IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 58, No. 3, March, 2025, pp. 501-510
Journal homepage: http://iieta.org/journals/jesa

Novel 3D Mesh Captured by LIDAR Texture Enhancement Using Neural Radiance Fields and Vision Transformers

Farooq Safauldeen Omar, Borkan Ahmed Al-Yaychili, Shayma Jaafar, Mohammed Safar*

Northern Technical University, Technical Engineering College Kirkuk, Department of computer Technology Engineering, Kirkuk 36001, Iraq

Corresponding Author Email: mohammed.sefer@ntu.edu.iq

https://doi.org/10.18280/jesa.580308

Received: 31 January 2025
Revised: 12 March 2025
Accepted: 20 March 2025
Available online: 31 March 2025

Keywords:
3D mesh, texture enhancement, Neural Radiance Fields, vision transformers, LIDAR

ABSTRACT

The quality of 3D mesh textures is imperative in applications related to gaming, virtual reality and digital content creation due to the importance of visual integrity. This paper presents a new pipeline for enhancing 3D mesh textures by incorporating Neural Radiance Fields (NeRF), Vision Transformers and self-supervised learning methodologies in order to enhance texture detail, consistency and mapping precision. This work proposes a new pipeline that combines NeRF, Vision Transformers (ViT) and self-supervised methodologies for enhancing 3D mesh textures with high fidelity while keeping the geometry of the underlying mesh intact. It generates multi-view synthetic images of the mesh using off-screen rendering and then trains a NeRF to get a radiance field that can generate higher-fidelity texture features. This results in a finer texture so with the further help of a Vision Transformer and a lightweight diffusion-based which can create globally coherent high-resolution edits. Experimental results indeed do not show any geometric distortion also as it already been suggested by the low Hausdorff distance and average distance metrics and whereas for texture evaluation using MSE and SSIM the visual quality increase is substantial.

1. INTRODUCTION

The rapid development of visual perception and processing in technologies has run parallel to a big change in how 3D models are generated from 2D images. So photogrammetry and laser scanning are some of the techniques that have enabled the derivation of complex 3D models from 2D images and it allowing better identification and representation of objects with accuracy for various industries. In architecture it can enhances the design visualization and its accuracy hence effective project implementation. The entertainment industry needs 3D models to help create immersive environments either in video games or film. However, these processes have normally been followed by a number of challenges including heavy costs and processes being very time-consuming also the requirement for specialized expertise hence raising the need to find effective methods. The latest breakthroughs in AI revealed new ways of creating 3D models notably with the use of Neural Radiance Fields (NeRF) [1].

NeRF is a method of synthesizing new views of complicated scenes by refining a volumetric scene function from input images using neural networks. Representing a considerable departure from more classical 3D reconstruction algorithms, and the approach results in substantial efficiency and accuracy improvements. Which enables NeRF to learn the continuous representation of a scene which is very powerful in a number of use cases where realism of visualization is imperative and such as virtual reality or the protection of cultural heritage [2]. The integration of NeRF with other advanced technologies like Vision Transformers (ViT) has greatly increased the capacity of 3D modeling. The ViT also known for competence in the compilation of global dependencies inside picture data enhancing the NeRF by enhancing texture information hence increasing the overall quality of the generated models [3].

This combination overcomes common problems in 3D graphic design such as poor texturing and inefficient rendering and leading to more accurate also a better-looking result. The application of self-supervised learning methods in this domain allows models to learn from unlabeled data reducing reliance on large labeled datasets and therefore streamlining model development. Despite such progress several challenges remain in the process of acquiring high-quality 3D reconstructions from 2D photographs. So, the details such as fluctuating lighting, occlusions and the intrinsic ambiguity in deducing depth from two-dimensional data can influence the precision of the produced models. Working out these challenges will take continued research and development in creating more robust algorithms which is capable of handling different real-world scenarios [4].

The continuing development in AI-driven approaches offers hope for overcoming these challenges and making 3D modeling more available and effective in a wide range of applications. The combination of NeRF and ViT through self-supervised learning heralds a promising route toward 3D model creation, dealing with some long-lasting issues in the

area. Thus those technologies will find themselves to be in a more important role across a range of applications from digital arts and entertainment to scientific research and cultural heritage preservations. This research will analyze these combined methodologies in terms of their effectiveness in improving the quality of 3D models and how they may affect future developments within the industry.

1.1 Problem statement and aim

Notwithstanding tremendous development in 3D model technology, challenges persist in high-quality refinement and texture reconstruction within complex scenarios; despite new techniques like NeRF—which demonstrate expertise in generating realistic 3D environments from multi-view photographs—such methods still struggle to represent intricate texture details and maintain processing efficiency for large-scale structures. ViT, while effective in extracting global spatial information from textures, face limitations in capturing high-frequency local textures and achieving computational efficiency in complex scenarios; furthermore, integrating NeRF and ViT frameworks for texture refinement poses significant challenges: resolving discrepancies in texture mapping, improving fidelity through higher resolution, and addressing occlusions and depth ambiguities—all of which restrict the widespread adoption of such technologies in autonomous systems, virtual reality, and cultural heritage preservation.

So this proposed work seeks to develop a coherent pipeline that integrates ViT and NeRF strengths for 3D model and 3D texture refinement and the 3D texture reconstruction improvement. The work contribution is to address gaps in spatio temporal consistency and also the texture fidelity through volumetric capabilities of NeRF in collaboration with ViT's self-attention mechanism for encoding spatial relations between those entities in a 3D environment. In this work try to attempts that is made towards high-performance 3D texture reconstruction and high-fidelity rendering through incorporation of state-of-the-art methodologies through hierarchical feature extraction context-aware feature extraction and also efficient mapping of textures with awareness for depth. It seeks to develop a strong and efficient model for use in many real-life scenarios such as high-fidelity 3D scene reconstruction complex refinement of textures and real-time scene understanding [5].

2. RELATED WORK

The development of 3D mesh texture improvement has seen significant improvements with NeRF, ViT and self-supervised training incorporated in its development. In this section relevant studies have been discussed which have developed high-fidelity techniques for enhancing textures.

