MNSRNet: Multimodal Transformer Network for 3D Surface Super-Resolution

Wuyuan Xie, Tengcong Huang
College of Computer Science and Software Engineering, Shenzhen University
wuyuan.xie@gmail.com

Miaohui Wang*
Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University
wang.miaohui@gmail.com

Abstract

With the rapid development of display technology, it has become an urgent need to obtain realistic 3D surfaces with as high-quality as possible. Due to the unstructured and irregular nature of 3D object data, it is usually difficult to obtain high-quality surface details and geometry textures at a low cost. In this article, we propose an effective multimodal-driven deep neural network to perform 3D surface super-resolution in 2D normal domain, which is simple, accurate, and robust to the above difficulty. To leverage the multimodal information from different perspectives, we jointly consider the texture, depth, and normal modalities to simultaneously restore fine-grained surface details as well as preserve geometry structures. To better utilize the cross-modality information, we explore a two-bridge normal method with a transformer structure for feature alignment, and investigate an affine transform module for fusing multimodal features. Extensive experimental results on public and our newly constructed photometric stereo dataset demonstrate that the proposed method delivers promising surface geometry details compared with nine competitive schemes.

1. Introduction

With the increasing improvements of the capability and demand in the sensing and analyzing of real-world objects, more and more 3D vision-based applications require the input of high-quality object surface [11, 43]. However, most current 3D acquisition devices do not provide high-quality 3D data. In view of this practical difficulty, it is desirable to develop low-cost computer vision methods to enhance the acquisition quality for 3D data collectors.

Intuitively, the most straightforward way to improve the quality of an acquired 3D surface data is to directly perform the up-sampling operation in 3D domain. The existing studies can be classified as *voxel*-based, *point cloud*-based, and *mesh*-based methods according to the representation of a 3D surface. 1) The *voxel*-based methods [6] have been used in 3D surface processing for many years, which commonly have high requirements for equipment and computation. 2) *Point cloud* is the most simple way to represent a 3D object, which has been directly up-sampled [26, 40, 44] based on a special convolutional neural network (CNN) structure [29]. Due to the intrinsic irregularity of *point cloud*, it is difficult to achieve dense and high-quality 3D surface enhancement results. 3) *Mesh*-based methods, as the most wildly-used 3D representation, have been studied based on mesh subdivision and vertex interpolation [3]. With the development of deep neural networks, mesh-based CNN structures [12, 14, 32] have inspired several data-driven methods for the up-sampling operation on mesh-based 3D surfaces [24]. Nevertheless, these traditional schemes can only optimize some mathematical properties of the *mesh* data, while learning-based methods face the problem of large amounts of insufficient data.

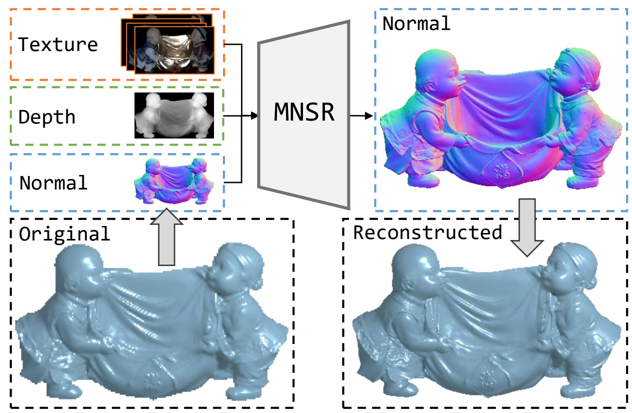

Figure 1. **Illustration of the proposed multimodal transformer framework for 3D surface super-resolution.** The texture, depth, and normal modalities are jointly investigated to perform 3D surface super-resolution in 2D domain.

Due to the aforementioned difficulties in improving surface quality in 3D domain, some preliminary investigations have aimed to enhance the surface quality in 2D domain. By

*This work was supported in part by the National Natural Science Foundation of China (No. 61902251 and No. 61701310), in part by Natural Science Foundation of Shenzhen City (No. 20200805200145001 and No. JCYJ20180305124209486), and in part by Natural Science Foundation of Guangdong Province (No. 2019A1515010961 and No. 2021A1515011877). (*Corresponding author*: *Miaohui Wang*)

representing 3D surface in 2D domain using normals and displacements in the field of physical cloth enhancement [19], the related 3D surface has been indirectly up-sampled through 2D image super-resolution (SR) algorithms [45]. This kind of strategy can avoid a high computational complexity, which is also benefited from well developed 2D image SR techniques. However, these existing methods in 2D domain usually only explore a single modality, which is lack of utilizing the multimodal attributes of 3D objects to further improve the performance of up-sampling.

Inspired by the above discussions, we present a multi-modal transformer network for 3D surface super-resolution by jointly considering the texture, depth, and normal modalities as shown in Fig. 1. More specifically, the texture, depth, and normal data are obtained from a low-resolution 3D object surface. Then, the texture and depth modalities are firstly aligned by a transformer network to the normal modality, and the related side features are fused into the main SR backbone network. Finally, a fine-grained 3D object surface is reconstructed by the enhanced normal map. To sum up, there are three main contributions compared with the previous approaches:

- To better utilize the modality information acquired by camera sensors, we investigate a novel multimodal-driven surface super-resolution network (denoted as "MNSRNet") to fuse the texture and depth modalities so as to enhance a 3D object surface in 2D domain.
- To capture the auxiliary modality information more easily, the original texture photographs are divided into hierarchical texture representations in multimodal pre-processing stage (MPS). Further, we design a new cross-modality transformer alignment (cmTA) module to align auxiliary modality information, and explore a cross-modality affine fusion (cmAF) module based on affine transform mechanism to fuse the intermediate features.
- Due to the lack of multimodality training data, we have also established a new *photometric stereo* dataset[1] which consists of 400 objects. Extensive experimental results on public and our newly constructed datasets demonstrate that the proposed method achieves superior performance compared with 9 competitive methods.

