3D Surface Super-resolution from Enhanced 2D Normal Images: A Multimodal-driven Variational AutoEncoder Approach

Wuyuan Xie[1] , **Tengcong Huang**[1] and **Miaohui Wang**[1,2,3*]
[1]Shenzhen University, Guangdong Key Laboratory of Intelligent Information Processing
[2]State Key Laboratory of Radio Frequency Heterogeneous Integration (Shenzhen University)
[3]Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS)
{wuyuan.xie, tengconghuangcs, wang.miaohui}@gmail.com

Abstract

3D surface super-resolution is an important technical tool in *virtual reality*, and it is also a research hotspot in *computer vision*. Due to the unstructured and irregular nature of 3D object data, it is usually difficult to obtain high-quality surface details and geometry textures via a low-cost hardware setup. In this paper, we establish a multimodal-driven variational autoencoder (mmVAE) framework to perform 3D surface enhancement based on 2D normal images. To fully leverage the multimodal learning, we investigate a multimodal *Gaussian* mixture model (mmGMM) to align and fuse the latent feature representations from different modalities, and further propose a cross-scale encoder-decoder structure to reconstruct high-resolution normal images. Experimental results on several benchmark datasets demonstrate that our method delivers promising surface geometry structures and details in comparison with competitive advances.

1 Introduction

With the rapid development of computer hardware and software technology, the requests for 3D surface reconstruction, perception and analysis are becoming more and more popular in various immersive applications [Feng *et al.*, 2019], such as metaverse. High-quality 3D data is one of the fundamental carriers for satisfying these application scenarios. However, on one hand, the production speed of high-resolution (HR) 3D content suffers due to the high cost of 3D data-acquisition sensors. On the other, the utilization of existing legacy 3D objects with low-resolution (LR) needs to be enhanced. Considering these facts, it is necessary to develop a low-cost method for the surface enhancement of 3D objects.

Several methods for 3D surface super-resolution have been developed in the past few years, which can be roughly divided into two categories depending on the computation domain: 1) 3D domain-based methods and 2) 2D domain-based methods. In the 3D domain, there are voxel-based methods [Xie *et al.*, 2020], point-cloud-based methods [Luo *et al.*, 2021], distance-functions-based methods [Chen *et al.*, 2021b] and mesh-based methods [Schult *et al.*, 2020]. In these methods, those traditional ones can only optimize some surface mathematical properties and lead to a smooth result, while learning-based methods face the problem of the lack of data and high computational overhead due to the complexity of 3D data representations [Hanocka *et al.*, 2019].

In the 2D domain, existing studies usually use normal images [Chen *et al.*, 2018], [Zhang *et al.*, 2021] or depth images generated by projecting the surface to 2D domain [Schwarz *et al.*, 2018], [Voynov *et al.*, 2019] to enhance a 3D surface. Surface super-resolution in the 2D domain has two significant advantages: 1) the computational complexity will be greatly reduced, and 2) many well-evaluated approaches in the image processing field can be reused directly, such as image super-resolution (SR). Due to the fact that a normal map carries more micro geometric shapes and details but a depth map only contains range or distance information, it is a better choice to perform 3D surface SR in the 2D normal domain.

However, existing SR methods in the 2D domain usually only focus on a single image modality, which rarely utilizes the complementary information available in different modalities of a 3D object. Recently, some image synthesis methods [Esser *et al.*, 2021], [Huang *et al.*, 2021] and image SR methods [Yang *et al.*, 2020], [Liu *et al.*, 2021] have taken multimodal information into account and improved the performance of network. Besides, [Xie *et al.*, 2022] developed a preliminary multimodal transformer framework for 3D surface super-resolution (MNSRNet), which considered three different modalities including depth, RGB and normal images.

Inspired by MNSRNet, we develop a multimodal-driven variational autoencoder (mmVAE) framework to perform 3D surface super-resolution based on enhanced 2D normal images. There are two significant differences between mmVAE and MNSRNet: Firstly, MNSRNet introduced a depth modality branch to preserve large surface structure correctness. However, since 3D surface SR focuses on recovering high-frequency information, introducing a whole branch is informatively redundant. Therefore, mmVAE discards the

[*]This work was supported in part by the National Natural Science Foundation of China under Grants 61902251, in part by the Natural Science Foundation of Shenzhen City under Grants JCYJ20220809160139001, and in part by the Natural Science Foundation of Guangdong Province under 2021A1515011877 and 2022A1515011245. (*Corresponding author*: *Miaohui Wang*)

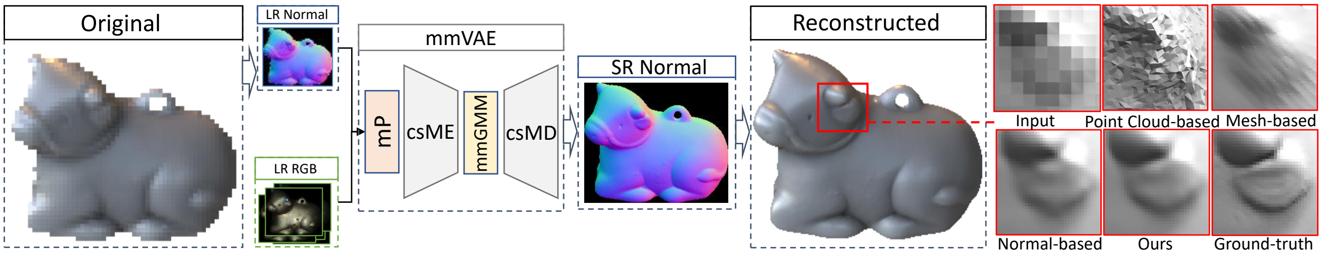

Figure 1: **Illustration of the proposed multimodal VAE framework for 3D surface super-resolution.** The texture and normal modalities are jointly investigated to perform 3D surface super-resolution based on enhanced 2D normal images. The 3D reconstruction results on the right-hand side show the superiority of our method.

depth modality but introduces a cross-scale structure to preserve large surface structures, which can also reduce the computational overhead. Secondly, MNSRNet introduced a multimodal transformer to perform feature alignment and fusion. However, due to the limitations of the transformer structure, the training is usually unsatisfactory when the amount of data is insufficient, and the lack of direct constraints on the alignment module leads to limited performance. To address these problems, mmVAE investigates a VAE-based structure to reduce the requirements and improve the constraints on the modality alignment and fusion process by a feature reparameterization structure. However, since the classic VAE architecture is limited by the standard normal distribution in latent variable reparameterization, resulting in limited generation ability, we further propose a *Gaussian* mixture model and the corresponding loss for the multimodal alignment to improve its expression ability.

