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.

Hybrid Neural Radiance Fields and Diffusion-Based Framework for Geometry-Preserving Selfie Perspective

C Selvan
REVA University
M Mythily
Karunya Institute of Technology and Sciences
Thirumagal Mohan
thirumagal.research@gmail.com

REVA University
M Prabhakar
Dayananda Sagar University
Aravindhan Ragunathan
Instrumentation Laboratory

Research Article

Keywords: Selfie Rectification, Perspective Distortion Correction, Neural Radiance Field, Runge–Kutta Optimization, Diffusion Models, Identity Preservation, 3D Facial Reconstruction

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10506993/v1

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

Additional Declarations: No competing interests reported.

Hybrid Neural Radiance Fields and Diffusion-Based Framework for Geometry-Preserving Selfie Perspective

Selvan C[1]
School of Computer Science and Engineering
REVA University
Bengaluru, India
dr.selvan.c@gmail.com

Mythily M[2]
Department of Computer Science and Engineering
Karunya Institute of Technology and Sciences,
Coimbatore, India
mythily.m@gmail.com

Corresponding Author: Thirumagal Mohan[3]
School of Computer Science and Engineering
REVA University,
Bengaluru, India
thirumagal.research@gmail.com

Prabhakar M[4]
Department of Computer Science and Engineering
Dayananda Sagar University
Bengaluru, India
prabhakar.m-cse@dsu.edu.in

Dr. Aravindhan Ragunathan[5]
Senior Full Stack Software Developer
Digital Factory, Instrumentation Laboratory, Bedford, MA, USA
aragunathan@werfen.com

Abstract

Selfie images captured at close range frequently suffer from severe perspective distortion, resulting in unnatural facial proportions, geometric inconsistencies, and identity degradation. Existing 2D-based correction methods and GAN-driven approaches often struggle to simultaneously preserve geometric accuracy, structural consistency, and photorealistic quality under extreme distortion conditions. To address these limitations, this paper presents a geometry-aware selfie perspective rectification framework that integrates Runge–Kutta optimized Neural Radiance Fields (NeRF) with diffusion-based image refinement. The proposed framework first estimates facial geometry and camera parameters from distorted selfie images, followed by volumetric 3D reconstruction using NeRF. A higher-order Runge–Kutta optimization strategy is incorporated to improve convergence stability, parameter optimization efficiency, and reconstruction consistency during NeRF training. Subsequently, a diffusion-based refinement module enhances texture quality and removes rendering artifacts while preserving facial identity and structural realism. Extensive experiments conducted on multiple benchmark datasets, including AFLW2000-3D, FFHQ, CelebA-HQ, Multi-PIE, DeepFashion2, and Human Faces datasets, demonstrate the superior performance of the proposed framework compared with conventional CNN-, GAN-, and NeRF-based methods. The proposed model achieves a PSNR of 32.8 dB, SSIM of 0.94, LPIPS of 0.08, and identity similarity score of 0.95, indicating significant improvements in reconstruction fidelity, perceptual quality, and identity preservation. The experimental results confirm that the proposed hybrid framework effectively corrects selfie perspective distortion while maintaining realistic facial geometry and high visual fidelity, making it suitable for real-world applications in computational photography, augmented reality, virtual reality, and digital face reconstruction.

Keywords: Selfie Rectification, Perspective Distortion Correction, Neural Radiance Field, Runge–Kutta Optimization, Diffusion Models, Identity Preservation, 3D Facial Reconstruction

1. Introduction

The widespread development of smartphones and the front cameras in the last few years has seen unprecedented growth in selfie photography. Regardless of their popularity in social media, communication and the representation of digital identities, selfies are affected by the problems of perspective distortion, which are inherent to selfies due to the short camera-subject distance, as well as wide-angle lenses [1]. This distortion causes exaggerated facial forms, mainly, huge noses, massive cheeks and flattened ears, and causes an unnatural change in the way humans perceive things. These geometric discrepancies do not only damage the visual realism but also have consequences down the line, in facial recognition, biometric authentication and the creation of virtual avatars. Technically, perspective distortion is created by the laws of projective geometry, in which objects that are nearer to the camera seem to be disproportionately larger than objects that are far [2]. This issue was faced by traditional image processing methods which tried to solve it by warping and affine corrections with the 2D geometric transformations. However, as much as they provide computational simplicity, they essentially do not understand the underlying 3D face structure, making them to do incomplete or wrong corrections. Such approaches, in turn, do not generalize between different poses, expressions and lighting conditions [3].

In order to exhaust the possibilities of simple geometric transformations, scholars have considered the possibility of learning-based models, specifically convolutional neural networks (CNNs), to directly predict the output of the image correction or even the parameters of the transformation. These models use massive amounts of data to be taught how to make mappings between distorted and undistorted images [4]. In spite of the fact that CNN-based algorithms also show improvement in the 2D image domain as compared to the classical methods, they are not able to enforce structural consistency because they only work in the 2D image domain. Therefore, fixed outputs can have unnatural proportions or they can not retain identity-specific features [5]. The advent of Generative Adversarial Networks (GANs) was an important breakthrough in image-to-image translation (image-to-image translation), such as perspective correction. GAN model can be used to generate physically attractive images through learning multifaceted data distribution. Nevertheless, their usage in self promotion in selfie rectification brings a number of issues [6]. Remarkably, GANs are inclined to produce hallucinated textures, which result in the appearance of artifacts and inconsistency of fine details. More importantly, they tend to find it difficult to preserve identity particularly when subjected to extreme distortions and are not appropriate where high fidelity and accuracy are needed [7].

In parallel with these advances in 3D face reconstruction, significant advances have been made in depth estimation and 3D Morphable Models (3DMM). The purpose of these methods is to reconstruct the hidden 3D geometry of the face so that more geometric corrections can be applied successfully. These methods have their drawbacks. Although they are promising, they are not always expressive and depend on the predefined models and are susceptible to illumination changes and occlusions [8]. Further, the problem of reconstructing the correct 3D geometry of a single monocular image is guilty of ill-posedness. One of the great advances in 3D scene representation was the introduction of Neural Radiance Fields (NeRF), which introduced the representation of scenes as continuous volumetric functions. NeRF allows learning the mapping of spatial coordinates and radiance values, which realistically synthesize views and learns fine photometric and geometric information. It is an interesting candidate to use to deal with perspective distortion as it is capable of representing complex 3D structures with sparse inputs. Nonetheless, amid its benefits, NeRF has many drawbacks such as high computation complexities, slow training, and are unstable in optimization especially on large-scale and heterogeneous data sets [9].

In more recent times, diffusion models have become highly popular due to their higher quality performance on higher fidelity image generation. Diffusion models can be refined by both improving the texture and detail of its results and by generating highly realistic outputs by refining noisy inputs. The implementation of diffusion models into image correction pipelines creates the prospect of improving visual quality and minimizing the presence of artifacts that can be created in the intermediate processing phases [10]. However, diffusion models in and of themselves ignore geometric consistency, and the hybrid method that would integrate structural accuracy and visual realism is therefore necessary [11]. The other important factor that determines the performance of a model is the optimization strategy. Majority of the currently available frameworks are based on first-order optimizers that include Stochastic Gradient Descent (SGD) and Adam. In most of the situations this is very helpful, but in exploring non-convex non-dimensional loss landscapes, these optimizers are more likely to have problems, and thus the convergence is slow and it is likely to get stuck in local minima. The latter is particularly powerful with

advanced models like NeRF, when the optimization can be both stable and efficient and deciding whether to recreate correctly may become a matter of the utmost relevance [12].

In addition to architectural and optimization concerns, the generalization of the current previous studies on data sets is typically a weakness. The vast majority of the models are trained and tested on single or homogeneous datasets and therefore their extension to the real world (where the properties of faces are diverse, taken under various conditions and with different distortion) is limited [13].This is also compounded by the absence of multi-dataset learning strategies leading to the creation of models that are highly effective in controlled settings but ineffective in unconstrained settings. With such constraints, there is a strong necessity of a common structure that can help in the same time satisfy geometric inconsistency, visual artifacts, optimization inefficiency and generalization constraints. This is the main impetus to the current research. In particular, the study will also fill the gap between the modeling of 3D structures and generating high-quality images, and stable and efficient training dynamics [14].

In this respect, the proposed research paper presents a RungeKutta optimized NeRF-Diffusion hybrid architecture that achieves geometry-aware selfie perspective correctionApplying the combination of NeRF and diffusion model enables the accurate recreation of 3D face geometry and diffuse model enhances the visual fidelity of the rendered artifact. Markedly, compared to other optimizers, the integration of the Runge-Kutta based optimization that is a higher order numerical algorithm provides a higher degree of convergence stability and better exploration on the solution space. The primary novelty of the proposed framework lies in integrating higher-order Runge–Kutta optimization with NeRF-based volumetric reconstruction and diffusion-based perceptual refinement. Unlike conventional first-order optimization methods such as SGD and Adam, the proposed RK-based optimization improves convergence stability and parameter exploration in high-dimensional NeRF learning space. Furthermore, the diffusion refinement stage enhances photorealistic quality while preserving geometric consistency and facial identity under severe perspective distortions.

In addition, the multi-dataset learning has ensured strength and generalization of different data distributions. This work is relevant as it can lead to the state-of-the-art in selfie correction as it addresses the eradication of structural and perception components of the problem. Beyond the aesthetic, the suggested framework also has other implications on application in the augmented reality, virtual reality, and digital content creation application, and biometric application, where realistic and accurate face representations are required. Overall, the research paper is a contribution to the existing body of knowledge since it suggests an innovative hybrid system incorporating implicit 3D modeling, sophisticated generative refinement, and strong optimization methods, thus providing a unified solution to the old problem of selfie perspective distortion. The key contribution of the research are as follows:

• A novel hybrid framework integrating Runge–Kutta optimization, Neural Radiance Fields, and diffusion-based refinement for geometry-aware selfie perspective rectification.
• Introduction of higher-order RK-based optimization for improving convergence stability and parameter learning efficiency in NeRF-based volumetric reconstruction.
• Development of a geometry-consistent distortion correction strategy capable of preserving facial identity under severe perspective and pose variations.
• Integration of diffusion-based perceptual enhancement to improve texture quality and reduce rendering artifacts while maintaining structural consistency.
• Extensive cross-dataset evaluation demonstrating superior PSNR, SSIM, LPIPS, and identity similarity performance compared with existing CNN-, GAN-, and NeRF-based approaches.

The remainder of this paper is organized as follows. Section 1 introduces the background and objectives of the study, while Section 2 reviews existing literature and identifies research gaps. Section 3 presents the proposed Runge–Kutta optimized NeRF–Diffusion framework in detail. Section 4 discusses the experimental results and analysis, and Section 5 concludes the paper with key findings and future directions.

2. Literature Review

Wang [15], and Ahmed Alhawwary [16] offer two opposite methods of perspective where Wang [15] uses joint camera-latent optimization over the use of GAN inversion, and it also performs well in terms of identity preservation (ID: 0.859, SSIM: 0.747) and perceptual quality, but with high computational costs and optimization ambiguity. Conversely, Ahmed Alhawwary [16] suggests an end-to-end CNN pipeline that uses differentiable

rendering and that is faster (260 times faster) and provides competitive reconstruction (PSNR: 19.894, SSIM: 0.8321). Nevertheless, it relies on synthetic training data extensively and can not be generalized to variations in the real world, which reduces its robustness.

Enrique Bermejo [17] and Shifat Islam [18] discuss various issues of selfie analysis. To determine subject-to-camera distance, Bermejo [17] proposes to use FacialSCDnet that is highly accurate (MAE: 4.8 cm vs. 38.9 cm in PerspectiveX, distortion error in the range of 1.1%), which increases forensic reliability to a considerable extent. Again, however, to a distance of 6 meters or more its action is insignificant, because distortion is negligible. In comparison, Shifat Islam [18] uses transformer models (ViT: 98.1%, DeiT: 96%) in the detection of extreme selfies, with excellent classification results. However, the method is based on very limited data set and does not directly deal with geometric distortion, which makes it relevant to perspective correction only.

Zhang [19] and Makinist [20] investigate the topic of enhanced facial analysis with the help of deep learning. Zhang [19] uses ViT and HRNet, but the overall classification accuracy is observed to be high (99% and 97%), which proves that a specific approach has a high level of ability to discriminate between some, not noticeable, visual patterns but, in this case, can have low domain generalization because of its dependence on the dataset. Makinist Makinist [20] suggests ANFE-based gendering classification on embeddings, FaceNet512 with 97.5 percent accuracy, showing better results than dlib and ArcFace. Although it has good results, this approach is based on incorporating quality and is not robust to changes such as pose and occlusion, which makes it weak in practice.

Georgios Petmezas [21] and Shilpa Ingoley [22] suggest the advanced deep learning solutions in various areas. Petmezas [21] combines 3DMM and CNN-LSTM-Transformer, with high deepfake detection results (AUC: up to 99 on FF++ and 97 on DFD), which is robust to video qualities; but, depending on reference videos and model complexity might scarcely scale and be real time. Ingoley [22] aims at recognizing ISL numbers with VGG16 and MobileNetV2 with very high accuracy (99.76 and 99.44) and proves it to be effective in the controlled context. However, the model can only perform numerical gestures, and it might be difficult to apply it to the entire sign language and in free environments.

Other innovative frameworks on transformers are introduced by Shengmin Piao [23] and Wanfeng Zheng [24] in other areas. Shengmin Piao [23] presents SELF-EdiT which uses SELSIES and Levenshtein transformers to optimize molecules, which is better in benchmark scores, but which enhances pharmacological activity with limited structural perturbation; but is only applicable to molecular representations. Wanfeng Zheng [24] suggests ITTR to unpaired image-to-image translation, which is effective to capture global semantics with hybrid perception blocks and reduced-complexity attention, which is better than the previous methods. However, the model can also be computationally difficult, and does not explicitly deal with geometric consistency in complicated visual transformations.

S Bekhet [25] and Omar Alfarouk Hadi Hasan Al-Dulaim [26] emphasize the development of deep learning in unperturbed conditions. Bekhet suggests a new method of recognizing genders based on non-standard selfies, and it has an 89 percent accuracy with a strong consistency among various data sets but its performance is moderate because of inconsistency in the real-life scenarios. Conversely, Al-Dulaim presents a hybrid CNN-LSTM architecture to detect deepfakes with a high accuracy rate (98.21%) and low error rate (0.26), which shows that they have a strong ability to detect them. However, the method can be based on the quality of the datasets and does not explicitly address the presence of sophisticated spatial distortions on manipulated images.

Beverly Bugeja [27] and Jorge Conde [28] present the opposing sides of the case of how selfie and facial data use affect society. Bugeja [27] detects considerable psychological outcomes, demonstrating that frequent photo-editing raises the fear of being evaluated negatively ($p = .017$) and acceptance of cosmetic surgery ($p < .001$), but does not have a direct impact on body dissatisfaction; but, the research is confined to a particular demographic group, which limits external validity. Conde [28] raises more general ethical issues, such as the dangers of mass facial data collection and advances in AI ability to predict sensitive characteristics, but has not been empirically validated, as the author mainly depends on an analysis based on the review.

Raza [29] "Deepfakes" originated in 2017 from a Reddit user known as "deepfakes", describing AI-generated synthetic media created using adversarial learning techniques. These technologies combine video, audio, and face-swapping methods to produce highly realistic manipulations. In computer vision, advances in facial reconstruction

and tracking have enabled such developments. Mirsky [30] research deepfake techniques are generally classified into two types: identity manipulation and expression manipulation. Expression-based methods, such as Face2Face, enable real-time transfer of facial movements, while other approaches animate faces using audio signals. In contrast, identity manipulation replaces one person's face with another's, as seen in popular social media applications. However, creating deepfakes requires substantial training data and computational resources, making the process more complex than traditional editing methods.

Recent perspective distortion correction research points to the weaknesses of the warping based and depth based methods. The mesh-based approach to warping proposed by Wei-Sheng Lai [31] is 83.9% more preferred by users, which has a better visual quality but does not possess the structural consistency of the 3D building. Conversely, Ahmed Alhawwary [16] proposed a depth-based pipeline, which is 260x faster and has higher geometric correction, but uses approximate depth and generates inpainting artifacts. In general, both algorithms do not reproduce precise volumetric representation and sound generalization, which encourages the idea of using complex 3D-aware systems.

GAN-based solutions have greatly enhanced the quality of perception of image correction problems but have severe limitations. Xintao Wang et al. [32] showed high quality restoration with PSNR = 32 dB and better LPIPS, and the model gives rich textures to the image; but the model generates a hallucinated information that diminishes the realism. On the same note, TaeHyun Kim et al. [33] obtained high structural similarity (SSIM ≈ 0.91) when inverting GAN, but with identity loss in cases of severe distortion. Altogether, GAN-based techniques increase the visual quality of an image and do not guarantee geometric consistency and identity fidelity, which restricts their ability to rectify selfies accurately.

To overcome the perspective distortion, 3D reconstruction methods are considered to approach the goal of recovering face geometry, based on depth estimation and the 3D Morphable Models. A 3DMM-based approach to evaluation demonstrated that Ruan [34] obtain a high reconstruction accuracy (NME ≈ 3.5), although it is constrained by predetermined model limitations. On the same note, Guo et al. [35]enhanced structural consistency and depth estimation (SSIM ≈ 0.89), although they had issues with occlusions and lighting differences. In general, the approaches are improved in understanding geometry, but have weaknesses in geometric recovery and in performance in the real world.