2. Related Work

In this section, we briefly review some representative image-based SR methods, including single image super-resolution (SISR) and multimodal image super-resolution (MISR), because the proposed 3D surface super-resolution framework is mainly conducted on 2D normal image.

[1] https://drive.google.com/file/d/1At34c7LrIQ_qcJLtFZqbjotngk_cQNeB/view?usp=sharing

2.1. Single Image Super-resolution

CNN-based SISR method [9] has been widely developed in the past few years. By introducing the residual learning, Kim *et al*. introduced VDSR [17] and DRCN [18] to ease the training difficulty. Lim *et al*. proposed EDSR [23] to cut some unnecessary CNN modules, and established a deeper network. To handle unknown degradation, Shocher *et al*. developed a zero-shot learning network [35]. With the success of self-attention mechanism in the field of natural language processing (NLP), the transformer-based structure has been studied [5, 42]. Besides, some other useful modules have been also introduced in SISR, such as Laplacian pyramid structure [20], dense residual structure [47], generative adversarial network (GAN) [21, 39], attention mechanism [7, 27, 46], dual regression network [13], *etc*.

The existing SISR methods have achieved promising results on natural RGB images. However, the normal image is totally different from the RGB image, where a normal pixel represents the geometry information. For instance, two adjacent normal pixels may be completely different, and there is lack of the characteristics of smooth magnitude changes in an RGB image. In view of this, it is necessary to develop new approaches for 3D surface super-resolution in 2D normal domain.

2.2. Multimodal Image Super-resolution

The idea of combining multimodal information, (*e.g*., different view-points, different sensors, different domains), is a popular research topic in computer vision [36, 48]. In MISR, some researchers have employed multimodal information to enhance the reconstruction performance. For instance, Almasri *et al*. [2] adopted a high-resolution image information to up-sample a thermal image obtained by the thermal camera. Wang *et al*. [38] utilized the image segmentation map as a prior information to improve the learning performance of the GAN model. Li *et al*. [22] employed a normal image to guide the super-resolution of the texture image. Deng *et al*. [8] introduced two images with different exposures to perform the SR task.

It demonstrates that MISR has been studied in some preliminary investigations, exploration of these methods for 3D object surface super-resolution based on multimodality is still in its infancy, partly due to the difficulty of identifying suitable multimodal descriptors to represent the distinct features of 3D surface. To our knowledge, there are few methods to consider multimodal information in up-sampling a 3D surface in 2D domain. This is the fundamental motivation of this study.

3. Proposed Method

3.1. Overview

Problem formulation. Our goal is to up-sample a surface normal map, and then reconstruct it into an enhanced 3D

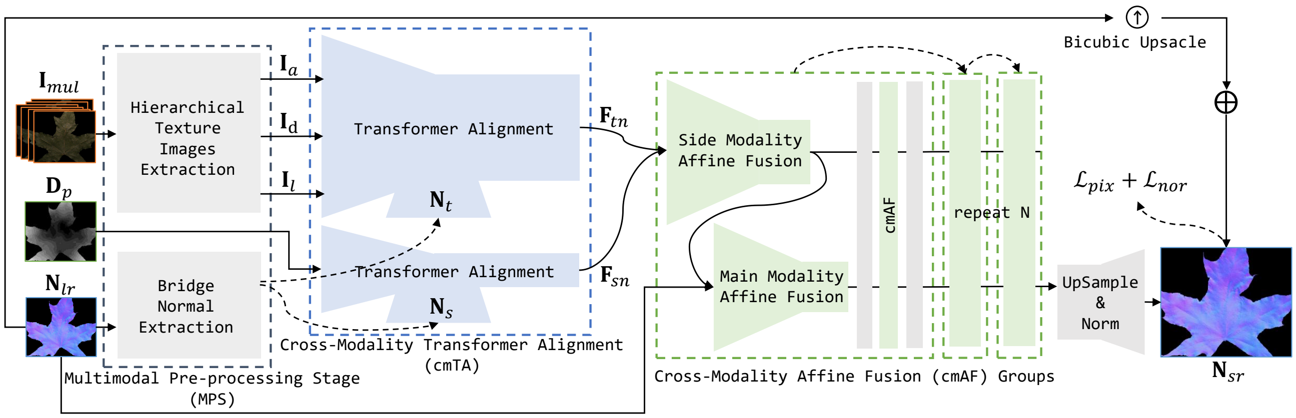

Figure 2. **Overview of the proposed multimodal super-resolution network for 3D object surface in 2D domain.** It mainly consists of the multimodal pre-processing stage, cross-modality transformer alignment, and cross-modality affine fusion. The SR normal map is reconstructed as an enhanced 3D surface by *surface-from-normal* in *photometric stereo*.

object surface. Since we propose to represent the 3D geometry surface in 2D normal domain, it becomes a normal image super-resolution problem. Therefore, the overall task can be formulated as the optimization of minimizing a specially-designed distance between the SR normal map $\mathbf{N}_{sr}$ and the ground-truth normal map $\mathbf{N}_{gt}$.

$$\min_{\mathbf{N}_{sr}} \ (\mathcal{L}_{overall}(\mathbf{N}_{sr}, \mathbf{N}_{gt})), \tag{1}$$

where $\mathcal{L}_{overall}(\cdot)$ represents the special distance which can be expressed as a weighted sum of the normal pixel loss $\mathcal{L}_{pix}$ and the normal angle loss $\mathcal{L}_{nor}$. Then, we have

$$\begin{aligned} &\mathcal{L}_{overall}(\mathbf{N}_{sr}, \mathbf{N}_{gt}) \\ &= \lambda_{pix}\mathcal{L}_{pix}(\mathbf{N}_{sr}, \mathbf{N}_{gt}) + \lambda_{nor}\mathcal{L}_{nor}(\mathbf{N}_{sr}, \mathbf{N}_{gt}) \\ &= \frac{\lambda_{pix}}{h \times w \times c}|\mathbf{N}_{sr} - \mathbf{N}_{gt}|_1 + \frac{\lambda_{nor}}{h \times w}\sum_{i,j}(1 - \mathbf{n}_{i,j}^{\top}\tilde{\mathbf{n}}_{\mathbf{i},\mathbf{j}}) \end{aligned}, \tag{2}$$

where (h, w, c) represents the height, width, and channel of a predicted normal image.

