Software Engineering and Applications 软件工程与应用, 2025, 14(1), 86-93
Published Online February 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.141009

Hans汉斯

基于跨域引导扩散模型的高保真3D人脸纹理生成

张思状*，成　浩

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年1月21日；录用日期：2025年2月14日；发布日期：2025年2月26日

摘　要

从单张野外图片生成高保真3D人脸纹理是一项具有挑战的工作，现有方法在颜色和光照恢复方面已经取得了显著的进展，但是却仍然无法较好地重建中高频纹理细节。其主要原因在于真实面部UV纹理数据集的匮乏，现有模型大多基于合成UV纹理图训练模型，由于缺少真实标签的监督其与真实UV纹理相比必然存在较大的差异，从而导致模型学习到错误的纹理分布。基于以上思考，我们尝试使用原始图片空间中的细节纹理来引导UV空间中UV纹理图的生成，并提出两阶段训练方式以缓解仅使用合成UV纹理图训练模型带来的个性化细节缺失问题。此外，借助于扩散模型在图像生成任务中的卓越性能，我们还设计了一款跨域引导扩散模型，其将空间域和频率域中的细节信息编码为高级语义条件，用来引导扩散模型的生成过程，从而实现近乎精确的重建。最后，我们将跨域引导扩散模型作为UV纹理生成器嵌入到三维重建框架中，用于重建高保真的3D人脸纹理。实验结果表明本文提出的跨域引导扩散模型能较好地生成中高频纹理细节，并在定量和定性分析中明显优于其他3D人脸纹理生成工作。

关键词

跨域引导扩散，小波变换，3D人脸纹理，UV纹理图

High-Fidelity 3D Facial Texture Generation Based on Cross-Domain Guided Diffusion Model

Sizhuang Zhang*, Hao Cheng

School of Optical-Electrical Computer Engineering, University of Shanghai for Science & Technology, Shanghai

Received: Jan. 21st, 2025; accepted: Feb. 14th, 2025; published: Feb. 26th, 2025

*通讯作者。

文章引用: 张思状, 成浩. 基于跨域引导扩散模型的高保真 3D 人脸纹理生成[J]. 软件工程与应用, 2025, 14(1): 86-93.
DOI: 10.12677/sea.2025.141009

Abstract

Generating high-fidelity 3D facial textures from a single outdoor image is a challenging task. While existing methods have made significant progress in color and lighting recovery, they still struggle to accurately reconstruct mid-to-high frequency texture details. This is primarily due to the lack of real facial UV texture datasets. Most models are trained using synthetic UV texture maps, which inherently differ from real UV textures due to the absence of ground truth supervision, leading to inaccurate texture distribution learning. In light of this, we attempt to use detailed textures from the original image space to guide the generation of UV texture maps in the UV space. We propose a two-stage training approach to alleviate the loss of personalized details caused by training models solely on synthetic UV texture maps. Additionally, leveraging the exceptional performance of diffusion models in image generation tasks, we design a cross-domain guided diffusion model. This model encodes detailed information from both spatial and frequency domains into high-level semantic conditions to guide the diffusion process for near-accurate reconstruction. Finally, we integrate the cross-domain guided diffusion model as a UV texture generator into a 3D reconstruction framework to reconstruct high-fidelity 3D facial textures. Experimental results demonstrate that our proposed cross-domain guided diffusion model effectively generates mid-to-high frequency texture details and significantly outperforms other 3D facial texture generation methods in both quantitative and qualitative analyses.

Keywords

Cross-Domain Guided Diffusion, Wavelet Transform, 3D Face Texture, UV Texture Map

1. 引言

具有高保真纹理的 3D 人脸在许多领域被广泛应用，如人脸识别[1] [2]、面部动画[3] [4]、面部重演[5] [6]、智能医疗[7]等。然而，由于缺乏足够的先验信息，从单张 2D 人脸图像重建高保真的 3D 人脸纹理仍然是一个具有挑战性的问题。传统上，三维人脸纹理通常通过回归 3DMM (3D Morphable Model)系数来获取[8]。然而，这种方法只能概括性地描述纹理的整体特征，难以细致地恢复个性化的纹理细节和光照效果。