A work has produced high-fidelity textures out of sparsely observable photos via restoration of occluded areas in 3D radiance field. ViT are leveraged in the method for effectively processing complex semantic and spatial compositions in a scene. And with volumetric grid of NeRF considered as an input the model attains a high level of detail and homogeneity in textures. The work it has established a masked self-supervised training which can boost representational efficiency in NeRF immensely and offering a high fidelity in textures with no additional label information. And the experimental evaluation confirms that such a method can produce real 3D textures with fewer defects in comparison with traditional NeRF-based approaches [5].

DeLiRa employs a generalist Vision Transformer (ViT) to jointly learn scene radiance fields through lighting and depth estimation. In contrast to traditional NeRF models—which explicitly encode scene geometry to achieve high-fidelity 3D texture reconstruction—DeLiRa enables uniform illumination control and refinement-aware depth optimization while maintaining computational efficiency. The work has introduced its method outperforms state-of-the-art state in both sharpness and uniformity of textures for NeRF-based techniques and contribution comes in its capacity for both photometric and geometric information extraction out of unorganized sets of images supporting continuous 3D texture reconstruction. And constitutes a basis for 3D realism improvement in virtual reality and gaming environments [6].

A self-supervised model for 3D neural field learning (N3F) leverages 2D image feature extraction through knowledge distillation, utilizing a pre-trained Vision Transformer (ViT) to capture high-semantic-level image features. In contrast to conventional NeRF techniques that rely solely on pixel-wise color values for 3D radiance field reconstruction, N3F enhances texture continuity and sharpens boundary definitions, thereby generating high-fidelity 3D textures. The study demonstrates that integrating semantically enriched 2D features into 3D neural fields produces textures with improved aesthetic quality and contextual relevance. By employing self-supervised training to eliminate dependency on labeled datasets, N3F provides an efficient solution for large-scale 3D texture synthesis [7].

This work improves upon NeRF by tackling aliasing in texture rendering through a multiscale model. Unlike conventional point sampling techniques, it employs conical frustums to integrate image details across multiple scales. The 3D textures are generated using Mip-NeRF, resulting in reduced aliasing and a more sophisticated structure compared to standard NeRF implementations. The study demonstrates that this approach significantly enhances 3D texture reconstruction accuracy, particularly in scenes with intricate textures such as foliage, high-detail surfaces, and complex lighting conditions. Mip-NeRF is specifically designed to handle high-resolution textures efficiently, meeting demanding performance requirements [8].

Another study employs a GAN to enhance low-fidelity NeRF outputs, improving texture sharpness and reducing noise. By integrating a super-resolution model, the approach refines NeRF-generated textures without requiring additional high-fidelity input images. Experimental results demonstrate that this method produces visually appealing textures with enhanced perceptual fidelity and structural integrity. The technique is particularly valuable for gaming, virtual reality, and computer-generated content creation, where high-fidelity textures play a crucial role in delivering immersive experiences [9].

Another work utilizes ViT, which outperform conventional CNN-based techniques in handling complex spatial structures and maintaining overall texture cohesion, according to this study. So with a mechanism of self-attention ViTs has potentially learn long-term 3D textures dependencies and then gain an improvement in and accuracy in terms of reconstruction. The work claims that ViTs enable improved feature fusion in NeRF-based techniques, leading to enhanced texture and geometry cohesion. The results show a significant

information regarding deep architectures contribution in 3D mesh textures quality and an idea that a merge between NeRF and Vision Transformers holds a high future development in 3D rendering is proposed [10].

Following work has optimized integration of a single-scene sampled voxel grid with a CNN so such system can counteract discrete artifact of a low resolution voxel grid or predict a temporally variable and an animation variable voxel grid. While such volumetric approaches have seen tremendous success for novel view synthesis generalizability to high resolution images is naturally limited in such discrete spatio-temporal samples and high resolution image creation necessitates a more complex 3D sampling [11].

Recent research in view synthesis has demonstrated groundbreaking advancements in generating high-quality, realistic renderings through innovative approaches in neural representations a differentiable rendering and multi-view optimization. Below is an extended overview of additional significant contributions to the field:

One novelty lies in the approach suggested in a study that combines the attention mechanism with the convolutional networks to handle some of the challenges brought about by sparse input views in novel view synthesis. So improves the perceptual quality of the rendered scenes through an integration of multi-scale depth features within the network of the project while using a structural unit with adaptive channel weighting algorithms. Whereas traditional methods usually require explicit 3D supervision and model's joint attention mechanism its end-to-end training capability ensure robustness for object-centric and scene-level renderings. The work delivers substantial improvements for tasks involving monocular images with limited viewpoints, where such quality degradation commonly occurs. Transparent objects have consistently presented challenges in rendering pipelines due to their complex light transport properties [12].

Another work introduces Transparent Neural Surface Refinement, which employs Snell's law to achieve physically accurate tracing of both refracted and reflected light. The major novelty is the differentiable optimization that sends the photometric evidence into the surface model. TNSR improves both the geometry estimation and novel view synthesis for transparent objects by adding bending and reflection to volume rendering. These improvements are a significant step toward applying neural rendering frameworks to complex materials like glass and water. This work addresses the main limitations of real-time 3D reconstruction in traditional SLAM systems and demonstrates a new approach by incorporating 3D Gaussian splatting with depth priors into a state-of-the-art competitive baseline. It enables dense 3D reconstruction through a differentiable optimization process regularized by both photometric and geometric losses [13].

The depth priors add an additional level of regularisation to improve the accuracy in posture estimation and scene reconstruction. Real-time rendering and its optimization using CUDA enable the implementation to achieve an optimal balance between real-time performance and high geometric fidelity, making it suitable for practical robotics and AR/VR applications. Those scenarios with few input views and traditional methods often cannot guarantee the quality of rendering. This work has extended a 3D Gaussian splatting to incorporate depth-sensitive constraints through monocular depth estimation and scale-invariant loss functions [14].

Another work has the model to avoid overfitting by resorting to spherical harmonics to represent a color to maintain the low-opacity splats that other methods get rid of therefore it enables a better reconstruction of the scene. The result is a significantly improved perceptual quality so with metrics such as PSNR, SSIM and perceptual similarity showing striking improvements. This is an innovation that really underlines the importance of depth priors in overcoming few-shot limits The work integrates a pipeline in code that combines NeRF, mesh rendering, and ViT to enhance 3D object texturing, drawing substantial benefits from insights in the relevant literature [15].