To summarize, this work has the following contributions: (1) establishing a new multimodal 3D surface super-resolution framework based on the VAE framework in the 2D normal domain, (2) investigating a *Gaussian* mixture model in fusing the normal and texture modalities for 3D objects, which is a more comprehensive multimodal fusion approach to exploit auxiliary modality information, and (3) developing a new cross-scale modality VAE network structure, which is able to simultaneously preserve large surface structure as well as fine-grained surface geometry.

2 Related Work

Because our mmVAE is based on enhanced 2D normal images, we briefly review some closely representative image-based SR methods, including single image super-resolution (SISR) and multimodal image super-resolution (MISR).

2.1 Single Image Super-resolution

In the SISR task, following the research by [Dong *et al.*, 2014], a number of CNN-based approaches have been proposed such as VDSR [Kim *et al.*, 2016], EDSR [Lim *et al.*, 2017], and RCAN [Zhang *et al.*, 2018b]. Recently, with the success of the self-attention mechanism, transformer-based network structure has been also studied for SISR [Chen *et al.*, 2021a]. Besides, some other useful modules such as VAE structure [Liu *et al.*, 2020], [Chira *et al.*, 2022], generative adversarial network (GAN) [Chen *et al.*, 2022], and dual regression network [Emad *et al.*, 2021] have been used in SISR.

2.2 Multimodal Image Super-resolution

The combination of multimodal information (*e.g.*, different viewpoints, sensors, or domains), especially homogeneous multimodality that means different types of modality information obtained from the same type of sensor for the same object, is a hotspot for MISR to enhance image reconstruction performance [Yao *et al.*, 2021]. For instance, [Wang *et al.*, 2018] introduced the image segmentation map as the prior information to improve the learning performance of GAN. [Liu *et al.*, 2021] proposed a CVAE structure by using reference images to enhance the restoration performance. [Li *et al.*, 2019] employed a normal image to guide the super-resolution of a texture image. [Xie *et al.*, 2022] proposed a transformer-based multimodal network to learn from different modality images for surface enhancement.

Frankly speaking, MISR has been investigated in some preliminary investigations, but the exploration of multimodal-based 3D object surface enhancement is still in its infancy. One of the reasons is the difficulty of constructing descriptors that can effectively represent information between different modalities. And another important reason is how to integrate and fusing these descriptors effectively is also challenging. In the light of the above considerations, exploring mechanisms for more efficient multimodal exploitation of 3D surface modality information through VAE structures is the most direct motivation for our paper.

3 Multimodal Variational AutoEncoder

3.1 Overview

Problem Formulation. Our goal is to enhance the surface of a 3D object via the assistance of RGB modality. To simplify this problem, we propose to represent a 3D surface in the 2D normal domain, which converts 3D surface SR into a normal image super-resolution task. Generally speaking, SR aims to restore the spatial resolution of an input LR normal image $\mathbf{N}_{lr} \in \mathbb{R}^{H\times W\times 3}$ by scale s, which can be obtained by minimizing the loss function $\mathcal{L}_{overall}$ between the upscaled normal image $\mathbf{N}_{sr} \in \mathbb{R}^{sH\times sW\times 3}$ and the ground-truth (GT) normal image $\mathbf{N}_{gt} \in \mathbb{R}^{sH\times sW\times 3}$.

Architecture. Previous studies have witnessed the positive effect of multimodal data in deep learning [Yao *et al.*, 2021] and the power of VAE structure. In light of this, the proposed multimodal variational autoencoder (mmVAE) super-resolution network takes the RGB and normal modalities si-

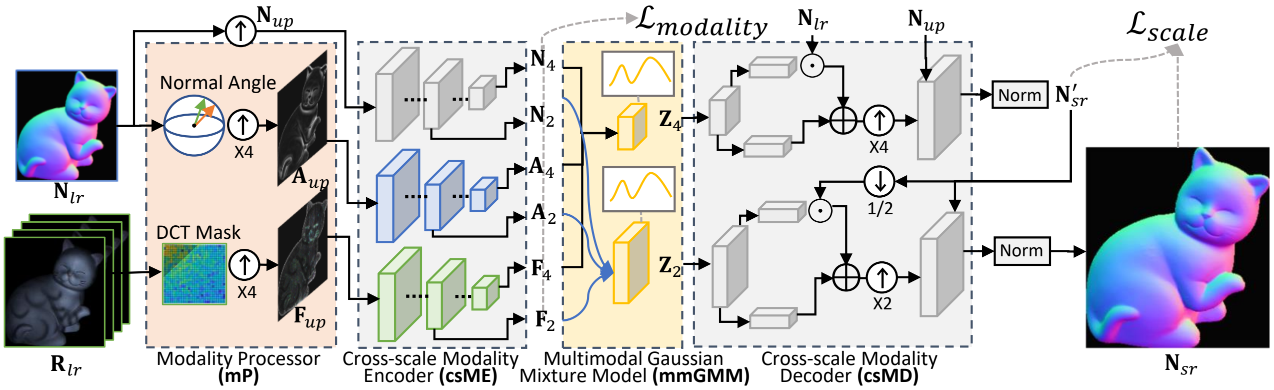

Figure 2: **Overview of the proposed multimodal super-resolution network for 3D object surface in the 2D normal domain.** It mainly consists of a modality processor (mP) module, a cross-scale modality encoder-decoder (csME & csMD) module, and a multimodal *Gaussian* mixture modal (mmGMM) module.

multaneously, and adopts a cross-scale multimodal encoder-decoder structure as a backbone. The overview of mmVAE is depicted in Figure 2, which can be formulated as

$$\mathbf{N}_{sr} = \mathbf{csMD}(\mathbf{mmGMM}(\mathbf{csME}(\mathbf{mP}(\mathbf{N}_{lr}, \mathbf{R}_{lr})))), \quad (1)$$

where $\mathbf{R}_{lr}$ represents multiple LR images obtained under different lighting conditions at the same view, and the LR normal image $\mathbf{N}_{lr}$ is reconstructed by [Xie *et al.*, 2014]. It can be seen that our mmVAE consists of three main components: 1) a modality processor denoted by $\mathbf{mP}(\cdot)$, 2) a multimodal *Gaussian* mixture model denoted by $\mathbf{mmGMM}(\cdot)$, and 3) a cross-scale modality encoder-decoder structure denoted by $\mathbf{csME}(\cdot)$ and $\mathbf{csMD}(\cdot)$, respectively.