为了克服这些局限性，研究人员开始关注超越 3DMM 系数的 UV 纹理图生成任务。UV 纹理图能够提供更加精细的 3D 纹理细节，并更好地反映个体特征和环境光照。现有基于生成对抗网络(GAN)的方法已经取得了显著的结果[9]-[11]，然而 GAN 训练过程中的不稳定性和使用合成数据训练带来的偏差问题会极大地影响其生成精细细节的能力。最近，成等人[12]提出了一种基于扩散模型的 UV 纹理生成方法，其在恢复光照信息和纹理颜色方面取得了较为逼真的效果，但是却仍然无法恢复中高频细节信息，例如法令纹和鱼尾纹等。

究其原因，我们发现由于真实 UV 纹理数据集的匮乏，大多数模型依赖于合成 UV 纹理数据进行训练，而这些合成数据缺乏真实纹理的监督，导致模型无法准确学习到真实 UV 纹理的细节分布，从而在生成过程中出现纹理失真和细节缺失的问题。我们认为在没有真实标签监督的情况下，利用原始图片空间中真实的细节纹理分布来引导 UV 空间中 UV 纹理图的生成，能在一定程度上缓解仅使用合成 UV 纹

理图训练模型所带来的个性化细节缺失问题。此外，鉴于扩散模型在图像生成任务中展现出的卓越性能，我们设计了一种跨域引导扩散模型。该模型将空间域和频率域中的细节信息编码为高级语义条件，用于引导模型的扩散过程，从而实现近乎精确的纹理重建。最后，我们将预训练好的跨域引导扩散模型作为UV纹理生成器，嵌入到FFHQ-UV三维重建框架中[11]，用于高保真3D人脸纹理的重建。

我们的主要贡献如下：

(1) 我们使用原始图片空间中的真实纹理分布来引导UV空间中UV纹理图的生成，并提出两阶段训练方式，缓解了仅使用合成UV纹理数据集训练模型时带来的个性化细节缺失问题。

(2) 引入跨域细节先验特征，将空间域中的像素级信息和频域中不同方向的高频细节信息编码为高级语义条件，用来引导去噪扩散隐式模型(ddim)捕获并重建出更加精细的纹理细节。

2. 相关工作

在三维人脸纹理生成领域，国内外学者进行了广泛的研究。Deep3d使用编码器回归3DMM (3D Morphable Model)系数来恢复三维人脸纹理[8]，但是由于低维参数空间的限制，这种方法只能概括性地描述纹理的整体特征，难以细致地恢复个性化的纹理细节和光照效果。Lee等人提出了一种不确定性感知的编码器和一个结合图卷积神经网络与生成对抗网络的全非线性解码器，以有效地重建高质量的3D人脸形状和纹理[9]。Rai等人提出了一种3D可控生成式人脸模型，通过结合2D生成模型与语义人脸操作，利用可微渲染技术在无3D监督下训练高质量的反照率和精确的3D形状，实现了细节丰富的3D人脸渲染编辑[10]。Bai等人利用StyleGAN实现了多视角标准化人脸图像生成，经过精细化的UV纹理提取、校正和补全过程，生成了高质量的UV映射，并且构建了一个包含超过50,000个高质量纹理UV映射的大规模面部UV纹理数据集[11]。成浩等人提出了一种基于扩散模型的UV纹理生成方法，实现了较为逼真的环境光照信息和颜色信息[12]。

综合来看，虽然已有的研究在颜色和光照的恢复方面取得了一定的成果，但在高保真3D人脸纹理生成，尤其是中高频细节的捕捉方面，依然存在较大的提升空间。因此，本文设计了一款跨域引导扩散模型，专注于从单张2D人脸图像重建高保真的3D人脸纹理，并且取得了具有竞争力的结果。