The addition of attention techniques a ViTRefiner model parallels other methodologies that increase feature extraction and rendering quality from sparse viewpoints, such as multiscale-depth transformers. Integration of depth priors and the use of differentiable optimization-as described in at least one study on 3D Gaussian splatting and transparent object modeling-can further enhance texture fidelity and reduce artifacts in the process of multi-view rendering. These enhancements will further improve the proposed NeRF framework's ability to transfer photometric knowledge for capturing finer texture and surface details, enabling higher-quality and more consistent rendering of complex geometries. Moreover, these techniques - similar to those explored for depth-aware optimization in Gaussian splatting for novel view synthesis - will be integrated with the model's ray sampling and positional encoding. Such approaches particularly benefit texture refinement in high-variance regions, consequently improving performance in poorly observed scenarios.

Integrating transparent object modeling techniques will enhance proposed work volume rendering workflow for complex surfaces especially glass or those with reflective materials. With these advances system can gain the enhancement of authentic textures and broaden into a wider selection of real 3D object applications.

3. METHODOLOGY

NeRF is a framework that represents 3D scenes using a fully connected multi-layer perceptron (MLP) neural network. Designed for novel view synthesis, it achieves state-of-the-art photorealistic rendering from continuous viewpoint inputs. A key advantage of NeRF is its ability to train effectively with fewer input images compared to other methods, while maintaining robust performance even when handling dynamic viewpoints. The system demonstrates strong capabilities in generating scenes from view representations. [16]. The main NeRF approach take's a scene as neural volume which it described by the weights of MLP and it uses 5D: (x, y, z, θ, Φ) as inputs of MLP which it configured from 3D position of view x= (x, y, z) and 2D direction of view d= (θ, Φ) corresponding to point along with ray of camera. Where the output of the MLP resulted from three major color channels c= (r, g, b) and volume density (σ) for pixel surface of 2D image at that point of view. The MLP feed forward network can be written like FΘ: (x; d) → (c; σ). MLP can be optimized by differentiable function of volume rendering and trained on a set of real images and their viewing directions must be known by evaluating variances between the actual pixel color and predicted pixel color from the volume rendering process the loss function can be selected [17]. To understand how NeRF approach works in detail, Figure 1 shown below from original paper that can be very useful to come with NeRF approach concept [18].

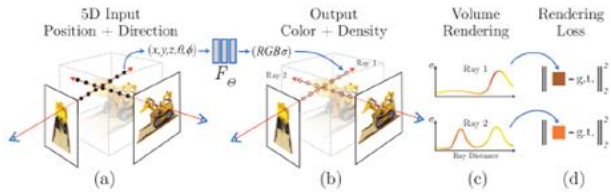

Figure 1. NeRF representation of scene pipeline. (a) Passing of 5D input by feed forward. (b) Mapping the output of 4D in the space of 2D. (c) Volume rendering by ray marching. (d) Optimization by rendering loss

NeRF become absolutely necessary in this regard for the task of texture prediction and enhancement by using their own capability of producing intricate and lifelike textures from multi-view images. NeRF-Texture has introduced a new style of synthesizing textures by separating mesostructural details from base forms and inputting them into a NeRF decoder in order to obtain view-dependent textures so this implicit representation in this work easily allows the creation of fine textures on the curved and planar surface and removing artifacts of latent-space inconsistencies by some clustering restrictions. Besides that a hybrid sampling approach and adaptive scene decomposition are done in MDNeRF for better textures in larger scenes without typical issues like blurred and missing details of textures and due to its innovative features the use of spiral sampling with modular sub-scene decomposition together with distinct texture quality enhancement helps the software achieve high performance in large-scale scene renderings [19].

Meanwhile NeRF-VPT relied on a cascade view prompt tuning framework and iteratively performing the refinement of textures through RGB data obtained in previous rendering processes and using them as visual prompts. In such a way baseline has gained an enhanced level of performance in the effective recovery of proper texture details and adaptation in conditions of sparse input-view setups making this approach highly desired for practical use in the most stringent manner. This is furthered by the incorporation of multi-resolution hash grid characteristics along with multi-view priors making it better to rebuild such a huge open landscape that exhibits intricate details and is highly geometrically accurate [20]. A dual-branch architecture captures more features for MM-NeRF. These bring better PSNR results with fewer issues relating to underfitting while allowing NeRF to yield finer performance on different conditions concerning better texture predictions as mentioned. These together reveal that NeRF makes a revolutionary impact on generating realistic view-consistent textures for many applications [21].

Vision Transformer (ViT), investigated by the architecture of the model consists from many components as shown in Figure 2. The components work together by following very brief steps. Slice an image into fixed –size of patches. Resulted patches then flattened to get lower-dimensional linear embeddings. Positional embeddings added. Providing the sequence to a conventional transformer encoder as an input. The model pertrained with image labels [22].

Focusing on picture categorization using the downstream dataset.

Steps of processing images and calculation of Liner Projection of flattened patches in ViT model Processing Input Image:

Input: Raw image taken as an input.

Patching: The raw image is divided into patches.

For instance, if raw image size= (224*224) and each patch size = (16*16) then the number of total patches will be (224/16) = (14*14) =196 patch as shown in Figure 3.