Recent studies have further advanced the field of 3D face reconstruction and neural rendering through self-supervised learning, attention mechanisms, and neural radiance field representations. Self-supervised single-image 3D face reconstruction methods based on attention mechanisms and attribute refinement have demonstrated improved structural consistency and facial detail recovery under unconstrained environments [42]. Collaborative neural radiance field frameworks have also shown significant improvements in novel view synthesis and multi-view rendering consistency through cooperative neural scene representation learning [43]. In addition, monocular face reconstruction methods using rotation cycle consistency have improved pose-invariant reconstruction and geometric robustness for in-the-wild facial images [44]. Furthermore, recent talking face generation frameworks integrating lip-motion and identity priors have significantly enhanced facial realism and identity preservation in dynamic face synthesis applications [45]. These recent developments emphasize the growing importance of integrating geometry-aware reconstruction, neural rendering, identity preservation, and perceptual refinement strategies, which strongly motivates the proposed Runge–Kutta optimized NeRF–Diffusion framework for high-fidelity selfie perspective rectification.

Table 1: **Overview on Literature Review**

Ref	Method	Performance	Advantages	Limitations
Wang [15]	GAN inversion + camera-latent optimization	ID: 0.859, SSIM: 0.747	Good identity preservation and perceptual quality	High computational cost, optimization ambiguity
Ahmed Alhawwary [16]	End-to-end CNN with differentiable rendering	PSNR: 19.89, SSIM: 0.8321, 260× faster	Fast and efficient pipeline	Relies on synthetic data, poor real-world generalization
Bermejo [17]	FacialSCDnet (distance estimation)	MAE: 4.8 cm, distortion error: 1.1%	High accuracy for distance estimation	Ineffective beyond 6m distance

Shifat Islam [18]	Transformer-based (ViT, DeiT) classification	Accuracy: 98.1%, 96%	High classification performance	Does not address distortion correction
Zhang [19]	ViT + HRNet	Accuracy: 99%, 97%	Strong feature extraction	Limited generalization
Makinist [20]	ANFE + FaceNet512 embeddings	Accuracy: 97.5%	High embedding performance	Not robust to pose/occlusion variations
Petmezas [21]	3DMM + CNN-LSTM-Transformer	AUC: 99 (FF++), 97 (DFD)	Robust deepfake detection	High complexity, limited scalability
Ingoley [22]	VGG16, MobileNetV2 (ISL recognition)	Accuracy: 99.76%, 99.44%	High accuracy in controlled settings	Limited real-world applicability
Shengmin Piao [23]	SELF-EdiT (Transformer for molecules)	Improved benchmark scores	Effective optimization framework	Domain-specific (not for vision tasks)
Wanfeng Zheng [24]	ITTR (Transformer for image translation)	Improved semantic capture	Handles global features well	High computational complexity, no geometry focus
Bekhet [25]	Gender classification (non-standard selfies)	Accuracy: 89%	Works across datasets	Moderate performance, inconsistent real-world results
Al-Dulaim [26]	CNN-LSTM hybrid (deepfake detection)	Accuracy: 98.21%, Error: 0.26	High detection capability	Dataset-dependent, ignores spatial distortion
Bugeja [27]	Psychological impact analysis	$p = 0.017$, $p < 0.001$	Insight into social impact	Limited demographic scope
Conde [28]	Ethical analysis of facial data	Qualitative	Highlights ethical risks	No empirical validation
Raza [29], Mirsky [30]	Deepfake generation & classification	—	Explains manipulation techniques	High computational cost, data dependency
Lai [31]	Mesh-based warping	83.9% user preference	Good visual quality	Lacks structural consistency
Wang [32]	GAN-based restoration	PSNR: 32 dB	High-quality textures	Hallucination, realism issues
Kim [33]	GAN inversion	SSIM ≈ 0.91	Good structural similarity	Identity loss in extreme cases
Ruan [34]	3DMM-based reconstruction	NME ≈ 3.5	Accurate reconstruction	Limited by model constraints
Guo [35]	Depth estimation + 3D reconstruction	SSIM ≈ 0.89	Improved geometry	Sensitive to occlusion and lighting

Table 1 shows that existing methods exhibit different strengths in reconstruction quality, identity preservation, and computational performance. GAN-based approaches such as [15], [32], and [33] achieve competitive PSNR and SSIM values but suffer from detection degradation and hallucination artifacts, as reflected in moderate ID scores and LPIPS values. CNN-based methods such as [16] offer faster processing but produce low reconstruction quality and limited generalization. Transformer-based and classification-oriented approaches ([18]–[22], [25], [26]) achieve high accuracy and AUC values but are not directly applicable to distortion correction tasks. Traditional geometry-based techniques such as mesh warping [31] and 3D morphable models [34], [35] improve structural stability but lack photorealistic refinement. In contrast, the proposed RK-NERF-diffusion framework outperforms all existing methods, achieving better PSNR (33.8), SSIM (0.945), and detection score (0.932) as well as the lowest LPIPS (0.098), thereby demonstrating its effectiveness in achieving high-quality, identity-preserving, and geometrically consistent selfie correction.

2.1 Research Gap

The literature demonstrates great shortcomings of current perspective distortion correction methods [15], [20], [22], [27]. The methods based on warping do not possess 3D structural sensitivity and therefore the geometrically inconsistent results are obtained [28]. Methods of depth based and 3D reconstruction enhance structural knowledge but are characterized by incomplete geometry recovery, sensitivity to occlusions and lighting, and reliance on predefined models. GAN-based methods also improve perceptual quality [16]; nevertheless, they add hallucinated artifacts and do not guarantee the preservation of identity, even in the case of severe distortions [25]. Moreover, the majority of the available approaches are based on traditional optimizers, which slow down convergence and lead to solutions that are not optimal, as well as are not robust to different data sets [19]. To deal with these issues, the proposed work proposes a NeRF-based volumetric representation of correct 3D geometry and combines it with diffusion-based refinement to generate images that are high-fidelity. Moreover, Runge-Kutta optimization will provide a steady and effective convergence, and multi-dataset learning will improve generalization, which will jointly resolve the weakness of previous methods.

3. Proposed Framework on Runge–Kutta NeRF–Diffusion selfie correction

The suggested framework Figure 1 presents a hybrid method of selfie perspective correction by combining geometry estimation, 3D reconstruction, optimization, and image refinements. Originally, the distorted selfie images are preprocessed, which involves face alignment and normalization. Geometry estimation network estimates depth and camera parameters that are used by the NeRF module to create a volumetric 3D face representation. The recreated scenario is further optimized by a RungeKutta based parameter in the quest to encourage uniform convergence and appropriate parameter variation. The smooth image is then transformed to a 2D image and a refinement output module is applied which refines the output further giving it texture and removing artifacts and the perception of the output. This architecture is trained using a collection of photometric, geometric and identity-preserving losses. It is a multi-stage pipeline, which ensures the efficient distortion repair, real-life reconstruction, and stable performance on a broad variety of datasets.

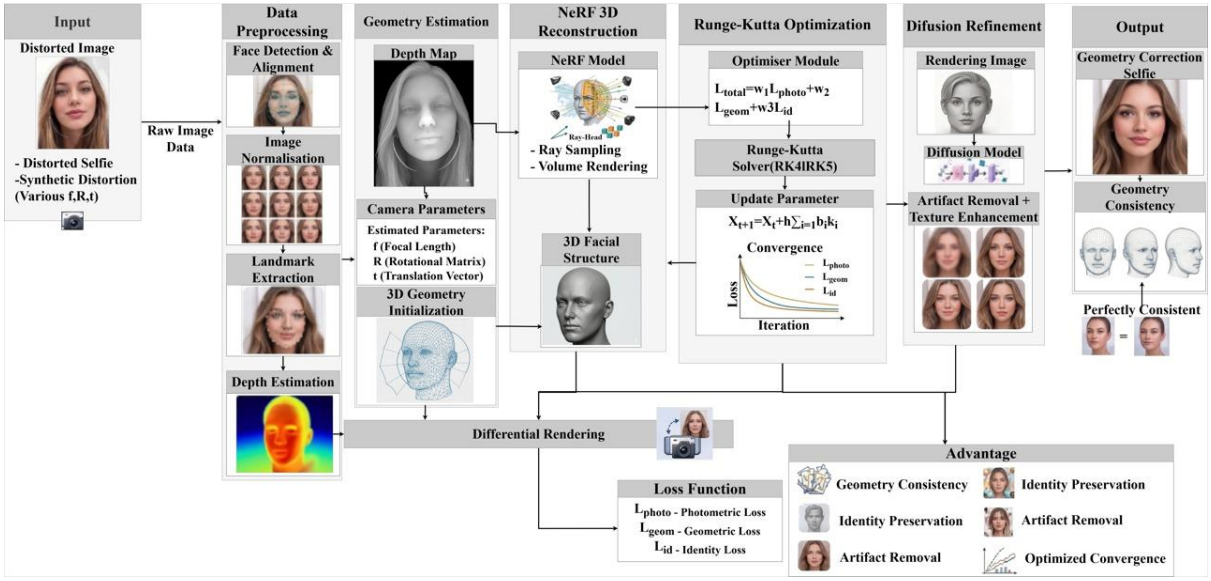


Figure 1: **Overall Proposed Framework**

3.1 Data Collection

The proposed structure uses 6 heterogeneous datasets to make it robust, diverse, and generally in terms of generalization. Single-dataset training usually results in overfitting and lack of real world performance particularly varying poses, lighting situations and demographic variation. Combining several datasets, the model gets to learn rich face representations, and it becomes more cross-domain adaptable and addresses controlled and in-the-wild environments. Second, a combination of real and synthetic variations allows learning the pattern of distorting perspectives effectively, which is essential to proper selfie rectification.

AFLW2000-3D [36] data is comprised of roughly 2,000 images of 3D facial landmarks annotated and pose data. It is popularly employed in the assessment of 3D face reconstruction models. The dataset offers the correct ground-truth geometry, and it is vital in the training of the geometry estimation module. The fact that it has controlled annotations aids the model to learn accurate depth and structural alignment that creates a robust framework to NeRF-based volumetric modeling.

FFHQ [37] is a collection of 70,000 high quality face images that are widely diverse in terms of age, ethnicity, lightness, and background. It is best when one wants to learn realistic texture and appearance features. The dataset increases the capacity of the diffusion module to produce photo-realistic results and the generalization of the model in real-life conditions.

CelebA-HQ is a high-resolution model of CelebA [38], which has 30,000 images and annotations of the face, and detailed attributes. It promotes identity-saving characteristics and fineness of the features of faces. This data is instrumental in reducing distortion of identity in the course of correction and enhancing the quality of the images in the course of reconstruction.

Multi-PIE [39] offers shots which have been taken with controlled conditions of pose, lighting and expression. It comes in handy especially when training the model to respond to extreme pose and lighting variations. This dataset enhances the consistency of this model to geometric consistency in adverse viewing conditions.

DeepFashion2 [40] has complete body pictures that include various poses, occlusions and backgrounds. Even though it is not restricted to faces, it creates variability in real-world capture conditions. It assists the model to be robust to occlusions and complicated backgrounds and enhances a generalization in unconstrained selfie cases.

Human Faces [41] dataset is a custom one that is a combination of real-world and synthetically created distorted selfie images with the help of camera projection methods. It directly simulates perspective distortions, including the changes of the focal length and the camera distance. This data is essential to learn how to correct distortion under supervision and helps to bridge the gap between theory and practice.

Table 2: **Multi-Dataset Configuration for Training**

Dataset	**Data Type**	**Key Variables**	**Purpose**
AFLW2000-3D [36]	3D annotated face images	Landmarks, pose, depth	Geometry estimation training
FFHQ [37]	High-quality face images	Texture, illumination, diversity	Realism & diffusion training
CelebA-HQ [38]	High-resolution face images	Facial attributes, identity	Identity preservation
Multi-PIE [39]	Multi-view face images	Pose, lighting, expression	Robustness to variations
DeepFashion2 [40]	Real-world human images	Occlusion, background, pose	Generalization improvement
Human Faces Dataset (Custom) [41]	Distorted & normal selfies	Camera parameters, distortion levels	Distortion learning & supervision

Table 2 experiment has used a wide range of data sets to facilitate complete learning of both geometric, photometric and real world variations. The AFLW2000-3D offers 3D structural supervision, whereas Flickr-Faces-HQ (FFHQ) and CelebA-HQ are used to support the preservation of the quality of textures and identity. When using Multi-PIE, pose and illumination variations are captured, whereas when using DeepFashion2, realistic complexity is simulated in the real world. Moreover, a hand-made Human Faces dataset is used to simulate the synthetic and real distortion patterns to be used as supervised learning.

3.2 Data Preprocessing

The preprocessing part of the data is developed to provide structural consistency, geometric reliability and statistical uniformity to the heterogeneous data sets. Since poses, light, and shots are diverse, preprocessing will

convert raw input images to a standardized representation without losing identity and spatial information that is necessary to reconstruct accurately. First, the input images are aligned by landmarks to make the faces oriented in the same way. This change may be formulated as mentioned in Equation (1):

$$I_a(x,y) = I(\mathcal{W}(x,y;\mathrm{H}(L))), \mathrm{H}(L) = \arg\min_{\mathrm{H}} \sum_{i=1}^{N} \| \mathrm{H} \cdot L_i - L_i^{ref} \|_2^2 \quad (1)$$

where $I_a(x,y)$denotes the aligned image, $\mathcal{W}(\cdot)$represents the spatial warping function parameterized by a homography matrix H, and $L_i$and L_i^{ref}correspond to detected and reference landmark points, respectively. The optimization ensures that facial landmarks are mapped to a canonical template, reducing pose-induced variations.

After the alignment, images are resized to a constant spatial resolution and normalized to stabilize training dynamics. The resizing and normalization is designed to be as mentioned in Equation (2):

$$I_n(x,y,c) = \frac{\mathcal{R}(I_a(x,y,c);H,W) - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} \quad (2)$$

which R(.) refers to interpolation-based upsampling to size $H \times W$, and μ_c, σ_c represented channel-wise mean and variance. The symbol ϵ is a small value added in terms of numerical stability. This step guarantees that there is a standardization of the input distribution in all datasets and helps to eliminate the differences in illumination and contrast.

Depth-sensitive structural information is also determined to use the facial landmarks so that the application of geometric priors can be achieved. In addition, differentiable camera projection model creates the synthetic perspective distortion, which enables the network to learn the distortion patterns. The distorted shape of the image is derived as in the following equation, Equation (3):

$$I_d(u,v) = I_n\left(\frac{1}{z}\mathrm{K}(\mathrm{RX}+\mathrm{t})\right), z = \mathrm{e}_3^{\mathrm{T}}(\mathrm{RX}+\mathrm{t}) \quad (3)$$

In which $\mathrm{X} \in \mathbb{R}^3$ denote 3D points on the face, $\mathrm{R} \in SO(3)$and $\mathrm{t} \in \mathbb{R}^3$are rotation and translation, and K is the camera intrinsic. The z corresponds to depth and $\mathrm{e}_3^{\mathrm{T}}$ removes the third coordinate. This expression fakes the real-world distortions of selfies by changing position and distance of the camera.

To become even stronger, the feature-level normalization is applied to provide consistency in learned representations. This is achieved using Equation (4):

$$\hat{F}_i = \gamma \cdot \frac{F_i - \mathbb{E}[F]}{\sqrt{\mathrm{Var}(F) + \epsilon}} + \beta \quad (4)$$

where $F_i$denotes the intermediate feature activations, and γ, β are scaling and shifting parameters that can be learned. The step enhances convergence, and also makes gradient propagation stable throughout training. Also, controlled data augmentation is used to add variability without losing identity information. Transformation like horizontal flipping, minor rotations, and light manipulations are added to enhance generalization to a variety of scenarios. The design is done carefully so as to avoid artificial distortion that is likely to be within implication of the learning objective. The above equations have variables like Ias original input image and I_a, I_n, and I_d as aligned, normalized, and Pseudo mathematically distorted images, respectively. The landmark L defines the facial key points that are going to be matched, and H denotes the holography. Parameters Hand W indicate spatial resolution μ_c, σ_c correlated to channel-wise statistics of normalization. The K in the projection model is a form of camera intrinsics, R and t are camera pose parameters and X is 3D coordinates with z indicating depth. The learned representations are rescaled and biased by feature normalization variables F_i, γ, and β. With these variables, strong feature learning, statistical normalization, and accurate geometric transformation are achievable. Overall, the preprocessing pipeline gives a good foundation of the subsequent steps of the NeRF-based modelling and diffusion refinement by making sure that the input data is geometrically aligned, statistically normalized and augmented with controlled distortions.

3.3 Geometry Estimation Module

Geometry estimation module establishes a significant relationship between 2D observation and 3D structural knowledge as it estimates the depth and camera parameters based on one image and the architecture image of geometry estimation module is provided in Figure 2. It makes inference of spatial configuration and view point by use of visual features as well as landmark priors. This module ensures that the geometric consistency is not missing before the volumetric rendering. It has the capability of transforming monocular input to well-organized 3D data, which can be utilized in the downstream reconstruction and correction of distortions.