3. 整体设计

3.1. 模型架构

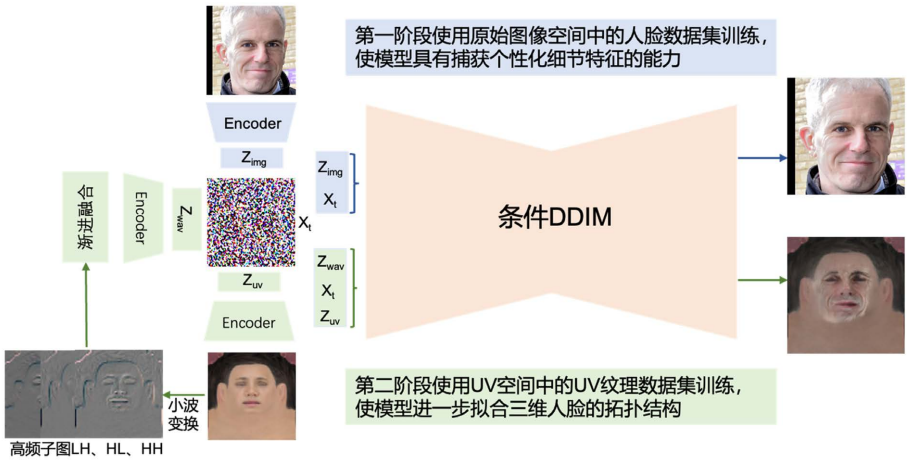

Figure 1. The architecture of cross-domain guided diffusion model
图 1. 跨域引导扩散模型架构

本文提出了一种跨域引导扩散模型，用来从单张野外图片生成具有高保真度的 UV 纹理图，并将模型嵌入到 FFHQ-UV 框架中，以重建高保真 3D 人脸纹理。由图 1 可知，我们主要解决了两个关键问题：1) 使用原始图片空间中的真实纹理分布来引导 UV 空间中 UV 纹理图的生成，具体的实现方法是设计了一种两阶段训练方式，分别使用原始图片空间中的人脸图片和 UV 纹理图训练模型，缓解了仅使用合成 UV 纹理数据集训练模型时带来的个性化细节缺失问题；具体来讲，我们在第一阶段使用原始图像空间中的人脸数据集训练模型，使模型学习到纹理的真实分布，增强模型捕获个性化细节特征的能力；由于第一阶段中模型已经学习到了真实纹理的分布情况，那么在第二阶段中，我们冻结了大部分参数，然后使用 UV 空间中的 UV 纹理数据集训练剩余参数，使模型进一步拟合三维人脸的拓扑结构，以更好的适应现阶段的任务。2) 引入跨域细节先验特征，将空间域中的像素级信息和频域中不同方向的高频细节信息编码为高级语义条件，用来引导扩散模型捕获并重建出更加精细的纹理细节；即将空间域中的原始图片和经过小波分解后的频域信息作为显式的细节引导，并经过 encoder 编码为语义信息作为引导扩散的条件先验，其中空间域中的原始图片提供了像素级别的语义信息，在频率域中我们引入中高频细节语义信息以实现精细细节的扩散引导。下面将详细介绍所提出的解决方法。

3.2. 基于原始图片域中真实纹理引导的 UV 纹理生成

现有开源 UV 纹理数据集大多是学者们使用模型人为创造出来的，由于缺少真实标签的监督其与真实 UV 纹理相比必然存在较大的误差，从而导致模型学习到错误的纹理分布。例如成等人[12]，仅使用人造 UV 纹理数据集训练扩散模型，使得模型无法捕捉到真实纹理中的细微差异，导致模型生成的结果缺乏个性化细节。基于以上观察，我们提出使用原始图片空间中的细节纹理来引导 UV 空间中 UV 纹理图的生成，并设计了一种两阶段的训练方式，分别在原始人脸数据集和合成的 UV 纹理数据集中学习真实纹理的非线性分布以及对应拓扑结构下 UV 纹理图的整体分布。具体来讲，在第一阶段我们借鉴 diffae [13]构建了一个自动编码器，并在 FFHQ 人脸数据集[14]下进行训练，然后将原始图片编码为 512 维度的语义向量 Z_{img}，并将其作为扩散先验条件与随机采样得到的噪声图 X_t 一起输入去噪扩散隐式模型(ddim) [15]执行逆向去噪过程，进而生成由语义向量 Z_{img} 引导生成的人脸图片。我们的条件 DDIM 解码器接受输入 $z=\left(Z_{img}, x_T\right)$ 以生成输出图像，它通过以下逆向(生成)过程建模 $p_\theta\left(x_{0:T}\middle|Z_{img}\right)$，具体的参数设置和公式推导请参考 diffae [13]：

$$p_\theta\left(x_{0:T}\middle|Z_{img}\right)=p\left(x_T\right)\prod_{t=1}^{T}p_\theta\left(x_{t-1}\middle|x_t, Z_{img}\right) \tag{1}$$