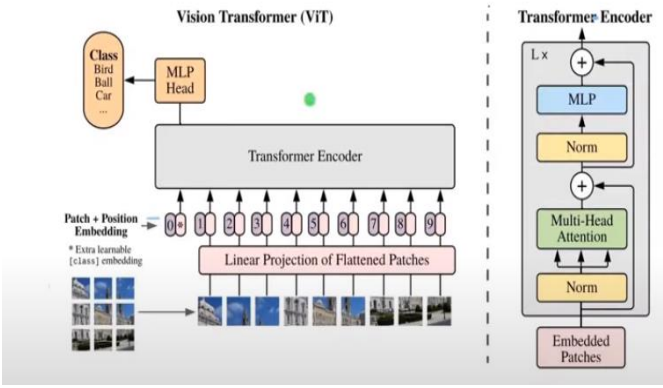

Figure 2. Architecture of ViT model

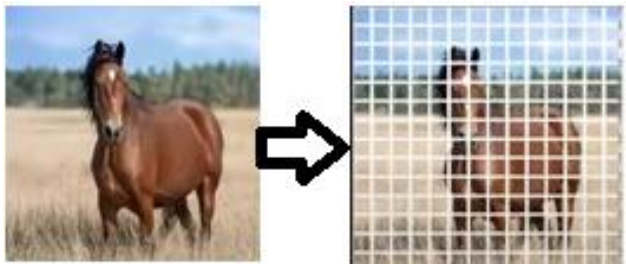

Figure 3. Patching raw image

Flattening: A 1D vector is created by flattening each patch. Each flattened patch will contain 16*16*3=768 elements due to the image has three color channels (RGB) where the first channel represents red color and the second channel represents green color and the third channel represents blue color. And 16 × 16 pixels is the patch size as shown in Figure 4.

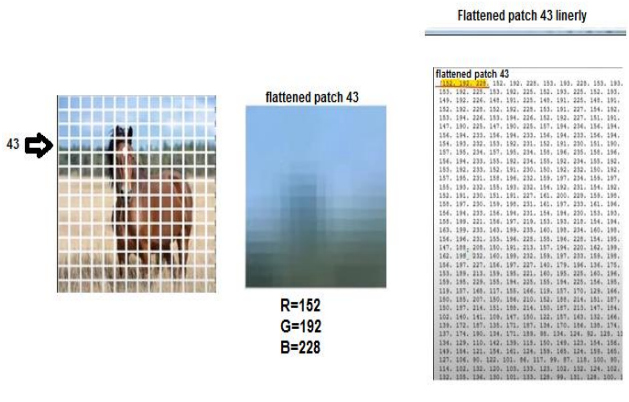

Figure 4. Flattening patch 43 as an example

Linear Projection: in this step the number of items in the flattened patches going to be reduced. For example, if the number of flatting such patches is 768 item and we want to reduce it to 512 items as an example under condition of maintaining the important information of image and the resolution at the acceptable level. This process will lead to decrease occupied memory space which it result to increase speed of processing after reducing number of items and this can be done by fully connected neural network MLP through

following formula y=xw+b where y is desirable number of items for each patch, x is base number of items for each patch, w is weight matrix (y, x) and b is bias value [23].

The proposed pipeline has integrated both NeRF and ViT by generating some sets of views of the 3D mesh via volumetric modeling first then effectively capturing a complex geometric and textural details. NeRF's volumetric method that represents the scene as a radiance field sampled along rays and yielding high-quality 2D renderings of a set of views. So the ViT then refines these renderings with the self-attention mechanism to capture global relationships between image patches that especially help for smoothing texture seams and fixing the currents errors. Together the self-attention layers enable the model to learn non-local dependencies so that the refined texture is consistent across the whole surface of the mesh. The integration of volumetric view generation and the patch-based global refining results in higher-quality textures without compromising geometry and introducing undesirable artifacts.

An analysis demonstrates how the self-supervised learning enhances the texture consistency and completes partially observed areas. The pipeline has created several synthetic views of a wall mesh which has been extracted from full room mesh then masks out some areas in the photos so that the model learns to fill the missing details. The process promotes a consistency by comparing overlapping patches between different views so providing consistent textures across camera viewpoints. Experimental results demonstrate that the absence of self-supervision results in slightly increased MSE and more jagged texture transitions. The self-supervised strategy is thus essential to smooth out transitions to maintain local details and enhance the overall visual fidelity of the resulting 3D model.

Improving texture requires the contribution of ViT can employ the self-attention mechanisms in the extraction of spatial relationships and contextual information from texture images. Unlike CNNs which it often biased toward local textures and ViTs adaptively change their receptive field to capture both local and global patterns. For instance, ViTs might effectively represent the high-frequency component and texture variation by recalibrating frequency information with techniques like Laplacian pyramids as shown in the case of Laplacian-Former. These enhance attention to texture and edge features by offering significant enhancement toward applications such as texture-based segmentation and reconstruction [24].

ViTs can be found to be more robust to occlusions and domain shifts because their multi-head self-attention mechanism effectively captures complex spatial relationships among image patches. This allows for accurate texture amplification by focusing on intricate details and wider contextual relationships. And research indicates that ViTs surpass conventional models in recognising intricate patterns and textures without requiring pixel-level supervision rendering them particularly suitable for texture refinement tasks in noisy or ambiguous datasets [25]. Through the integration of these characteristics ViTs provide a more thorough comprehension of spatial and textural patterns enhancing the accuracy of synthesised and reconstructed textures in applications spanning medical imaging to 3D modelling [25].

Refining UV maps such as that through self-supervised learning which can bring huge potential contributions toward better quality in 3D textures of both games and virtual reality and digital content creation. Some of the latest improvements that have been spotted involve the application of self-supervised systems in solving the challenge of reconstructing higher-quality UV maps without manual annotation [26]. propose a "Map and Edit" framework that integrates 2D generative models with 3D priors to complete the missing parts in multi-view facial photos while maintaining their texture-rich UV maps intact. And the approach minimizes the domain gap between real and synthesized data by using synthesized multi-view images that maintain the same textural details constantly. Likewise, a self-supervised "StyleGAN" approach clearly modifies the attributes or generates high-resolution UV textures that are capable of preserving intricate features with coherence in identity [27]. And the integration of NeRF and ViT within this self-supervised framework provides better global coherence of texture to improved geometry precision higher quality enhanced texture fidelity according to metrics such as MSE and SSIM so it will thus enable realistic generation of textures with superior performance for downstream tasks including unsupervised domain adaptation of person reidentification [28].

It develops this work through updating the UV map with self-supervised learning in order to enhance texture detail and reduce geometric distortion, hence being more adaptable to various domains. The presented solutions address directly the pipeline that has been described in your question and introduce a novel answer to the issue of integrity and consistency of texture in 3D.