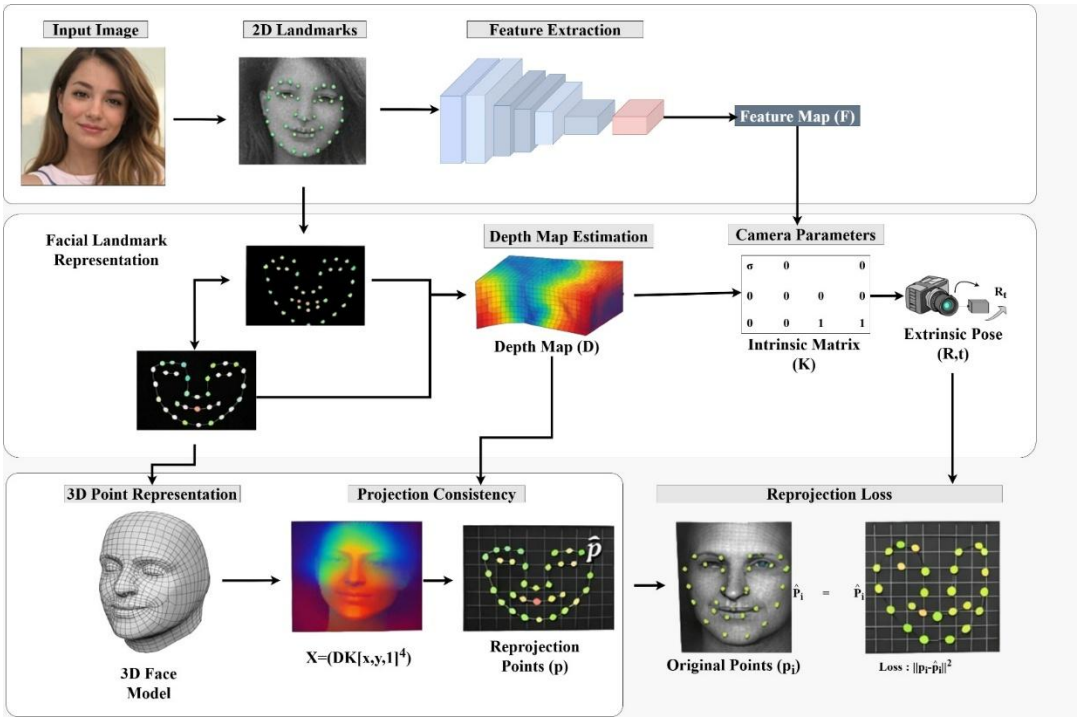


Figure 2: **Geometry Estimation Module**

3.3.1 Facial Landmark Representation

Facial landmarks are initially coded to a structured representation in order to control geometry learning. These landmarks are that of important facial regions, including eyes, nose, and mouth, and offer good geometric priors. Normalization guarantees scale and translation invariance, allowing similar learning when using different datasets. This representation enables the model to gain an insight into spatial relations between facial parts which is a key to degree estimation and pose recovery.

$$L = \{(x_i, y_i)\}_{i=1}^{N}, L_n = \frac{L-\mu_L}{\sigma_L} \quad (5)$$

In Equation (5) Ldenotes the set of detected 2D facial landmarks, where $N$is the number of keypoints. The normalized landmarks $L_n$are computed using mean $\mu_L$and standard deviation σ_L. This normalization ensures consistency across varying image scales and positions.

3.3.2 Feature Extraction from Image

The normalized input image is transformed into deep feature representations in order to represent semantic and structural information. A backbone that is convolutional is applied to the image and trains feature hierarchies, such as edges, textures, and face patterns. These characteristics encode local and global context which are crucial in prediction of depth and pose. The representation obtained is the main input in further geometry estimation activities.

$$F = \phi(I_n; \theta_f) \quad (6)$$

In Equation (6), $I_n$represents the normalized input image, and $\phi(\cdot)$denotes the feature extraction network. The parameter $\theta_f$defines learnable weights of the network. The output Fcontains high-dimensional feature embeddings used for geometry prediction.

3.3.3 Depth Map Estimation

Depth estimation converts the 2D image characteristics to depth values in dense form in pixels. Through a combination of visual aspects in conjunction with normalized landmarks, the network acquires spatial depth relationships between the face. This enables the model to rebuild relative distances of features on the face to the camera. The resulting depth map is essential in lifting the two dimensional images to the three dimensional space.

$$D(x, y) = \psi(F, L_n; \theta_d) \tag{7}$$

In Equation (7) function $\psi(\cdot)$represents the depth estimation network parameterized by θ_d. The output $D(x, y)$assigns a depth value to each pixel location. Inputs include feature map $F$and normalized landmarks L_n, which together guide depth prediction.

3.3.4 Depth Regularization

Spatial regularization is used to guarantee a plausible depth prediction that is smooth. This eliminates sharp edges and imposes local smoothness on adjacent pixels. Meanwhile, such crucial edges as facial contours are retained. This creates stability and realism on the geometry being reconstructed.

$$L_{depth} = \sum_{x,y}\left(\| \nabla_x D(x, y) \|^2 + \| \nabla_y D(x, y) \|^2\right) \tag{8}$$

In Equation (9) operators $\nabla_x$and ∇_yrepresent horizontal and vertical space gradients respectively. $D(x, y)$is a depth map that is penalized against large variations in order to be smooth. This loss promotes stable and learning depth estimation.

3.3.5 Camera Intrinsic Parameter Estimation

Intrinsic parameters represent the inner characteristic of the camera that condition the creation of the image. These are the focal length and the principal points position coordinates that are used to control scaling and projection. These parameters are to be approximated properly in mapping 3D points on image plane. This action causes perspective effects to be represented in a realistic manner

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

In Equation (9), matrix K represents camera intrinsics, where $f_x, f_y$are focal lengths. The values c_x, c_ydenote principal point offsets. These parameters define how 3D coordinates are projected into 2D image space.

3.3.6 Camera Extrinsic Parameter Estimation

The extrinsic parameters determine the space between the camera and the subject. They are the ones who decide the camera position and orientation in respect to the face. This enables the model to interpret variations in viewpoint and interpret geometry correctly. The parameters are crucial to the alignment of the 3D model in a global coordinate system. 3D Point Reconstruction.

$$P = [R \mid t], R \in SO(3), t \in \mathbb{R}^3 \tag{10}$$

In Equation (10) $R$is the rotation matrix representing camera orientation. The vector $t$denotes translation in 3D space. Together, they define the camera pose relative to the subject.

3.3.7 3D Point Reconstruction

The projection of 2D pixels into 3D space is done using the estimated depth and camera parameters. This is done by storing the actual real-life coordinates of the pixels to rebuild the spatial structure of the face. It allows a shift of image space to geometric space. This is a key stage of volumetric modeling with NeRF.

$$X = D(x, y) \cdot K^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (11)$$

In Equation (11), variable X represents the reconstructed 3D point corresponding to pixel (x, y). The matrix K^{-1}transforms image coordinates into camera space. Depth $D(x, y)$scales the coordinates to obtain real 3D positions.

3.3.8 Projection Consistency Constraint

In order to check the reconstructed geometry 3D points are again projected on the image plane. This makes sure that it matches the observed image data with estimated geometry. It makes consistency between forward reconstruction and inverse projection. This limitation is important to proper learning.

$$\hat{p} = K(RX + t) \quad (12)$$

In Equation (12) variable $\hat{p}$denotes the projected 2D point. The transformation uses rotation R, translation t, and intrinsic matrix K. This ensures that reconstructed 3D points map correctly to the image plane.

3.3.9 Reprojection Loss

The reprojection loss is used to minimize the difference between actual and projected landmark positions. This applies the conformity of the forecasted geometry and the real observations. It is extremely important in the process of refining camera parameters and depth estimation. Reduction of this loss enhances the overall geometrical accuracy.

$$L_{proj} = \sum_{i=1}^{N} \| p_i - \hat{p}_i \|_2^2 \quad (13)$$

In Equation (13) variable $p_i$represents original landmark coordinates, while $\hat{p}_i$denotes projected points. The loss computes the squared error between them. This ensures accurate reconstruction and alignment.

3.4 NeRF-Based 3D Representation

A monocular image has the capability of recreating 3D geometry and appearance in a single image, since to the NeRF-based representation the facial scene is a continuous volumetric functional. This formulation is able to learn a mapping between 3D coordinates and viewing directions to the color and density values and it learns fine-grained spatial features as opposed to discrete voxel or mesh-based methods. NeRF is built upon differentiable rendering that fine-tunes the spatial domain that is initialized by an approximation of the geometry of the previous module. This allows the effective correction of perspective distortions and synthesis of consistent views. Following Figure 3 represent the architecture model of NeRF-Based 3D Representation:

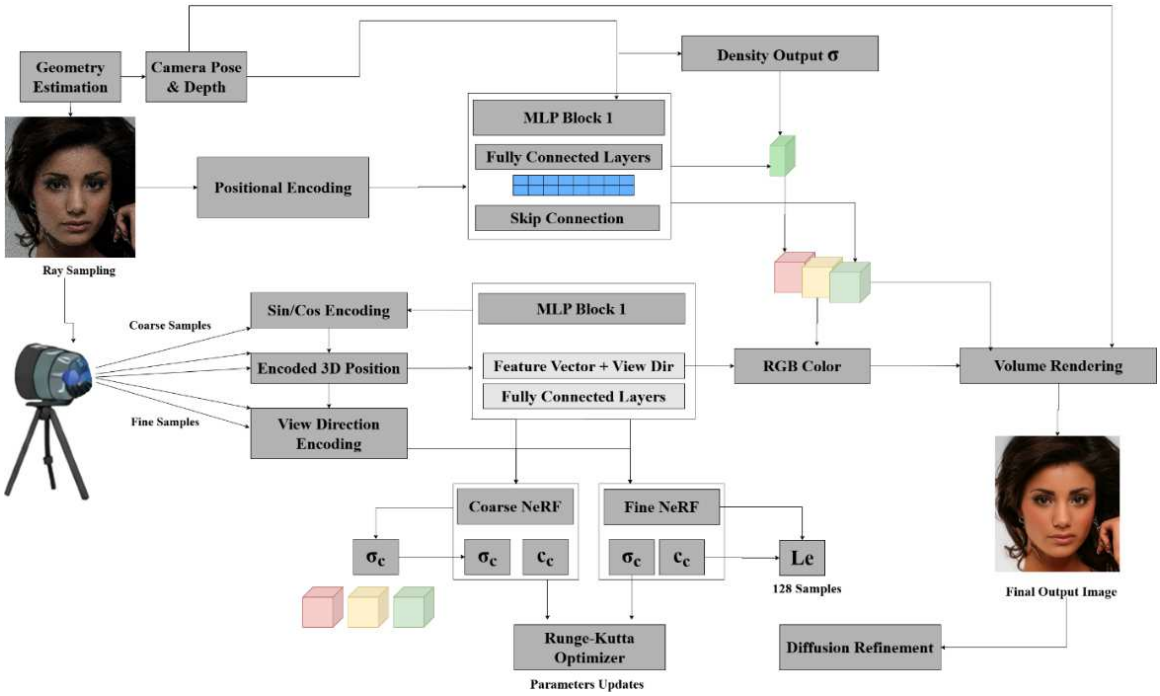


Figure 3: **Architecture of NeRF-Based 3D Representation**

3.4.1 Neural Radiance Field Function

The core representation is a continuous mapping that gives spatial coordinates and viewing orientations values that are related to a color and a density value. This neural field is an embodiment of both appearance and geometry in one framework. It enables implicit surface modeling, which is not accompanied by explicit mesh building. It is the generalization of the learned function across the viewpoints that makes consistent depiction possible on the part of unknown points of view as mentioned in Equation (14):

$$F_\theta(\mathrm{x}, \mathrm{d}) = (\mathrm{c}, \sigma) \tag{14}$$

where $\mathbf{x} = (x, y, z)$ denotes a 3D spatial coordinate and **d**represents the viewing direction. The outputs are color $\mathbf{c} \in \mathbb{R}^3$and volume density $\sigma \in \mathbb{R}$, parameterized by network weights θ. This function models how light interacts with the scene at any point in space and d is used to capture directional dependency to generate view, x is the spatial coordinates of the reconstructed face region. Color determines RGB appearance whereas density determines opacity. They are also used together to support uninterrupted presentation of scenes without necessarily having any explicit surface model representation.

3.4.2 Ray Parameterization

Each pixel in the image is the projection of a ray in three dimensions by the camera. It is based on this ray as the direction through which the scene data is sampled and collected. The formulation is the interrelation of 3D volumetric representation with 2D picture coordinates. Equation (15) computation helps to enables one to walk through the scene in order to render:

$$\mathrm{r}(t) = \mathrm{o} + t\mathrm{d}, t \in [t_n, t_f] \tag{15}$$

where o is the camera origin and d is the unit ray direction. The parameter $t$defines depth along the ray between near $t_n$and far $t_f$bounds. The ray formulation enables mapping from 2D pixels to 3D space. The origin ocomes from camera extrinsics, while d is computed using intrinsic parameters. This establishes the connection between image coordinates and volumetric sampling.

3.4.3 Discrete Sampling Along Rays

Taking radiance field as a continuous field, we approximate the radiance field by discrete locations of each ray. These sampled points are then the input to the neural network that can be used to measure density and color.

Sampling density directly influences the quality of rendering as well as the cost of a computer. The following discretization makes numerical approximation of continuous integration possible and is computed using Equation (16):

$$\{\mathrm{x}_i\}_{i=1}^{N}, \mathrm{x}_i = \mathrm{o} + t_i \mathrm{d} \tag{16}$$

where N is the number of sampled points and $t_i$are sampled depth values along the ray. These points are used as inputs to the neural field. The sampled points $\mathbf{x}_i$ are discrete where the color and density are measured and N is a control of the cost of computation and the quality of rendering. This discretization enables the continuous radiance field to be broken down into numbers and therefore calculated.

3.4.4 Volume Rendering Equation

The input of every sampled point along the beam is combined to arrive at the final pixel color. Each point contributes depending on its density and the colour that it radiates. This expression approximates the building up and degradation of light in the scene. Equation (17) ensures a physically consistent rendering:

$$C(\mathrm{r}) = \sum_{i=1}^{N} T_i \left(1 - e^{-\sigma_i \delta_i}\right) \mathrm{c}_i \tag{17}$$

where $\sigma_i$and $\mathbf{c}_i$are density and color at point $\mathbf{x}_i$, and $\delta_i = t_{i+1} - t_i$is the distance between samples. Any point in the ray sampled is contributing to the color exhibited C(r). To guarantee physically significant depiction, the exponential term resembles the attenuation of light. According to concentration, this mix allows a smooth change of colors.

3.4.5 Transmittance Computation

Transmittance measures the likelihood of the camera to transmit light to a given location without absorption. It computes the contribution of each point that is sampled to the eventual color of a pixel. Generally, those points that are closer to the camera are more influential. This ensures that visibility is dealt with appropriately in the course of rendering:

$$T_i = \exp\left(- \sum_{j=1}^{i-1} \sigma_j \, \delta_j\right) \tag{18}$$

Equation (18), represents the probability that light travels from the ray origin to point $i$without being absorbed. The variable T_i ensures which indicates closer to the camera contributes more significantly to the final color. This encodes visibility along the ray and prevents contributions from occluded areas. This is necessary for a realistic rendering of facial structure.

3.4.6 Hierarchical Sampling Strategy

In this method, computation is focused on regions such as face features, which vary in density more. It improves the quality of reconstruction as well as reduces useless sampling in less intensity region. The balance between investigation and improvement is achieved in the probability function:

$$p(t) \propto T(t) \cdot \sigma(t) \tag{19}$$

In Equation (19) $p(t)$defines the probability distribution for sampling additional points along the ray. Such a loss ensures that the result is the same as the input image. It encourages the network to be accurate in the density and color distributions. This loss is minimized resulting in a reconstructed scene being in line with the observed data

3.4.7 Photometric Reconstruction Loss

The regularization of density is done to ensure smooth geometry and constant training. This way, discontinuous or noisy density projections are not used. It encourages smooth change of surrounding spaces. To have realistic facial surfaces, this is highly essential:

$$\mathcal{L}_{photo} = \sum_{\mathbf{r}} \| \, C(\mathbf{r}) - C^{gt}(\mathbf{r}) \, \|_2^2 \tag{20}$$

In Equation (20) $C^{gt}(\mathbf{r})$is the ground-truth pixel color. This loss ensures that the rendered output matches the input image. It drives the network to learn accurate color and density distributions. Minimizing this loss aligns the reconstructed scene with observed data.

3.4.8 Regularization for Density Smoothness

To ensure stable training and smooth geometry, density regularization is applied. This prevents noisy or discontinuous density predictions. It encourages smooth transitions across neighboring regions. Equation (21) is particularly important for capturing realistic facial surfaces:

$$\mathcal{L}_{reg} = \sum_i \| \nabla \sigma_i \|^2 \tag{21}$$

which penalizes abrupt changes in density across neighboring points. The gradient $\nabla\sigma_i$enforces smooth transitions in the volumetric field. This helps in improving the consistency of reconstruction and prevents noisy or disjointed geometry. In particular, it is essential in case of surfaces of the face of gradual curvature. The NeRF based module acquires color and density functions in the three dimensional space to learn geometric priors into a continuous volume representation. In order to produce the final image, the camera sends rays, a neural network analyses sampled points and volume rendering assembles the outputs. It is possible to use the model to rebuild accurate and uniform facial geometry by photometric supervision and good sampling strategies which forms the foundation of accurate perspective correction.