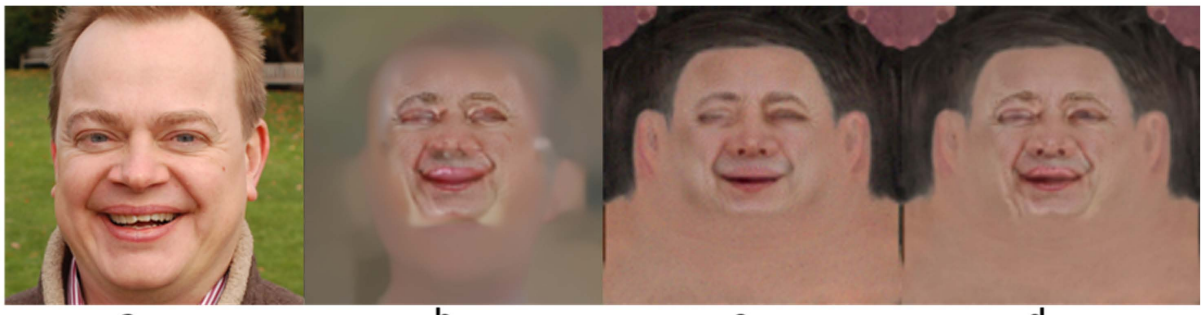

Figure 2. The architecture of cascaded dilated convolutional network
图 2. 不同训练策略下生成的 UV 纹理图之间的比较

然后我们将训练好的 ddim 嵌入到 FFHQ-UV 框架中作为 UV 纹理生成器得到结果图 2(b)，我们惊奇

的发现虽然生成的结果无法很好的拟合三维人脸的 UV 拓扑结构，但是却能捕捉并生成更为精细且个性化的细节特征，且这种细节特征与仅使用人造 UV 纹理数据训练的 UV 纹理生成器所得到的结果(图 2(c))相比具有明显的优越性。鉴于以上实验结果的观察，我们引入了第二阶段的训练。具体来讲，我们冻结第一阶段中的部分参数，并使用人造 UV 纹理数据集训练剩余的参数，以期在保留模型生成精细细节的同时进一步拟合对应三维人脸拓扑结构下的 UV 纹理分布。最终结果如图 2(d)所示，其既保留了捕捉和生成精细纹理的能力又能较好的拟合三维人脸的 UV 拓扑结构，有效的证明了原始图片空间中的细节纹理能较好的引导 UV 空间中精细纹理的生成。

3.3. 跨域细节先验引导扩散的 UV 纹理生成

Figure 3. Visual effects comparison
图 3. 视觉效果比较

由于训练资源和可用数据集的限制，我们无法使用较高分辨率(例如 1024*1024 或者 512*512)的数据集训练所提出的模型，导致模型无法很好的捕捉和生成中高频纹理细节，这极大的影响了 UV 纹理图的保真度，所以需要输入更多的中高频细节特征增强模型对精细细节的敏感度。我们设计了一种跨域信息提取模块，可提取不同目标域(空间域和频域)中的细节特征，进而引导更加精细的细节生成。如图 1 所示，基于空间域中像素级信息的引导，我们将图片输入到 encoder 中，并将其编码为 512 维度的语义向量 Z_{uv}。基于频率域中不同方向高频细节信息的引导，我们将小波变换嵌入网络结构，并提取三个方向的高频子图作为高频细节信息，并且使用 sumpooling 将三个高频子图融合为原始高频信息，然后我们使用 encoder 将其编码为 512 维度的语义向量 Z_{wav}。对于提取到的细节语义信息 Z_{uv} 和 Z_{wav}，我们将其作为扩

