Textured Mesh Quality Assessment: Large-scale Dataset and Deep Learning-based Quality Metric 复现报告

摘要

常见的数字资产包括文本、图片、音频和视频等，近年来，随着三维重建技术的快速发展，获得高质量高精度的 3D 模型已经成为一件非常容易的事情，越来越多的创作者可以通过现有的三维重建方法获得自己的 3D 模型，甚至出现了类似于“文生图”的方式，通过一段提示词就能获得高精度的模型，建模效率大大提升，这也让 3D 模型的规模和复杂性呈爆炸式增长，对网络实时传输造成很大的负担。因为某些应用程序和设备的约束，所以需要对高精度模型进行简化和或有损压缩，这可能会降低其视觉质量。为了确保最佳体验质量，必须评估视觉质量以准确驱动压缩并在视觉质量和数据大小之间找到正确的折中方案，于是本文主要专注于纹理的 3D 网格的主观和客观质量评估，作者首先建立了一个大规模数据集，该数据集包含超过 343k 个扭曲失真的模型，收集了 4,500 多名参与者的 148,929 个主观质量判断。作者还利用这个数据集，提出了一种基于学习的 3D 图形质量指标。

关键词： 网格；三维模型；质量评估

1 引言

在现实生活中，我们有很多种方式获得 3D 模型，这些模型在建筑设计、影视特效、游戏开发以及虚拟现实等多个领域都发挥着重要作用。传统的建模方法包括手工建模和扫描技术，手工建模因为依赖于人工，一般是通过使用专业软件如 Blender、Maya 或 3ds Max 来创建精细的三维结构，所以速度较慢。随着采集技术的成熟，现在更多的使用 3D 扫描技术，这种方法利用激光扫描或结构光等技术，将物体表面的细节转化为数字数据，从而生成高保真的三维模型，通过 3D 扫描，我们可以从现实世界中获取对象的精确几何形状。此外，随着计算机视觉和深度学习的快速发展，基于图像的建模方法也逐渐兴起，通过分析多张二维图像来重建三维场景，这种方法不仅提高了建模的效率，也降低了对专业技术的依赖，获取高质量高精度的模型的门槛越来越低，方式越来越多，过程越来越简单。无论是通过手工建模、扫描还是自动化生成，3D 模型的获取方式多样化，使得设计师和艺术家能够在各种应用中实现创意的可视化，推动了各行业的数字化转型和创新发展。

过去几年计算机的计算能力开始出现爆炸式的增长，之前被大家广为接受的一个观点是网格越细，产生的结果可能越准确 [4]，这也导致生成的网格的元素数量成倍增加，然而仅仅

增加元素的数量并不足以保证准确的结果，这种增加可能只能带来很小的视觉效果提升 [6]，。为了降低 3D 内容的复杂性，避免因传输而导致不必要的延迟，对模型进行简化和压缩是不可避免的，这一点类似于视频的传输。这些操作减少了数据量的大小，降低了处理、存储和传输的成本的同时，也不可避免地导致了用户体验质量 (QoE) 的下降。因此，我们必须找到数据量大小与 QoE 的折中方案。感知质量可以使用主观研究和客观指标进行评估。[5]

基于以上背景，作者提出了 3D 图形的公共质量评估指标和数据集，该数据集填补了带有颜色和纹理的网格数据集的空白。该数据集包括通过 55 个源模型的压缩生成的超过 343k 扭曲失真的网格，作者从这 343k 个失真网格中选取了 3000 个具有挑战性的模型，并且在 4500 名参与者的大规模众包主观实验中对这些扭曲模型进行了注释。该数据集是迄今为止与主观评分相关的最大的纹理网格质量评估数据集。然后作者基于注释后的数据集，提出了一种基于图像的 3D 图形深度学习质量指标，称为 Graphics-Learned Perceptual Image Patch Similarity (LPIPS)。LPIPS 对渲染生成图片进行处理，将图片切分成小块，然后将这些小块输入到卷积神经网络中，网络在顶部学习权重以提取特征。接着，这些特征被融合和汇总，以预测每个小块的质量，图片的整体质量评分即对局部小块质量计算平均值得出。作者提出的 Graphics-LPIPS 指标在纹理网格数据集上优于其他图像质量评估指标。