The $\mathbf{mP}(\cdot)$ module firstly extracts two auxiliary modality guidances: 1) *normal angle* guidance and 2) *RGB frequency* guidance. These two modality guidances have a similar distribution, which will be helpful to be fused in the subsequent network structure. Then, these two modalities with the LR normal image are upscaled to generate the corresponding modality inputs $\mathbf{A}_{up}$, $\mathbf{F}_{up}$ and $\mathbf{N}_{up}$, respectively. Next, similar to a classical encoder-decoder structure, these upscaled modalities are fed into the encoder module $\mathbf{csME}(\cdot)$ to generate an internal cross-scale modality feature $\mathbf{X}_s$, where s denotes a downsampling factor and $\mathbf{X}$ indicates their source modalities. After that, the $\mathbf{mmGMM}(\cdot)$ module utilizes $\mathbf{X}_s$ in the same scale to align and fuse the related modality information. Therefore, $\mathbf{mmGMM}(\cdot)$ will output latent probability variables $\mathbf{Z}_s$ with different scales. Finally, $\mathbf{Z}_s$ will be fed into the decoder module $\mathbf{csMD}(\cdot)$ to restore an auxiliary normal image $\mathbf{N}^{'}_{sr}$ and the final normal image $\mathbf{N}_{sr}$. In Sec. 3.3 and Sec. 3.4, we provide the details of $\mathbf{mmGMM}$ and our $\mathbf{csME}$-$\mathbf{csMD}$ structure.

Loss Function. In order to train the proposed mmVAE in the manner of multimodal features and cross-scale structure, we devise a loss function combined with the modality loss and scale loss, which can be formulated as Eq. (2).

$$\mathcal{L}_{overall} = \mathcal{L}_{modality} + \lambda_1 \times \mathcal{L}_{scale}, \quad (2)$$

where $\mathcal{L}_{modality}$ represents a modality loss, $\mathcal{L}_{scale}$ represents a scale loss, and λ_1 is a positive scaling factor.

Inspired by the *Kullback-Leibler* (**KL**) divergence [Kullback and Leibler, 1951] and the VAE paradigm [Minnen *et al.*, 2018], we design a two-part modality loss function $\mathcal{L}_{modality}$: 1) The first part adopts the paradigm of the VAE architecture converging the probability variables before the GMM module to $\mathcal{N}(0,1)$ to obtain a more generalized *Gaussian* mixture model for each modality; 2) The second part directly minimizes the probability variables distance between the target normal and the other modalities after the GMM module, thus providing an explicit multimodality alignment. As a result, $\mathcal{L}_{modality}$ is defined by

$$\mathcal{L}_{modality} = \sum_{s}^{2,4} \sum_{x}^{A,F} (\mathbf{KL}(\mathcal{N}(\mu_s^x, \sigma_s^x), \mathcal{N}(0,1)) + \lambda_2 |\mathbf{z}_s^N - \mathbf{z}_s^x|), \quad (3)$$

where s denotes a feature down-scale factors, and x denotes different modalities. μ and σ represent the mean and variance of a *Gaussian* distribution model of these probability variables, respectively. The details of them can be found in Sec.3.3. λ_2 is a positive scaling factor.

Then, we design a scale loss function for mmVAE as

$$\mathcal{L}_{scale} = \mathcal{L}_{content}(\mathbf{N}_{sr}) + \lambda_3 \times \mathcal{L}_{content}(\mathbf{N}^{'}_{sr}), \quad (4)$$

where the content loss $\mathcal{L}_{content}(\cdot)$ is used to constrain $\mathbf{N}_{sr}$ and $\mathbf{N}^{'}_{sr}$ simultaneously. λ_3 is a positive scaling factor. And the detail of the content loss is defined in Eq. (5).

$$\mathcal{L}_{content}(\mathbf{N}) = \frac{1}{H \times W} \sum((1 - \mathbf{cos}(\mathbf{N}, \mathbf{N}_{gt})) + \lambda_4(1 - \mathbf{cos}(\downarrow \mathbf{N}, \mathbf{N}_{lr}))), \quad (5)$$

where (H, W) represents the height and width of a predicted normal image. $\mathbf{cos}(\cdot,\cdot)$ denotes the element-wise cosine operator calculating the difference between two normal images, and $\downarrow$ represents a downsampled normal map with the same size as $\mathbf{N}_{lr}$. This back-projection error $\mathcal{L}_{prj}$ inspired by [Haris *et al.*, 2018] is developed to constrain the downsampled SR normal modality close to LR one , which is used to preserve large surface structures. λ_4 denotes a positive scaling factor.

3.2 Modality Processor (mP)

As aforementioned, due to the differences across different modalities, multimodal-based methods usually encounter negative influence in combining the latent features from different modalities. Because the SR task is focusing on reconstructing the high-frequency information of the input data, in order to reduce the data distribution differences and distill the valuable features from the normal image and RGB images, we represent them as a *normal angle* guidance $\mathbf{A}_{lr}$ and an *RGB frequency* guidance $\mathbf{F}_{lr}$. It is worth noting that both $\mathbf{A}_{lr}$ and $\mathbf{F}_{lr}$ represent the high-frequency information of an input 3D surface.

Normal Angle. The extraction of the *normal angle* guidance consists of two steps: 1) generating a low-frequency normal image called a shape normal image $\mathbf{N}_{shape}$, and 2) calculating the vector angle between the shape normal and the original normal.

Firstly, the generation of $\mathbf{N}_{shape}$ can be formulated as

$$\mathbf{N}_{shape} = conv(\mathbf{N}_{lr}, \kappa_{ave}), \tag{6}$$

where $conv(\cdot)$ represents a convolution operation and κ_{ave} denotes an average filter kernel.

Secondly, we obtain the angle between $\mathbf{N}_{shape}$ and $\mathbf{N}_{lr}$ by

$$a_{lr} = arccos(\mathbf{n}_{lr} \cdot \mathbf{n}_{shape}), \tag{7}$$

where the normal angle $a_{lr} \in \mathbf{A}_{lr}$ for each pixel is the arccosine result of the dot-product of $\mathbf{n}_{lr}$ from $\mathbf{N}_{lr}$ and the corresponding vector $\mathbf{n}_{shape}$ from $\mathbf{N}_{shape}$.

Based on above steps, the original normal image will be converted from $\mathbf{N}_{lr} \in \mathbb{R}^{H \times W \times 3}$ to $\mathbf{A}_{lr} \in \mathbb{R}^{H \times W \times 1}$ representing the pixel-wise relief of a 3D object surface. It is noted that smaller values indicate smoother surface, while larger values represent the high frequency and contour information of a 3D surface. For instance, a toy normal angle map is provided in Figure 2.

