1] https://huggingface.co/SG161222/Realistic_Vision_V5.1_noVAE

modalities (RGB face with hair, hair-removed RGB face, sketch face with hair, and hair-removed sketch face), supporting diverse forensic matching scenarios and enabling systematic evaluation of skull-face association under realistic appearance changes.

- Additionally, we have conducted experiments on retrieval of face from the gallery of face images for the given query (i.e., reconstructed face image) from the skull using three face recognition backbones (i.e., FaceNet, ArcFace and VGGFace) and evaluated their retrieval performance using Recall@k, mean Average Precision (mAP@k) and Mean Reciprocal Rank (MRR@k) metrics.

The rest of the paper is organized as follows. In Section 2, we introduce the latest research on Craniofacial Identification using reconstruction. In Section 3, we present the details of the dataset proposed in this paper. In Section 4, we introduce the methodology for Craniofacial reconstruction. Implementation details, experimental results and a detailed analysis are discussed in Section 5. The conclusion is drawn in Section 6, followed by Declaration and Acknowledgment in Section 7 and Section 8, respectively.

2 Related work

Traditional Craniofacial Reconstruction. Craniofacial reconstruction has traditionally been performed manually by forensic experts, who superimpose facial information over a skull using anatomical guidelines and tissue depth markers information [13, 14]. These approaches are labour-intensive, require significant expertise, and lack scalability. Moreover, they offer limited evaluation due to the absence of reliable ground truth data. To address these limitations faced by traditional methods, early computational methods use statistical learning techniques such as least squares support vector regression [15], latent root regression [16], and partial least squares regression [17]. Although these approaches are used to model global relationships between skull and facial features which are heavily on predefined landmarks and struggle to capture fine-grained local details. Additionally, their performance is constrained by limited training data and weak generalization across populations.

Deep learning-based approaches. With the recent advancement in deep learning models and their application in the field of generative AI, the task of image generations has been significantly improved. GAN-based frameworks [18], particularly Pix2pix (cGAN) [19] and cycleGAN [20], have been widely used for cross-domain image-to-image synthesis due to their ability to learn mappings between unpaired domains. However, these approaches in craniofacial reconstruction [7, 8] often struggle to effectively preserve structural attributes and are difficult to leverage multimodal information (like text and images), especially in scenarios involving large geometric and structural variations between skulls and facial appearance. Hence, GAN-based methods remain limited by their training instability and tendency to produce blurred or structurally inconsistent results.

Later on very few studies were done on diffusion-based approaches [9, 10] to reconstruct face from the given skull. These methods use stable diffusion model [11], conditioned with text prompts for more controllable and improved face generation.

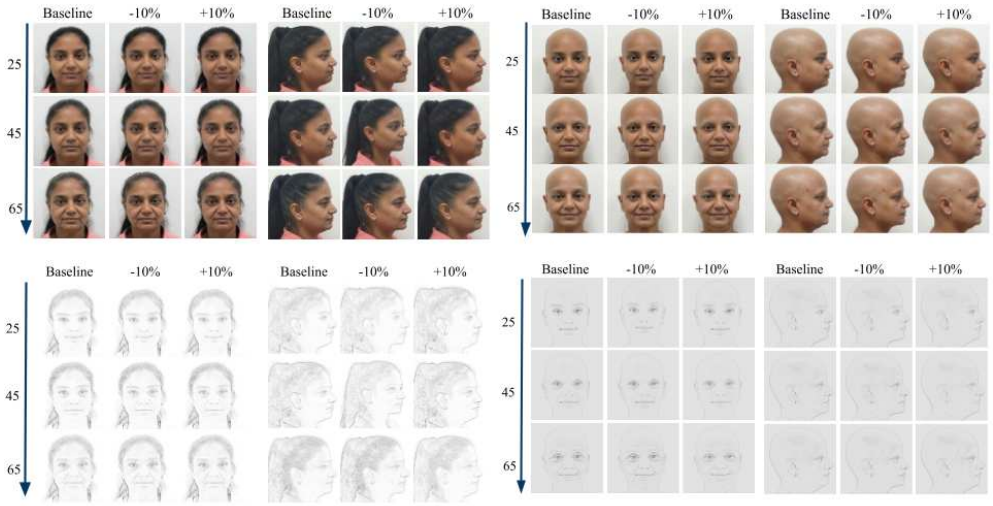


Fig. 2 Sample images of Subject ID 21 from the curated dataset across different modalities (RGB and sketch), with different facial appearances (with and without hair) showing variation across age (25, 45, 65) and BMI (-10%, baseline, +10%) for frontal and lateral view face image.

However, these studies are conducted on 3D CT scans of Chinese individuals, which are very difficult to get due to high cost and high radiation exposure. Notably, these diffusion-based models are still under exploration for craniofacial reconstruction.