2 相关工作

在本节中，将简要介绍以往的 3D 模型质量评估方法，包括网格、体素和点云等，而我复现的这篇文章主要专注于带有颜色和纹理属性的网格形式存储的 3D 模型。3D 领域的质量评估一般是由 2D 领域发展推广而来，所以 3D 质量评估与图片质量评估 (IQA) 高度相关，关于 IQA 的内容也会有提及。

2.1 主观质量评估

理想情况下，图像的视觉质量通过在受控环境中进行的主观实验来进行评估，实验者可以是专家，也可以是普通用户，通过他们的主观判断获得平均意见分数 MOS。MOS 是对图像感知质量的直接衡量，对于现有技术的改进具有重要的指导意义 [2]。主观质量评估的方法可以分为几种类型，包括绝对质量评估和相对质量评估。在绝对质量评估中，观察者对单张图像的质量进行评分，而在相对质量评估中，观察者则比较两张或多张图像，判断哪一张质量更好。常见的评分尺度包括五分制或十分制，这些尺度可以帮助量化观察者的感知。

主观评估直接基于人类的视觉反馈，具有较高的准确性，但其实施成本较高且耗时较长，因此在现实应用中通常还是使用客观质量评估。此外，观察者的个体差异和评估环境的变化也可能影响结果的可靠性。为了克服这些问题，研究者们逐渐引入了更为系统化的评估方法，例如使用标准化的评估协议和多样化的观察者群体。

2.2 客观质量评估

传统图像质量评估: 图片质量评估经历了多个阶段的演变，从早期简单化的方法过渡到如今复杂的模型，逐渐形成了系统的评估框架。峰值信噪比 (PSNR) 是最早用于图片质量评估的指标之一，用于计算重建视频与原始视频之间的最大可能误差。尽管 PSNR 很简单，但它具有显着的缺点，因为 PSNR 指标假设像素独立，所以它不足以评估图像的高层次结构，无

法充分反映人类视觉系统 (HVS) 感知的视觉内容的感知质量。随着对 HVS 的深入了解，许多感知驱动的距离度量方法被提出，例如 SSIM [7]，MSSIM [8]，FSIM [9] 等. 其中比较经典的是于 2004 年被引入的结构相似度指数 (SSIM)。SSIM 旨在通过测量亮度、对比度和结构信息的相似性来评估图片质量。SSIM 取得了显著的成功，并成为后续研究的基石。

基于学习的质量评估指标： LPIPS [10] 提出是因为在当时广泛使用的 L2/PSNR、SSIM、FSIM 指标在判断图片的感知相似度时给出了与人类感知相违背的结论，也就是说，这些经典方法在某些情况下无法真实反映人类观察图像时的感觉和判断。而相比之下，基于学习的感知相似度度量要更符合人类的感知, 因为它们不仅仅依赖于低层次的像素信息，而是提取了更高层次的特征。LPIPS 提供了一种更为准确和有效的工具，用于衡量图像之间的感知相似度，进而推动了图像处理领域的发展。

2.3 网格的质量评估

与图像相比，3D 内容的质量评估领域，特别是那些具有颜色属性（以纹理图或顶点/点颜色的形式）的领域，仍然可以被认为处于早期阶段。

在早期，与 3D 模型相关的主观质量测试最早是在网格上进行的，更准确地说是在不带颜色属性的纯几何体模型上，这种测试可以评估简化、平滑、水印和压缩所引起的伪影。涉及点云的研究在近年来也变得普遍，早期是针对于无色点云，之后也出现了彩色点云的质量评估。虽然不带颜色属性的网格质量评估开始得很早，但是带颜色属性的网格的研究较少。这些研究尝试了在 IQA 中常用的方法，包括上文提到的绝对质量评估和相对质量评估，一些研究 [1] 指出两个模型之间进行比较的方法最适合用于评估 3D 图形的质量。