散先验条件与随机采样得到的噪声图 X_t 一起输入 ddim 执行逆向去噪过程，进而生成由跨域细节先验引导生成的高保真 UV 纹理图。我们的条件 DDIM 解码器接受输($z=\left(Z_{uv},Z_{wav},x_T\right)$ 以生成输出图像，它通过以下逆向(生成)过程建模 $p_\theta\left(x_{0:T}\middle|Z_{uv},Z_{wav}\right)$，具体的参数设置和公式推导请参考 diffae [13]：

$$p_\theta\left(x_{0:T}\middle|Z_{uv},Z_{wav}\right)=p\left(x_T\right)\prod_{t=1}^{T}p_\theta\left(x_{t-1}\middle|x_t,Z_{uv},Z_{wav}\right) \tag{2}$$

4. 实验设计与验证

4.1. 实施细节

Figure 4. Visual effects comparison
图 4. 视觉效果比较

我们的模型在两个公共数据集上进行了训练，分别为原始图片域中的 2D 人脸图片数据集 FFHQ [14] 和基于此数据集人为合成的 UV 空间中的 UV 面部纹理数据集 FFHQ-UV [11]。其中，第一阶段使用 FFHQ 数据集进行训练，将其中 80%的数据作为训练集，其余作为测试集便于后续测试实验结果，我们采用和 diffae 相同的参数设置，batch_size 为 16，时间步长设置为 T = 15，epoch 为 1000，学习率为 0.1，采用

ADAM 优化器进行优化；第二阶段使用 FFHQ-UV 数据集进行训练，为了尽可能保留模型对精细细节的捕捉和生成能力，我们冻结部分参数后继续训练，并且降低了迭代次数，此时的 epoch 设置为 100，其他参数不变。我们所有的训练过程均在配备了 NVIDIA 3090 GPU 的计算机上进行。

4.2. 定性分析

我们将所提出的方法与目前最先进的方法进行了比较，包括 FFHQ-UV [11]和成等人提出的最新工作[12]。为了进行公平的比较，我们使用了他们公开发布的模型和代码，并在 FFHQ 测试集上进行了实验。图 3 和图 4 展示了基于单张 2D 人脸图像重建的 3D 人脸纹理的比较结果。在图 3 中，a 为原始人脸图片，d 为使用本文方法的重建结果，其中 d3 为生成的 UV 纹理图，d2 为渲染后的三维人脸，d1 为将渲染后的三维人脸投影到二维图片中的结果展示，用于和原始输入图片进行更为直观的视觉对比。由图可以得出，我们的方法相较于 FFHQ-UV (图 3b)和成等人的方法(图 3c)，显著的捕捉和生成了具有中高频细节的 UV 纹理图和 3D 面部纹理，特别是在恢复一些精细细节方面，我们的方法明显优于现有的 SOTA 方法。另外，由于引入了高频细节特征，我们的模型增强了对高光信息等高频细节的感知能力，使得我们的模型生成了更加逼真的光照效果，例如图 4 第四行中 d1 和 d2 的额头部分，我们的结果更加接近输入图像中的光照结果。

4.3. 定量分析

为了评价人脸纹理的重建精度，我们使用 SSIM [16]、PSNR 和 LPIPS [17]作为评价指标，与最先进的方法进行比较。为了进行公平的比较，我们在 FFHQ 测试集中随机挑选出 300 张无遮挡的人脸图片用于定量评估，并使用其公开发布的模型和代码计算相关分数。表 1 展示了在 FFHQ 测试集上进行纹理重建的定量比较结果。可以观察到，我们的方法在所有指标中均处于领先地位，这进一步证明了所提方法在高保真 3D 面部纹理生成方面的优越性能。

Table 1. Comparison of the proposed model with other methods
表 1. 所提出的模型与其他方法对比

Method	SSIM ↑	PSNR ↑	LPIPS ↓
FFHQ-UV (CVPR 2023)	0.8892	34.7146	0.0441
Cheng 等(2024)	0.9320	36.4925	0.0368
Ours	0.9412	37.1963	0.0287

5. 总结

本文提出了一种基于跨域引导扩散模型的高保真 3D 人脸纹理生成方法。通过在空间域和频率域中利用细节信息，模型在生成中高频纹理细节方面表现出色。利用原始图片域中的真实纹理分布来引导 UV 纹理生成，并结合两阶段训练策略，有效缓解了合成 UV 纹理图时个性化细节缺失的问题。实验结果显示，该方法在定量和定性分析中均优于现有方法，尤其是在纹理细节和光照恢复方面表现突出。