$\mathcal{L}_{pix}$ represents a pixel-wise L_1 loss, which is commonly used in SISR to accelerate the training convergence. $\mathcal{L}_{nor}$ represents the cosine similarity to restrict the angle loss between the predicted normal $\mathbf{n}_{i,j}$ and the ground-truth normal $\tilde{\mathbf{n}}_{i,j}$. Training with the balance of these two loss measures, our model achieves the minimum reconstruction error in practice.

Architecture. Previous studies have witnessed the positive effect of multimodal data in the SR task [2]. In light of this, we adopt three modalities in *photometric stereo*, including the texture, depth, and normal images. The texture and depth images are obtained under different lighting conditions at the same view.

The overall of the proposed network architecture is shown in Fig. 2, and formulated as

$$\mathbf{N}_{sr} = \mathcal{M}_{SR}(\mathbf{N}_{lr}, \mathcal{M}_{EX}(\mathbf{I}_{mul})). \tag{3}$$

where $\mathbf{I}_{mul}$ represents the raw multimodal information, including multi-lighting texture images, depth image, and low-resolution (LR) normal image. Generally, Eq. (3) can be decomposed into two sub-tasks: multimodal feature extraction $\mathcal{M}_{EX}$, and multimodal super-resolution $\mathcal{M}_{SR}$.

The multimodal feature extraction stage, $\mathcal{M}_{EX}$, consists of the MPS and cmTA modules. $\mathbf{I}_{mul}$ is firstly processed by the MPS module to produce side modality features used to bridge the related normal maps. Then, these resulted features are fed to the cmTA module, M_{cmTA}, which has a transformer structure acted as a feature encoder, aligning and extracting intermediate features from different modalities. $\mathcal{M}_{EX}(\cdot)$ can be represented by

$$\mathbf{F}_{tn}, \mathbf{F}_{sn} = \mathcal{M}_{cmTA}(\mathcal{S}_{MPS}(\mathbf{I}_{mul}, \mathbf{D}_p, \mathbf{N}_{lr})), \tag{4}$$

where $\mathbf{F}_{tn}$ and $\mathbf{F}_{sn}$ are the aligned side modality features. $(\mathbf{I}_m, \mathbf{D}_p, \mathbf{N}_{lr})$ represents the multi-lighting photographs, depth image, and LR normal image, respectively.

After this cross-modality alignment, several cmAF blocks (formed a cmAF sequence) are employed to fuse the side modality features and the main modality features together. Subsequently, the fused feature maps are fed to an up-sampling module, which consists of one upscale block, two 3×3 convolution layers, one vector normalization module, and one Bicubic interpolation module connected from the beginning for the residual learning. $\mathcal{M}_{SR}(\cdot)$ can be formulated as

$$\mathbf{N}_{sr} = \phi(\mathcal{M}_{UP}(\mathcal{M}_{cmAFs}(\mathbf{F}_{lr}, \mathbf{F}_{tn}, \mathbf{F}_{sn}), \tag{5}$$

where $\mathbf{F}_{lr}$ denotes a main modal feature starting with the shallow features of the LR normal map extracted by three convolutional layers. $\mathcal{M}_{cmAFs}(\cdot)$ denotes a cmAF sequence, and $\mathcal{M}_{\mathcal{UP}}(\cdot)$ represents a upscale block. $\phi(\cdot)$ denotes the combination of convolutional layer, vector normalization, and Bicubic interpolation. The vector normalization layer limits the output normal to the unit length, and

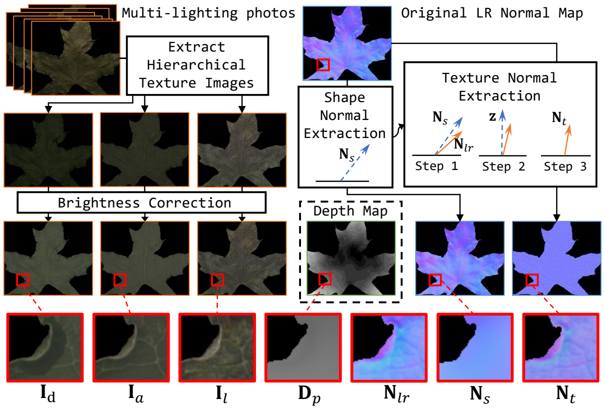

Figure 3. **Multimodal pre-processing stage (MPS).**

the up-sampling module enlarges a feature map to the output size. There are some choices in the up-sampling module, such as *deconvolution* [28] and *pixel shuffle* [34]. For a simple demonstration, we adopt [34] as our upscale block in the experiments.

3.2. Multimodal Pre-processing Stage (MPS)

In the MPS module, we focus on two main problems: 1) how to reduce the data distribution differences of the side modalities between different datasets, and 2) how to establish a correlation relationship between the side modality and main modality. For the first problem, we extract the hierarchical texture representations from multiple lighting photographs. For the second problem, we extract two bridge normal maps to connect the side modality information in normal domain. The overall pipeline of the proposed MPS is shown in Fig. 3.

Hierarchical texture. Due to the diversity of the target object materials and surface geometry structures, uncertainty of sensor, and 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 misleading noise data also contains useful information. To fully utilize this kind of information, we first calculate a pixel-wise darkest texture $\mathbf{I}_d$ to capture self-obscuring structures and under-exposure textures. Then, the lightest image $\mathbf{I}_l$ is extracted to capture the non-*Lambertian* reflection and over-exposure texture information. Finally, a pixel-wise average image $\mathbf{I}_a$ is extracted to represent the texture modality less affected by those unfavorable issues.