The UV mapping optimization in proposed pipeline is indicated in this work does consists of iterated over mesh faces by considering boundary distances in UV space and merging or splitting UV shells according to a user-defined threshold. This operation minimizes or eliminates visible seams and that enforces a more uniform texture layout. After the refined UV layout is determined and the system refined texture data onto the 3D surface by blending color information in overlapping seam regions. While those changes eliminate many patch inconsistencies and some complex cases such as very high-frequency details or large gaps in the mesh of the dataset that may still needs additional refinement. And by consolidating textural alignment at the UV level so the pipeline enhances visual coherence and geometric accuracy in the final rendered result and thereby its suitability to industrial design or cultural heritage application scenarios requiring high-fidelity texture integrity.

4. IMPLEMENTATION

This paper proposes a novel approach in improving the texture quality of 3D meshes through the fusion of advanced methodologies in trimesh processing amd pyrender offscreen rendering also complex neural rendering with NeRF using ViT. It is in this unique fusion that each technique contributes to producing spectacular visual results.

The used dataset for these experiments in this work has comprises primarily a LiDAR-scanned 3D mesh of a room and then extracted a wall from it and the selected due to the intricate surface details by including fine cracks and non-uniform color. Synthetic 2D images as seen in the Figure 5 are rendered from the same large-scale mesh at various angles and simulating a multi-view setup. Approximately 80% of these images are employed to train the NeRF and ViT modules and the remaining 20% are reserved for testing and validation. This strategy allows the model to be inspected from various

angles that enabling the pipeline to learn both extensive surfaces and fine texture details. This dataset while focusing on a single structure realistically reflects real-world texture and shape variations to providing a convincing representation of common architectural or cultural heritage scans.

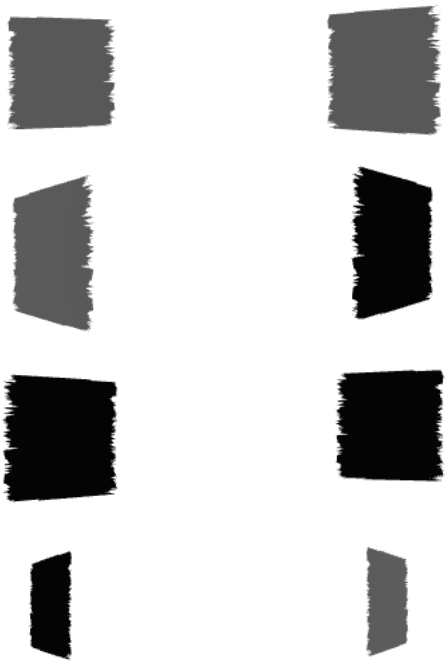

Figure 5. The synthetic 2D images

Preparing the 3D mesh by capturing the object using lidar device as is it shown the Figure 6.

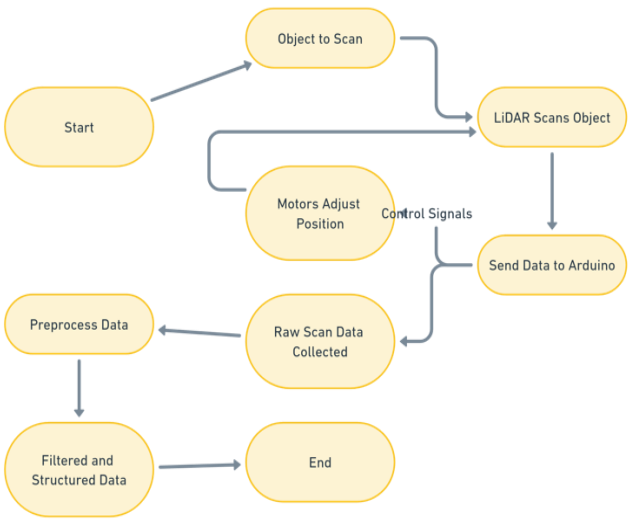

Figure 6. Lidar setup

Firstly, the pipeline imports a 3D object mesh from a specified directory using the trimesh library which can handle many 3D file formats. This step includes testing whether a mesh has been loaded properly before proceeding can reflected in an assertion check that will stop execution if the file is not present. Once loaded the mesh can be visualized by a custom function interfacing with Plotly for interactive three-dimensional charting it can be seen in Figure 1. And it takes the vertices and faces from the mesh constructs a 3D representation and renders it in a web-compatible format. This step is very important for the initial assessment of the geometry of the mesh and also for the identification of any flaws that need to be resolved during the refinement process. It can be seen in Figure 7.

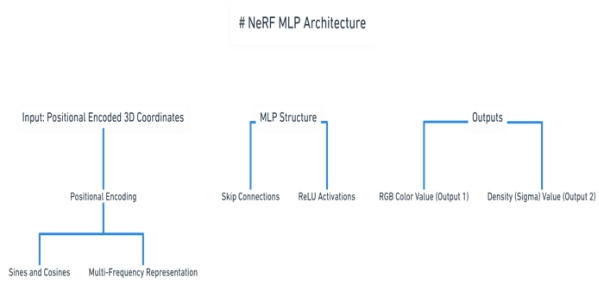

Figure 7. NeRF MLP architecture

Pyrender helps create the final renderings or synthetic views from different angles under different conditions of light and lighting in a manner similar to other methods. One does this in Pyrender by setting up a virtual world with a mesh and a directional light and then does the offscreen rendering with several camera views around it by emulating natural environmental interactions of the object in the scene. These will become very important steps later in the pipeline for the texture details to be consistent across different angles. As it shown in Figure 8.

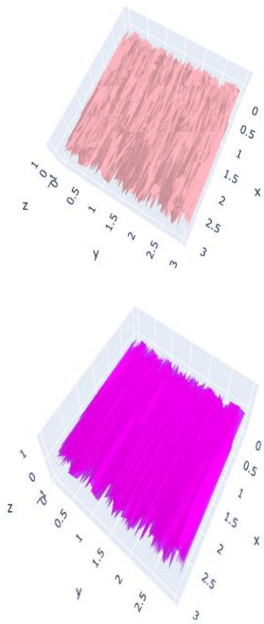

Figure 8. The loaded mesh

The original mesh was captured from lidar scanned room as shown in Figure 9 while Figure 10 is the only wall that has been tested in this work. Figure 11 shows the mesh with the texture on it. Figure 12 shown the enhanced and added points.