对于客观质量评估来说，人们提出了许多基于图像的方法来评估 3D 数据的质量，所以网格质量评估整体滞后于 IQA，也经历了从纯几何距离到感知驱动距离的过程。然后随着机器学习的兴起，卷积神经网络（CNN）被研究并推广用于评估网格和点云的质量，许多研究者使用 CNN 从多个视图中提取特征并得出分数。目前来说，对于 3D 内容，开发基于深度学习的质量指标仍然具有挑战性，主要是由于缺乏大型且丰富的 3D 对象数据集，特别是那些具有颜色属性的数据集。

3 本文方法

3.1 本文方法概述

正如在相关工作部分所说，目前缺乏针对具有颜色属性的 3D 图形的质量指标，尤其是基于深度学习的方法。所以作者基于他们创建的大规模纹理网格数据集（其中 3000 个样本进行了主观评分），提出了一个深度学习质量指标。他们参考了 LPIPS 指标 [10]，LPIPS 指标利用深度神经网络评估两个图像补丁之间的感知相似性，作者将其适配至 3D 数据，然后利用他们的数据集进行了训练。选择 LPIPS 的原因在于其在许多场景下都表现良好 [3], 是现在广泛使用的基于学习的评价指标。这种方法相当于提取了更高层面的特征而不仅仅停留在像素上。

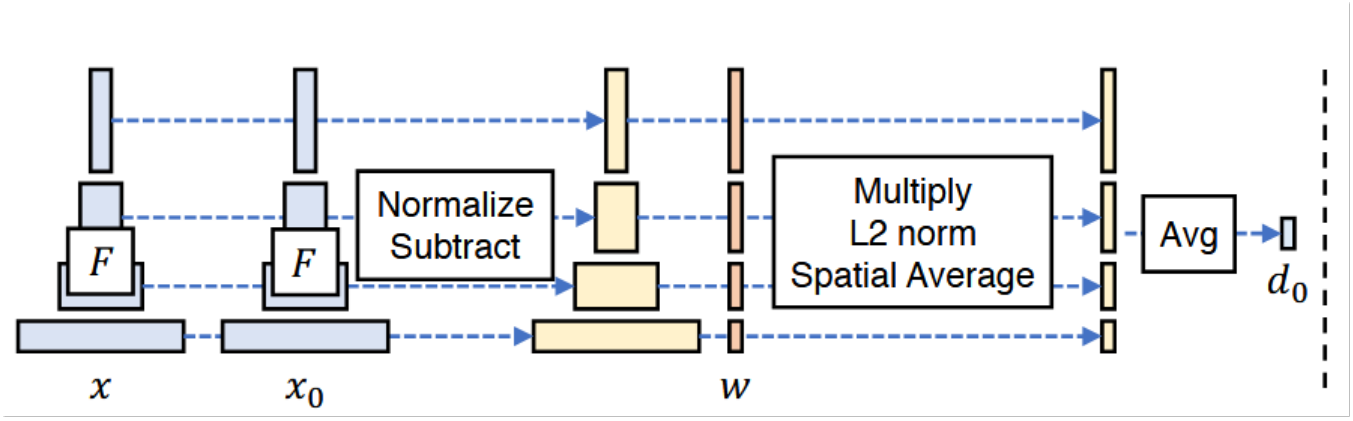

图 1. 作者改进后的 LPIPS 网络结构示意图，将两个补丁 x 和 x_0 输入给定的网络 F, 将提取到的特征使用 L2 范数进行空间平均，最终输出距离值 d_0

3.2 特征提取模块

LPIPS 的基本原理是使用预先训练的神经网络从两个图像补丁 x 和 x_0 中提取深层特征，将两幅图像输入到两个具有相同权重的 CNN 网络 F 中，F 可以是简单的卷积神经网络 (Alex,VGG 等)，将得出的特征向量进行比较，通过 w 层给与权重，降维后得到最终的结果，这个过程将两个块之间的特征差异 $(F(x) - F(x_0))^2$ 映射到了感知差异。与原始的 LPIPS 相比，Graphics-LPIPS 添加了一个权重为 w_0 的 1×1 卷积层，这使得模型的校准能力更强。最后，将一张图片所有补丁获得的分数 $d(x, x_0)$ 相加求均值，即选择简单平均池化来获得每个图像的最终分数, 整体图像质量 $\hat{Q}_I$ 计算为：

$$\hat{Q}_I = \frac{1}{N}\sum_{i=1}^{N} \text{LPIPS}(X_i, \tilde{X}_i), \tag{1}$$