Since the brightness of these hierarchical textures can vary greatly and is not friendly used in the model training, we propose to adjust the brightness to the same value as much as possible. The brightness correction in Eq. (6) is done by calculating a shifting bias, and then the brightness is aligned to the maximum value without overflowing the maximum pixel magnitude.

$$\mathbf{I} = \mathbf{I}' + \max(\min(\beta, 1 - \max(\mathbf{I}')), -\min(\mathbf{I}')), \quad (6)$$

where $\beta = \mu - \mathbf{I}'$ denotes a shifting bias, μ denotes the overall average value in the training dataset. $\mathbf{I}'$ and $\mathbf{I}$ represents the hierarchical texture maps before and after the correction, respectively.

Bridge normal maps. As aforementioned, a surface normal image is very different from a natural RGB image. In such a case, the hierarchical texture images may contain some unfavorable information, which indicates that the side modalities may be inconsistent or misaligned to the main modality. Thus, we propose to use the texture normal map $\mathbf{N}_t$ and the shape normal map $\mathbf{N}_s$ as bridges between depth and normal, and texture and normal, respectively.

Inspired by the observation that a depth image is lack of the detail information but it contains a rough shape information and the position relationship of a given surface, we generate a shape normal map $\mathbf{N}_s$ by average filtering the normal map with the window size 3×3 and 100 times. The shape normal map can be reconstructed as a blurry surface, which is used to represent a rough object shape. Since the depth and shape normal maps have the similar structures, we use the shape normal map $\mathbf{N}$ in Eq. (7) as a guidance to align features from a depth image to the normal modality in the following cmTA module.

$$\mathbf{N}_s = conv(\mathbf{N}, \kappa_{ave}), \quad (7)$$

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

Similarly, we are looking forward to a hierarchical texture that can represent the pure texture information without the shape interference. To obtain the texture normal map $\mathbf{N}_t$, we propose to compute a directional bias between the original normal and the shape normal. This computation is illustrated in Fig. 3, and formulated as

$$\mathbf{N}_t = rot(\mathbf{N}_{lr}| < \mathbf{N}_s, \mathbf{z} >), \quad (8)$$

where $rot(\cdot)$ denotes a rotate manipulation, $< \cdot, \cdot >$ denotes an element-wise rotation, and $\mathbf{z} = [0, 0, 1]^t$ represents the z-axis direction. The texture normal map $\mathbf{N}_t$ contains less shape information, which flattens the reconstructed surface. $\mathbf{N}_t$ represents the high-frequency detail information of a given surface, which is similar to the extracted texture image without the shape information. Consequently, we propose to use the texture normal map as a guidance to align the texture modality from the RGB domain to the normal domain.

3.3. Cross-modality Transformer Alignment (cmTA)

To align the above cross-modality information, we further design a cross-modality transformer alignment (cmTA) module as shown in Fig. 4. Before cmTA, all the input multimodalities are passed through three 3×3 convolutional layers to extract the related shallow feature $\mathbf{F}_\mathrm{x}$ (*i.e.*, $\mathrm{x} = \{a, l, d\}$). In other words, $\mathbf{I}_a$, $\mathbf{I}_l$, and $\mathbf{I}_d$ will be mapped

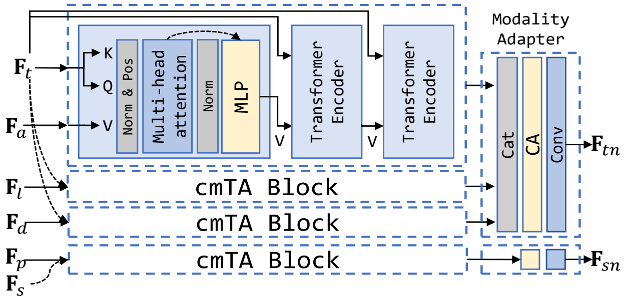

Figure 4. **Cross-modality transformer alignment (cmTA).**

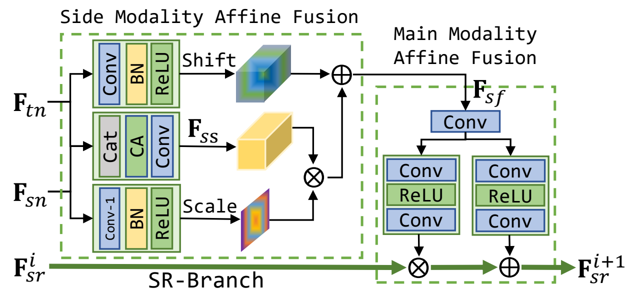

Figure 5. **Cross-modality affine fusion (cmAF).**

to $\mathbf{F}_a$, $\mathbf{F}_l$, and $\mathbf{F}_d$, respectively. Since these extracted hierarchical texture features may contain rich color and structural information, their high-frequency information is more complicated than the original normal map. By introducing the texture normal map as a bridge, the cmTA module can capture more texture features, and project them to the normal domain. Similarly, for the depth feature $\mathbf{F}_p$, since a depth image not only carries the shape information but also carries the reconstructed position information, it can assist in reconstructing the low-frequency of a 3D surface in the shape normal map domain.

Firstly, a cmTA module can be considered as the combination of several cmTA blocks and two modality adapters. Each cmTA block uses the transformer structure to align one modality feature with a bridge normal feature, and produces an aligned feature (*e.g.*, average image feature $\mathbf{F}_a$ aligned with the texture normal feature $\mathbf{F}_{an}$).

$$\mathbf{F}_{\mathtt{x}n} = \mathcal{B}_{cmTA}(\mathbf{F}_{\mathtt{x}}, \mathbf{F}_{\delta}), \tag{9}$$

where $\mathtt{x}$ denotes one of the candidate modality feature, and δ denotes the corresponding bridge normal feature. It is noted that when $\mathtt{x}$ denotes a texture image, δ means a texture normal feature. Similarly, when $\mathtt{x}$ denotes a depth image, δ means a shape normal feature.