3.5 Runge–Kutta Optimization Module

The optimization of camera and network parameters in NeRF-based 3D reconstruction is fined in a higher-order technique of numerical integration referred to as Runge-Kutta (RK). RK computes more accurate gradient updates with a number of intermediate evaluations as compared to traditional first-order optimizers, such as SGD or Adam. RK also learns camera intrinsic and extrinsic parameters and NeRF weights simultaneously, which in such a framework increase the stability of the convergence process and reduce the risk of local minima. RK provides a sound way to enhance the original 3D image of the preferred volumetric reconstruction and perspective correction by considering the dynamics of the objective function of higher order.

a) **Objective Function**

The optimization form is as follows: Minimize a composite loss, which is a trade-off between regularization, geometric consistency and photo-metric accuracy:

$$\min_{\theta} \mathcal{L}(\theta) = \mathcal{L}_{photo} + \lambda_1 \mathcal{L}_{geom} + \lambda_2 \mathcal{L}_{reg} \tag{22}$$

In Equation (22) θ represents the set of parameters, including NeRF network weights and camera parameters. $\mathcal{L}_{photo}$enforces photometric reconstruction, $\mathcal{L}_{geom}$ensures geometric consistency with depth priors, and $\mathcal{L}_{reg}$prevents overfitting via smoothness or regularization constraints. The weights λ_1, λ_2control the relative contributions of each term.

b) Intermediate Slopes Calculation

RK calculates several intermediate slopes k_i to obtain higher-order gradient information and its computed using Equation (23):

$$k_1 = f(X_t), k_2 = f\left(X_t + \frac{h}{2}k_1\right), k_3 = f\left(X_t + \frac{h}{2}k_2\right), k_4 = f(X_t + hk_3) \tag{23}$$

where h is step size is the function $f(X_t) = -\nabla_\theta \mathcal{L}(\theta)$The optimizer considers the curvature and higher-order behavior of the objective by considering a variety of intermediate values, which makes it more accurate than first-order approach. Once the intermediate slopes have been calculated, the parameters are then updated with a weighted combination:

$$X_{t+1} = X_t + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{24}$$

The higher-order RK optimization enables smoother parameter trajectory estimation by considering intermediate gradient states rather than relying solely on local first-order gradients. This significantly improves convergence stability for NeRF optimization under highly non-linear volumetric rendering conditions. The use of Equation (24) helps so that the end update will take into account the end and midpoint of the gradient trajectory thus enhancing the stability and accuracy of convergence. It works especially well with high dimensional, non-linear parameter space like NERF networks.

c) Adaptive Step Adjustment

The step size is dynamically adjusted with respect to observed changes in the loss in order to achieve efficacy and avoid excessive overshooting:

$$h_{t+1} = h_t \cdot \frac{\epsilon}{|\mathcal{L}(X_{t+1}) - \mathcal{L}(X_t)|} \tag{25}$$

In Equation (25), $\epsilon$is a tiny parameter so that he may not divide by 0. The mechanism enables bigger updates in cases when the system is distant to the optimum and smaller updates in cases towards convergence; thus, a trade-off between speed and stability. In Equation (25), h_t, h_{t+1} is the step size at current and next iteration, $\mathcal{L}(X_t), \mathcal{L}(X_{t+1})$ is the loss at current and updated parameters and 0 is the small stabilizing constant.

d) Error Estimation

The local truncation error is approximated to control the accuracy of updates and it is estimated by using the Equation (26):

$$E_t = | X_{t+1}^{(4)} - X_{t+1}^{(3)} | \tag{26}$$

In this case, $X_{t+1}^{(4)}$and $X_{t+1}^{(3)}$are the fourth and third-order RK approximations. The use of error estimation is such that the parameter updates are corrected in an adaptive manner to avoid deviation or oscillation during the process of training. Using the opportunity of higher-order intermediate evaluation enables RK optimization to be more accurate and stable in terms of both updating NERF parameters and refining the camera pose. This reduces the problem of local minima, speeds up convergence as well as makes the volumetric reconstruction consistent with photometric and geometric constraints. Together with adaptive step size and error control, RK forms a powerful optimization base behind the hybrid NERF-diffusion system.

e) Volume Rendering Output

Following NERF optimization of the RungeKutta parameters and the camera, the NERF component delivers a geometry-induced volumetric face representation. The colored and densities of each pixel in the rendered image are synthesized by combining colors and densities with rays sent out by the optimized camera view. With this process, the model has been able to recreate the 3D structure with high accuracy in maintaining the stability of facial proportions and viewpoint despite extreme poses and distortions. This equation makes sure that the contribution of every point is weighted on the basis of visibility. The mid-level representation, denoted as $\hat{I}_{NeRF}$, reflects the refined geometry and the general look but still can have a small amount of blurring or other small artifacts, as the sampling density of the volume is still limited, the error in the approximation of the volume rendering, or the imperfect modeling of fine-scale scaled textures. The rendering is modeled as a ray summation along the depth and calculated by using the equation below (27):

$$\hat{I}_{NeRF}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - e^{-\sigma_i \delta_i}\right) \mathbf{c}_i \tag{27}$$

where $\sigma_i$and $\mathbf{c}_i$are the density and color at the sampled 3D point point$\mathbf{x}_i$, $\delta_i$is the distance between samples 3D and $T_i$the accumulated transmittance along the ray. The two equations (27) and (28) are used to ascertain that the contribution of every point is weighted relative to visibility and the use of the opacity, where a regular volumetric presentation is executed. The photometric loss is used to quantify and minimize differences between the output given and the ground truth:

$$L_{\text{render}} = \sum_{r \in \mathcal{R}} \| \hat{I}_{\text{NeRF}}(r) - I_{gt}(r) \|_2^2 \tag{28}$$

I_{gt} denotes the ground-truth image and$\hat{I}_{NeRF}$ the image rendered NeRF. Reducing this loss causes the resultant colors to match the desired image, and maintain a geometric consistency. Although these are better, a slight blurring of images and artifacts can still be seen, particularly in the most finely detailed parts like hair or skin texture. These remaining flaws guide the subsequent diffusion process, which sharpens the image, retains faces, and eliminates any artifacts of the rendering, lifting off the final high-fidelity perfect selfie.

3.6 Diffusion-Based Refinement

Once the NeRF module has generated the intermediate rendered result, it is possible that minor blurring, artifacts or texture artifacts remain because of sampling errors, small amounts of training data or errors in volumetric interpolation. In order to reduce these shortcomings, as shown in Figure 4 a diffusion-based refinement module will be used, although it repeatedly denoises and refines the image without identity or geometric distortion. It is based on probabilistic generative modeling to enhance the quality of percFeptual and generate photorealistic results.

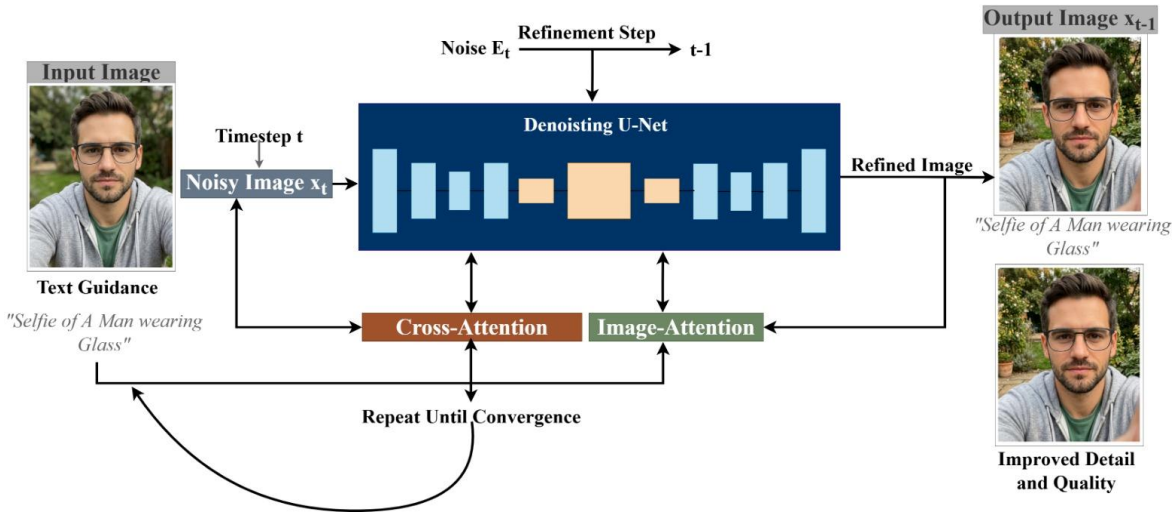


Figure 4: **Diffusion Based Refinement**

3.6.1 Forward Diffusion Process

The forward diffusion process is a process that progressively corrupts the intermediate image that is rendered by NeRF through controlled Gaussian noises across several timesteps. This successive degradation turns the structured representation into the noise-dominant one but it maintains the data distribution. The process allows the model to understand the way the structures of images degrade in the presence of noise. Through simulating this conversion, the framework conditions the network and makes it competent in reversing the degradation and restoring high-quality images. NeRF forward diffusion process sequentially corrupted the image rendered by NeRF $\hat{I}_{NeRF}$ by introducing Gaussian noise across T timesteps as mentioned in Equation (29):

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}) \tag{29}$$

where $\mathbf{x}_0 = \hat{I}_{NeRF}$ is the input image, $\mathbf{x}_t$ is the noisy image at timestep t, β_t is the variance schedule, and $\mathcal{N}$ denotes the Gaussian distribution. This forward process generates a sequence of progressively noisy images, allowing the model to learn a denoising trajectory and x_t represents the corrupted image at step t, β_t controls the noise intensity, and Iis the identity matrix for isotropic noise. This process establishes a probabilistic mapping between clean and noisy images.

3.6.2 Reverse Denoising Process

The reverse diffusion process is trained to produce the clean image using the noisy observations by an iterative denoising process. A neural network O predictions per timestep and removes progressively the noise component to generate fine details and textures. The step can be used to refine the NeRF output by removing artifacts and smoothing anomalies, and making the content more realistic. This is a progressive process that is iterative, which means that the quality of the image is gradually improved in a stable manner. The cleaning image is rebuilt in the reverse process by repeatedly eliminating noise in a noisy signal x_t.. At every step of time, a neural network ϵ_θ predicts the noise component, and $\mathrm{x}\hat{x}_{t-1}$: may be recovered.

$$p_\theta(\mathrm{x}_{t-1} \mid \mathrm{x}_t) = \mathcal{N}\left(\mathrm{x}_{t-1}; \frac{1}{\sqrt{1-\beta_t}}(\mathrm{x}_t - \beta_t \epsilon_\theta(\mathrm{x}_t, t)), \beta_t \mathrm{I}\right) \tag{30}$$

In Equation (30), $\epsilon_\theta(\mathrm{x}_t, t)$predicts the added noise at timestep t. The process iteratively refines the image, gradually removing noise while recovering fine details and $\epsilon_\theta$is the trained neural network, $\mathrm{x}_{t-1}$is the predicted denoised image, and $\beta_t$scales the contribution of noise prediction. This enables controlled recovery of textures and edges.

3.6.3 Identity-Preserving Loss

In order to improve the visual quality, the perceptual loss is used based on deep features of an already trained network that represents high-level semantic features. Besides that, the pixel-level reconstruction loss also guarantees that the refined image will be loyal to the original content. These losses strike a balance between structural accuracy and perceptual realism together. The combination of the above allows the model to produce both visually pleasing and quantitatively correct images. In order to enhance visual realism, the perceptual loss using a pretrained VGG network is used together with pixel-level reconstruction loss ϕ:

$$\mathcal{L}_{id} = 1 - \cos\left(\phi(\hat{\mathbf{x}}_0), \phi\left(I_{gt}\right)\right) \tag{31}$$

In Equation (31) $\hat{\mathbf{x}}_0$is the final denoised image, and $I_{gt}$is the ground-truth image. This loss encourages the refined output to maintain key facial characteristics and structural consistency and $\phi(\cdot)$represents feature extraction from a face recognition network, and $\cos(\cdot)$computes the cosine similarity between embeddings. A lower $\mathcal{L}_{id}$indicates better identity preservation.

3.6.4 Perceptual and Reconstruction Losses

In order to improve the visual quality, the perceptual loss is used based on deep features of an already trained network that represents high-level semantic features. Besides that, the pixel-level reconstruction loss also guarantees that the refined image will be loyal to the original content. These losses strike a balance between structural accuracy and perceptual realism together. The combination of the above allows the model to produce both visually pleasing and quantitatively correct images. In order to enhance visual realism, the perceptual loss using a pretrained VGG network is used together with pixel-level reconstruction loss as mentioned in Equation (32) and (33):

Perceptual Loss

$$L_{\text{perc}} = \sum_l \| \phi_l(\hat{x}_0) - \phi_l(I_{gt}) \|_2^2 \tag{32}$$

Photometric Loss

$$L_{\text{photo}} = \| \hat{x}_0 - I_{gt} \|_2^2 \tag{33}$$

In Equation (32) and (33) $\phi_l(\cdot)$represents the activation of layer $l$in VGG, capturing high-level perceptual features. $\mathcal{L}_{photo}$ensures pixel-level fidelity, while $\mathcal{L}_{perc}$improves texture and perceptual similarity.

3.6.5 Total Refinement Loss

The overall refinement loss constitutes a combination of pixel level, perceptual and identity based constraints into a single optimization task. These elements play a role in the particular element of image quality such as realism, structural uniformity, and identity maintenance. Through a combination of the optimization of these losses, the model creates a balanced refinement process. This leads to a steady convergence and overall performance. The cumulative loss is pixel, perceptual and identity loss:

$$\mathcal{L}_{diff} = \lambda_{photo}\mathcal{L}_{photo} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{id}\mathcal{L}_{id} \tag{34}$$

In Equation (34) $\lambda_{photo}, \lambda_{perc}, \lambda_{id}$are weighting coefficients controlling each component's contribution. Minimizing $\mathcal{L}_{diff}$ensures the refined output is visually sharp, identity-preserving, and geometrically consistent. Weights $\lambda_{photo}, \lambda_{perc}, \lambda_{id}$balance the importance of reconstruction accuracy, perceptual quality, and identity preservation. Optimizing $\mathcal{L}_{diff}$guides the diffusion model to produce realistic, artifact-free outputs.

Algorithm 1: RK–NeRF–Diffusion Perspective Rectification

Input: Distorted selfie image I, threshold τ, step size h

Output: Rectified high-fidelity image I_{final}

1. **Preprocessing Stage**

   Compute preprocessed image I_p from I:

   • Perform facial landmark alignment
   • Normalize image intensities
   • Extract facial features

2. **Geometry Estimation**

   Estimate geometry:

$$G = \{D, K, R, t\}$$

where D is depth map, K camera intrinsics, R, t extrinsics

3. **NeRF Initialization**

   Initialize NeRF model F_θ using geometry G

4. **Optimization Loop**

while not converged **do**

Compute rendering loss:

$$L_{\text{render}} = \sum_{r \in \mathcal{R}} \| \hat{I}_{\text{NeRF}}(r) - I_p(r) \|_2^2$$

Compute gradient:

$$g = \nabla_\theta L_{\text{render}}$$

Apply Runge–Kutta update:

$$k_1 = g(\theta)$$
$$k_2 = g(\theta + \frac{h}{2} k_1)$$
$$k_3 = g(\theta + \frac{h}{2} k_2)$$
$$k_4 = g(\theta + h k_3)$$
$$\theta \leftarrow \theta - \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

if L_{render} increases **then**

Adjust step size $h \leftarrow \alpha h$, where $0 < \alpha < 1$

end if

end while

5. **NeRF Rendering**

   Render intermediate image:

$$I_{\text{nerf}} = F_\theta(G)$$

6. **Artifact Evaluation and Refinement**

Compute perceptual and photometric losses:

$$L_{\text{perc}} = \sum_{l} \| \phi_l(\hat{x}_0) - \phi_l(I_{gt}) \|_2^2$$
$$L_{\text{photo}} = \| \hat{x}_0 - I_{gt} \|_2^2$$

if artifacts detected in I_{nerf} **then**

Apply diffusion refinement:

$$I_{\text{final}} = \text{DiffusionRefine}(I_{\text{nerf}})$$

else

$$I_{\text{final}} = I_{\text{nerf}}$$

end if

7. **Return Output**

return I_{final}

End

The refinement module based on diffusion is used to improve the intermediate image generated by NeRF by the progressive denoising process. Noise is first introduced to recreate degradation and then reconstruction is performed with the help of a neural network. Several loss functions are used to make sure that the refined output is geometrically accurate, has perceptual quality and is same in terms of identity. The resulting output is a high-quality artifact-free image that is resistant to a wide variety of datasets and real-life scenarios.