3 Dataset

We have obtained S2F dataset as mentioned in [8]. We extended the dataset to 120 subjects following the same acquisition procedures and protocols. Required ethical approval has been taken from the institute ethical committee for the collection of required dataset. Also, the required volunteer consent form has been signed by the applicant for contributing to this dataset and research outputs. The proposed Skull-to-Face (S2F) dataset contains paired 2D skull X-ray and facial images captured from both frontal and lateral views. To focus on craniofacial structural information, the facial images were preprocessed to remove hair-related attributes, including hair, beard, and mustache, using the Google Gemini tool [21], thereby retaining only the craniofacial landscape. We use the prompt " age, gender, body mass, face orientation" to model craniofacial variations. Then, the facial images were synthesized across three age groups (25, 45, and 65 years) and three body mass index (BMI) variations (+10%, baseline, and -10%). This results in $9(= 3 \times 3)$ face variations for each view. As we have two views, hence $18(= 2 \times 9)$ total variations per subject. This process contributes 2,160 $(= 18 \times 120)$ paired skull-face samples from frontal and lateral views. Furthermore, to increase cross-view skull-face associations, both frontal and lateral skull images were paired with facial images from both views, contributing 36 skull-face pairs. Finally, we have a total of $4,320(= 36 \times 120)$ paired samples. To further support multi-modal

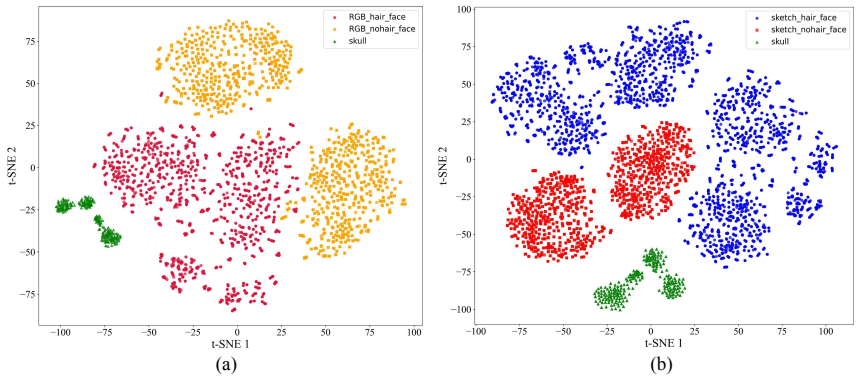


Fig. 3 2D t-SNE visualization of skull and face embeddings across different facial modalities. (a) shows embeddings of RGB faces with hair and without hair together with skull embeddings, while (b) shows embeddings of sketch faces with hair and without hair together with skull embeddings. Best viewed in color.

forensic analysis, the same pairing protocol is applied to four facial modalities: RGB faces, hair-removed faces, sketch faces, and hair-removed sketch faces. To get sketches from the RGB face images, we used the Canny-derived facial edge detector filter [22]. This results in a total of 17,280 multi-modal skull-face associations. Sample images from the prepared dataset are illustrated in Figure 1 and Figure 2. The dataset was divided into 90% training and 10% testing subsets. To further improve the diversity of the training data, nine different data augmentation techniques were applied, increasing the training set to 38,880 paired skull-face samples per modality. Here, we want to clarify that each synthetically generated image is cross-verified and validated by four distinct volunteers to ensure the identity of individuals in the proposed dataset. Additionally, we also visualize the distribution of our curated dataset in 2D t-SNE plot in Figure 3.

4 Methods

In this section, we introduce *Cranio-Diff*, a diffusion-based framework for skull-conditioned facial reconstruction that integrates structural biometric conditioning with soft semantic guidance, as depicted in Figure 4. During training, a pretrained Stable Diffusion model [12] combined with ControlNet [11] is employed to establish skull-craniofacial correspondence and generate structurally consistent facial reconstructions. A frozen text encoder [23] provides auxiliary semantic control through standardized biometric text prompts encoding age, gender, pose (i.e., Frontal or Lateral), and body fat (i.e., BMI). The framework is supervised using perceptual and structural preserving objectives to ensure visual fidelity and biometric consistency.