In Eq. (9), $\mathcal{B}_{cmTA}(\cdot)$ denotes a cmTA block. Inspired by the efficiency of a self-attention mechanism [37] in capturing global information, the proposed cmTA module consists of several transformer encoders similar to [10]. Since we expect that the proposed deep network can capture the cross-modality features and map them to the bridge normal features, the cmTA block is organized in the following recursive structure, and defined in Eq. (10).

$$\mathbf{F}^{i+1}_{\mathtt{x}n} = \begin{cases} \mathcal{B}^i_{cmE}(q, k=\mathbf{F}_\delta, v=\mathbf{F}_{\mathtt{x}}), & i=1 \\ \mathcal{B}^i_{cmE}(q, k=\mathbf{F}_\delta, v=\mathbf{F}^i_{\mathtt{x}n}), & n \geq i > 1 \end{cases}, \tag{10}$$

where $\mathcal{B}^i_{cmE}(q,k,v)$ denotes the i^{th} cross-modality encoder (cmE), and (q,k,v) denotes the self-attention layer paradigms (query, key, and value). Inside a cmE, each feature map is cropped to 9 patches with the position embedding. Subsequently, using the self-representation in term of self-attention, we treat the bridge normal and modality features as one, and adopt a multi-head attention mechanism to learn the representation between them. By using multiple recursive structures in Eq. (10), the bridge normal feature is repeatedly connected to the k and q inputs of the cmE module via the skip connection structure. In this case, the proposed network expects to capture the relevant information between different modalities, and gradually align the related information in normal domain.

After the cmTA module, each modality feature pair would jointly generate a cross-modality feature. However, as mentioned above, a hierarchical texture map contains some unfavorable information, thus we further adopt two modality adapters to reduce the relevant unfavorable information. The modality adapters use the channel attention (CA) [16] followed by one convolution block to model the importance relationship between different channels. Specifically, it will produce three texture modalities ($\mathbf{F}_{an}$, $\mathbf{F}_{ln}$, and $\mathbf{F}_{dn}$) from a cmTA module. Then, we concatenate them together as $\mathbf{F}_{xn} \in \mathbb{R}^{3f\times h\times w}$, and use a CA layer followed by a 1×1 convolution block to distill the most important texture information $\mathbf{F}_{tn} \in \mathbb{R}^{f\times h\times w}$. After the process of these two modality adapters, two aligned texture feature $\mathbf{F}_{tn}$ and shape feature $\mathbf{F}_{sn}$ are generated, namely side-modality features.

3.4. Cross-modality Affine Fusion (cmAF)

After the cross-modality alignment, side-modality features are fused into the main-modality to assist the target SR feature representation. Inspired by the spatial feature transform mechanism [38] which takes a segmentation probability map as a prior, we propose the cmAF module, $\mathcal{M}_{cmAF}$, to fuse the extracted texture and shape features into the backbone of the SR network step by step. As shown in Fig. 5, cmAF can be separated into the side-modality affine fusion stage and the main-modality affine fusion stage.

In the side-modality affine fusion stage, the CA layer firstly is used to fuse $\mathbf{F}_{sn}$ and $\mathbf{F}_{tn}$, which aims to produce the side-stream feature $\mathbf{F}_{ss}$, representing the joint guidance combined both the texture and shape information. As mentioned earlier, the shape texture is lack of the detail information, but it has a strong relationship with the surface vertex position and structure. We then adopt a 3×3 convolution block to distill a general shape feature from $\mathbf{F}_{sn} \in \mathbb{R}^{f\times h\times w}$ to $\mathbb{R}^{1\times h\times w}$. Based on the fact that a texture feature map contains more information than a shape feature map, the texture feature map can provide more detailed shifting information, and should not be distilled. Thus, we use a 3×3 convolution block to obtain a general shifting feature from $\mathbf{F}_{tn} \in \mathbb{R}^{f\times h\times w}$ to $\mathbb{R}^{f\times h\times w}$. Finally, $\mathbf{F}_{ss}$ will be pixel-wisely multiplied by a scaling map, and then added to the shifting map. The side-modality affine fusion in Fig. 5 can be formulated as

$$\mathbf{F}_{sf} = \mathbf{F}_{ss} \otimes \mathcal{C}_E^1(\mathbf{F}_{sn}) \oplus \mathcal{C}_E^f(\mathbf{F}_{tn}), \tag{11}$$

where $\mathbf{F}_{sf}$ denotes the side feature. $\otimes$ and $\oplus$ refer to the element-wise multiplication and addition, respectively. $\mathcal{C}_E^x(\cdot)$ denotes a convolution module consisting of one BN layer and one ReLU activation layer, which aims to make the output stable and easier to use.

After obtaining the side feature $\mathbf{F}_{sf}$, we further employ a convolution module to learn another affine transform to fuse different modality features. Since the information of $\mathbf{F}_{sf}$ has been heavily distilled, we use two convolution blocks with one ReLU layer to prepare the corresponding affine shifting and scaling maps for the main-modality feature $\mathbf{F}_{sr}$ which is defined in Eq. (12).

$$\mathbf{F}_{sr}^{i+1} = \mathbf{F}_{sr}^{i} \otimes \mathcal{C}_F(\mathbf{F}_{sf}) \oplus \mathcal{C}_F(\mathbf{F}_{sf}) \tag{12}$$

where $\mathbf{F}_{sr}^{i}$ denotes the i^{th} main-modality feature in different stages of the SR side branch, whose dimension is the same as $\mathbf{F}_{sf}$. $\mathcal{C}_F(\cdot)$ denotes a linear structure of two convolution layers and one ReLU layer. As shown in Fig. 2, this cmAF module will be repeated several times in order to fully fuse and leverage the cross-modality information. Finally, the resulted features will be fed into our upscale module as the rest part of Eq. (5).

4. Experimental Results

4.1. Experimental Protocols

Dataset descriptions. The training of MNSRNet requires high-resolution labels. Currently, the most widely-used *photometric stereo* datasets do not have enough images for the multimodality training, such as the *DiLiGenT* dataset (10 objects) [33] and the *Gourd & Apple* dataset (3 objects) [1]. Therefore, we have established a new *photometric stereo* dataset, namely WPS (wonderful photometric stereo). WPS contains 400 different objects, including *butterfly wings*, *leaves*, *oil paintings*, *handicrafts*, *etc*. Each object is captured under 18 predefined lighting conditions as shown in Fig. 6.