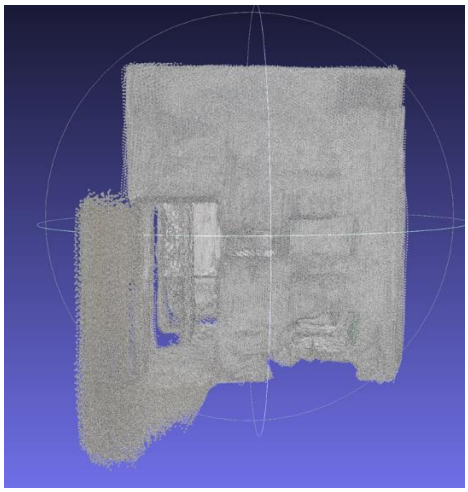

Figure 9. The full scanned room

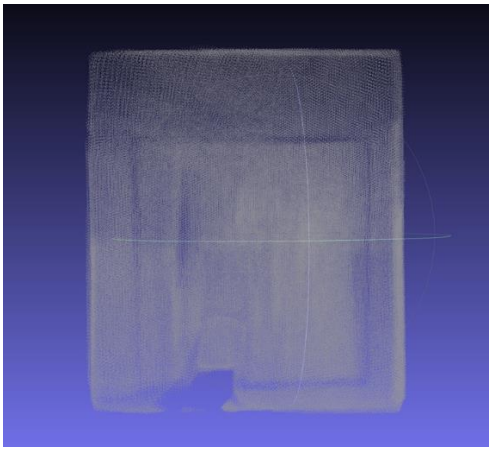

Figure 10. The wall mesh

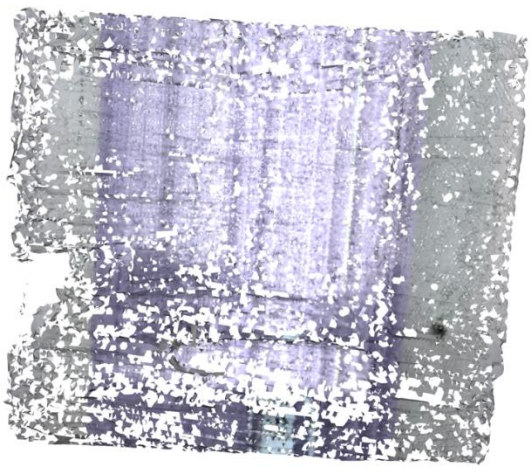

Figure 11. Wall with the texture

Key to the enhancement in texture an improved version of the NeRF uses an MLP with skip connections and positional encoding of the input coordinates which make a prediction about color and density sigma at many points along rays through 3D space and rebuilding an image with highly detailed textures it can be seen in Figure 13. The positional encoding function allows the model to capture high-frequency details through the sinusoidal changes applied to the input coordinates, which efficiently helps the model learn small details across the surface of the mesh.

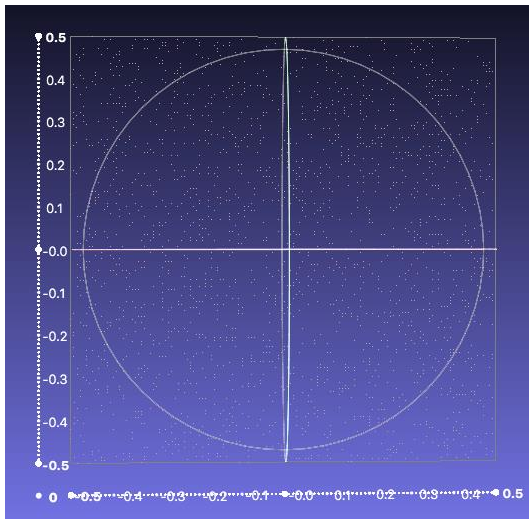

Figure 12. The points that been added to enhance the mesh

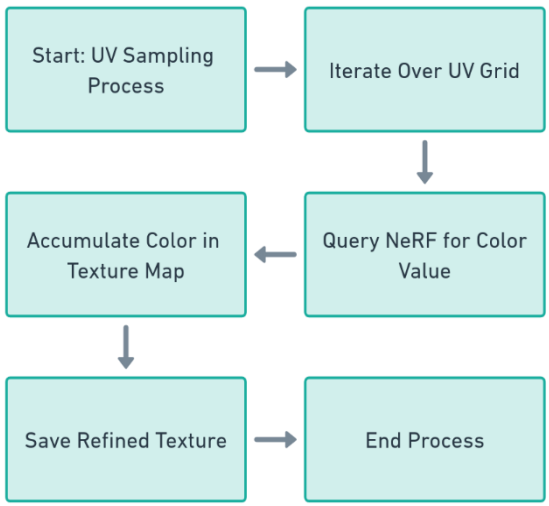

Figure 13. The sampling processes

After that the NeRF-based modification is followed by a Vision Transformer to enhance the consistency of mesh surface texture. So the ViT here processes features extracted from the generated images by using its attention mechanisms to focus on areas that require such textural enhancements. This step will help ensure that the enhanced textures become both a detailed and well-distributed across the entire mesh, hence eliminating any form of visual discontinuity.

The pipeline is completed with the generation of processed textures which are reapplied to the original geometry. and diffusion model can be optionally included to denoise and enhance texture details. The finalized augmented mesh is then exported as a new OBJ file so possibly with its enhanced texture files ready for use in applications requiring high-quality 3D representations.

The pipeline in Figure 14 is unique, as it integrates multiple advanced technologies for 3D mesh augmentation. These technologies were selected based on their specific capabilities in geometric data processing, combining rendering techniques with machine learning-driven texture enhancement. The integration of these technologies increases the visual quality of the meshes and automates the enhancement process, making it scalable for the large datasets used in gaming and virtual reality.