RGB frequency. Due to the diversity of the object materials and surface geometry structures, the uncertainty of camera sensors or lighting conditions, the raw multi-lighting photographs may contain many unfavorable issues, such as exposure errors, shadows from self-obscuring, speculator reflection, and uneven brightness due to different reflection intensities. However, those outliers also contain useful information. To fully utilize the input data, we develop a two-step scheme to denoise the RGB modality images and extract their valuable information: 1) extracting three brightness level images from $\mathbf{R}_{lr}$, and 2) performing a masked DCT-transform to separate and combine their high-frequency features.

For the first step, we calculate a pixel-wise brightest image $\mathbf{R}_l$, a darkest image $\mathbf{R}_d$, and an average image $\mathbf{R}_a$ from the original RGB modality images, respectively. This operation aims to denoise and capture the different levels of surface material detail. For the second step, we convert these images from the spatial domain into the frequency domain. Then we mask the low $1/8$ part on the frequency map, and leave the high-frequency part. By performing the invert-DCT, we can reconstruct a feature map with only the high-frequency information. Finally, these three features are concatenated to get the final RGB frequency guidance $\mathbf{F}_{lr}$. The RGB frequency guidance can be formulated as

$$\mathbf{F}_{lr} = \mathcal{T}^{'}_{dct}(mask(\mathcal{T}_{dct}(\mathbf{R}_l, \mathbf{R}_d, \mathbf{R}_a))), \tag{8}$$

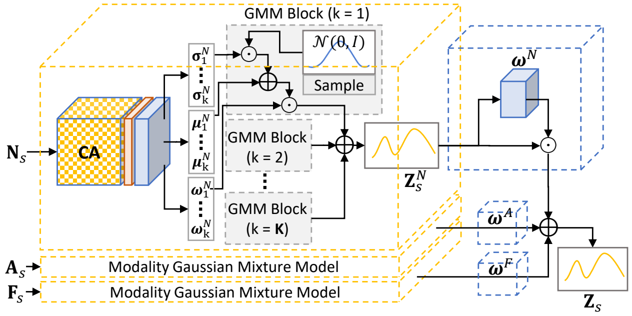

Figure 3: **Detail of multimodal *Gaussian* mixture model (mmGMM).** The internal modality features $\mathbf{N}_s$, $\mathbf{A}_s$ and $\mathbf{F}_s$ are firstly generated by a sequence of masked `Conv`, `ReLU` and convolutional layers to generate parameters which can be used to construct a *Gaussian* model. Then, the modality latent variable $\mathbf{Z}_s^x$ is sampled and generated from these *Gaussian* models. Finally, the fused latent variable $\mathbf{Z}_s$ is generated by using a convolutional block as weight.

where $\mathcal{T}_{dct}(\cdot)$ and $\mathcal{T}^{'}_{dct}(\cdot)$ represent the forward-DCT and invert-DCT respectively. The $mask(\cdot)$ manipulation removes the top-left $1/8$ DCT frequency map which represents the low-frequency texture information. As a result, $\mathbf{A}_{lr}$ and $\mathbf{F}_{lr}$ represent high-frequency information from the surface normal and the texture structure of an input object, respectively. The related complement information will be upscaled and used as the auxiliary modality guidance to improve the subsequent network.

3.3 Multimodal *Gaussian* Mixture Model (mmGMM)

Mathematically, a classical variational autoencoder (VAE) network can be formulated as

$$\mathbf{Y} = \mathbf{Decoder}(\mathbf{Z}), where \quad \mathbf{Z} = \mathbf{P}(\mathbf{Encoder}(\mathbf{X})), \tag{9}$$

where $\mathbf{X}$ denotes the input of VAE and $\mathbf{Y}$ denotes the output of VAE. Based on the VAE structure, a high-dimensional latent variable $\mathbf{Z}$ is fed into the decoder to generate a target result. This latent variable $\mathbf{Z}$ is a variable that is sampled from a probability model parameterized by the encoded data. This paradigm of encoding the input data and then sampling it from a probability model can reduce the noise of the latent features and further improve the generation capability of a learned model.

Generally, most VAEs adopt a *Gaussian* model $\mathcal{N}(\mu, \sigma)$ [Pu *et al.*, 2016] as the probability modeling of latent variables. Each dimension of the latent variable represents an implicit feature of the target probability distribution, which can finally fit the target distribution in the form of a Gaussian mixture model by decoders. However, due to the fact that the ability of these models to express features is relatively low, and they rarely considers the multimodal information, which makes them increase the number of feature layers but it is still difficult to fit the ideal potential distribution of the decoder. This further can limit their generation performance. Based on these considerations, we propose to use a multimodal *Gaussian* mixture model (mmGMM) as shown in Figure 3, which is used as the probability model for our mmVAE to generate the encoded feature more accurately. Due to the fact that a *Gaussian* mixture model can be expressed as a combination of multiple *Gaussian* models efficiently, we design the **mmGMM** module, and it can be formulated as a weighted summary of several *Gaussian* models

$$\mathbf{Z}_s = \mathbf{P}(\mathbf{A}_s, \mathbf{F}_s, \mathbf{N}_s) = \sum\nolimits_{x}^{A,F,N} \omega^x \mathbf{Z}_s^x, \tag{10}$$

where $\mathbf{Z}_s^x = \sum_{k=1}^{K} \omega_k^x \mathcal{N}(\mu_k^x, \sigma_k^x)$, x from $\{A, F, N\}$ denotes the normal angle, RGB frequency, and normal modality, respectively. s denotes a downsampling scalar.

Firstly, we adopt a channel attention block [Zhang *et al.*, 2018a], which extracts the started feature from **csME**. Then with a `ReLU` layer followed by a convolution layer, we convert the related features to represent the mean μ_k^x, variation σ_k^x, and weights ω_k^x of a GMM distribution, where x denotes the corresponding modality and k represents the number of *Gaussian* components in a combined *Gaussian* mixture distribution. Each set of parameters will be used to resample a standard *Gaussian* distribution by a reparameterization trick [Kingma *et al.*, 2015] and to calculate the weighted summary as a modality latent variable $\mathbf{Z}_s^x$. This part is equivalent to compressing a large number of *Gaussian* distributions to obtain a lower-dimensional *Gaussian* mixture feature. Finally, with another convolution block, each $\mathbf{Z}_s^x$ will generate a modality weight ω^x and calculate another weighted summary to obtain the corresponding fused latent variable $\mathbf{Z}_s$ for the following **csMD**.

3.4 Cross-scale Modality-based Encoder-Decoder

To more effectively utilize the information in different scales of the fused modal features on the **mmGMM** structure, we propose a cross-scale network structure to drive the whole VAE operation. The cross-scale modality encoder (csME) encodes the input modal data into latent parametric features which control the probability distribution of **mmGMM**. Then, the latent features are resampled by **mmGMM** and decoded by the cross-scale modality decoder (csMD).