To fairly evaluate the performance of the proposed MNSRNet, the testing dataset is composed of *DiLiGenT*, *Gourd & Apple*, and 80 objects selected from WPS. Note that the rest of objects in WPS are only used for training (*e.g.*, 9(training):1(validation)). 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 inputs.

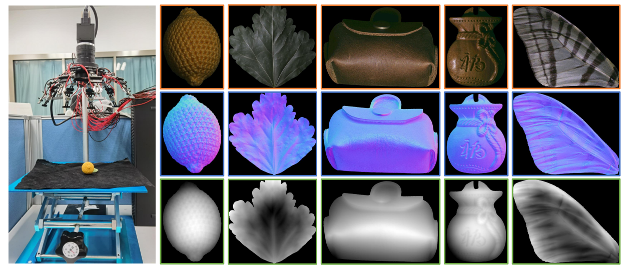

Figure 6. **Illustration of the new wonderful photometric stereo (WPS) dataset**. The first left image shows our set-up for establishing the WPS dataset, which provides 18 lighting directions. The right side shows some multimodal examples in WPS: from the top to the bottom are the original texture modality, normal modality, and depth modality, respectively.

3D surface reconstruction. After obtaining the normal map $\mathbf{N}_{sr}$, we can integrate it to obtain the final reconstructed object surface, which is also called *surface-from-normal* (SfN) [31]. Specifically, we have adopted the public available discrete geometry-based SfN method [41] to reconstruct a SR 3D surface.

Evaluation metrics. For quantitative comparisons, we adopt various quality measurements to evaluate the performance as comprehensively as possible. In image domain, we take the commonly-used indicators in the SISR task, such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

Besides, two widely used metrics are used to quantitatively measure the 3D reconstruction results [31, 41], including mean angular error (MAE) and mean relative depth error (MRDE).

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

where $\tilde{\mathbf{n}}_{i,j}$ and $\mathbf{n}_{i,j}$ denotes the predicted normal and the ground-truth normal, respectively. $\|\mathbf{N}\|$ represents the total number of input normal pixels. In normal domain, five statistical indicators are computed in terms of MAE, including *Mean*, *Median*, 5°, 10°, and *Variation*.

MRDE is used to evaluate the accuracy of the estimated vertexes.

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

where $\tilde{\mathbf{p}}_{i,j}$ and $\mathbf{p}_{i,j}$ denote the vertex position of the reconstructed surface by [41] and the ground-truth surface, respectively.

To sum up, the first two indicators (PSNR and SSIM) assess the prediction accuracy, and the higher the better. The next three indicators (MEAN, MID, and VAR) capture the mean, median, and variation of the angular error, and the lower the better. The sixth and seventh indicators (5° and 10°) represent the percentage of pixels within 5- or 10-degree angular error, and the higher the better. The last indi-

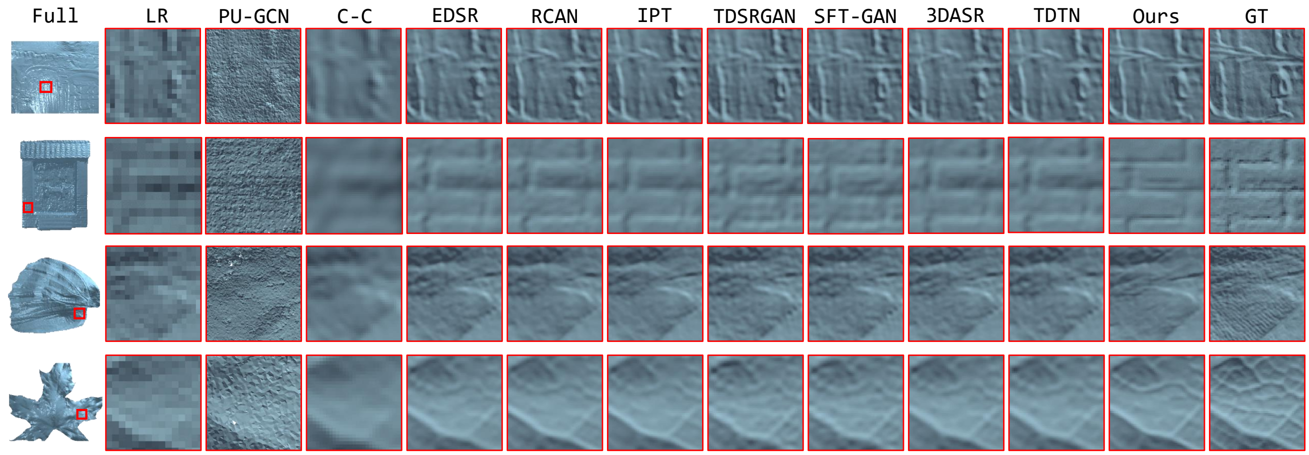

Figure 7. **Visual comparisons of 3D surface super-resolution** between 10 methods under the $\times 4$ setting. For a better comparison, the region in the red box is zoomed in the 2nd-13th columns. "Full" means the original surface, "LR" means the down-sampled object surface, and "GT" means the ground-truth. **Please zoom in the electronic version for better details.**

cator (MRDE) assess the reconstructed quality of 3D object surface, and the lower the better.

Implementation details . MNSRNet has been implemented in *PyTorch*, and the *Adam* optimizer is used with default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). For the SR branch, we use 20 groups of cmAF. We have trained MNSRNet using a mini-batch size of 8 for 1000 epochs with an *Nvidia Tesla A100* GPU, which takes about two days and nights. Since the transformer module needs a fixed input size, all the input images are adaptively cropped. For example, in $\times 4$ scale, the HR and LR image patches are 196×196 and 48×48, respectively. All of the trained weights for each layer are initialized by the *Kaiming* distribution [15], and the bias is initialized as a constant. We do not apply any special data augmentation methods except for the random rotation (90°, 180°, and 270°) and the horizontal flip.