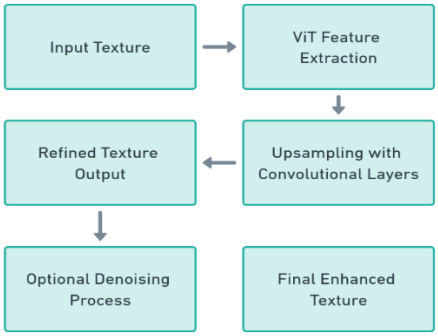

Figure 14. The proposed pipeline

The computational demands of proposed work can be significant especially for densely on meshes. The training time scales linearly with the number of rays sampled in NeRF and the size of the ViT while the utilization of volumetric representations demands large GPU memory. Scaling to larger input images or more transformer layers deep yields higher quality texture but this quality is at the expense of significantly longer convergence times. The work profiling shows that optimizing batch size and utilizing more efficient sampling and hashing-based on accelerations can reduce runtime. And providing actual numbers in terms of average frames per second and total training time on a standard hardware configuration will allow the future workers to assess the feasibility of applying this method for real-time or near-real-time applications.

5. RESULTS

As such both the original and enhanced mesh possess 39,974 vertices with 78,877 faces each, pipeline maintaining the overall topology.

Both meshes have volumes approaching zero, indicating that the geometries are likely thin or open.

The Hausdorff distance between the original and enhanced meshes is 0.181420. whereas the average point-to-mesh distance ranges from 0.065471 to 0.069484. Values are small so which in return means that the geometric integrity has been preserved and modifications were mostly related either for surface texture or minor geometric adjustments.

The Hausdorff distance is measured about 0.181420 which indicates that the pipeline preserves the original mesh geometry with minimal deformation and prove that changes are largely made up of textural improvements. So in this contrast to pure image-based up sampling or editing methods that may unintentionally distort surface geometry by the NeRF–ViT pipeline which restricts most changes to color correction and UV alignments. Such strict keeps the baseline shape to ensures that the output model preserves its geometric fidelity and a valuable consideration in applications like architectural visualization and historic site preservation. So by coupling Hausdorff distance with metrics such as MSE and SSIM provides an overall picture of the fidelity of geometric and textural preservation or enhancement.

MSE= 0.032866 and SSIM = 0.665241 are measures of texture similarities. A lower MSE means that the original and refined textures are better matched in a pixelwise sense. On the other hand a structural similarity above 0.65 means that the appearances of such kind of textures would be moderately similar and with essential spatial characteristics remaining maintained.

These quantitative measurements confirm that the pipeline updates the texture in an efficient way without distorting much in geometry and on the other hand it aligns well the new texture according to the original structural details.

To demonstrate the advantages of the proposed NeRF-ViT pipeline, we conduct a quantitative comparison with two state-of-the-art NeRF-based texture enhancement methods: Mip-NeRF and NeRF-SuperResolution. While those technique was originally tested on its intended dataset so the reported MSE and SSIM values provide approximate benchmarks. This work's pipeline results in a mean squared error of 0.032866 and a structural similarity index of 0.665241 which reflecting lower pixel-level error and higher structural similarity with respect to the baseline data. The performance difference highlights that the volumetric modeling with NeRF has coupled with Vision Transformers of global context extraction and results in superior texture reconstruction and consistent appearance throughout the 3D surface. Also the comparison results illustrate the pipeline's capability in retaining complex features and preserving higher geometric accuracy while improving texturing and making it a suitable choice for real-world scenarios where durability and realism are crucial. Table 1 can show a comparison with other results.

Table 1. Comparison with other works

Method	SSIM (↑)
Proposed (NeRF + ViT)	0.665241
Mip-NeRF [8]	~0.961
NeRF-SR [9]	~0.824

6. DISCUSSION

How well NeRF can be combined with a refinement step using a ViT architecture that leading to the final output that gives detailed texture without compromising on the integrity of the original mesh geometry. Volumetric representations allow for the aggregation of information coming from multiple synthetic viewpoints in NeRF and thus addressing gaps or inconsistencies in the raw texture particularly when this texture is incomplete or of poor quality in some meshes.

One salient observation here is that mesh volume, vertex count, and face count remain unchanged also this would support the method working mostly on a per-vertex or per-texel color and does not change the underlying geometry. And a low Hausdorff distance augurs well and suggests that if the position of vertices was altered then such changes are spatially local and small. MSE and SSIM values suggest an increase in fidelity w.r.t texture. The SSIM value of 0.665241 assures a good degree of similarity and even if it is not perfect. It can suggest that the structural coherence in the refined texture is good. Also artefacts or differences in seams can still appear due to poor UV sampling or in areas which are not adequately covered by synthetic viewpoints.

These results also suggest that designers or content creators within practical applications can efficiently enhance and refine textures of extensive or intricate 3D objects using the pipeline. In future work one might elaborate on further advanced UV-mapping methods such as barycentric interpolation or more advanced diffusion models for denoising and further development of 3D-aware transformers that are able to directly manipulate geometric features.

7. CONCLUSION

This paper proposes a full pipeline for enhancing textures by exploiting multi-view data via NeRF with Vision Transformer-based refinement to improve the consistency of the textures. And the developed approach retains the geometric mesh features as identified by low distance metrics and increases the resolution of the textures with fewer visual artifacts so proven by moderate MSE and SSIM scores. These results constitute further evidence that learning-based approaches can enhance 3D assets for gaming, simulation and digital content creation.

The analysis shows a numerous weakness and recommended pipeline improvements. Too few views constitute a significant limitation especially when the 3D mesh doesn't have rendered views in some areas so the pipeline can produce blurry or imprecise textures. So to solve this problem by using adaptive sampling or intentionally choosing the new viewpoints to guarantee sufficient coverage of complicated surfaces. The computationally demanding nature of NeRF and ViT training constitutes a problem that particularly for big meshes with a high number of polygons. The strategies are to reduce computing expenses that could involve employing a more computationally efficient neural architecture and for instance the lightweight attention mechanisms or hashing-based radiance fields is used to speed up convergence. This work has expectation of good UV mapping thar can leave behind lingering seam artifacts at the intersection of texture patches and recommending the application of sophisticated seam detection or local patch blending. Although those weaknesses may not diminish overall usefulness by recognizing of them sets the stage for future refinements that will enhance efficiency and precision.

Further refinement in this pipeline might be related either to sophisticated UV mapping or adaptive sampling techniques for minimizing artifacts while it far more intrinsic integration with the generative model for advanced realism in the textures. Bu the proposed system represents a tractable and fast approach for improved 3D content production suitable for both high-quality rendering and interactive applications.