其中 N 代表图片切分成的补丁数量，X 代表参考图像，$\tilde{X}$ 代表失真图像。

Graphics-LPIPS 的网络架构如图 1 所示。根据 LPIPS 作者的建议，Graphics-LPIPS 使用具有固定权重的预训练 AlexNet 作为特征提取器，并且仅学习顶部卷积层的权重 w 和 w_0。

3.3 损失函数定义

Graphics-LPIPS 采用和 LPIPS 相同的损失函数，训练网络的损失函数计算方式如下：

$$E_I = (\hat{Q}_I - MOS_I)^2, \tag{2}$$

其中 E_I 是针对一副图像的损失，反映了模型预测的质量评分与实际评分之间的差距。

4 复现细节

4.1 与已有开源代码对比

Graphics-LPIPS 代码基于 LPIPS, 数据集和源代码都是开源的, 项目地址 https://github.com/MEPP-team/Graphics-LPIPS. 原文最终的网络是使用预训练 AlexNet 提取特征的 LPIPS

网络。在源代码中，作者给出了基于 AlexNet、VGG 和 Squeeze 提取特征的 LPIPS 网络，我在源代码的基础上，补充了 ResNet18 的代码，并且采用预训练的 ResNet18 进行特征提取，与原文的实验结果进行比较。我并没有使用更复杂的网络，因为 LPIPS 的效果与作为特征提取器的网络的深度没有必然联系。

4.2 实验环境搭建与复现

环境配置：实验环境为 windows10，cuda11.3，python 环境配置：pytorch=1.12.0,torchvision=0.13.0,numpy=1.25.0,scipy=1.13.1,scikit-image=0.24.0,opencv-python=4.10.0. 以上环境配置均符合作者给出的要求。

训练：原作者的最终模型训练了 10 个轮次，前 5 个周期使用初始学习率 10^{-4}，后 5 个周期采用线性衰减，每个批次包含 4 幅图像 (图像由 3D 模型主视角渲染生成)，每幅图像由 150 个补丁组成，每个补丁的大小为 64 × 64，补丁中保证包括实际模型而不是空白背景，对于补丁数量少于 150 的图像，将补丁重复，直到达到这个数量。而我修改后的基于 ResNet 的 LPIPS，初始学习率选择 4×10^{-4}，其他参数不变的情况下训练 14 个轮次。运行特定的 python 文件即可开始训练，控制台中会实时输出训练情况以供参考，如图 2 所示。

测试：原论文中采用五折交叉验证，每个折叠包含从 11 个源模型获得的约 600 个扭曲模型，最终根据相关性评估了指标的性能，相关性度量选择了皮尔逊线性相关性 (PLCC) 和斯皮尔曼等级相关性 (SROCC)，计算原 MOS 与预测的 MOS 之间的相关性。原文中还就分类能力进行了评估。

```
-------------------------------------------------------------------------------------------
SHUFFLINGGGGGGGGGGG!!!
dataroot: ['.\\dataset\\Trainset_shuffled_1.csv']
Column names are Model, stimulus, MOS, VP1, VP2, VP3, VP4, TotalPatches
Processed 2398 lines (distorted stimuli).
Total nb of patches to load: 359550.0
These patches correspond to 2397 stimuli (with repetions) = 2397 optimizations
nb of stimulusID to load 2397.0
saving the model at the end of epoch 2, iters 720000
nb batch 600.0
Epoch Loss 0.029672
----------------------------------------TEST----------------------------------------
./dataset/TexturedDB_20%_TestList_withnbPatchesPerVP_threth0.6.csv: 100%|████████████████████
Testset Total 90450.000
Testset val step = nb batches =  151.000
Testset Loss 0.026
Testset MSE 0.026
SROCC 0.830
End of epoch 2 / 10      Time Taken: 261 sec
-------------------------------------------------------------------------------------------
```