4.2. Performance Comparisons

Comparison methods. We have compared our MNSRNet with 9 representative methods, which can be categorized into four groups: *mesh*-based method (denoted by "*Mesh*"), *point cloud*-based method (denoted by "*Points*"), SISR-based methods (denoted by "*SISR*"), and MISR-based methods (denoted by "*MISR*").

For *Mesh* methods, Catmull-Clark subdivision (C-C) [25] has been the most widely-used mesh subdivision method. It can efficiently up-sample a triangular mesh by the heuristic algorithms. We have used the implementation version built in *Blender* for comparison.

For *Points* methods, we choose the PU-GCN network [30] to represent the SR task for *point clouds*. In the experiments, we convert the related meshes into *point clouds* for PU-GCN, perform the up-sampling, and re-convert it into meshes for comparison [4].

For *SISR* methods, we choose EDSR [23] to represent a residual learning structure, RCAN [46] to represent a convolutional attention structure, and IPT [5] to represent a self-attention structure. For these methods, we have fine-tuned the corresponding models on WPS to show their best performance.

For *MISR* methods, we choose TDSRGAN [2] to represent a early fusion method, SFT-GAN [38] and 3DASR [22] to represent a hybrid fusion method, and TDTN [8] to represent a hybrid fusion method with the *self-attention* structure. It is noted that our task cannot fully provide the modalities needed in the original methods, and we have adjusted the above methods to fit for our WPS benchmark.

Qualitative results. Fig. 7 demonstrates the visual comparisons of some representative 3D objective surfaces. For *SISR*, benefiting from the powerful natural image pretraining model, some methods still can perform well after fine-tuning on our WPS dataset. However, since these SISR methods have not considered the cross-modal information, they are not sufficient to obtain the best visual quality. For *MISR*, they may not take full advantage of the additional multimodal information. As a result, they may even have negative effects (*e.g.*, heavily aliased surfaces) due to intermodal differences. Visually, the proposed method achieves a promising subjective quality with enough surface details and geometry structures.

Quantitative results. Table 1 summarizes the detailed average results on the hybrid testing dataset, including *DiLiGenT*, *Gourd & Apple*, and WPS. Specifically, our method achieves all 8 of the first-best results in terms of PSNR, SSIM, MEAN, MID, VAR, 5°, 10°, and MRDE on the $\times 2$ setting, and achieves 6 of the first-best results and 2 of the second-best results on the $\times 4$ setting. Experiments show that better results are obtained without the negative effects of instability caused by the multimodal inputs and information confusion between multimodalities. MNSRNet outperforms the existing methods in most cases. The main reason can be that our method can employ more cross-modality

Table 1. **The average comparison results between 10 state-of-the-art methods** on the hybrid testing datasets. "[+]" means the higher the better, and "[-]" means the lower the better. The first-best is highlighted by ***bold***, and the second-best is highlighted by *underline*.

Scale	Type	Algorithm	PSNR[+]	SSIM[+]	MEAN[-]	MID[-]	VAR[-]	5°[+]	10°[+]	MRDE[-]
×2	Points	PU-GCN [30]	18.5747	0.7382	16.0982	12.3674	252.8209	0.3369	0.6083	16.5069
	Mesh	C-C [25]	22.1881	0.9010	9.0866	4.7667	162.4632	0.5934	0.7741	11.0118
	SISR	EDSR [23]	27.5593	0.9522	5.1324	2.0439	157.7664	0.7845	0.8942	5.1496
		RCAN [46]	27.7209	0.9535	5.0834	1.9327	153.1603	0.7819	0.8908	5.1578
		IPT [5]	27.9756	0.9545	4.9149	1.7689	166.2459	0.8076	0.9008	4.7418
	MISR	TDSRGAN [2]	26.1982	0.9434	6.3183	2.6441	182.8687	0.7019	0.8577	5.6872
		SFT-GAN [38]	27.3630	0.9509	5.3260	2.1304	162.4754	0.7721	0.8898	5.5245
		3DASR [22]	28.3017	0.9581	4.6702	1.7837	149.0149	0.8133	0.9069	5.2261
		TDTN [8]	27.8263	0.9555	4.7814	1.7344	158.1208	0.8111	0.9030	5.5026
		Ours	**28.7662**	**0.9605**	**4.4277**	**1.6312**	**146.0815**	**0.8303**	**0.9123**	**4.5849**
×4	Points	PU-GCN [30]	15.8404	0.5917	24.3387	19.5136	443.0975	0.1390	0.3607	20.5933
	Mesh	C-C [25]	20.9022	0.8539	11.8487	6.5019	352.5511	0.4860	0.6836	13.7369
	SISR	EDSR [23]	23.0609	0.8909	9.1407	3.7643	323.7065	0.6302	0.7951	7.0694
		RCAN [46]	23.6058	0.9024	8.7591	3.5196	344.9552	0.6542	0.8120	6.8314
		IPT [5]	23.6268	0.9041	8.2695	**2.8598**	335.1193	**0.7063**	0.8275	6.6668
	MISR	TDSRGAN [2]	22.4461	0.8819	10.5153	4.7397	345.8229	0.5370	0.7277	10.5188
		SFT-GAN [38]	22.7683	0.8816	10.0431	4.7215	330.2237	0.5551	0.7606	9.6531
		3DASR [22]	22.9138	0.8901	9.2341	3.6517	349.1082	0.6473	0.8010	6.9161
		TDTN [8]	22.6384	0.8861	9.4434	3.4755	371.0740	0.6524	0.7993	7.3051
		Ours	**23.7961**	**0.9053**	**7.9945**	2.9255	**302.6197**	0.6993	**0.8307**	**6.4827**

Table 2. **Ablation experiments** on the proposed cross-modality alignment and fusion methods.