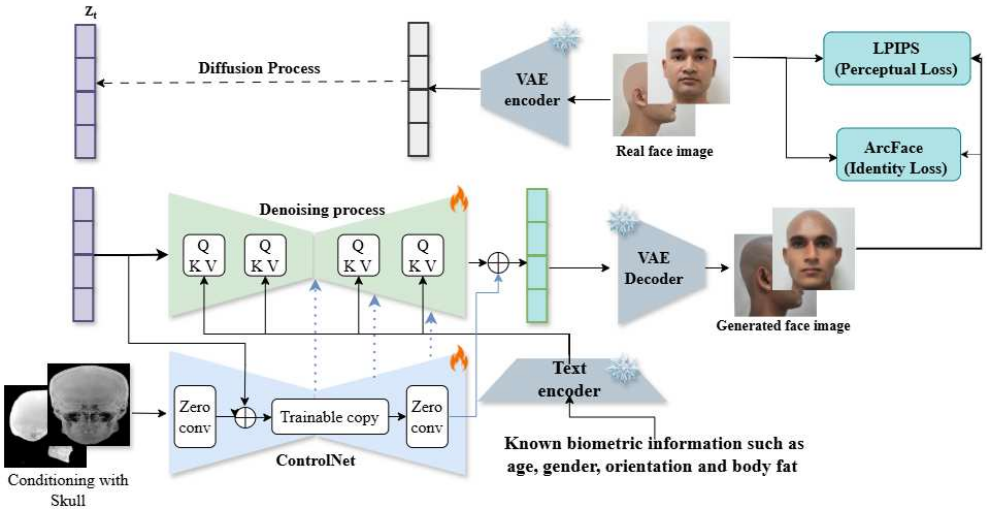


Fig. 4 Proposed *Cranio-Diff* framework: In this framework, first a frozen VAE encoder maps the real face image into latent space, where forward diffusion produces a noisy latent z_t. A trainable ControlNet branch conditions the denoising UNet on skull morphology and the UNet is jointly guided by biometric text embeddings (age, gender, orientation, body fat) from a frozen text encoder. The denoised latent is decoded by a frozen VAE decoder to synthesize the final face image and is supervised by LPIPS perceptual loss and ArcFace loss against the real face.

4.1 Overview

The proposed framework consists of three major components. **First**, skull-based structural conditioning via ControlNet. **Second**, biometric text-guided diffusion generation and **third**, perceptual and structural-preserving refinement. During training, only the ControlNet branch and the denoising UNet are fine-tuned, while the VAE encoder, VAE decoder, and text encoder remain frozen. This helps in preserving the generative capacity of the pretrained diffusion backbone with minimal training parameters.

4.2 Skull-Conditioned Structural Conditioning

Given a skull image $X_s \in \mathbb{R}^{H\times W\times 3}$, the proposed framework employs a ControlNet branch to extract and inject structural anatomical conditioning into the denoising UNet. The skull image is first processed through a zero-initialized convolution layer, ensuring that at initialization the conditioning signal introduces no perturbation to the pretrained UNet weights:

$$F_s^{(0)} = \mathcal{Z}_{in}(X_s), \quad F_s^{(0)} \in \mathbb{R}^{C\times H\times W} \tag{1}$$

where $\mathcal{Z}_{\text{in}}(\cdot)$ denotes the input zero convolution. The projected skull features are subsequently combined with intermediate UNet encoder activations via element-wise

addition and propagated through a trainable copy of the UNet encoder $\mathcal{E}_{\text{ctrl}}(\cdot)$:

$$F_s = \mathcal{E}_{ctrl}(F_s^{(0)} \oplus F_{unet}), \quad F_s \in \mathbb{R}^{C\times H\times W} \tag{2}$$

Here, F_{unet} denotes the feature map extracted from the encoder block of the denoising UNet. A second zero-initialized convolution $\mathcal{Z}_{\text{out}}(\cdot)$ then projects these features they are injected into the corresponding UNet blocks through residual feature fusion.

$$C_s = \mathcal{Z}_{out}(F_s), \quad C_s \in \mathbb{R}^{C\times H\times W} \tag{3}$$

This design allows the ControlNet to learn rich skull-to-face structural mappings while leaving the pretrained generative backbone intact. The zero convolution initialization ensures training stability by preventing early-stage noise propagation into the denoising UNet.

4.3 Biometric Text Conditioning

To incorporate soft semantic priors alongside structural skull conditioning, known biometric attributes, which include subject age, gender, head orientation, and body fat index. These attributes are encoded as standardized textual descriptions. Given a text prompt $\mathcal{T}$ encoding these attributes, the frozen text encoder $E_\tau(\cdot)$ produces semantic embeddings:

$$\tau = E_\tau(\mathcal{T}), \quad \tau \in \mathbb{R}^{L\times d} \tag{4}$$

where L denotes the token sequence length and d the embedding dimension. These embeddings are injected into each cross-attention block of the denoising UNet as keys and values, while queries are derived from intermediate noisy latent features. The cross-attention operation is formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{5}$$

where $Q = z_t W_Q$, $K = \tau W_K$, $V = \tau W_V$, and $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ are learnable projection matrices within the denoising UNet. This mechanism enables the model to modulate facial generation according to biometric priors, complementing the hard structural constraints provided by the ControlNet branch.

4.4 Skull-conditioned Diffusion Generation

Given a real facial image x_0, the VAE encoder $\mathcal{E}(\cdot)$ maps it into a compact latent representation z_0 . The forward diffusion process progressively corrupts this latent with Gaussian noise over T timesteps:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}\, z_{t-1},\, \beta_t I) \tag{6}$$

At timestep t, the noisy latent is expressed as:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon \tag{7}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. The denoising UNet, jointly conditioned on the skull conditioning map C_s and text embeddings τ, predicts the injected noise as $\epsilon_\theta(z_t, t, C_s, \tau)$. The diffusion training objective is:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, C_s, \tau)\|_2^2\right] \tag{8}$$

The denoised latent is subsequently decoded by the frozen VAE decoder $\mathcal{D}(\cdot)$ to produce the synthesized facial image $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$. Here, $\hat{z}_0$ is the latent representation estimated by UNet.

4.5 Structural-Preserving Refinement

Although the diffusion objective ensures structural skull correspondence, additional supervision is required to preserve perceptual quality and biometric structural consistency in the generated facial images. Accordingly, two complementary loss functions are employed.