For the **csME** module, as shown in Figure 2, the input modality guidance is passed through a three-3×3-convolution layer to extract the related shallow angle feature, shallow frequency feature, and shallow normal feature from $\mathbf{A}_{up}$, $\mathbf{F}_{up}$, and $\mathbf{N}_{up}$, respectively. Then, to allow the network to pay attention to both large structures and small details of a 3D object, we use several scaling blocks, which are composed of 2 convolution layers connected to a `ReLU` activation layer followed by a pooling layer, to resize the input features by the scale $\frac{1}{2}$. This conversion will generate several shallow features in different scales. For instance, $Scale = 4$, it will be the $\mathbf{X}_{up} \in \mathbb{R}^{4H \times 4W \times C} \to \mathbf{X}_2 \in \mathbb{R}^{2H \times 2W \times C}$ and $\mathbf{X}_2 \in \mathbb{R}^{2H \times 2W \times C} \to \mathbf{X}_4 \in \mathbb{R}^{4H \times 4W \times C}$, where $\mathbf{X}$ denotes one of the candidate modality features (*i.e.*, $\mathbf{X} = \{\mathbf{A}, \mathbf{F}, \mathbf{N}\}$). These features will be fed into **mmGMM** to generate the corresponding fused latent variables $\mathbf{Z}_2$ and $\mathbf{Z}_4$.

For the **csMD** module, due to the fact that the effectiveness of a gradual refinement based on the hierarchical structure has been proven [Cai *et al.*, 2019], we combine the high scaling and low scaling latent variables which aim to capture different levels of object details together to help the mmVAE focus both geometry shape and detail comprehensively, and thus obtain a better reconstruction result.

Firstly, for high scaling latent variables $\mathbf{Z}_4$ in **csME**, we employ two `Conv` blocks (consisting of three convolutional layers and one batch-normalization layer with one `ReLU` activation layer) to decode two latent features as the scalar and bias information. Then with a 3×3 convolution layer, the original input normal image $\mathbf{N}_{lr}$ is scaled and added by these features. After this feature enhancement, using a scale s upsampling block $\mathbf{UP}_s(\cdot)$ [Shi *et al.*, 2016], an upscaled high-frequency feature is generated $\mathbf{F}_{sr}^{'}$. Followed by a 3×3 convolution layer and added to the upscaled normal image $\mathbf{N}_{up}$ with a vector normalization operation, an SR normal image denoted as $\mathbf{N}_{sr}^{'}$ directly from scale ×4 information is generated. This process can be formulated by Eq. (11)

$$\mathbf{N'}_{sr} = \gamma(\mathcal{C}_o(\mathbf{UP}_4(\mathbf{N}_{lr}) \otimes \mathcal{C}_s(\mathbf{Z}_4) \oplus \mathcal{C}_b(\mathbf{Z}_4)) + \mathbf{N}_{up}), \tag{11}$$

where γ denotes the vector normalization operation. $\otimes$ and $\oplus$ refer to the element-wise multiplication and addition, respectively. $\mathcal{C}_s(\cdot)$ and $\mathcal{C}_b(\cdot)$ denote the convolution module generating two latent features, and $\mathcal{C}_o(\cdot)$ denotes three output convolutional layers. After that, we adopt larger latent variables $\mathbf{Z}_2$ to refine the normal image generated by $\mathbf{Z}_4$. This process formulated by Eq. (12) will generate the final output upscaled normal image $\mathbf{N}_{sr}$. The whole network is trained by optimizing the loss function in Sec. 3.1.

4 Experiments

In this section, we implement our mmVAE and compare it with recent 3D surface SR methods, including mesh-based, point-cloud-based, depth-based, normal SISR, and normal MISR methods.

$$\mathbf{N}_{sr} = \gamma(\mathcal{C}_o(\mathbf{UP}_2(\mathbf{N}_{lr}) \otimes \mathcal{C}_s(\mathbf{Z}_2) \oplus \mathcal{C}_b(\mathbf{Z}_2)) + \mathbf{N'}_{sr}). \tag{12}$$

4.1 Experimental Protocols

Implementation Details. Our mmVAE has been implemented in *PyTorch*, and the Adam optimizer is used with default parameters (β_1 = 0.9 and β_2 = 0.999) and K = 4 in **mmGMM** module. The detailed hyper-parameter settings of loss function are $\lambda_1 = 5$, $\lambda_2 = 0.2$, $\lambda_3 = 0.2$, $\lambda_4 = 0.1$. The initial learning rate is 3*e*-4 We have trained mmVAE by using a mini-batch size of 4 for 3000 epochs with one *Nvidia Tesla P100* GPU, which takes about two days and nights. All the input images for training are adaptively cropped, randomly rotated (90°, 180°, and 270°), and horizontally flipped. For instance, in the ×4 scale, the HR and LR image patches are 196×196 and 48×48, respectively. All the trained weights are initialized by the *Kaiming* distribution, and the bias is initialized as a constant. In addition to the above operations, we have not introduced additional training skills.

Datasets. Due to the fact that mmVAE is a data-driven SR scheme, it requires a lot of high-resolution normal image pairs. However, the most widely-used *Photometric Stereo* datasets, such as the *DiLiGenT* dataset (10 objects) [Shi *et al.*, 2019] and the *Gourd & Apple* dataset (3 objects) [Alldrin

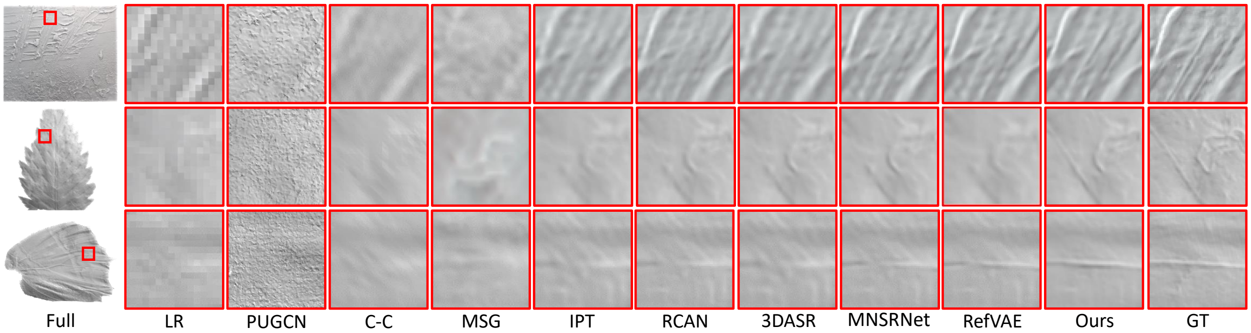

Figure 4: **Visual comparisons of 3D surface SR between 9 methods in the ×4 setting**. For a better comparison, the region in the red box is zoomed in the 2nd-12th columns. "Full" means the original surface, "LR" means the down-sampled object surface, and "GT" means the ground-truth in the red box. Please zoom in the electronic version for better details.