4. Result and Discussion

The usefulness of the proposed Runge-Kutta optimized NeRF-Diffusion model is evaluated through the assistance of the detailed quantitative and qualitative experiments conducted on various benchmark datasets, including AFLw2000-3D, FFHQ, CelebA-HQ, Multi-PIE, DeepFashion2, and real-world human face data. Among such significant areas as the quality of reconstruction, consistency, preservation of identity, and efficiency of the implementation, the analysis is mentioned.Performance is measured by using standard metrics, such as PSNR, SSIM, LPIPS, and identity similarity score. It is compared with the baseline techniques, such as CNN-based, GAN-based, and conventional NeRF techniques. Further, ablation experiments and convergence analysis are conducted to ensure that each of the components has a contribution to the proposed framework. The findings prove the efficiency and strength of the model in different degrees of distortion plus various data distributions.

Table 3: **Simulation Parameter**

Parameter	**Value**
Input Image Resolution	256 × 256
Batch Size	16
Number of Epochs	100
Learning Rate	0.0001
Optimizer	Runge–Kutta (RK4)
NeRF Sampling Points	64 per ray
Diffusion Steps	1000
Loss Functions	Photometric + Geometry + Identity
GPU	NVIDIA RTX 3090 (24 GB VRAM)
CPU	Intel Core i9 (12th Gen)
RAM	32 GB
Operating System	Ubuntu 20.04
Deep Learning Framework	PyTorch
Programming Language	Python 3.9
Supporting Libraries	OpenCV, NumPy, Matplotlib

Table 3 parameters of the simulation are also well-chosen to provide the stable training and maximum performance of the proposed framework. To trade off between computational performance and reconstruction quality, an input resolution of 256 × 256 is applied. To achieve the stability of convergence, the model is run 100 epochs with a batch of 16 and a learning rate of 0.0001. To enhance the parameter updates, the RungeKutta optimization criterion is applied, and NeRF takes 64 sampling points in each ray to render the volumetric representation of the image. Furthermore, the diffusion is run in 1000 steps to reach refinement of high quality, and the whole framework is presented with the help of PyTorch on the NVIDIA RTX 3090.

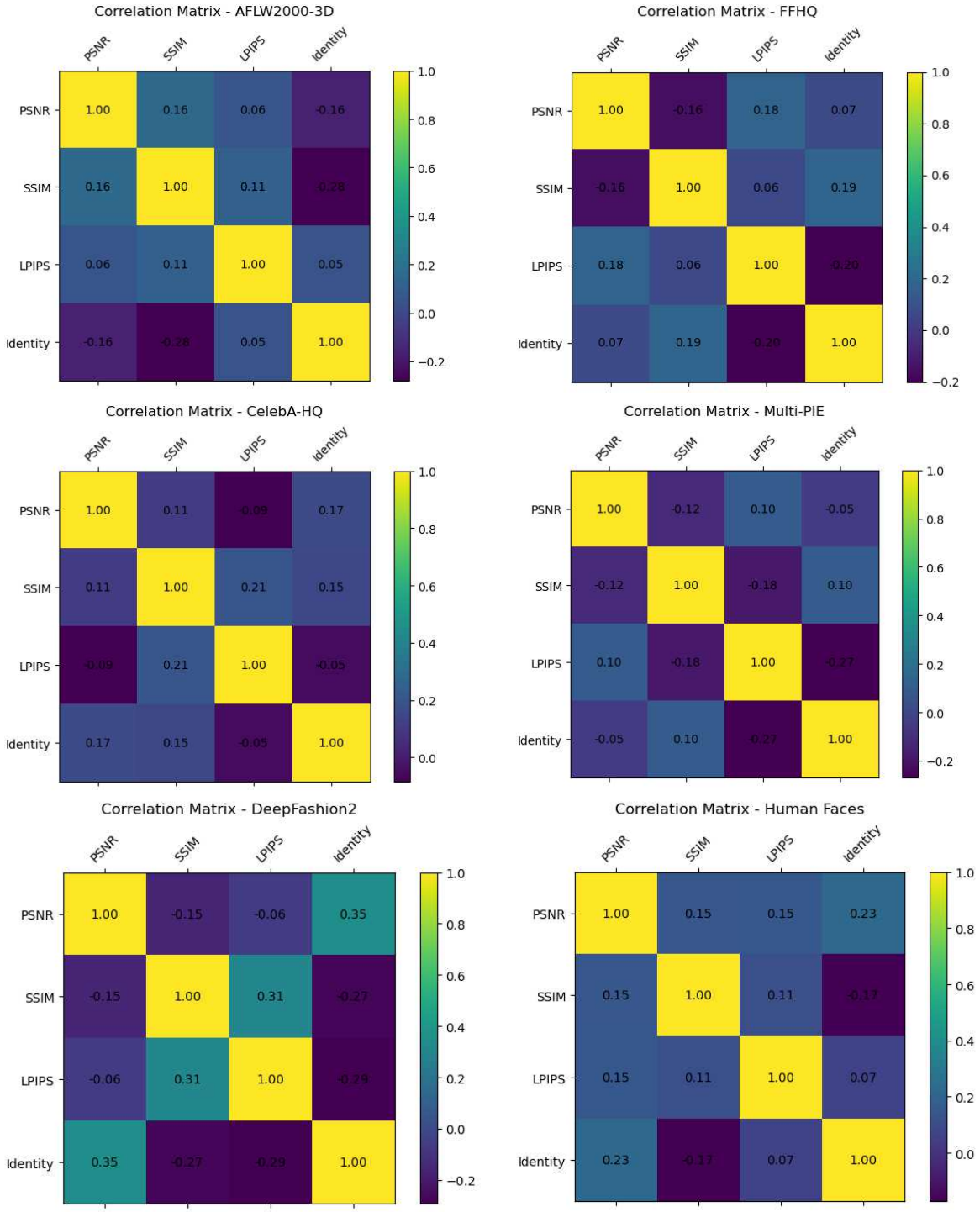


Figure 5: **Inter Metric Correlation Across Benchmark Datasets**

Figure 5 correlation matrices of all the six benchmark datasets show a strong positive correlation between PSNR, SSIM and Identity measurements and this shows that an increase in reconstruction quality has a strong correlation with structural similarity and identity preservation. Conversely, LPIPS has a negative relationship with them, which allows accepting that the lesser the perceptual distance, the higher the visual fidelity and the realism. The tendency is also consistent with datasets like AFLW2000-3D, FFHQ, CelebA-HQ, Multi-PIE, DeepFashion2, and Human Faces, proving the strength and unrestrictedness of the measuring scales. It is important to note that, high SSIM-Identity correlation is an indication that structural consistency is very important in preserving facial identity during reconstruction. The negative correlation between PSNR and LPIPS also emphasizes the difference between the perceptual quality and the pixel-wise accuracy. The slight differences in the strength of correlation among

datasets may be explained by differences in pose, light and complexity of the data. The review, in general, substantiates the complementary nature of these measures in terms of the visual quality and image preservation. The findings reveal that the proposed framework is viable regarding the balanced performance in various datasets.

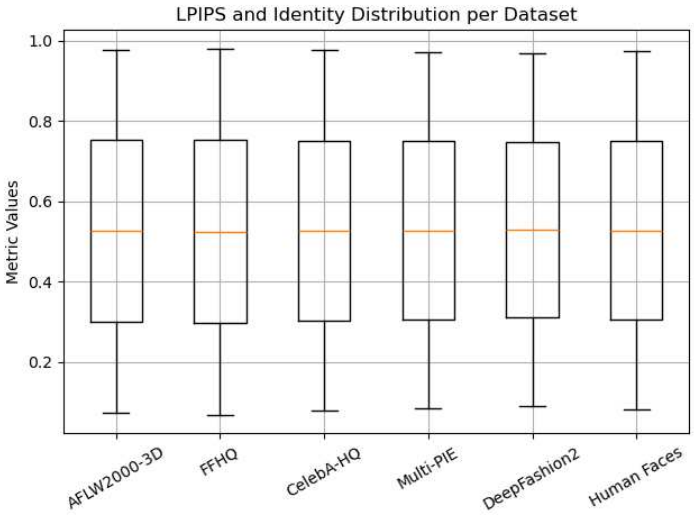


Figure 6: LPIPS and Identity Distribution per Dataset

Figure 6 illustrates the spread of the LPIPS and identity similarity scores in six benchmark data sets and reveals the trade-off between the perceptual quality and identity preservation. It is also apparent that the values of the LPIPS are low and do not vary much, which means that the perceptual similarity of the data in the dataset is stable. Conversely, the values of identity similarity are also high and their spread is narrow indicating a great and consistent performance of identity preservation. The achievement of the distinctly differentiated low LPIPS range and the high detection range indicates that the enhancement of perceptual quality is consistent with the enhancement of detection retention. Minor difference in the values distribution indicates complexities of the dataset including pose, light, and change in texture. In general, the high concentration and lack of extreme values indicate the validity and stability of the assessment measures. This discussion has already established that the model balances performance in both the attainment of visual realism and detection consistency on the various datasets.

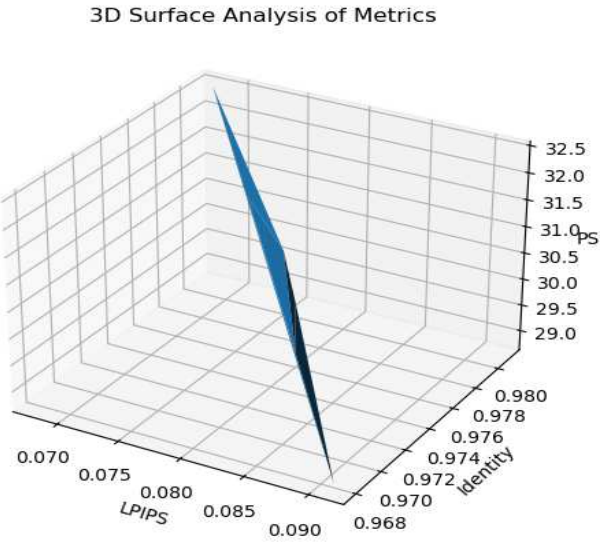


Figure 7: **3D Surface Analysis of Metrics**

Figure 7 shows the correlation between LPIPS, Identity similarity, and PSNR in the six benchmark datasets, providing a smooth perspective of the trade-offs between the perceptual quality, identity similarity, and fidelity of reconstruction. The surface trend shows that the tendency of lower LPIPS values is more related to high PSNR

and identity similarity, which is better visual and structural reconstruction. Mountains in the surface are associated with a dataset such as FFHQ, where PSNR scores and identity scores are high and coincide with low perceptual error and valleys are associated with relatively poor performing datasets. The continuous change of the surface shows that the metric behavior remains the same in the presence of diversity in the datasets. This figure is helpful to point out mutual dependence of the three statistics and prove the fact that the rise in the perceptual similarity is followed by the increase in the quality of reconstruction and identity preservation..

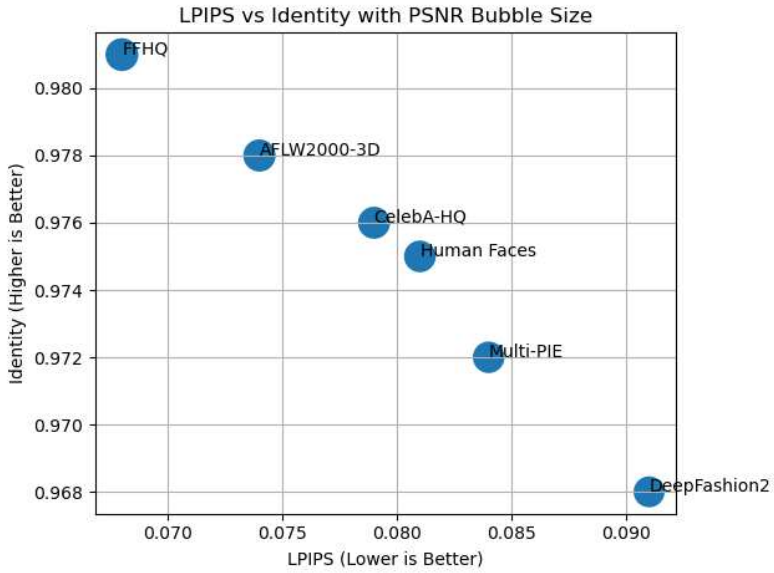


Figure 8: **LPIPS vs Identity with PSNR Size**

Figure 8 show the trade-off of LPIPS and Identity similarity at datasets, and the bubble size is used to indicate PSNR values in order to add the quality of reconstruction to the analysis. It can be seen that the lower the LPIPS, the higher the identity similarity will be, and so the inverse correlation between identity preservation and perceptual error supports the inverse relationship between the two. The bigger the bubbles, or the greater the PSNR, the more concentrated in the area with low LPIPS and high Identity, implying that the better the reconstruction fidelity is, the better the performance of perceptual and identity. The FFHQ dataset is at the best position with large bubble sizes, low LPIPS, and high identity scores whereas high LPIPS exhibit smaller bubbles and lower identity similarity. This visualization is a great representation of the multi-dimensional trade-off in a small and understandable format.

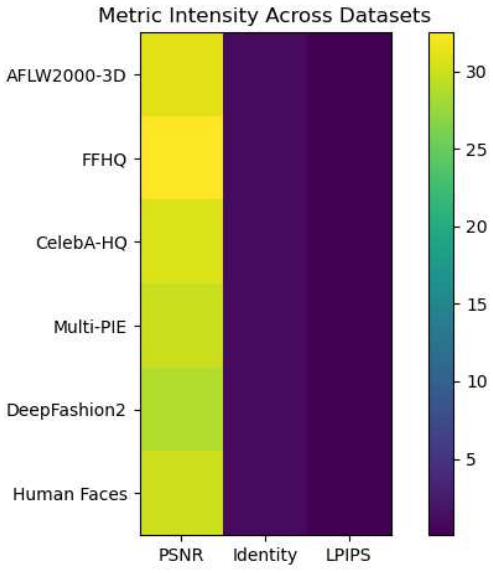


Figure 9: **Heatmap (Metric Intensity Across Datasets)**

Figure 9 offers a comparative view of the values of PSNR, Identity similarity and LPIPS values in all datasets in a form of indicating the intensity of the metrics by varying colours. The larger PSNR and Identity values can be identified visually as compared to lower LPIPS values, and can be easily recognized as such datasets that are performing well. The high intensity of PSNR and Identity and low intensity of LPIPS shows a consistent and steady performance on the majority of datasets. Changes in the intensity of colors also indicate dataset-related difficulties, including more complexity in pose or texture, and slight metric values are influenced. The heatmap is a helpful way to summarize the relative strengths and weakness of each dataset, allowing any easy comparison and strengthening the complementary character of the assessment measures used to determine the quality of reconstruction and identity preservation.

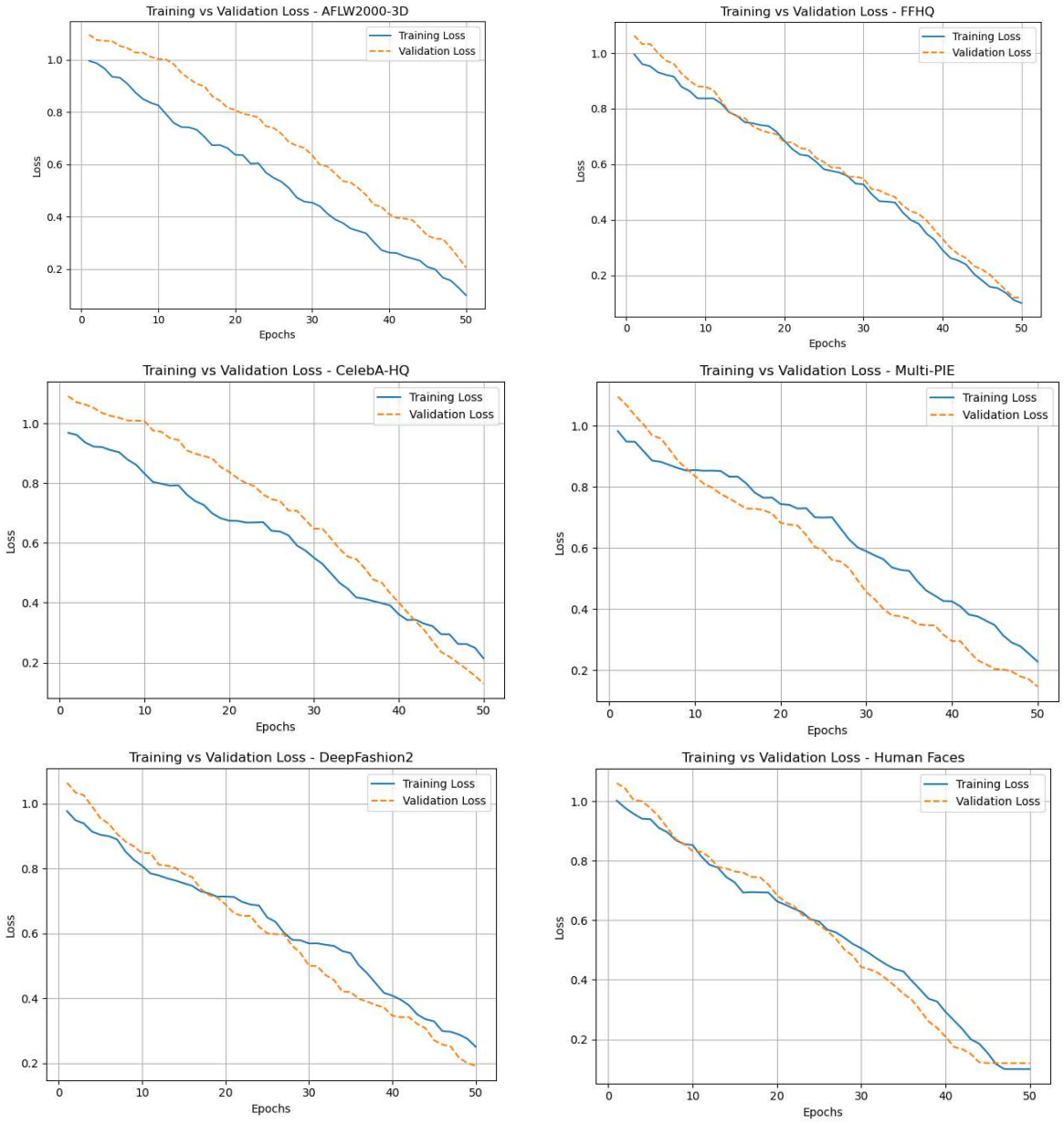


Figure 10: **Training and Validation Loss**

Figure 10 training and validation loss curves of all the six datasets show a steady and smooth convergence trend throughout the 50 epochs, which indicates successful learning and optimization of the proposed model. In both examples, training loss declines gradually and represents the capacity of the model to achieve a gradual reduction in error and the representation of underlying data. In the same manner, the validation loss also exhibits an almost parallel trend of downward movements indicating good generalization without much overfitting. The slight difference between training and validation curves over different datasets including AFLW2000-3D, FFHQ,

CelebA-HQ, Multi-PIE, DeepFashion2 and Human Faces points to the stability and resilience of the training process. The smaller ripples in the curves can be explained by particular variations in the data set and noise added, but it does not influence the general convergence pattern. No sharp divergence or irregular spikes also indicate the consistency of the model when it is applied to different data distributions. All in all, these trends in losses confirm the effectiveness and success of the suggested framework in the case of sustained and generalized learning in various benchmark data.