Perceptual Loss. To improve perceptual fidelity between the generated and real facial images, LPIPS perceptual loss [24] is applied:

$$\mathcal{L}_{\text{LPIPS}} = \sum_{l} \|\phi_l(\hat{x}_0) - \phi_l(x_0)\|_2^2 \tag{9}$$

where $\hat{x}_0$ denotes the generated facial reconstruction and $\phi_l(\cdot)$ represents deep feature activations extracted from a pretrained perceptual network at layer l.

Structural Loss. To enforce biometric structural consistency, the generated image is projected into the ArcFace embedding space. The embeddings of the generated and real facial images are computed as $z_{\text{gen}} = F_{\text{arc}}(\hat{x}_0)$ and $z_{\text{gt}} = F_{\text{arc}}(x_0)$, where $z_{gen}, z_{gt} \in \mathbb{R}^d$ respectively, and the structural consistency loss [25] is formulated as:

$$\mathcal{L}_{\text{id}} = 1 - \frac{z_{\text{gen}}^{\top} z_{\text{gt}}}{\|z_{\text{gen}}\|\|z_{\text{gt}}\|} \tag{10}$$

Total Objective. The final optimization objective jointly minimizes structural, perceptual, and structural-preserving losses:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_1 \mathcal{L}_{\text{LPIPS}} + \lambda_2 \mathcal{L}_{\text{id}} \tag{11}$$

where λ_1 ($= 0.20$) and λ_2 ($= 0.20$) control the relative contributions of perceptual and structural-preserving regularization, respectively. The values of λ_1 and λ_2 are selected empirically. During optimization, gradients are propagated through both the ControlNet branch and the denoising UNet, ensuring training efficiency and stability.

5 Experiments

The primary objective of forensic craniofacial reconstruction is to identify unknown deceased individuals; thus, the effectiveness of reconstructed faces in recognition algorithms is essential. Hence, we tested various baselines with benchmark S2F dataset

with different evaluation metrics. We also evaluate the retrieval performance of our proposed model with various face recognition backbones on different retrieval metrics.

5.1 Implementation Details

The proposed *Cranio-Diff* method is implemented in Pytorch and trained on Nvidia H200 GPU with 141 GB VRAM, designed for complex deep learning computations. We trained our proposed framework for 350 epochs. We utilized Realistic Vision v5.1 (fine-tuned version of Stable Diffusion v1.5), as the backbone of generative framework. All images are resized to 512 x 512 spatial resolution. We selected the best hyper-parameters empirically: a learning rate of 0.0001, a batch size of 14, a weight decay of 0.00001 using the AdamW optimizer. For training our proposed model with S2F dataset, we partitioned this dataset at subject level (to prevent identity leakage) into training (108 subjects) and test (12 subjects) with the split ratio of 90:10 respectively. Total nine types of data augmentation techniques are applied to improve the generalization and robustness of the proposed method, which includes contrast jitter ($\pm$20%, 0.8$-$1.2), brightness jitter ($\pm$15%, 0.85$-$1.15), horizontal flipping (probability = 1.0), small-angle rotation with Gaussian noise ($\pm 5^{\circ}$, $\sigma = 0.01$), and random erasing (maximum erased area = 5%). Furthermore, composite augmentations were also employed, including contrast + rotation ((0.8$-$1.2) + $\pm 5^{\circ}$), brightness + noise ((0.85$-$1.15) + $\sigma = 0.01$), flip + contrast (flip + (0.8$-$1.2)), and erase + brightness + rotation (combined augmentation). All frontal and lateral views, synthetic age and BMI variations, and various augmentations are derived within that subject's split and all results are reported based on the test set.

5.2 Experimental results

Quantitative comparison. The proposed framework is evaluated on S2F dataset, we have used Fréchet Inception Distance (FID) [26], Structural Similarity Index scores (SSIM) [27], LPIPS [24]. Table 1 provides quantitative results on different baseline methods and our model achieves the improved performance across majority of mentioned metrics, with FID of 58.89, SSIM of 0.82, LPIPS of 0.24, which indicates that the generated face has improved photorealistic appearance and semantically alignment with its corresponding skull. We have also evaluated retrieval performance using three backbones (i.e., FaceNet, ArcFace and VGGFace) using three retrieval metrics (i.e, recall@k, mAP@k and MRR@k) with different gallery sizes, as shown in Table 3. From this table, we have observed that out of three backbones, VGGFace is performing better with recall@10 of 69.23, mAP@10 of 17.93, MRR@10 of 32.00 on gallery size of 100. It is also observed that when gallery size increase from 100 to 200, there is small reduction in the values of mentioned retrieval metrics.

Qualitative comparison. Figure 5 shows the qualitative comparisons between different generative models with our proposed framework. From this figure, it is observed that our FCR method reconstructs photorealistic face images. To show the effectiveness of *Cranio-Diff* across different craniofacial variations, Figure 6 and Figure 7 show results on generated face images with different variations of age and BMI

Table 1 Quantitative comparison of different methods on S2F dataset. Here, FID, LPIPS and SSIM scores are reported.

