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

---

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. 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.

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
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 multimodal 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 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.

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
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 \( N_{sr} \) and the ground-truth normal map \( N_{gt} \).

\[
\min_{N_{sr}} \mathcal{L}_{overall}(N_{sr}, N_{gt}), \tag{1}
\]

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

\[
\mathcal{L}_{overall}(N_{sr}, N_{gt}) = \lambda_{pix} \mathcal{L}_{pix}(N_{sr}, N_{gt}) + \lambda_{nor} \mathcal{L}_{nor}(N_{sr}, N_{gt}) = \frac{\lambda_{pix}}{h \times w \times c} \| N_{sr} - N_{gt} \|_1 + \frac{\lambda_{nor}}{h \times w} \sum_{i,j} (1 - \frac{\mathbf{n}_{i,j} \cdot \mathbf{n}_{i,j}}{\mathbf{n}_{i,j} \cdot \mathbf{n}_{i,j}})), \tag{2}
\]

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 \( \mathbf{n}_{i,j} \). 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

\[
N_{sr} = \mathcal{M}_{SR}(N_{lr}, \mathcal{M}_{EX}(I_{mul})). \tag{3}
\]

where \( I_{mul} \) 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. \( I_{mul} \) 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, \( \mathcal{M}_{cmTA} \), 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

\[
F_{tn}, F_{sn} = \mathcal{M}_{cmTA}(S_{MPS}(I_{mul}, D_p, N_{lr})), \tag{4}
\]

where \( F_{tn} \) and \( F_{sn} \) are the aligned side modality features. \( (I_{mr}, D_p, N_{lr}) \) 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 \( 3 \times 3 \) 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

\[
N_{sr} = \phi(\mathcal{M}_{UR}(\mathcal{M}_{cmAFs}(F_{lr}, F_{tn}, F_{sn}))), \tag{5}
\]

where \( F_{lr} \) 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 \( \mathcal{M}_{UR}(\cdot) \) represents a upscale block. \( \phi(\cdot) \) denotes the combination of convolutional layer, vector normalization, and Bi-cubic interpolation. The vector normalization layer limits the output normal to the unit length, and
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, specular 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 the maximum pixel magnitude. 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.

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

where $\beta = \mu - \mathbf{I'}$ denotes a shifting bias, $\mu$ 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 $3 \times 3$ 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.

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

where $\text{conv}(:)$ represents the convolution operation, and $\kappa_{\text{ave}}$ 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

$$\mathbf{N}_t = \text{rot}(|\mathbf{N}_t| < \mathbf{N}_s, \mathbf{z} >),$$

where $\text{rot}(:)$ denotes a rotate manipulation, $\mathbf{z} = [0, 0, 1]^T$ 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 $3 \times 3$ convolutional layers to extract the related shallow feature $\mathbf{F}_x$ (i.e., $x = \{a, l, d\}$). 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).

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 ($F_{tn}$, $F_{sn}$, and $F_{dn}$) from a cmTA module. Then, we concatenate them together as $F_{sn} \in \mathbb{R}^{3f \times h \times w}$, and use a CA layer followed by a $1 \times 1$ convolution block to distill the most important texture information $F_{tn} \in \mathbb{R}^{f \times h \times w}$. After the process of these two modality adapters, two aligned texture feature $F_{tn}$ and shape feature $F_{sn}$ 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, $M_{cmAF}$, 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 $F_{sn}$ and $F_{tn}$, which aims to produce the side-stream feature $F_{ss}$, 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 $3 \times 3$ convolution block to distill a general shape feature from $F_{sn} \in \mathbb{R}^{f \times h \times w}$ 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 $3 \times 3$ convolution block to obtain a general shifting feature from $F_{tn} \in \mathbb{R}^{f \times h \times w}$ to $\mathbb{R}^{f \times h \times w}$. Finally, $F_{ss}$ 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

$$F_{sf} = F_{ss} \otimes C_E^f(F_{sn}) \otimes C_E^f(F_{tn}),$$

where $F_{sf}$ denotes the side feature. $\otimes$ and $\oplus$ refer to the element-wise multiplication and addition, respectively. $C_E^f(\cdot)$ denotes a convolution module consisting of one BN and one ReLU activation layer, which aims to make the element-wise multiplication and addition, respectively.

The next three indicators (MEAN, MID, and VAR) capture the statistical indicators are computed in terms of MAE, including mean angular error (MAE) and mean relative depth error (MRDE).

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

where $\tilde{n}_{i,j}$ and $n_{i,j}$ denotes the predicted normal and the ground-truth normal, respectively. $||N||$ 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.

$$\text{MRDE} = \frac{1}{||N||} \sum_{i,j} ||\tilde{p}_{i,j} - p_{i,j}||,$$

where $\tilde{p}_{i,j}$ and $p_{i,j}$ 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-
Full LR PU-GCN C-C EDSR RCAN IPT TDSRGAN SFT-GAN 3DASR TDTN Ours GT

Figure 7. Visual comparisons of 3D surface super-resolution between 10 methods under the ×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 (β₁ = 0.9 and β₂ = 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 ×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 pre-training 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 inter-modal 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 ×2 setting, and achieves 6 of the first-best results and 2 of the second-best results on the ×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
Ablation experiments on different modalities (×2 setting). The modality selected (not selected) is represented by ✓ (×).

<table>
<thead>
<tr>
<th>Normal Nlr</th>
<th>Texture Itxt</th>
<th>Depth Dp</th>
<th>PSNR</th>
<th>SSIM</th>
<th>MEAN</th>
<th>MRDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>27.9376</td>
<td>0.9536</td>
<td>4.9883</td>
<td>5.0325</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>28.3151</td>
<td>0.9564</td>
<td>4.6942</td>
<td>4.8478</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>28.3816</td>
<td>0.9585</td>
<td>4.5611</td>
<td>4.7091</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>28.7662</td>
<td>0.9605</td>
<td>4.4277</td>
<td>4.5849</td>
</tr>
</tbody>
</table>

As seen, both the texture and depth modalities can effectively improve the performance of the surface super-resolution.
References