et al., 2008], do not have enough objects for training. Therefore we use the latest wonderful photometric stereo dataset *WPS* for training [Xie *et al.*, 2022], which contains 400 high-resolution multimodal samples. To fairly evaluate the performance of these methods, we use *DiLiGenT*, *Gourd & Apple* and select $\frac{1}{6}$ of the *WPS* dataset as the testing set. The rest of the *WPS* dataset is used as the training (validation) set. Both the training and testing data are down-sampled with the Bicubic (BI) degradation by $\times\frac{1}{2}$ and $\times\frac{1}{4}$ to generate the LR images as the network input. To perform a fair evaluation, all the compared methods are trained from scratch using the official settings and the same training iterations as mmVAE.

Evaluation Metrics. For quantitative comparisons, we adopt four quality measurements to evaluate the super-resolution performance. In the 2D normal image domain, we take two commonly-used indicators in SISR for evaluation which are Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). Besides, two widely-used metrics are used to quantitatively measure the 3D reconstruction results [Quéau *et al.*, 2018], including Mean Angular Error (MAE) and Mean Relative Depth Error (MRDE). The final object surface is reconstructed by the Surface-from-Normal (SfN) method [Xie *et al.*, 2014], [Cao *et al.*, 2021]. Specifically, we have adopted the public available discrete geometry-based SfN method [Xie *et al.*, 2014] to reconstruct an enhanced 3D surface. The MAE result is calculated by

$$MAE = \frac{1}{||\mathbf{N}||}\sum_{i,j} arccos(\tilde{\mathbf{n}}_{i,j}\cdot\mathbf{n}_{i,j}), \tag{13}$$

where $\tilde{\mathbf{n}}_{i,j}$ and $\mathbf{n}_{i,j}$ denote the predicted normal and the ground-truth normal, respectively. $||\mathbf{N}||$ represents the total number of the input normal pixels.

In addition, MRDE can be computed by

$$MRDE = \frac{1}{||\mathbf{N}||}\sum_{i,j} ||\tilde{\mathbf{p}}_{i,j} - \mathbf{p}_{i,j}||, \tag{14}$$

where $\tilde{\mathbf{p}}_{i,j}$ and $\mathbf{p}_{i,j}$ represent the vertex position of the reconstructed surface and the ground-truth surface, respectively.

4.2 Baseline Settings

We have compared our mmVAE with 8 representative methods, which can be categorized into five groups: mesh-based method (denoted by "Mesh"), point-cloud-based method (denoted by "Points"), depth-based methods(denoted by "Depth"), SISR-based methods (denoted by "SISR"), and MISR-based methods (denoted by "MISR"). For mesh-based methods, Catmull-Clark subdivision (C-C) [Loop and Schaefer, 2008] is one of the most widely-used mesh subdivision methods, which can efficiently upscale a triangular mesh by a heuristic algorithm. For point-cloud-based methods, we choose the PU-GCN network [Qian *et al.*, 2021] to represent the SR task of point clouds. In the experiments, we convert the related meshes into point clouds for PU-GCN, upscale, and re-convert them into meshes for comparison [Bernardini *et al.*, 1999]. For depth-based methods, we choose the MSG network [Voynov *et al.*, 2019] to represent the SR task in depth images. For normal SISR-based methods, we choose RCAN [Zhang *et al.*, 2018a] to represent a convolutional attention structure and IPT [Chen *et al.*, 2021a] to represent a self-attention structure. For normal MISR-based methods, we choose 3DASR [Li *et al.*, 2019] to represent the hybrid fusion method, RefVAE [Liu *et al.*, 2021] to represent a hybrid fusion method with the VAE structure, and MNSRNet [Xie *et al.*, 2022] to represent the multimodal normal-based method with the transformer structure. The detailed experimental setting is consistent with [Xie *et al.*, 2022].

4.3 Performance Comparisons

Figure 4 demonstrates the visual comparisons of some representative 3D objective surfaces. For SISR, due to the small size of the existing 3D training dataset, the CNN-based structure as RCAN showed better results than the transformer structure such as IPT. For MISR, they may not take full advantage of the additional multimodal information. The simple structure of 3DASR cannot fuse the modality information well and hence suffers from the negative effects of instability and confusion across multimodalities. MNSRNet with the transformer structure suffers from the size of the training set. As seen, RefVAE uses both VAE structure and multimodal information to obtain the best result in the baseline methods.

Table 1 provides the detailed average results on all testing datasets, including *DiLiGenT*, *Gourd & Apple*, and *WPS*. Specifically, our method achieves 11 of the first-best results and 1 of the second-best result in terms of PSNR, SSIM,