Method	RGB Face			Hair-Removed Face			Sketch Face			Hair-Removed Sketch		
	FID↓	LPIPS↓	SSIM↑	FID↓	LPIPS↓	SSIM↑	FID↓	LPIPS↓	SSIM↑	FID↓	LPIPS↓	SSIM↑
CycleGAN [20]	206.91	0.53	**0.57**	93.82	0.48	0.60	293.82	0.68	**0.70**	379.64	0.62	0.63
Pix2Pix [19]	400.00	0.74	0.11	178.62	0.47	0.70	285.22	0.66	0.68	294.48	0.65	**0.69**
CUT [28]	203.28	0.52	0.53	86.42	0.45	0.64	326.59	0.52	0.47	336.33	0.60	0.51
ControlNet [11]	90.55	0.59	0.30	107.82	0.43	0.49	182.11	0.75	0.16	110.21	0.42	0.60
ICCR-Diff [10]	77.13	0.46	0.45	72.97	0.40	0.56	**74.21**	0.50	0.23	101.18	0.40	0.50
Cranio-Diff (Ours)	**60.51**	**0.39**	0.56	**58.89**	**0.24**	**0.82**	80.45	**0.48**	0.19	**64.47**	**0.26**	0.45

Table 2 Ablation studies on Loss functions and text prompts. Here, w/o means without.

Configuration	FID↓	SSIM↑	LPIPS↓
Cranio-Diff (w/o text prompts)	64.46	0.78	0.31
Cranio-Diff (w/o ArcFace loss)	64.29	0.80	0.28
Cranio-Diff (w/o LPIPS loss)	**53.28**	0.81	0.25
Cranio-Diff (with text+both losses)	58.89	**0.82**	**0.24**

Table 3 Retrieval performance comparison under different gallery sizes on the S2F dataset. Higher values indicate better retrieval performance.

Method	Backbone	Gallery Size = 100						Gallery Size = 200					
		Recall		mAP		MRR		Recall		mAP		MRR	
		@10	@25	@10	@25	@10	@25	@10	@25	@10	@25	@10	@25
ICCR-Diff [10]	FaceNet [29]	36.67	61.67	5.71	8.38	11.55	13.08	21.39	41.67	2.78	4.33	6.19	7.48
	ArcFace [25]	41.11	66.94	9.54	13.35	15.76	17.36	27.78	47.58	4.85	6.91	8.08	9.32
	VGGFace [30]	30.28	51.11	5.52	8.36	9.82	11.12	17.78	35.56	2.92	4.39	5.55	6.62
Cranio-Diff (Ours)	FaceNet [29]	61.97	86.75	11.54	14.23	23.20	23.50	39.74	67.95	6.46	8.31	13.72	15.52
	ArcFace [25]	64.53	**98.17**	13.54	16.35	28.49	28.74	46.15	72.65	7.33	9.55	17.08	18.78
	VGGFace [30]	**69.23**	91.88	**17.93**	**21.25**	**32.00**	**32.20**	**48.72**	**72.65**	**10.76**	**13.11**	**20.36**	**21.94**

with high fidelity and semantic details. Figure 8 illustrates ROC-AUC curve comparison with three different backbones on S2F dataset with the gallery size of 100. From this ROC-AUC curve, it is observed that our proposed method *Cranio-Diff* shows promising results compared to the state-of-the-art method (i.e., ICCR-Diff) with the best backbone of VGGFace, achieving an ROC-AUC of 73.79%. Figure 9 shows qualitative results on retrieval of top-10 faces from the gallery for the given generated face images across different facial modalities. For retrieval of faces, we have utilized our best performing backbone, which is VGGFace. Here, the similarity of the generated face with the faces in the gallery is calculated on the basis of the Euclidean distance between the embeddings of images. For retrieval, the gallery contains multiple images of the correct identity for the given query image.

5.3 Ablation Studies

To evaluate the contribution of each component in the proposed framework, we conduct ablation studies on both biometric text conditioning and loss functions. The effect of biometric text prompts and loss functions is analyzed in Table 2. For loss ablation, two variants are considered: (i) removing the ArcFace loss while retaining the LPIPS perceptual loss, and (ii) removing the LPIPS loss while retaining the ArcFace loss.

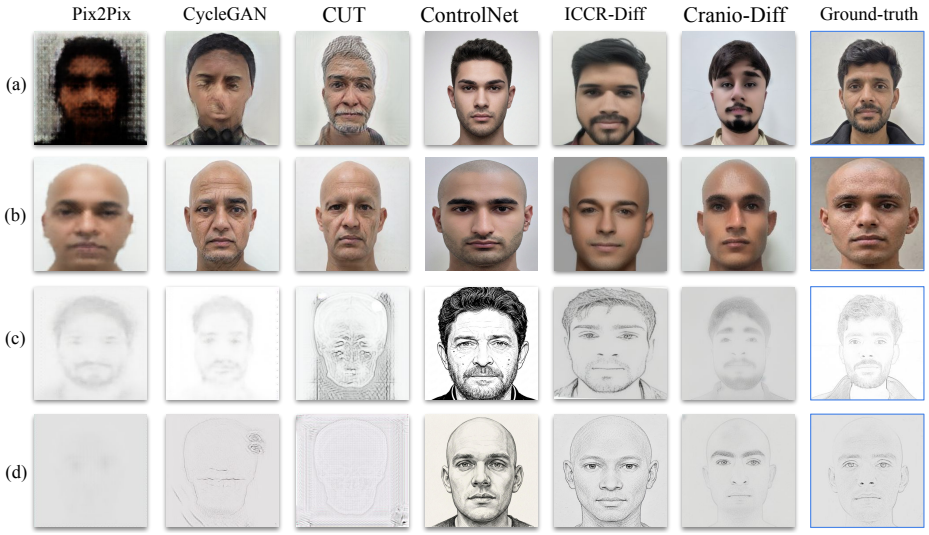


Fig. 5 Qualitative results on generated images with ground truth images across different modalities for different generative methods. Here (a), (b), (c) and (d) show results on different face modalities.