图 2. 训练过程演示

5 实验结果分析

作者在已有的开源代码中给出了训练好的模型，使用该模型得出的训练结果如图 3 所示，复现结果接近于原文数据。在此基础上，我使用源代码中的 VGG 网络重新训练得出的结果和我添加的 ResNet 网络的结果如图 4 所示，MSE 整体趋势相近，所以只给出了训练过程中

SROOC 的变化趋势。对训练好的网络与原文和复现的结果进行比较，结果如表 1 所示，其中 Graphics-LPIPS 是原文宣称的最好结果，AlexNet 是对原文的复现，VGG 是源代码中包含的模型，ResNet 是本人修改的模型。可以看到，AlexNet 的结果实际上与原文的结果完全一致，与原文中最好结果有一定的差距是因为原实验在渲染图片时还施加了其他的条件（如材质和光照），结果会随着条件的改变而改变，最好结果也是在这个基础上产生的，但是复现实验并未复现这些条件，所以只获得了原始条件下的结果，当然，这个结果与原文是一致的。同样的，在不考虑其他条件的情况下，使用 ResNet18 作为特征提取器的 Graphics-LPIPS 在 PLCC 和 SROCC 上都超过了原指标。而基于 VGG 的网络无论是训练还是测试时的结果都不是很好，猜测可能与超参数设置有关。