Scale	Type	Algorithm	*DiLiGenT*	*Gourd & Apple*	*WPS*
			PSNR↑/SSIM↑/MAE↓/MRDE↓	PSNR↑/SSIM↑/MAE↓/MRDE↓	PSNR↑/SSIM↑/MAE↓/MRDE↓
×2	Points	PU-GCN [Qian *et al.*, 2021]	22.7884/0.8019/10.0552/9.3777	24.1669/0.7876/7.2156/1.5350	23.9856/0.7251/7.5096/5.8443
	Mesh	C-C [Loop and Schaefer, 2008]	24.5128/0.9105/5.0687/6.3240	26.5517/0.8771/0.8916/0.8736	26.4602/0.7999/4.0879/3.4845
	Depth	MSG [Voynov *et al.*, 2019]	28.7298/0.9554/4.3926/5.9665	33.2658/0.9890/0.6083/0.6457	34.0693/0.9639/2.5894/2.4289
	SISR	IPT [Chen *et al.*, 2021a]	28.7556/0.9625/3.8207/5.8153	33.7279/0.9873/1.4558/0.5557	36.7309/0.9775/2.0263/2.1598
		RCAN [Zhang *et al.*, 2018a]	30.7655/0.9759/2.8982/3.8772	34.5141/0.9899/1.2404/0.5130	38.3461/0.9782/2.1502/1.6714
	MISR	3DASR [Li *et al.*, 2019]	29.6974/0.9664/3.4351/5.1201	34.4950/0.9891/1.2877/0.5387	37.9963/0.9786/2.2238/1.7798
		MNSRNet [Xie *et al.*, 2022]	30.7720/0.9763/2.8853/4.4479	34.1527/0.9896/1.2996/0.5769	37.9990/0.9789/**2.0050**/1.7727
		RefVAE [Deng *et al.*, 2021]	30.5138/0.9730/3.0885/4.0017	34.5274/0.9897/1.2235/0.4702	38.4997/0.9780/2.0230/1.6969
		Ours	**30.7969/0.9766/2.3334/3.7977**	**34.5723/0.9902/1.1634/0.4476**	**38.5276/0.9793**/2.0231/**1.6326**
×4	Points	PU-GCN [Qian *et al.*, 2021]	18.5747/0.7382/13.2567/10.3674	22.1672/0.7611/9.3709/2.5241	21.7603/0.6727/11.6703/6.0484
	Mesh	C-C [Loop and Schaefer, 2008]	24.1655/0.8585/8.6984/7.4588	26.9087/0.8777/2.9155/1.1320	24.1063/0.7967/5.8517/3.7967
	Depth	MSG [Voynov *et al.*, 2019]	24.8738/0.8937/7.8564/7.3132	29.1214/0.9705/2.1031/1.0129	29.9963/0.9185/4.2786/3.4201
	SISR	IPT [Chen *et al.*, 2021a]	25.2410/0.9136/6.5397/6.6953	30.6276/0.9695/2.7039/0.9519	31.8643/0.9358/3.9953/2.5922
		RCAN [Zhang *et al.*, 2018a]	27.1337/0.9304/6.1913/4.7991	32.8232/0.9808/1.8830/0.9292	32.5700/0.9441/3.6573/2.5588
	MISR	3DASR [Li *et al.*, 2019]	26.6126/0.9176/7.0156/5.8078	32.0332/0.9723/2.5102/0.8520	31.7228/0.9322/4.1421/2.7914
		MNSRNet [Xie *et al.*, 2022]	27.3644/0.9320/5.9156/4.2861	33.0666/0.9814/1.8183/0.7351	32.6350/0.9426/3.7705/**2.0440**
		RefVAE [Deng *et al.*, 2021]	27.2967/0.9323/5.8756/4.5097	33.3056/0.9826/1.7447/0.7752	32.6443/0.9446/3.6538/2.4731
		Ours	**27.6255/0.9342/5.7964/4.2827**	**33.4820/0.9833/1.6881/0.6052**	**33.1782/0.9482/3.5907**/2.1511

Table 1: Quantitative comparison results. The average comparison results between 9 methods on three different datasets. "↑" means the higher the better, while "↓" means the lower the better. The first-best is highlighted in **bold**, and the second-best is highlighted in underline.

mP	mmGMM	Cross-scale	PSNR	SSIM	MAE	MRDE
×	×	×	26.9277	0.9208	6.8913	5.9609
√	×	×	27.2726	0.9298	6.1635	4.6789
√	√	×	27.3557	0.9313	6.0589	4.4791
√	×	√	27.4190	0.9335	5.9589	4.4532
√	√	√	**27.6255**	**0.9342**	**5.7964**	**4.2827**
Hyperparameter K = 1			27.4521	0.9332	5.9074	4.4511
Hyperparameter K = 2			27.5905	0.9330	5.8404	4.3991
Hyperparameter K = 3			27.6210	0.9339	5.8021	4.3513
Hyperparameter K = 4			**27.6255**	**0.9342**	**5.7964**	**4.2827**
Hyperparameter K = 5			27.6247	0.9340	5.8009	4.3006

Table 2: Ablation experiments on **mP**, mmGMM, cross-scale encoder-decoder and the hyperparameter K of **mmGMM**.

MAE, and MRDE on both the ×2 and ×4 settings. As seen in Table 1, our method performs better than other methods in upsampling small images.

In addition, we have conducted the time and space comparison experiments on the DiLiGenT dataset. In the ×4 setting, MNSRNet takes a total of 20.9647s and 11445MB, while mmVAE takes only 8.2965s and 4265MB. It validates the effectiveness of the proposed cross-scale structure.

4.4 Ablation Study

mmVAE contains three main modules for multimodal learning, including **mP**, **mmGMM**, and **cross-scale encoder-decoder** structure. To verify the effectiveness of these modules, we further conduct additional experiments on the *DiLiGenT* dataset with the ×4 setting. Five independent experiments are conducted as shown in Table 2, where the related modules selected (not selected) are denoted by the symbol "√" ("×"). For the mmGMM module, we also performed ablation experiments for different hyperparameter K. It can be seen that with the increase of K, the performance has improved. However, this also has marginal effects, and too large K does not bring significant gains. Meanwhile, it is worth noting that when K = 1, the properties GMM will be lost.

In the ablation experiments, the replacement of **mP** is to use the lightest RGB images as the auxiliary guidance. The replacement of **mmGMM** is to directly concatenate all the related modality information, and then use three 3×3 convolution layers and one 1×1 convolution layer to shrink the intermediate channels similar to 3DASR. The replacement of the cross-scale structure is to use the results for scale $\frac{1}{4}$ as the final outputs. In Table 2, the ablation result of **mmGMM** shows better than MNSRNet. That is because *DiLiGenT* has smaller image size, so the reconstruction results are more easily influenced by low-frequency information. However, as the image size increases, the crossmodal information derived from mmGMM plays a more important role. For instance, mmVAE without the mmGMM module is less effective than MNSRNet on the *WPS* dataset, where the related PSNR/SSIM/MAE/MRDE results are 32.5947/0.9420/3.8024/2.2555, respectively.

5 Conclusion

In this paper, we have established a multimodal-driven variational autoencoder super-resolution framework to enhance 3D surfaces in the 2D normal domain. More specifically, we jointly considered the RGB and normal modalities to restore high-quality surface details as well as preserve fine-grained geometry structures. To effectively utilize the multi-modality information, we extracted two guidance features as auxiliary. To effectively fuse the related auxiliary information, we further investigated a multimodal *Gaussian* mixture model for 3D surface super-resolution. Moreover, we developed a cross-scale encoder-decoder structure to fully utilize the cross-modality information. Finally, we reconstruct an enhanced 3D object surface from the recovered high-resolution normal image. Experimental results on different benchmark datasets demonstrate the effectiveness of the proposed approach qualitatively and quantitatively.

References

[Alldrin *et al.*, 2008] Neil Alldrin, Todd Zickler, and David Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[Bernardini *et al.*, 1999] Fausto Bernardini, Joshua Mittleman, Holly Rushmeier, Claudio Silva, and Gabriel Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 5(4):349–359, 1999.

[Cai *et al.*, 2019] Lei Cai, Hongyang Gao, and Shuiwang Ji. Multi-stage variational auto-encoders for coarse-to-fine image generation. In *SIAM International Conference on Data Mining*, pages 630–638, 2019.