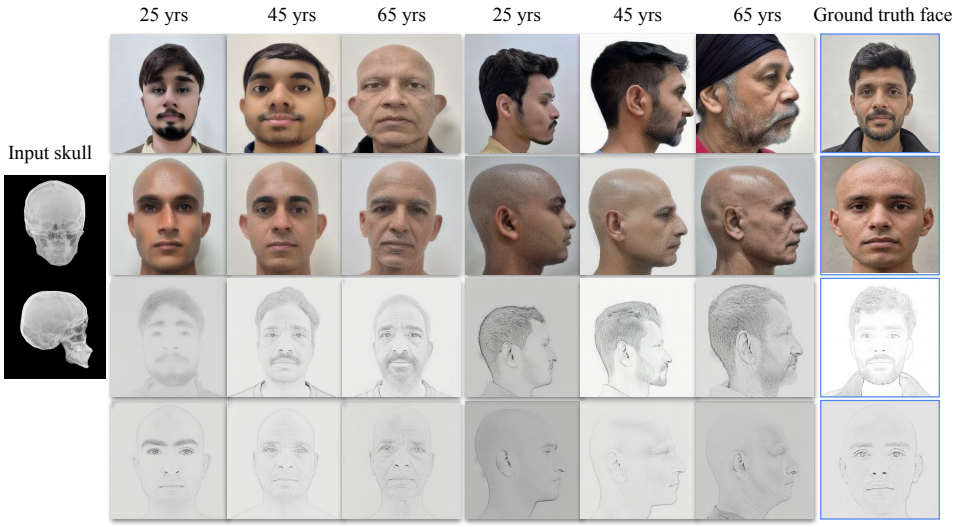


Fig. 6 Qualitative results on generated images with ground truth images across different modalities for age variations (25 yrs, 45 yrs and 65 yrs) of Cranio-Diff method when input skull image is given.

The results demonstrate that both losses contribute positively to the reconstruction quality and structural preservation of the generated facial images. When both losses

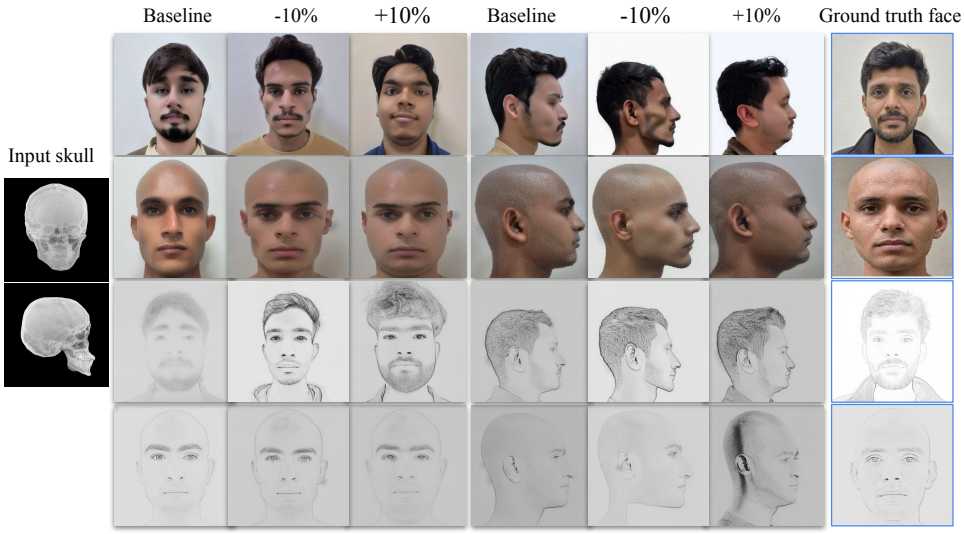


Fig. 7 Qualitative results on generated images with ground truth images across different modalities for BMI variations (-10%, baseline and +10%) of Cranio-Diff method when input skull image is given.