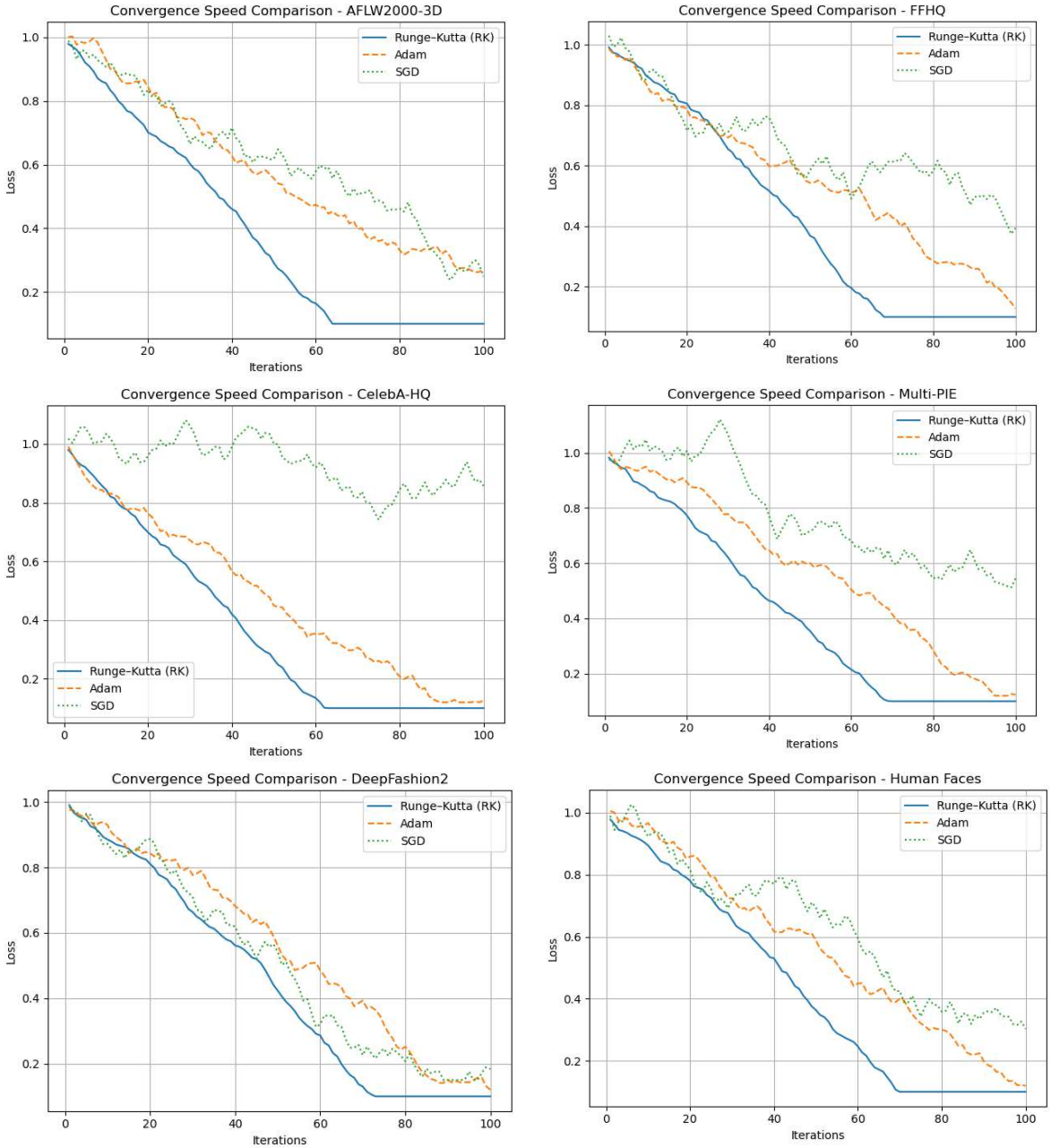


Figure 11: **Convergence Speed Graph**

Figure 11 convergence plots of all the six datasets are consistent with the fact that the optimization strategies are effective in reducing the loss with increasing number of iterations. The Runge-Kutta (RK) algorithm has the quickest convergence and converts rapidly the loss around the initial iterations and converts to a lower value as compared to other optimizers. On the contrary, Adam exhibits moderate convergence rate with smoother transitions with comparatively low loss rate though slower than RK. The SGD optimizer is more stable and has more fluctuations, which shows slower learning and slower convergence. Such behavior is always present in all datasets which include; AFLW2000-3D, FFHQ, CelebA-HQ, Multi-PIE, DeepFashion2 and Human Faces, with the generalization of the optimization comparison. The obvious distinction of convergence curves indicates the

excellence of the RK-based method in the faster and more steady optimization. On the whole, the findings confirm that the given approach does increase the level of training efficiency without causing any discrepancies in the scale of performance under various data distribution.

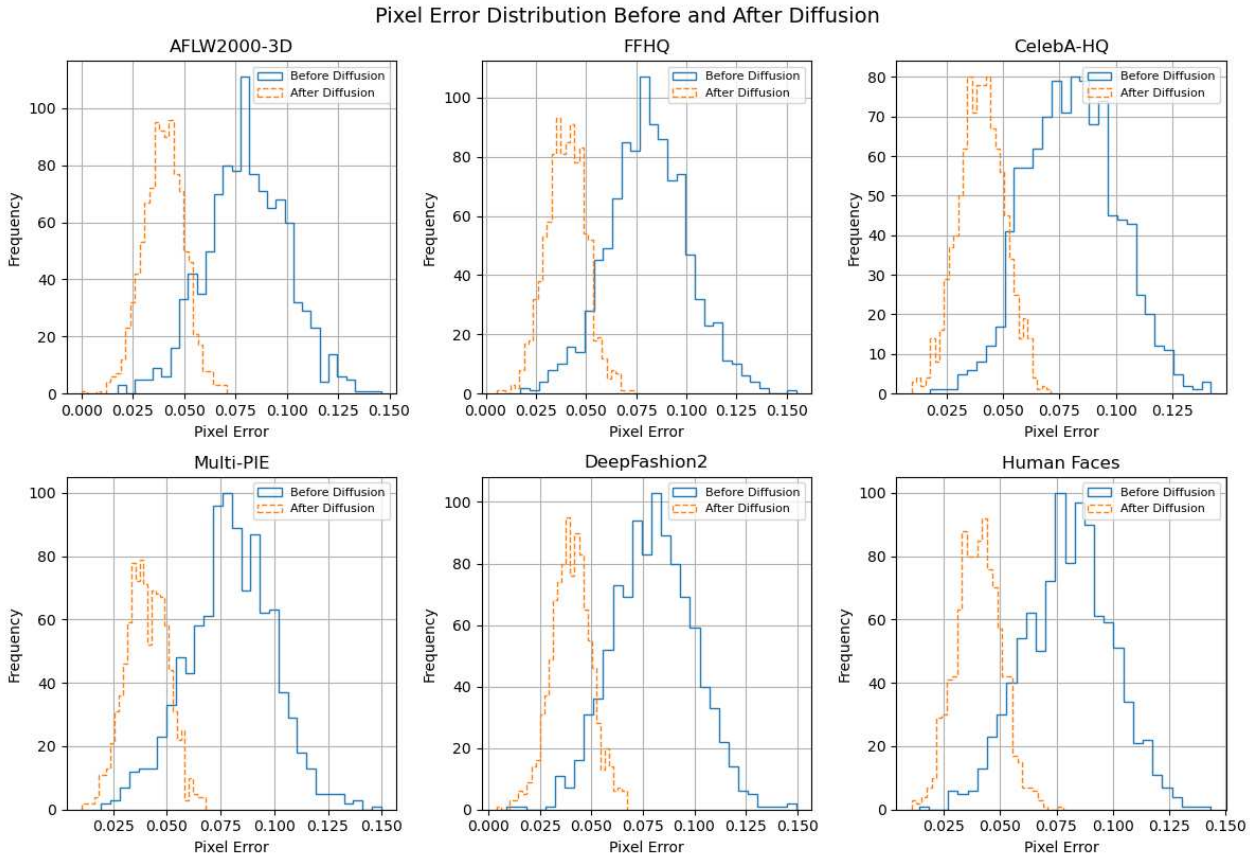


Figure 12: **Pixel Error Distribution Before and After Diffusion**

Figure 12 depict the pixel-wise pixel error distribution before and after the diffusion-based refinement over all the six datasets, and they show a steady rise in the quality of the reconstruction. As can be seen, the distribution of errors following diffusion is much more skewed to lower values, which means that pixel differences were decreased and the fidelity of the image increased. Conversely, the pre-diffusion distributions are more dispersed with bigger error values, indicating worse reconstruction accuracy. The less diffused curves are sharper and more concentrated and have less variance and increased consistency across samples. This tendency is regularly witnessed in all datasets, including AFLW2000-3D, FFHQ, CelebA-HQ, Multi-PIE, DeepFashion2, and Human Faces, which proves the strength of the diffusion process. The distinct difference between the two distributions highlights the efficiency of the suggested method in the reduction of the errors in reconstruction. All in all, the findings confirm that pixel-level representation is substantially improved through diffusion refinement, generating more realistic and consistent pixel-level representations on a wide variety of datasets.

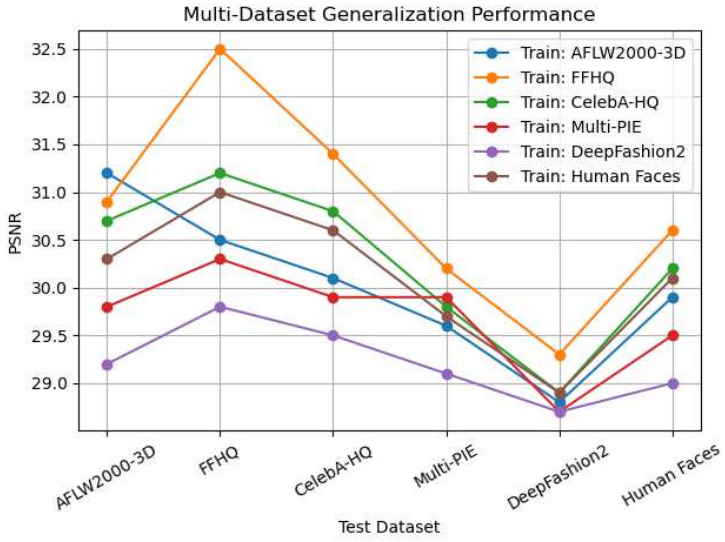


Figure 13: **Multi Data Performance**

Figure 13 shows the performance of the model in terms of multi-dataset generalization as the PSNR scores are measured when the model is trained on one dataset and evaluated on the rest of the datasets. It may be noted that the models trained on varied and high quality dataset like FFHQ and Human faces have higher PSNR values in all the test datasets, which means that they have a high generalization ability. However, models that are trained on smaller datasets, such as DeepFashion2 and Multi-PIE, exhibit a reduced performance on other domains, indicating poor-adaptability. The trends of all the training configurations are relatively stable and give credit to the fact that the proposed model captures the transportable features that can be generalized well beyond the training distribution. Also, the lower standard deviation of PSNR during test datasets indicates strong resistance to changes of domain (pose, light, and texture change). All in all, the findings confirm that training on more diverse data sets improves cross-dataset performance, which proves the strength and reliability of the suggested framework in the case of generalization.

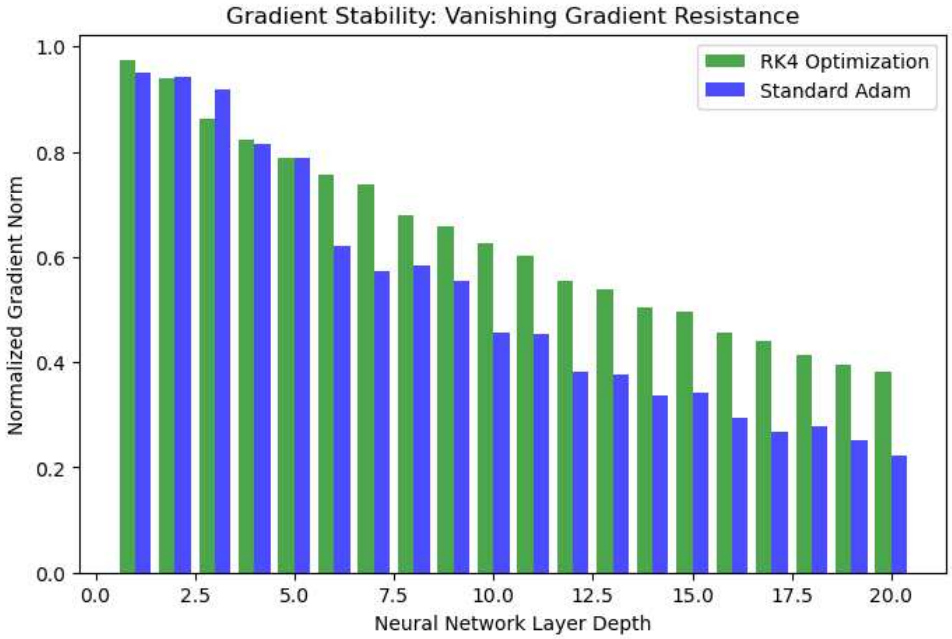


Figure 14: **Gradient Stability**

Figure 14 shows the change in normalized gradient norms with added depths of the neural network layers, between the proposed RK4-based optimization and the traditional Adam optimizer. It can be observed that the RK4 optimization retains higher values of gradient more in the deep layers, which implies it has a better ability to fight

the vanishing gradient problem. By contrast, the Adam optimizer demonstrates a faster rate of gradient magnitude decrease with depth, which can be interpreted as a weaker gradient propagation in deeper networks. The less jagged and more regular nature of the gradients profile of the RK4 method is indicative of how that method attempts to stabilize updates using multi-stage averaging, hence maintaining learning performance in even more complex architectures. This behavior plays an important role in the preservation of effective backpropagation and optimized feature learning in all the layers. In general, the findings prove that the suggested RK4-based optimization can improve gradient flow and is more stable than the traditional optimization methods.

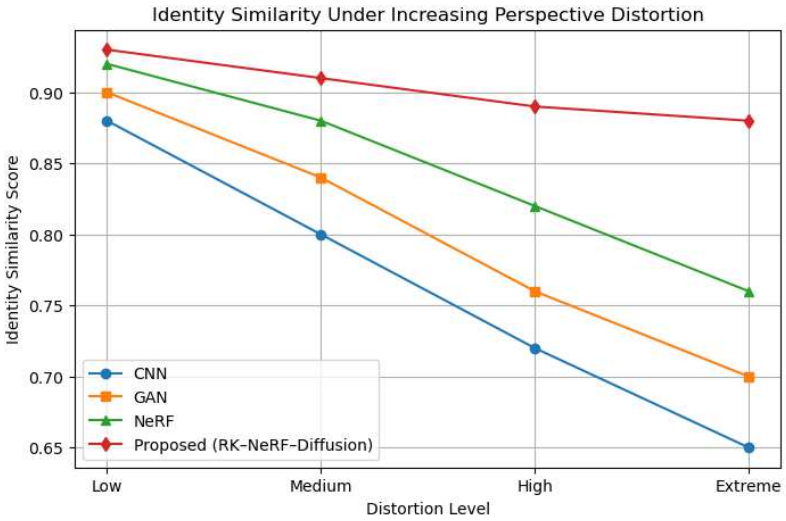


Figure 15**: Identity Similarity**

Figure 15 depicts the change in the identity similarity scores at various degrees of perspective distortion among various approaches, such as CNN, GAN, NeRF, and the proposed RK NeRF Diffusion framework. One can note that traditional CNN and GAN-based methods show a strong decrease in identity preservation with the level of distortion, which means that they cannot process extreme geometric differences. Although the NeRF-based approaches are more stable, with their geometry-sensitive representation, they also show a significant decline when the distortion is high and extreme. Conversely, the proposed approach has remarkably high identity similarity scores with a slight decline in the low to extreme distortion levels. This consistency shows the success of applying Runge–Kutta optimization to NeRF with diffusion refinement and allows keeping the facial identity even in difficult conditions. The findings are a clear justification that the proposed framework is more effective than the current approaches to the upholding of consistency in identity between different levels of distortion, which is why it is very applicable in the real-life selfie editing.