受到扩散模型固有局限性的影响，本文所提方法的生成速度较慢，因此我们计划进一步优化模型的效率，并探索更多的细节恢复策略，以提高 3D 人脸纹理重建的质量和应用范围。

参考文献

[1] Liu, F., Zhu, R., Zeng, D., Zhao, Q. and Liu, X. (2018) Disentangling Features in 3D Face Shapes for Joint Face Reconstruction and Recognition. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City,

18-23 June 2018, 5216-5225. https://doi.org/10.1109/cvpr.2018.00547

[2] Liu, F., Zhao, Q., Liu, X. and Zeng, D. (2020) Joint Face Alignment and 3D Face Reconstruction with Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 664-678. https://doi.org/10.1109/tpami.2018.2885995

[3] Cao, C., Wu, H., Weng, Y., Shao, T. and Zhou, K. (2016) Real-Time Facial Animation with Image-Based Dynamic Avatars. *ACM Transactions on Graphics*, **35**, 1-12. https://doi.org/10.1145/2897824.2925873

[4] Cao, C., Bradley, D., Zhou, K. and Beeler, T. (2015) Real-Time High-Fidelity Facial Performance Capture. *ACM Transactions on Graphics*, **34**, 1-9. https://doi.org/10.1145/2766943

[5] Chaudhuri, B., Vesdapunt, N. and Wang, B. (2019) Joint Face Detection and Facial Motion Retargeting for Multiple Faces. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 9711-9720. https://doi.org/10.1109/cvpr.2019.00995

[6] Chaudhuri, B., Vesdapunt, N., Shapiro, L. and Wang, B. (2020) Personalized Face Modeling for Improved Face Reconstruction and Motion Retargeting. *Computer Vision—ECCV* 2020, Glasgow, 23-28 August 2020, 142-160. https://doi.org/10.1007/978-3-030-58558-7_9

[7] Tu, L., Porras, A.R., Morales, A., Perez, D.A., Piella, G., Sukno, F., *et al*. (2019) Three-Dimensional Face Reconstruction from Uncalibrated Photographs: Application to Early Detection of Genetic Syndromes. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*, Shenzhen, 17 October 2019, 182-189. https://doi.org/10.1007/978-3-030-32689-0_19

[8] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y. and Tong, X. (2019) Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Long Beach, 16-17 June 2019, 285-295. https://doi.org/10.1109/cvprw.2019.00038

[9] Lee, G. and Lee, S. (2020) Uncertainty-Aware Mesh Decoder for High Fidelity 3D Face Reconstruction. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 6099-6108. https://doi.org/10.1109/cvpr42600.2020.00614

[10] Rai, A., Gupta, H., Pandey, A., Carrasco, F.V., Jason Takagi, S., Aubel, A., *et al*. (2024) Towards Realistic Generative 3D Face Models. 2024 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2024, 3726-3736. https://doi.org/10.1109/wacv57701.2024.00370

[11] Bai, H., Kang, D., Zhang, H., Pan, J. and Bao, L. (2023) FFHQ-UV: Normalized Facial UV-Texture Dataset for 3D Face Reconstruction. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 362-371. https://doi.org/10.1109/cvpr52729.2023.00043

[12] Cheng, H., Hui, Y., Jin, H. and Zhang, S. (2024) High-Fidelity Texture Generation for 3D Avatar Based on the Diffusion Model. 2024 16*th International Conference on Human System Interaction* (*HSI*), Paris, 8-11 July 2024, 1-6. https://doi.org/10.1109/hsi61632.2024.10613538

[13] Preechakul, K., Chatthee, N., Wizadwongsa, S. and Suwajanakorn, S. (2022) Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 10609-10619. https://doi.org/10.1109/cvpr52688.2022.01036

[14] Karras, T., Laine, S. and Aila, T. (2019) A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4396-4405. https://doi.org/10.1109/cvpr.2019.00453

[15] Song, J., Meng, C. and Ermon, S. (2020) Denoising Diffusion Implicit Models. arXiv: 2010.02502. https://doi.org/10.48550/arXiv.2010.02502

[16] Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/tip.2003.819861

[17] Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O. (2018) The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 586-595. https://doi.org/10.1109/cvpr.2018.00068