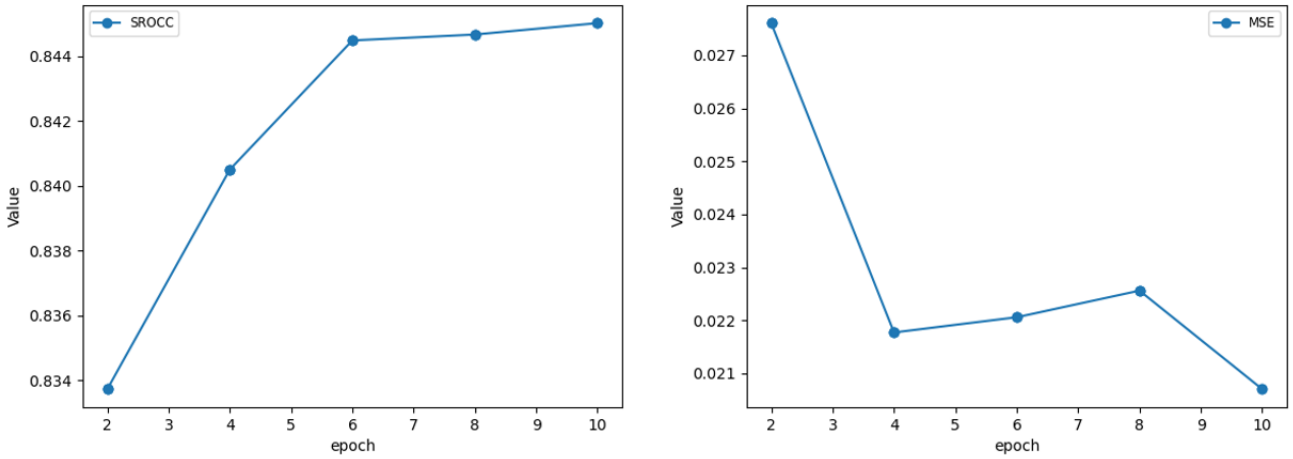

图 3. 复现实验的训练过程

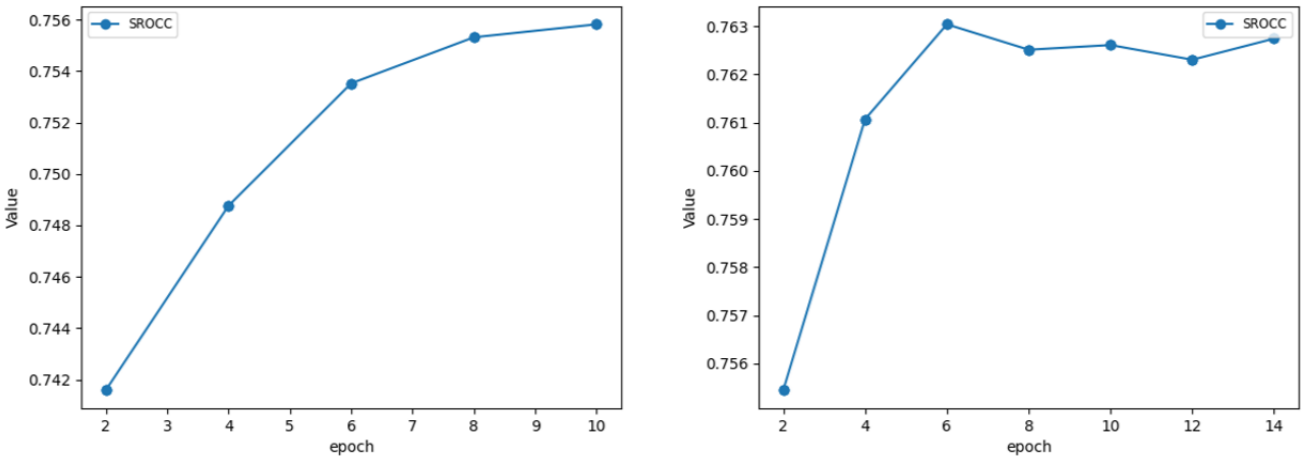

图 4. VGG（左）和 ResNet（右）的训练过程

	Graphics-LPIPS	**AlexNet**	**VGG16**	**ResNet18**
PLCC	0.89	0.854	0.808	0.870
SROCC	0.88	0.844	0.791	0.852

表 1. 测试结果

6 总结与展望

Graphics-LPIPS 的复现相对于很多三维相关的项目来说并不复杂，文章中还有很多关于数据集的实验和分析，因为不属于复现的内容，我并没有完全展示。LPIPS 也是一个现在正在广泛使用的方法，而我的改进也是站在前人肩膀上的简单的尝试。因为这篇文章的主要工作内容是带颜色的网格数据集的制作与主观实验，所以以计算机视觉、深度学习的角度来看，本文对 LPIPS 的改进并不复杂，甚至可以说很简单。Graphics-LPIPS 的实验设计也有一些不严谨的地方，比如主观实验中，受测者观察的是经过 Unity 渲染的视频，所以最终的评分结果可能受到渲染器质量的影响；还有文章考虑了不同条件（光照和材质等）下的 Graphics-LPIPS 的效果，但是与预测 MOS 进行对比的仍然是没有施加这些条件原 MOS，虽然理论上一个物体的质量是不会因为不同的渲染条件而改变的，但是在实际应用中，MOS 是有可能受到视角、光照和材质的影响的。

LPIPS 已经是六年前的方法了，现在也出现了很多关于 LPIPS 的改进算法，在 LPIPS 的基础上训练一个可以用于网格的评价方法也并不困难，所以本文对我最大的启发还是来源于这个高质量的网格数据集，我们可以用它来进行很多和三维表面质量评估有关的工作，而这也是我选择复现并且深入解读这篇文章的原因。同时，这篇文章在数据集制作流程上非常严谨，如果在未来我们需要制作一个数据集，无论是 2D 还是 3D，或者是用于任何研究方向，我相信这篇文章的研究思路与研究方法都将是一个非常有价值的参考。

参考文献

[1] Evangelos Alexiou and Touradj Ebrahimi. On the performance of metrics to predict quality in point cloud representations. In *Applications of Digital Image Processing XL*, volume 10396, pages 282–297. SPIE, 2017.

[2] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

[3] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] Ming-Jun Lai and George Slavov. On recursive refinement of convex polygons. *Computer Aided Geometric Design*, 45:83–90, 2016.

[5] Yana Nehmé, Johanna Delanoy, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. Textured mesh quality assessment: Large-scale dataset and deep learning-based quality metric. *ACM Transactions on Graphics*, 42(3):1–20, 2023.

[6] Tommaso Sorgente, Silvia Biasotti, Gianmarco Manzini, and Michela Spagnuolo. A survey of indicators for mesh quality assessment. In *Computer graphics forum*, volume 42, pages 461–483. Wiley Online Library, 2023.

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[8] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[9] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[10] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.