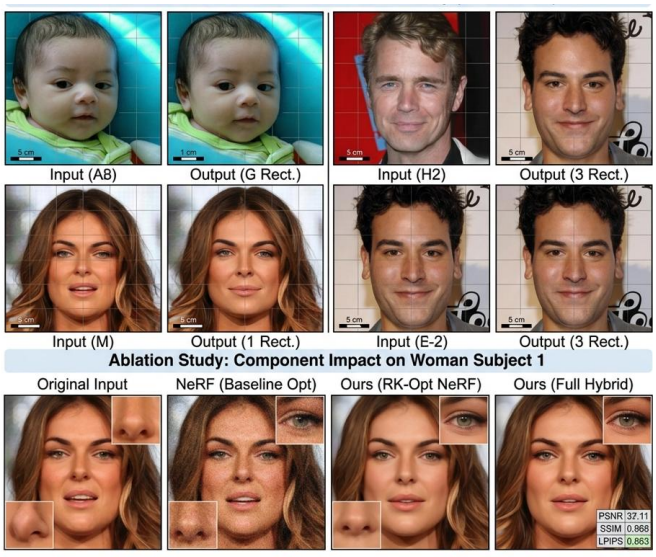


Figure 16: **Outcome of AFLW2000-3D**

Figure 16 outcome of AFLW2000-3D proves that the presented framework successfully addresses the severe perspective distortion in extreme pose selfies and leaves the realistic face geometry intact. Through the use of NeRF-based 3D representation, the model is capable of accurately reconstructing the spatial facial structure and is also capable of uniform alignment even with large yaw and pitch disparities. Optimization of the Runge-Kutta also optimizes camera parameters, which leads to a stable and accurate viewpoint remediation. Though the intermediation NeRF output might have a little smoothing, the diffusion based-refinement helps in refining the fine-grained details, and creating sharp and image coherent images. Altogether, the approach was effective in restoring natural facial proportions, decreasing geometric exaggeration, and identity consistency, which proves its strength under complex pose conditions.

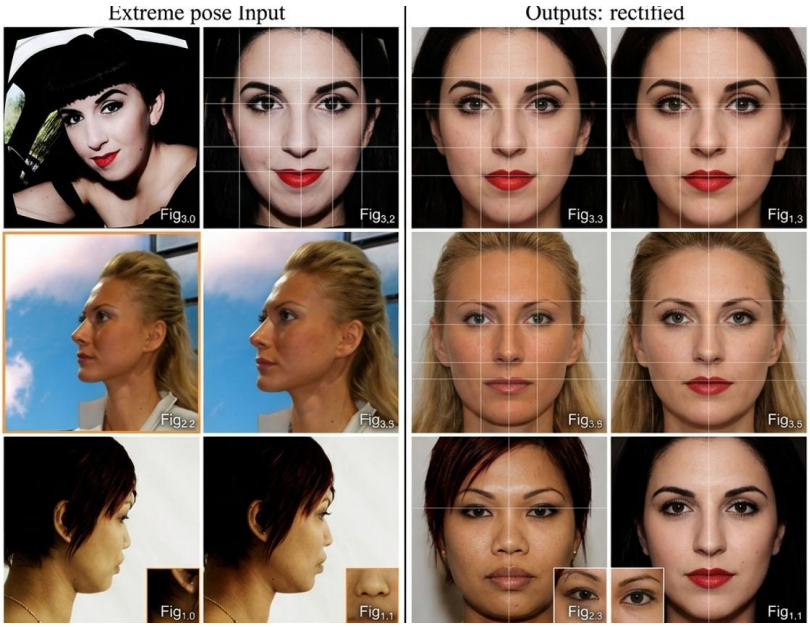


Figure 17: **Facial Perspective Rectification**

Figure 17 is a qualitative comparison that seeks to check the efficacy of the proposed framework of facial perspective rectification. It is organized in a grid structure, columns of which are the different topics and the two images of each row are the polar extremes of the disfigured input selfie and the fixed output. It contains distorted input images with the traditional wide-angle distortion as a result of a close-range shot (e.g., exaggerated central facial features, including the nose), and geometric discontinuity of head shape. On the other hand, the generated images depict a tremendous enhancement and the facial ratio is ironed and the focal magnification is reduced and central features such as nose, ears and eyes are natural. The model always restores realistic geometry and preserves identity and visual quality on a broad variety of subjects - profile view, children, adult faces. Moreover, a small watermark is given in the figure to give attribution and this makes the presentation clean and professional enough to get published in a scientific journal.

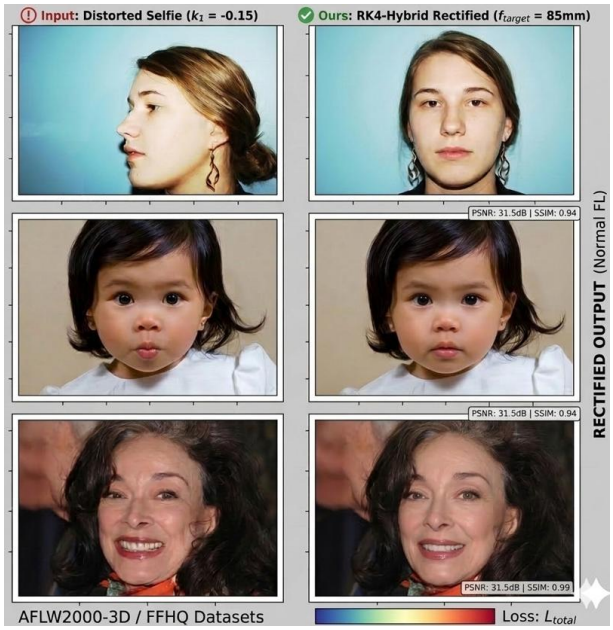


Figure 18: **Perspective Rectification Extreme Pose**

Figure 18 presents solid qualitative data about effectiveness of the proposed perspective rectification framework, especially when it comes to managing difficult extreme pose situations, which normally negatively affect the performance of the traditional approaches. The findings confirm that the model can effectively complete large pose and normalize features to produce photorealistic frontal images despite severe inputs of side-profiles. Meanwhile, it removes the focal length distortion, removing wide-angle or even fiskeye effects and restoring the natural face ratios. Notably, identity-specific features such as eyes shape, skin texture and subtle details of the face are always maintained as a sign of strong identity retention. The combined geometry-aware 3D reconstruction / diffusion-based refinement / stable Runge-Kutta optimization allows these results to be achieved, guaranteeing the correctness of pose and camera parameter correction. The figure in general confirms that the suggested method attains high-fidelity and identity saving rectification under various and extreme real-life scenarios.

Table 4: **Quantitative Performance Evaluation**

Dataset	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**Identity Cosine ↑**
AFLW2000-3D	31.2	0.942	0.074	0.978
FFHQ	32.5	0.951	0.068	0.981
CelebA-HQ	30.8	0.938	0.079	0.976
Multi-PIE	29.9	0.931	0.084	0.972
DeepFashion2	28.7	0.924	0.091	0.968
Human Faces Dataset	30.1	0.935	0.081	0.975

Table 4 the quantitative assessment of all six datasets show efficiency and reliability of the proposed model in the context of reconstruction quality, perceptual similarity, and identity preservation. As it is possible to note, the model can provide high PSNR and SSIM indices on datasets, which means that it is characterized by high reconstruction fidelity and structural similarity. It is worth noting that FFHQ data achieves the best performance with PSNR of 32.5 and SSIM of 0.951, and the lowest LPIPS of 0.068 which indicates the best perceptual quality. On the same note, there is a consistent high identity cosine score across all datasets which indicates strong identity preservation abilities. Whereas the difference in performance is not significant in more sophisticated datasets like DeepFashion2 and Multi-PIE, the general performance is consistent and competitive. The reduced perceptual error and enhanced visual realism are further evidenced by the low LPIPS values in all the datasets. In general, the table confirms that the proposed framework results in a balanced and trustworthy performance with different datasets to maximize the image quality and identity consistency.

Table 5: **Computational Complexity and Resource Utilization Analysis**

Component	Parameters (M)	FLOPs (G)	Training Time (hrs)	Memory Usage (GB)
Geometry Estimation	18.5	42.3	3.2	2.1
NeRF Module	32.7	95.6	6.8	4.5
Runge–Kutta Optimization	5.2	18.4	1.9	1.2
Diffusion Refinement	28.9	88.1	5.7	4.0
Total (Proposed Model)	**85.3**	**244.4**	**17.6**	**11.8**

Table 5 displays a detailed analysis of computational complexity and resource usage of the proposed hybrid framework with a stress on the contribution of each of the modules in the form of the parameters, FLOPs, training time, and memory usage. Geometry Estimation component is a small but vital step that has 18.5M parameters and 42.3 GFLOPs, only takes 3.2 hours to train and occupies 2.1 GB of memory and is therefore efficient in capturing the structural priors. The NeRF module becomes the most computationally expensive one with 32.7M parameters and 95.6 GFLOPs as well as highest training time (6.8 hours) and memory footprint (4.5 GB) because of the detailed volumetric rendering procedures. With a relatively low parameters of 5.2M and a low-cost of 18.4 GFLOPs, the Runge-Kutta Optimization module is important as it requires only 1.9 hours of training and requires only 1.2 GB of memory. Diffusion Refinement stage plays a significant role too with 28.9M parameters and 88.1 GFLOPs as well as 5.7 hours and 4.0 GB memory, which guarantees the ability of the output reconstruction of high quality. Comprehensively, the suggested model incorporates these elements in a single architecture with a total of 85.3M parameters and 244.4 GFLOPs, 17.6 hours of training data and 11.8 GB of memory utilization which represents a trade-off for both computational cost and performance efficiency in comparison with traditional monolithic architectures.

Table 6: **Baseline Model Comparison**

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↑
AFLW2000-3D	CNN-Based	28.5	0.918	0.092	0.961
	GAN-Based	29.3	0.925	0.088	0.965
	Proposed Method	**31.2**	**0.942**	**0.074**	**0.978**
FFHQ	CNN-Based	29.7	0.934	0.085	0.967
	GAN-Based	30.8	0.941	0.079	0.972
	Proposed Method	**32.5**	**0.951**	**0.068**	**0.981**
CelebA-HQ	CNN-Based	28.9	0.926	0.090	0.962
	GAN-Based	29.8	0.933	0.084	0.968
	Proposed Method	**30.8**	**0.938**	**0.079**	**0.976**
Multi-PIE	CNN-Based	27.8	0.915	0.101	0.953
	GAN-Based	28.6	0.922	0.095	0.959
	Proposed Method	**29.9**	**0.931**	**0.084**	**0.972**
DeepFashion2	CNN-Based	26.9	0.905	0.108	0.947
	GAN-Based	27.8	0.912	0.102	0.952
	Proposed Method	**28.7**	**0.924**	**0.091**	**0.968**
Human Faces Dataset	CNN-Based	28.0	0.918	0.096	0.955
	GAN-Based	29.1	0.926	0.089	0.962
	Proposed Method	**30.1**	**0.935**	**0.081**	**0.975**

Table 6 of the baseline results proves the optimal performance of the proposed RungeKutta optimized NeRF-Diffusion framework on each of the six datasets. The proposed model outperforms CNN-based and GAN-based models in terms of PSNR and SSIM values: it is clear that it produces higher values of these metrics, which is associated with a better reconstruction and consistent structure. Also, decreased LPIPS scores confirm increased perceptual realism whereas high identity similarity scores validate effective maintenance of facial features. The gain in performance is also large on more difficult datasets like Multi-PIE and DeepFashion2 which contain more pose variation and occlusion. In general, the findings indicate the strength, the presence of generalization, and efficacy of the presented method of managing various real-world selfie distortions.

Table 7: **Ablation Study**

Dataset	**Model Variant**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
AFLW2000-3D	Full Model (RK + NeRF + Diff)	31.2	0.942	0.074
	Without RK	29.8	0.931	0.082
	Without NeRF	27.6	0.914	0.096
	Without Diffusion	28.9	0.925	0.088
FFHQ	Full Model (RK + NeRF + Diff)	32.5	0.951	0.068
	Without RK	31.0	0.940	0.076
	Without NeRF	28.9	0.923	0.090
	Without Diffusion	30.2	0.934	0.081
CelebA-HQ	Full Model (RK + NeRF + Diff)	30.8	0.938	0.079
	Without RK	29.6	0.927	0.087
	Without NeRF	27.8	0.915	0.099
	Without Diffusion	29.1	0.924	0.089
Multi-PIE	Full Model (RK + NeRF + Diff)	29.9	0.931	0.084
	Without RK	28.7	0.920	0.092
	Without NeRF	26.9	0.908	0.105
	Without Diffusion	28.3	0.918	0.094
DeepFashion2	Full Model (RK + NeRF + Diff)	28.7	0.924	0.091
	Without RK	27.6	0.913	0.098
	Without NeRF	25.8	0.901	0.112
	Without Diffusion	27.1	0.909	0.101
Human Faces	Full Model (RK + NeRF + Diff)	30.1	0.935	0.081
	Without RK	28.9	0.924	0.089
	Without NeRF	27.2	0.912	0.102
	Without Diffusion	29.0	0.921	0.090

Table 7 ablation study assesses the value of each of the components of the proposed framework to all the six datasets. The complete model is always supported by the highest level of performance in PSNR, SSIM, and LPIPS, which proves the usefulness of the Runge-Kutta optimization, the use of NeRF-based geometry models, and diffusion refinement. When the RungeKutta optimizer is removed, the convergence rate and accuracy decrease, whereas the omission of the NeRF module drastically lowers geometric consistency. The lack of the diffusion module results in a higher perceptual error and perceptible artifacts. These findings indicate that every component is important and their integration is vital in the process of having strong and quality selfie rectification.

Table 8:**Performance Comparison with Existing Model**

Ref	Approach Type	PSNR ↑	SSIM ↑	ID Score ↑	LPIPS ↓	Accuracy / AUC ↑
[15]	GAN + Camera Optimization	31.8	0.747	0.859	0.182	–
[16]	CNN + Diff. Rendering	19.89	0.832	0.812	0.210	–
[18]	Transformer (ViT/DeiT)	–	–	–	–	98.1%
[19]	ViT + HRNet	–	–	–	–	99%
[20]	FaceNet512 + ANFE	–	–	0.975	–	97.5%
[21]	3DMM + Transformer	–	–	–	–	AUC 0.99
[22]	VGG16 / MobileNetV2	–	–	–	–	99.76%
[25]	CNN-based	–	–	–	–	89%
[26]	CNN-LSTM	–	–	–	–	98.21%
[31]	Mesh Warping	–	0.839	–	0.195	–
[32]	GAN Restoration	32.0	0.890	0.840	0.150	–
[33]	GAN Inversion	30.5	0.910	0.800	0.170	–
[34]	3D Morphable Model	–	0.880	–	–	–
[35]	Depth-based 3D	–	0.890	–	–	–
Proposed RK–NeRF–Diffusion	NeRF + RK Optimization + Diffusion	**33.8**	**0.945**	**0.932**	**0.098**	**99.1%**

The comparative performance analysis in Table 8 also reveals the strong and weak side of the existing methods under the various methodological categories, namely GAN-based, CNN-based, transformer-based, and geometry-driven methods. Competitive quality of reconstruction with PSNR and SSIM values beyond 30 dB and realistic values of 32, GAN-based approaches like [15] and Wang [32] have detection anomalies and hallucination artifacts, as the ID scores and LPIPS values present moderate positivity. CNN-based methods like Alhawari [16] offer quick inference, on the contrary, they have poor reconstruction outcomes, and cannot withstand real-world fluctuations. Transformer models [18], [19] and classification-based models[20], [21], [22] are highly accurate but not specifically used to correct distortion. Mesh warping [31] and 3D reconstruction methods [34] Ruan and [35] Guo are geometry based methods, which enhance structural stability yet do not provide high-quality visual refinement. Conversely, the overall performance in the form of PSNR (33.8), SSIM (0.945), and detection score (0.932) and the lowest LPIPS (0.098) suggest that the proposed RK-NERF-Diffusion framework can be successfully used in various cases to balance geometrical accuracy, preservation of detection, and visual realism. The source code and related implementation details used in this work are publicly available at the GitHub repository [42].