[Cao *et al.*, 2021] Xu Cao, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Normal Integration via Inverse Plane Fitting with Minimum Point-to-Plane Distance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2382–2391, 2021.

[Chen *et al.*, 2018] Lan Chen, Juntao Ye, Liguo Jiang, Chengcheng Ma, Zhanglin Cheng, and Xiaopeng Zhang. Synthesizing cloth wrinkles by CNN-based geometry image superresolution. *Wiley Computer Animation and Virtual Worlds*, 29(3-4):e1810, 2018.

[Chen *et al.*, 2021a] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12310, 2021.

[Chen *et al.*, 2021b] Zhiqin Chen, Vladimir G Kim, Matthew Fisher, Noam Aigerman, Hao Zhang, and Siddhartha Chaudhuri. Decor-gan: 3d shape detailization by conditional refinement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15740–15749, 2021.

[Chen *et al.*, 2022] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee Kenneth Wong. Deep photometric stereo for non-Lambertian surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):129–142, 2022.

[Chira *et al.*, 2022] Darius Chira, Ilian Haralampiev, Ole Winther, Andrea Dittadi, and Valentin Liévin. Image Super-Resolution With Deep Variational Autoencoders. *arXiv preprint arXiv:2203.09445*, 2022.

[Deng *et al.*, 2021] Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. Deep Coupled Feedback Network for Joint Exposure Fusion and Image Super-Resolution. *IEEE Transactions on Image Processing*, 30:3098–3112, 2021.

[Dong *et al.*, 2014] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Springer European Conference on Computer Vision (ECCV)*, pages 184–199, 2014.

[Emad *et al.*, 2021] Mohammad Emad, Maurice Peemen, and Henk Corporaal. DualSR: Zero-Shot Dual Learning for Real-World Super-Resolution. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1630–1639, 2021.

[Esser *et al.*, 2021] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. ImageBART: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:1–15, 2021.

[Feng *et al.*, 2019] Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. MeshNet: Mesh neural network for 3d shape representation. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 8279–8286, 2019.

[Hanocka *et al.*, 2019] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. MeshCNN: a network with an edge. *ACM Transactions on Graphics*, 38(4):1–12, 2019.

[Haris *et al.*, 2018] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1673, 2018.

[Huang *et al.*, 2021] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal Conditional Image Synthesis with Product-of-Experts GANs. *arXiv preprint arXiv:2112.05130*, 2021.

[Kim *et al.*, 2016] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[Kingma *et al.*, 2015] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems (NeurIPS)*, 28:1–9, 2015.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A Leibler. On information and sufficiency. *JSTOR The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[Li *et al.*, 2019] Yawei Li, Vagia Tsiminaki, Radu Timofte, Marc Pollefeys, and Luc Van Gool. 3D appearance super-resolution with deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9671–9680, 2019.

[Lim *et al.*, 2017] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 136–144, 2017.

[Liu *et al.*, 2020] Zhi-Song Liu, Wan-Chi Siu, Li-Wen Wang, Chu-Tak Li, and Marie-Paule Cani. Unsupervised real image super-resolution via generative variational autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 442–443, 2020.

[Liu *et al.*, 2021] Zhi-Song Liu, Wan-Chi Siu, and Li-Wen Wang. Variational autoencoder for reference based image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 516–525, 2021.

[Loop and Schaefer, 2008] Charles Loop and Scott Schaefer. Approximating Catmull-Clark subdivision surfaces with bicubic patches. *ACM Transactions on Graphics*, 27(1):1–11, 2008.

[Luo *et al.*, 2021] Luqing Luo, Lulu Tang, Wanyi Zhou, Shizheng Wang, and Zhi-Xin Yang. PU-EVA: An Edge-Vector Based Approximation Solution for Flexible-Scale Point Cloud Upsampling. In *IEEE International Conference on Computer Vision (ICCV)*, pages 16208–16217, 2021.

[Minnen *et al.*, 2018] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in Neural Information Processing Systems (NeurIPS)*, 31:1–10, 2018.

[Pu *et al.*, 2016] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. *Advances in Neural Information Processing Systems (NeurIPS)*, 29:1–9, 2016.

[Qian *et al.*, 2021] Guocheng Qian, Abdulellah Abualshour, Guohao Li, Ali Thabet, and Bernard Ghanem. Pu-gcn: Point cloud upsampling using graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11683–11692, 2021.

[Quéau *et al.*, 2018] Yvain Quéau, Jean-Denis Durou, and Jean-François Aujol. Normal integration: a survey. *Springer Journal of Mathematical Imaging and Vision*, 60(4):576–593, 2018.

[Schult *et al.*, 2020] Jonas Schult, Francis Engelmann, Theodora Kontogianni, and Bastian Leibe. Dualconvmesh-Net: Joint geodesic and euclidean convolutions on 3d meshes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8612–8622, 2020.

[Schwarz *et al.*, 2018] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A Chou, Robert A Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, et al. Emerging MPEG standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):133–148, 2018.

[Shi *et al.*, 2016] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[Shi *et al.*, 2019] Boxin Shi, Zhipeng Mo Mo, Zhe Wu, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):271–284, 2019.

[Voynov *et al.*, 2019] Oleg Voynov, Alexey Artemov, Vage Egiazarian, Alexander Notchenko, Gleb Bobrovskikh, Evgeny Burnaev, and Denis Zorin. Perceptual deep depth super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5653–5663, 2019.

[Wang *et al.*, 2018] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 606–615, 2018.

[Xie *et al.*, 2014] Wuyuan Xie, Yunbo Zhang, Charlie CL Wang, and Ronald C-K Chung. Surface-from-gradients: An approach based on discrete geometry processing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2195–2202, 2014.

[Xie *et al.*, 2020] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Generative VoxelNet: learning energy-based models for 3D shape synthesis and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Xie *et al.*, 2022] Wuyuan Xie, Tengcong Huang, and Miaohui Wang. MNSRNet: Multimodal Transformer Network for 3D Surface Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12703–12712, 2022.

[Yang *et al.*, 2020] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5791–5800, 2020.

[Yao *et al.*, 2021] Chao Yao, Shuaiyong Zhang, Mengyao Yang, Meiqin Liu, and Junpeng Qi. Depth Super-Resolution by Texture-Depth Transformer. In *IEEE International Conference on Multimedia & Expo (ICME)*, pages 1–6, 2021.

[Zhang *et al.*, 2018a] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Springer European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[Zhang *et al.*, 2018b] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018.

[Zhang *et al.*, 2021] Meng Zhang, Tuanfeng Wang, Duygu Ceylan, and Niloy J Mitra. Deep detail enhancement for any garment. *Wiley Computer Graphics Forum*, 40(2):399–411, 2021.