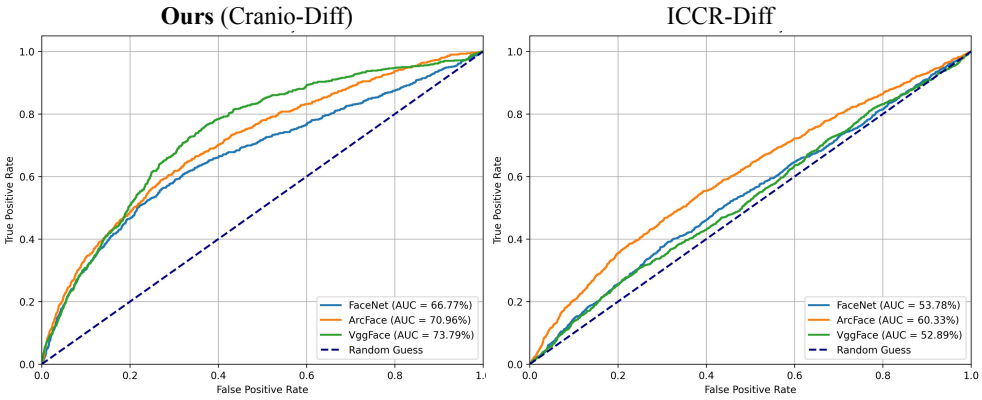


Fig. 8 Qualitative results showing ROC-AUC curve with gallery size of 100 on S2F dataset. Here left curve represents ours (*Cranio-Diff*) while right one represents ICCR-Diff. Best viewed in colors.

are jointly employed, the proposed *Cranio-Diff* framework achieves the best overall performance.

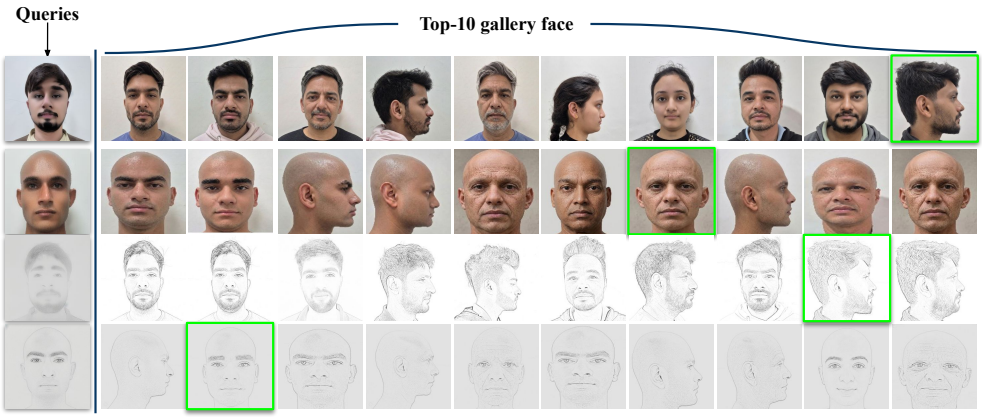


Fig. 9 Qualitative results on top-10 retrieval across different modalities. Here, queries are the generated image from different modalities and face image with green box are the ground truth images.

5.4 Limitations and Discussion

In forensic craniofacial reconstruction, due to absence of public benchmark dataset, this study is conducted on curated S2F dataset, which consists of only Indian individuals. Hence, due to this, model may capture and learn population specific features which may limits its generalizability across different ethnicity. Also, when comes to retrieving faces from the gallery database for a given generated query face image, it is observed that the reported results at different values of k and gallery size have high recall@k but low mAP@k. This indicates that when correct identity is retrieved, it is not consistently ranked at the top positions. Hence, future work could be in the direction of collecting more diverse dataset, leveraging multi-modalities and in building a methods for better feature disentanglement and representation learning. We want to emphasize that the reconstructed face image is a probabilistic approximation based on specific population data. Hence, the proposed *Cranio-Diff* framework can be seen as a scientific tool to assist forensic experts in craniofacial identification and anthropological investigations. However, it should not be considered as a replacement for established standard forensic identification methods rather, it may provide complementary evidence to assist expert assessment and guide further investigations.

6 Conclusion

This work presents a unified framework for craniofacial reconstruction using 2D frontal and lateral X-ray facial scans. It is easy to collect 2D X-ray-face data since getting a face X-ray scan is more affordable and convenient than other methods. Our proposed approach leverages ControlNet guided Stable Diffusion model (i.e, Realistic Vision v5.1) with dual-view X-ray conditioned map to preserve fine grained structural information. With more diverse face dataset across age and BMI, enables the model to learn variation in face across different age and body mass. Additionally, integrating corresponding text prompt via CLIP text encoder helps the model to incorporates semantic

details. By comparative quantitative and qualitative analysis of different generative models for skull to face image translation problem, we have found that our proposed framework shows improved performance over the other generative AI frameworks in cross-domain image translation which is generating realistic human faces from the given 2D X-ray skull images.

Data availability Due to ethical concerns, our skull-face data will be made available on request with a proper data use agreement. The dataset used in this study, which includes X-ray images and their corresponding face vectors (vector representation of face images), is made publicly available at: https://github.com/singh-ml/IIT_Mandi_S2F. However, the original faces with their corresponding skull images cannot be shared due to privacy and ethical concerns.

Code availability The code used in this study will be released once the manuscript is accepted.

Funding This research is funded by seed grant project No. IITM/SG/DIS-ROS-SPA/111, Indian Institute of Technology Mandi, Department of Higher Education, Ministry of Education, Government of India.