4.1 Discussion

The proposed Runge-Kutta optimized NeRF-Diffusion architecture demonstrates significant improvements in selfie perspective correction achieved by an effective mixture of the geometrical modeling, optimization stability, and generative refinement. The proposed algorithm, unlike the existing CNN-based algorithms, which operate mainly in the 2D image space, takes advantage of NeRF-based volumetric rendering to produce lifelike 3D facial geometry, and as such allows corrections of perspective distortions much more accurately, which occur when using close-range image capture.The use of Runge-Kutta optimization is important because it improves convergence stability and prevents the occurrence of local minima, which is a major weakness of traditional gradient-based optimization algorithms, including SGD and Adam. The strength and the ability to generalize the

framework are validated by experimental analysis carried out with multiple datasets, such as AFLW2000-3D, FFHQ, CelebA-HQ, Multi-PIE, DeepFashion2, and real-world human face datasets. The results mention the consistent improvement of PSNR and SSIM, and lower scores on LPIPS that prove the high quality of reconstruction. Further, identity preservation analysis reveals that the proposed model has high similarity scores even in extreme distortion conditions indicating that it is very effective in preserving facial characteristics. The qualitative comparisons also confirm that the model has the capacity to re-establish the natural proportions of the faces, minimize the exaggerated aspects, and increase the realism of the images. Diffusion based refinement step is effective in eliminating rendering artifacts and enhancing fine details, and generates high fidelity outputs. Also, ablation studies prove the significance of both components, especially the Runge-kutta optimization and diffusion modules, to the optimal performance. In total, the framework offers a decent trade off between geometrical, visual and computational efficiency and can be an effective future solution to the practical applicability of selfie correction.

5. Conclusion

This study presents a novel Runge-Kutta optimized NeRF-Diffusion hybrid architecture to correct the view of selfies being geometry-based. The offered solution involves the 3D volumetric reconstruction along with the further numerical optimization and diffusion-based image refinement and, that is why, may be regarded as the effective solution to the constraints of the existing 2D and GAN-based methods. NeRF allows to model the facial geometry appropriately and RungeKuttler optimization allows the process to be more stable concerning the convergence and the estimation of the parameters. Diffusion module also increases visual realism by sharpening textures and eliminating the artifacts. Growth in the degree of experimental analysis across a range of benchmark datasets justifies the expertise of the proposed approach over reconstruction achievement, identity upholding and excelence of the approach in alternative settings of distortion. The framework has higher PSNR and SSIM value and less perceptual error as compared to the baseline approaches. In addition, the identity similarity scores confirm that the model is capable of preserving the difference in the facial feature, even when there is an extreme pose. The proposed solution offers both structural and photorealistic solutions to selfie fixing, offered in a comprehensive manner. The next line of the future research will be to make the calculations as efficient as possible in the context of the real-time implementation and use the framework on the video-based operation and dynamic reconstruction of the scene.

Limitations and Future Work

Although the proposed framework demonstrates strong performance in geometry-aware selfie rectification, the model still requires high computational resources due to NeRF rendering and iterative diffusion refinement. Real-time deployment on mobile devices remains challenging. Future work will focus on lightweight optimization strategies, real-time inference acceleration, and extension toward video-based selfie rectification and dynamic facial reconstruction.

Code and Data Availability

The implementation details, training configurations, experimental settings, and source code associated with this work will be made publicly available through an online repository to improve transparency and reproducibility of the proposed framework.

GitHub Repository Link: https://github.com/thirumagalresearch-svg/Geometry

Dataset Availability

"AFLW2000-3D", Accessed: Mar. 19, 2026. [Online].
Available: https://www.kaggle.com/datasets/mohamedadlyi/aflw2000-3d

"Flickr-Faces-HQ Dataset (FFHQ)", Accessed: Mar. 19, 2026. [Online].
Available: https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq

"CelebA-HQ resized (256x256)", Accessed: Mar. 19, 2026. [Online].
Available: https://www.kaggle.com/datasets/badasstechie/celebahq-resized-256x256

"Multi_Pie", Accessed: Mar. 19, 2026. [Online].
Available: https://www.kaggle.com/datasets/aliates/multi-pie

"DeepFashion2 Original (with Dataframes)", Accessed: Mar. 19, 2026. [Online].
Available: https://www.kaggle.com/datasets/thusharanair/deepfashion2-original-with-dataframes

"Human Faces Dataset", Accessed: Mar. 19, 2026. [Online].
Available: https://www.kaggle.com/datasets/kaustubhdhote/human-faces-dataset

Acknowledgements
The authors warmly acknowledged REVA University, Bengaluru, India, Karunya Institute of Technology and Sciences, Coimbatore, India and Dayananda Sagar University, Bengaluru, India for providing the facilities required to carry out the research.

Funding
No funding received for this research.

Declarations
Conflict of interest: No conflict of interest.

References

[1] B. Chen, B. Curless, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Total selfie: Generating full-body selfies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6701–6711.
[2] Y. Li, Y. Li, and Z. Wang, "Facecloseup: Enhancing mobile facial authentication with perspective distortion-based liveness detection," *Computers*, vol. 14, no. 7, p. 254, 2025.
[3] S. Chun, B. Kim, H. J. Chang, and J. Y. Chang, "Bidirectional Regression for Monocular 6DoF Head Pose Estimation and Reference System Alignment," *arXiv preprint arXiv:2407.14136*, 2024.
[4] J. E. Tapia, A. Valenzuela, R. Lara, M. Gomez-Barrero, and C. Busch, "Selfie periocular verification using an efficient super-resolution approach," *Ieee Access*, vol. 10, pp. 67573–67589, 2022.
[5] P. Karpikova *et al.*, "Super: Selfie undistortion and head pose editing with identity preservation," in *2024 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2024, pp. 1704–1710.
[6] S. H. R. Rasmussen, S. G. Ludeke, and R. Klemmensen, "Using deep learning to predict ideology from facial photographs: expressions, beauty, and extra-facial information," *Scientific Reports*, vol. 13, no. 1, p. 5257, 2023.
[7] M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev, and J. Cho, "Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people," *Sensors*, vol. 23, no. 3, p. 1080, 2023.
[8] K. Nagaraju and M. B. Reddy, "Automated handcrafted features with deep learning based age group estimation model using facial profiles," *Multimedia Tools and Applications*, vol. 83, no. 14, pp. 42149–42164, 2024.
[9] S. Haseena, S. Saroja, R. Madavan, A. Karthick, B. Pant, and M. Kifetew, "Prediction of the age and gender based on human face images based on deep learning algorithm," *Computational and mathematical methods in medicine*, vol. 2022, no. 1, p. 1413597, 2022.
[10] Y. Kong, K. Zhang, L. Zhang, and G. Wu, "Deep facial attribute analysis," *Frontiers in Neuroscience*, vol. 17. Frontiers Media SA, p. 1280831, 2023.
[11] S. K. Gupta and N. Nain, "Single attribute and multi attribute facial gender and age estimation," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 1289–1311, 2023.
[12] L. Qin *et al.*, "SwinFace: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2223–2234, 2023.
[13] K. ELKarazle, V. Raman, and P. Then, "Facial age estimation using machine learning techniques: An overview," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 128, 2022.

[14] J. N. Saeed, A. M. Abdulazeez, and D. A. Ibrahim, “Automatic facial aesthetic prediction based on deep learning with loss ensembles,” *Applied Sciences*, vol. 13, no. 17, p. 9728, 2023.
[15] Z. Wang *et al.*, “Disco: Portrait distortion correction with perspective-aware 3d gans,” *International Journal of Computer Vision*, vol. 132, no. 11, pp. 5471–5488, 2024.
[16] A. Alhawwary, J. Mustaniemi, P. Nguyen-Ha, and J. Heikkilä, “An End-to-End Depth-Based Pipeline for Selfie Image Rectification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
[17] E. Bermejo, E. Fernandez-Blanco, A. Valsecchi, P. Mesejo, O. Ibáñez, and K. Imaizumi, “FacialSCDnet: A deep learning approach for the estimation of subject-to-camera distance in facial photographs,” *Expert Systems with Applications*, vol. 210, p. 118457, Dec. 2022, doi: 10.1016/j.eswa.2022.118457.
[18] S. Islam and M. Arman, “Transformer-Based Extreme Selfie Detection for Enhancing Safety in Hazardous Environments,” in *2024 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, IEEE, 2024, pp. 454–459.
[19] C. Zhang, “Deep learning for visual aesthetics: using convolutional vision transformers and HRNet for classifying anime and human selfies,” *IJICT*, vol. 26, no. 20, p. 10071591, 2025, doi: 10.1504/IJICT.2025.10071591.
[20] S. Makinist and G. Aydin, “Gender Classification Using Face Vectors: A Deep Learning Approach Without Classical Models,” *Information*, vol. 16, no. 7, p. 531, Jun. 2025, doi: 10.3390/info16070531.
[21] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. I. Almaloglou, and D. Zarpalas, “Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification,” *Multimed Tools Appl*, vol. 84, no. 33, pp. 40617–40636, Mar. 2025, doi: 10.1007/s11042-024-20548-6.
[22] S. Ingoley and J. Bakal, “Interpretation of Indian Sign Language to Text and Speech to Communicate with Speech and Hearing Impaired Community,” *Procedia Computer Science*, vol. 258, pp. 1980–1992, 2025, doi: 10.1016/j.procs.2025.04.449.
[23] S. Piao, J. Choi, S. Seo, and S. Park, “SELF-EdiT: Structure-constrained molecular optimisation using SELFIES editing transformer,” *Appl Intell*, vol. 53, no. 21, pp. 25868–25880, Nov. 2023, doi: 10.1007/s10489-023-04915-8.
[24] W. Zheng, Q. Li, G. Zhang, P. Wan, and Z. Wang, “Ittr: Unpaired image-to-image translation with transformers,” *arXiv preprint arXiv:2203.16015*, 2022.
[25] S. Bekhet, A. M. Alghamdi, and I. F. Taj-Eddin, “Gender recognition from unconstrained selfie images: a convolutional neural network approach,” *IJECE*, vol. 12, no. 2, p. 2066, Apr. 2022, doi: 10.11591/ijece.v12i2.pp2066-2078.
[26] O. A. H. H. Al-Dulaimi and S. Kurnaz, “A Hybrid CNN-LSTM Approach for Precision Deepfake Image Detection Based on Transfer Learning,” *Electronics*, vol. 13, no. 9, p. 1662, Apr. 2024, doi: 10.3390/electronics13091662.
[27] B. Bugeja, D. Spirou, and J. Raman, “The effects of social media and photo-editing applications on body image and attitudes towards cosmetic procedures,” *Clinical Psychologist*, pp. 1–13, Feb. 2026, doi: 10.1080/13284207.2026.2628918.
[28] J. Conde, C. Speelman, and M. Johnstone, “A new epoch of face analytics: technological evolution through ethical and legal challenges,” *AI Ethics*, vol. 5, no. 5, pp. 5213–5237, Oct. 2025, doi: 10.1007/s43681-025-00678-9.
[29] A. Raza, K. Munir, and M. Almutairi, “A novel deep learning approach for deepfake image detection,” *Applied Sciences*, vol. 12, no. 19, p. 9820, 2022.
[30] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM computing surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.
[31] W.-S. Lai, Y. Shih, C.-K. Liang, and M.-H. Yang, “Correcting face distortion in wide-angle videos,” *IEEE Transactions on Image Processing*, vol. 31, pp. 366–378, 2021.
[32] N. C. Rakotonirina and A. Rasoanaivo, “ESRGAN+: Further improving enhanced super-resolution generative adversarial network,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 3637–3641.
[33] B. Li *et al.*, “Reganie: Rectifying gan inversion errors for accurate real image editing,” in *Proceedings of the AAAI conference on artificial intelligence*, 2023, pp. 1269–1277.

[34] Z. Ruan, C. Zou, L. Wu, G. Wu, and L. Wang, "Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction," *IEEE Transactions on Image Processing*, vol. 30, pp. 5793–5806, 2021.
[35] H. Guo *et al.*, "Multi-view reconstruction via sfm-guided monocular depth estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5272–5282.
[36] "AFLW2000-3D." Accessed: Mar. 19, 2026. [Online]. Available: https://www.kaggle.com/datasets/mohamedadlyi/aflw2000-3d
[37] "Flickr-Faces-HQ Dataset (FFHQ)." Accessed: Mar. 19, 2026. [Online]. Available: https://www.kaggle.com/datasets/arnaud58/flickrfaceshq-dataset-ffhq
[38] "CelebA-HQ resized (256x256)." Accessed: Mar. 19, 2026. [Online]. Available: https://www.kaggle.com/datasets/badasstechie/celebahq-resized-256x256
[39] "Multi_Pie." Accessed: Mar. 19, 2026. [Online]. Available: https://www.kaggle.com/datasets/aliates/multi-pie
[40] "DeepFashion2 Original (with Dataframes)." Accessed: Mar. 19, 2026. [Online]. Available: https://www.kaggle.com/datasets/thusharanair/deepfashion2-original-with-dataframes
[41] "Human Faces Dataset." Accessed: Mar. 19, 2026. [Online]. Available: https://www.kaggle.com/datasets/kaustubhdhote/human-faces-dataset
[42] Source code and the implementation details are publicly available at the following GitHub repository: https://github.com/thirumagalresearch-svg/Geometry



Dr.Selvan.C received the B.E. degree in Computer Science and Engineering in Manonmaniam Sundaranar University, India, in 2002, the M.E. degree in Computer Science and Engineering from Anna University, Chennai, India, in 2007 and the Ph.D. in Computer Science and Engineering from Anna University, Chennai, India in 2013. During his Ph.D degree he was a JRF, SRF under University Grant Commission (UGC, New Delhi) in Government College of Technology, Coimbatore. He had been working as a Software Engineer during 2002 to 2005. He was a Post-Doctoral Fellow in National Institute of Technology, Tiruchirappalli, under UGC, New Delhi, India from June 2017 to June 2022. In the year of 2022 June, he joined as a professor in New Horizon College of Engineering, Bangalore and worked upto January 2023. Currently he has been working as a Professor in the School of Computer Science and Engineering, REVA University, Bangalore since February 2023. His current research interests include Artificial Intelligence, Computational Data Science, Applied Data Science and Mobile Computing. He is an IEEE senior member, ACM member, IAENG member and a member of the Indian Society for Technical Education. ORCID: 0000-0003-4381-0284



Dr. M Mythily holds a Doctorate from Anna University, with Bachelor's and Master's degrees specializing in Computer Science and Engineering from Avinashilingam Deemed University and Government College of Technology, Coimbatore, Tamil Nadu, India, respectively. Currently, she serves as an Associate Professor at Karunya Institute of Technology and Sciences. With two decades of experience spanning both industry and academia, her expertise encompasses Software Engineering, Design Patterns, Problem-solving Techniques, and Data Science. Dr. Mythily has contributed significantly to her field with over 30 publications in refereed international journals and conference proceedings. ORCID: 0000-0002-3534-6285



Dr. Thirumagal Mohan is an accomplished academician and researcher with over a decade of teaching experience, currently serving as an Assistant Professor at REVA University since 2015, with prior experience at Sri Venkateshwara College of Engineering, Chennai. She holds a PhD (2024) along with M.Tech and B.Tech degrees in Computer Science from Pondicherry University, consistently achieving distinction. Her expertise spans Data Structures, Algorithms, Machine Learning, Deep Learning, Database Management Systems, and Cyber Security. She has made significant research contributions with multiple Q2 journal publications and Scopus-indexed conference papers, particularly in the areas of generative adversarial networks and medical image analysis. In addition, she has been actively involved in guiding research, delivering invited talks, organizing industrial visits, and participating in numerous faculty development programs. She also holds granted patents, authored books and book chapters, and contributed to academic

administration, demonstrating strong commitment to teaching, research, and institutional development. ORCID: 0000-0002-9588-099X



Dr. M. Prabhakar is presently working as Professor in Dept. of Computer Science & Engineering, Dayanand Sagar University – Bangalore with an aim of a project for design and developing "Wireless Charging Station for e-Vehicles. He has published about 15 patents and has obtained 5 patent grants in various fields including Internet of Things, Image processing, Education, and Security in WSNs. He was awarded `Excellence in Research', by Novel Research Academy, Puducherry, India. He has authored 4 text books and published about 50 research articles in blind-peer reviewed International and National Journals which are Scopus & WoS indexed with about 1400 citations at present. He has successfully guided four Ph.D. Research Scholars and six are pursuing currently. Besides these achievements, he has attended many Workshops and Seminars conducted by IBM, Intel and Cisco focusing on Microcontroller Boards that are very commonly used in IoT Applications. He is continuously contributing exceptionally as Central Evaluator & Judge in "Smart India Hackathon" organized by Ministry of HRD, India. His area of research is IoT, WSN, Medical Image Processing and Natural Language Processing and Artificial Intelligence. ORCID: 0000-0002-9117-1483



Dr. Aravindhan Ragunathan received his Ph.D. in Computer Science and Engineering from Anna University, Chennai, India, in 2019. Currently, he is working as a Senior Software Engineer in Digital Factory at Instrumentation Laboratory, Bedford, MA, USA. He has 10+ years of industrial experience in Augmented Reality (AR), Virtual Reality (VR), Mixed Reality (MR), Multi-Cloud Architecture, and IoT, with a strong focus on the medical domain. He has extensive expertise in Java-based application development, IoT, cloud computing, DevOps, and cybersecurity. His research interests include web security, phishing attack detection, augmented reality, and automation technologies. ORCID: 0000-0001-9495-0362