MPS	cmTA	cmAF	PSNR[+]	SSIM[+]	MEAN[-]	MRDE[-]
×	×	×	22.3302	0.8752	10.7182	11.7245
✓	×	×	23.2812	0.8949	8.9813	7.2617
✓	✓	×	23.4994	0.8995	8.7082	7.2021
✓	×	✓	23.6824	0.9043	8.2285	6.6091
✓	✓	✓	**23.7961**	**0.9053**	**7.9945**	**6.4827**

Table 3. **Ablation experiments on different modalities** (×2 setting). The modality selected (not selected) is represented by ✓(×).

Normal $\mathbf{N}_{lr}$	Texture $\mathbf{I}_{mul}$	Depth $\mathbf{D}_p$	PSNR	SSIM	MEAN[-]	MRDE[-]
✓	×	×	27.9376	0.9536	4.9883	5.0325
✓	✓	×	28.3151	0.9564	4.6942	4.8478
✓	×	✓	28.3816	0.9585	4.5611	4.7091
✓	✓	✓	28.7662	0.9605	4.4277	4.5849

information to help capture more comprehensive details, which is hard to learn using a single modality.

4.3. Ablation Study

MNSRNet contains three main modules for the multi-modality learning, such as MPS, cmTA, and cmAF. 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 module selected (not selected) is represented by the symbol "✓"("×").

In the experiment, the replacement of MPS is to use the lightest texture image as the texture modality, and the other two modalities remain unchanged. The replacement of cmTA is to simply 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. The replacement of cmAF is a combination of the CA layer and 1×1 convolution layer to fuse the side-modality to the main-modality. Experiments show that when all three modules are used, the best results can be achieved.

To demonstrate the effects of a single modality, we have conducted the additional experiments as provided in Table 3. As seen, both the texture and depth modalities can effectively improve the performance of the surface super-resolution.

5. Conclusion

In this paper, we have introduced a multimodal-based super-resolution network for 3D object surface in 2D normal domain. More specifically, we jointly considered the texture, depth, and normal modalities to restore high-quality surface details and preserve geometry structures. To effectively utilize the cross-modality information, we extracted two bridge normal maps as a cross-modality alignment guidance. Based on the texture and depth modalities, we have developed a cross-modality transformer alignment (cmTA) module to connect different modalities. In addition, we explored a cross-modality affine fusion (cmAF) module to fuse the features from the main network branch and the extracted side modalities. Finally, we reconstructed an enhanced 3D object surface from the recovered high-resolution normal map. Experimental results on different benchmark datasets demonstrate the effectiveness of the proposed approach in qualitatively and quantitatively.

References

[1] 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. 6

[2] Feras Almasri and Olivier Debeir. Multimodal sensor fusion in single thermal image super-resolution. In *Springer Asian Conference on Computer Vision (ACCV)*, pages 418–433, 2018. 2, 3, 7, 8

[3] Kosala Bandara, Thomas Rüberg, and Fehmi Cirak. Shape optimisation with multiresolution subdivision surfaces and immersed finite elements. *Elsevier Computer Methods in Applied Mechanics and Engineering*, 300:510–539, 2016. 1

[4] 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. 7

[5] 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. 2, 7, 8

[6] Ian Cherabier, Christian Häne, Martin R Oswald, and Marc Pollefeys. Multi-label semantic 3D reconstruction using voxel blocks. In *IEEE International Conference on 3D Vision (3DV)*, pages 601–610, 2016. 1

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11065–11074, 2019. 2

[8] 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. 2, 7, 8

[9] 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. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[11] 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. 1

[12] Lin Gao, Jie Yang, Tong Wu, Yu-Jie Yuan, Hongbo Fu, Yu-Kun Lai, and Hao Zhang. Sdm-net: Deep generative network for structured deformable mesh. *ACM Transactions on Graphics*, 38(6):1–15, 2019. 1

[13] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5407–5416, 2020. 2

[14] 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. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 7

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 5

[17] 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. 2

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016. 2

[19] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Springer European Conference on Computer Vision (ECCV)*, pages 667–684, 2018. 2

[20] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 624–632, 2017. 2

[21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017. 2

[22] 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. 2, 7, 8

[23] 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)*, pages 136–144, 2017. 2, 7, 8

[24] Hsueh-Ti Derek Liu, Vladimir G Kim, Siddhartha Chaudhuri, Noam Aigerman, and Alec Jacobson. Neural subdivision. *arXiv preprint arXiv:2005.01819*, 2020. 1

[25] Charles Loop and Scott Schaefer. Approximating catmull-clark subdivision surfaces with bicubic patches. *ACM Transactions on Graphics*, 27(1):1–11, 2008. 7, 8

[26] 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. 1

[27] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Springer European Conference on Computer Vision (ECCV)*, pages 191–207. Springer, 2020. 2

[28] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *IEEE international Conference on Computer Vision (ICCV)*, pages 1520–1528, 2015. 4

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 1

[30] 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. 7, 8

[31] 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. 6

[32] 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. 1

[33] 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. 6

[34] 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. 4

[35] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3118–3126, 2018. 2

[36] Jun Sun, Yakun Chang, Cheolkon Jung, and Jiawei Feng. Multi-modal reflection removal using convolutional neural networks. *IEEE Signal Processing Letters*, 26(7):1011–1015, 2019. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 5

[38] 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. 2, 5, 7, 8

[39] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Springer European Conference on Computer Vision (ECCV) workshops*, pages 1–16, 2018. 2

[40] Huikai Wu, Junge Zhang, and Kaiqi Huang. Point cloud super resolution with adversarial residual graph networks. *arXiv preprint arXiv:1908.02111*, 2019. 1

[41] 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. 6

[42] 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. 2

[43] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Ec-net: an edge-aware point set consolidation network. In *Springer European Conference on Computer Vision (ECCV)*, pages 386–402, 2018. 1

[44] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2799, 2018. 1

[45] Meng Zhang, Tuanfeng Wang, Duygu Ceylan, and Niloy J Mitra. Deep detail enhancement for any garment. In *Computer Graphics Forum*, volume 40, pages 399–411, 2021. 2

[46] 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. 2, 7, 8

[47] 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. 2

[48] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 465–476, 2017. 2