Author Contribution declaration Conceptualization, Ravi Shankar Prasad, Naresh Gurjar; Methodology, Ravi Shankar Prasad, Naresh Gurjar, Shashank Baghel and Dinesh Singh; Data curation, Ravi Shankar Prasad, Naresh Gurjar, Chirag; Original Draft Preparation, Ravi Shankar Prasad; Review & Editing, Ravi Shankar Prasad and Dinesh Singh.

7 Declaration

Ethical approval This study has been reviewed and approved by the Institute Ethical Committee, REF: IITM/IEC(H)/2024/DS/P1, IIT Mandi.

Consent to participate

Informed consent for participation was obtained from all the subjects who are involved in this study.

Consent for publication

The data used in this study were anonymized properly and do not contain personal information for subject identification.

Competing interests

The authors declare that they have no competing interests.

8 Acknowledgment

For creating this dataset, we sincerely want to thank the volunteers who dedicated their time and effort to help create the dataset called *S2F*. We also appreciate the cooperation and assistance from the Institute Health Centre during the data collection process and IIT Mandi for necessary support to this research. During the preparation of this work, the author(s) used Grammarly in order to correct the spelling and grammar. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] Claes P, Vandermeulen D, De Greef S, Willems G, Clement JG, Suetens P. Computerized craniofacial reconstruction: conceptual framework and review. Forensic science international. 2010;201(1-3):138–145.

[2] Huang J, Zhou M, Duan F, Deng Q, Wu Z, Tian Y. The weighted landmark-based algorithm for skull identification. In: International Conference on Computer Analysis of Images and Patterns. Springer; 2011. p. 42–48.

[3] Guerra R, Martos R, Ibáñez Ó, Valsecchi A, Bermejo E, De Luca S, et al. International validation study of AI-guided Craniofacial Superimposition in a contemporary population sample. Forensic Science International. 2025;p. 112628.

[4] Prasad RS, Singh D. Cross-domain identity representation for skull-to-face matching with benchmark dataset. Journal of Visual Communication and Image Representation. 2026;p. 104890.

[5] Prasad RS, Singh D. Spot-face: Forensic face identification using attention guided optimal transport. In: International Conference on Pattern Recognition. Springer; 2026. p. 553–566.

[6] Prasad RS, Sharma N, Singh D. Cranio-ID: Graph-Based Craniofacial Identification via Automatic Landmark Annotation in 2D Multi-View X-rays. arXiv preprint arXiv:251114411. 2025;.

[7] Li Y, Wang J, Liang W, Xue H, He Z, Lv J, et al. CR-GAN: Automatic craniofacial reconstruction for personal identification. Pattern Recognition. 2022;124:108400.

[8] Prasad RS, Singh D. FCR: Investigating Generative AI models for Forensic Craniofacial Reconstruction. arXiv preprint arXiv:250818031. 2025;.

[9] Zhang H, Wang X, Sun H, Sun Y, Jin Y, Li R, et al. CR-DM: A novel craniofacial reconstruction framework based on diffusion model. Multimedia Systems. 2025;31(1):20.

[10] Zhang M, Wu H, Zang Z, Wang J, Niu C, Li Y, et al. ICCR-Diff: Identity-preserving and controllable craniofacial reconstruction with diffusion models. Knowledge-Based Systems. 2025;330:114554.

[11] Zhao S, Chen D, Chen YC, Bao J, Hao S, Yuan L, et al. Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in neural information processing systems. 2023;36:11127–11150.

[12] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 10684–10695.

[13] Wilkinson C. Facial reconstruction–anatomical art or artistic anatomy? Journal of anatomy. 2010;216(2):235–250.

[14] Hwang HS, Park MK, Lee WJ, Cho JH, Kim BK, Wilkinson CM. Facial soft tissue thickness database for craniofacial reconstruction in Korean adults. Journal of forensic sciences. 2012;57(6):1442–1447.

[15] Li Y, Chang L, Qiao X, Liu R, Duan F. Craniofacial reconstruction based on least square support vector regression. In: 2014 IEEE international conference on systems, man, and cybernetics (SMC). IEEE; 2014. p. 1147–1151.

[16] Berar M, Tilotta FM, Glaunès JA, Rozenholc Y. Craniofacial reconstruction as a prediction problem using a Latent Root Regression model. Forensic science international. 2011;210(1-3):228–236.

[17] Jia B, Zhao J, Xin S, Duan F, Pan Z, Wu Z, et al. Craniofacial reconstruction based on heat flow geodesic grid regression (HF-GGR) model. Computers & Graphics. 2021;97:258–267.

[18] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27.

[19] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1125–1134.

[20] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2223–2232.

[21] Google.: Gemini. Large language model developed by Google DeepMind. https://gemini.google.com/.

[22] Canny J. A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence. 1986;(6):679–698.

[23] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PmLR; 2021. p. 8748–8763.

[24] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF conference on computer vision and pattern recognition. IEEE; 2018. p. 586–595.

[25] Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 4690–4699.

[26] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems. 2017;30.

[27] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004;13(4):600–612.

[28] Park T, Efros AA, Zhang R, Zhu JY. Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer; 2020. p. 319–345.

[29] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 815–823.

[30] Parkhi O, Vedaldi A, Zisserman A. Deep face recognition. In: BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association; 2015. .