REFERENCES

[1] Balusa, B.C., Chatarkar, S.P. (2024). Bridging deep learning & 3D models from 2D images. Journal of the Institution of Engineers (India) Series B. https://doi.org/10.1007/s40031-024-01176-y

[2] Croce, V., Caroti, G., De Luca, L., Piemonte, A., Véron, P. (2023). Neural Radiance Fields (nerf): Review and potential applications to digital cultural heritage. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48: 453-460. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-453-2023

[3] Mittal, A. (2023). Neural radiance fields: Past, present, and future. arXiv preprint arXiv:2304.10050. https://arxiv.org/html/2304.10050v2

[4] Gao, K., Gao, Y., He, H., Lu, D., Xu, L., Li, J. (2022). Nerf: Neural radiance field in 3d vision, a comprehensive review. arXiv preprint arXiv:2210.00379. https://arxiv.org/pdf/2210.00379

[5] Irshad, M.Z., Zakharov, S., Guizilini, V., Gaidon, A., Kira, Z., Ambrus, R. (2024). Nerf-mae: Masked autoencoders for self-supervised 3d representation learning for neural radiance fields. In European Conference on Computer Vision, pp. 434-453. https://arxiv.org/abs/2404.01300

[6] Guizilini, V., Vasiljevic, I., Fang, J., Ambrus, R., Zakharov, S., Sitzmann, V., Gaidon, A. (2023). Delira: Self-supervised depth, light, and radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17935-17945. https://arxiv.org/abs/2304.02797

[7] Tschernezki, V., Laina, I., Larlus, D., Vedaldi, A. (2022). Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In 2022 International Conference on 3D Vision (3DV), pp. 443-453. https://arxiv.org/abs/2209.03494

[8] Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P. (2021). Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF international conference on computer vision, Montreal, QC, Canada, pp. 5855-5864. https://doi.org/10.1109/iccv48922.2021.00580

[9] Huang, X., Li, W., Hu, J., Chen, H., Wang, Y. (2023). RefSR-neRF: Towards high fidelity and super resolution view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8244-8253. https://doi.org/10.1109/CVPR52729.2023.00797

[10] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., Shah, M. (2022). Transformers in Vision: A survey. ACM Computing Surveys, 54(10s): 1-41. https://doi.org/10.1145/3505244

[11] Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y. (2019). Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751. https://doi.org/10.48550/arXiv.1906.07751

[12] Chen, X.J., Zhou, G., Liu, Y.J., Zhang, X., Tang, J. (2024). New view synthesis via multiscale-depth and transformers. In 2024 IEEE 25th China Conference on System Simulation Technology and its Application (CCSSTA), Tianjin, China, pp. 669-673. https://doi.org/10.1109/ccssta62096.2024.10691854

[13] Deng, W., Campbell, D., Sun, C., Kanitkar, S., Shaffer, M.E., Gould, S. (2024). Differentiable Neural Surface Refinement for Modeling Transparent Objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 20268-20277. https://doi.org/10.1109/cvpr52733.2024.01916

[14] Qu, Z., Zhang, Z., Liu, C., Yin, J. (2024). Visual slam with 3D Gaussian primitives and depth priors enabling novel view synthesis. arXiv preprint arXiv:2408.05635. https://doi.org/10.48550/arxiv.2408.05635

[15] Kumar, R., Vats, V. (2024). Few-shot novel view synthesis using depth aware 3D gaussian splatting. arXiv preprint arXiv:2410.11080. https://doi.org/10.48550/arxiv.2410.11080

[16] Lin, Y.C., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y. (2021). inerf: Inverting neural radiance fields for pose estimation. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1323-1330. https://doi.org/10.1109/IROS51168.2021.9636708

[17] Goli, L., Rebain, D., Sabour, S., Garg, A., Tagliasacchi, A. (2023). nerf2nerf: Pairwise registration of neural radiance fields. In 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, pp. 9354-9361. https://doi.org/10.1109/ICRA48891.2023.10160794

[18] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106. https://doi.org/10.1145/3503250

[19] Zhang, Y., Gao, Z., Sun, W., Lu, Y., Zhu, Y. (2024). MD-NeRF: Enhancing large-scale scene rendering and synthesis with hybrid point sampling and adaptive scene decomposition. IEEE Geoscience and Remote Sensing Letters, 21: 6017405. https://doi.org/10.1109/lgrs.2024.3492208

[20] Chen, L., Wang, G., Yuan, L., Wang, K., Deng, K. Torr, P.H.S. (2024). NeRF-VPT: Learning novel view representations with neural radiance fields via view prompt tuning. https://doi.org/10.48550/arxiv.2403.01325

[21] Dong, B., Chen, K., Wang, Z., Yan, M., Gu, J., Sun, X. (2024). MM-NeRF: Large-scale scene representation with multi-resolution hash grid and multi-view priors features. Electronics, 13(5): 844. https://doi.org/10.3390/electronics13050844

[22] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://arxiv.org/abs/2010.11929

[23] Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J. (2024). Reconstructing hands in 3d with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9826-9836. https://doi.org/10.1109/CVPR52733.2024.00938

[24] Azad, R., Kazerouni, A., Azad, B., Khodapanah Aghdam, E., Velichko, Y., Bagci, U., Merhof, D. (2023). Laplacian-former: Overcoming the limitations of vision transformers in local texture detection. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 736-746. https://doi.org/10.1007/978-3-031-43898-1_70

[25] Naseer, M.M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., Yang, M.H. (2021). Intriguing properties of vision transformers. In Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 23296-23308

[26] Li, Y., Kwak, J.G., Ku, B.H., Han, D., Ko, H. (2024). Towards high-fidelity facial UV map generation in real-world. Pattern Recognition Letters, 180: 68-74. https://doi.org/10.1016/j.patrec.2024.02.023

[27] Li, Y., Kwak, J.G., Han, D., Ko, H. (2022). Controllable Face Manipulation and UV Map Generation by Self-supervised Learning. arXiv preprint arXiv:2209.12050. https://doi.org/10.48550/arXiv.2209.12050

[28] Rehman, Z., Mahmood, A., Kang, W. (2024). Pseudo-label refinement for improving self-supervised learning systems. arXiv preprint arXiv:2410.14242. https://doi.org/10.48550/arxiv.2410.14242