# DEPTH SUPER-RESOLUTION BY TEXTURE-DEPTH TRANSFORMER

*Chao Yao*[1,4], *Shuaiyong Zhang*[2], *Mengyao Yang* [3], *Meiqin Liu* [2], *Junpeng Qi*[3]

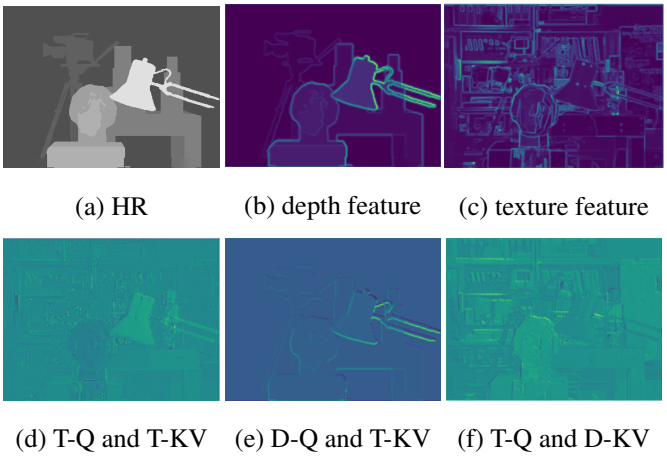
[1]University of Science and Technology, Beijing, 100083, China
[2]Beijing Jiaotong University, Beijing, 100044, China
[3]China Aerospace Academy of Systems Science and Engineering, Beijing, 100037, China
[4]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China.

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428393

## ABSTRACT

Depth maps have been still suffering from some non-negligible effects, resulting from the consumer-level sensors. The limited resolution of the acquired depth maps is one of these annoying issues. Many prominent researchers have recently made a lot of efforts, such as traditional filters, as well as the deep learning paradigms. However, depth super-resolution is still an open challenge. In this paper, we design a texture-depth transformer for depth super-resolution task, which can learn the corresponding structural information of the high-resolution texture images and the corresponding interpolated depth maps. Moreover, a multi-scale feature fusion strategy is exploited to further enhance the fusion feature. Complementary to a quantitative evaluation, we demonstrate the effectiveness of the proposed approach.

***Index Terms***— Transformer, Depth super-resolution, CNN, Residual learning



(a) HR (b) depth feature (c) texture feature

(d) T-Q and T-KV (e) D-Q and T-KV (f) T-Q and D-KV

**Fig. 1**. Visual analysis of transformer feature.

## 1. INTRODUCTION

Depth maps can accurately describe the structure information of scene due to each pixel value in depth maps represents the distance information corresponding to the scene. Nevertheless, due to the imaging limitation of depth sensors in practice, high quality and high resolution (HR) depth maps are difficult to be acquired directly. In the last decades, many depth super-resolution (SR) methods have been applied to recover the degraded depth maps, including traditional filter-based methods [1] [2], optimization-based methods [3] [4] and some latest convolutional neural network(CNN)-based methods [5] [6].

Early researchers concentrate on utilizing the local information to enhance the interpolated depth maps, especially for the edge information. Thus, some works based on low pass filter [1] [2] introduce edge guided information to interpolate the low resolution (LR) depth maps. However, depth boundaries are generally hard to reconstruct from LR depth maps and easy to lose sharpness particularly at large magnification factors due to the loss of spatial information. To refine the loss information, the optimization-based methods [3] [4] employed hand-craft optimization function and regularization term to constrain the edge structure of the interpolated depth maps, but accompanied by high computational cost and poor universality, which limits its application range.

With the rising of deep learning techniques, CNN has achieved most impressive performance in many computer vision tasks and recently has been applied to depth SR. As we observed, the backbone network of existing depth SR methods can be classified into two categories. On one hand, some researchers try to utilize low-to-high resolution networks to progressively extract features and raise the spatial resolution, which is similar as some classical single image super-resolution (SISR) methods [7] [8]. On another hand, the corresponding texture information is employed as guidance to recover the degraded depth maps [5] [6] [9]. Most of these approaches make the texture structural as the prior knowledge, however, texture discontinuities do not always coincide with depth maps, which results in texture blending in the reconstructed depth maps. Therefore, how to leverage texture information to help recover the depth maps and whether the texture images are required for all interpolation ratios,especially for

The corresponding author is Meiqin Liu(email:mqliu@bjtu.edu.cn)

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

the $2\times$ and $4\times$, still need to be developed and verified.

In this paper, we try to design a novel texture-depth transformer network (TDTN) for depth SR task, which aims to explore what texture information matters with depth SR. We introduce a transformer to learn the useful content information and structure information for depth SR task, from the interpolated depth maps and the corresponding texture images, respectively. Then, a multi-scale fusion strategy is exploited to improve the efficiency of texture-depth fusion. Experimental results show that the transformer indeed learn useful structural information from the corresponding texture images. Benefiting from the multi-scale fusion, the reconstruction quality of depth maps has been largely improved on subjective and objective evaluation.

## 2. RELATED WORK

The structural similarity between texture images and depth maps is the basis of texture-guided depth SR. *Hui et al.* [10] design a multi-scale guided convolutional network (MSG-Net) by complementing LR depth features with HR intensity features of texture images. *Zhao et al.* [11] present a texture-depth generative adversarial network (GAN) to learn the geometry structural similarity of texture-depth. *Guo et al.* [5] present a named DepthSR-Net to fuse the texture-depth features in different scales, which is built on a residual U-Net deep network architecture. In their work, the hierarchical features are extracted by encoder-decoder structure of U-Net and the final HR depth map is achieved by adding the learned residual to the interpolated depth map. Different from above methods, *Lutio et al.* [12] propose an alternative interpretation of guided SR, which tries to find a transformation from the guide to the target, particularly a pixel-wise mapping from one image domain to another.

In fact, transformer has been proved the success in various natural language processing (NLP) tasks. Recently, many attempts are made to explore the benefits of transformer in computer vision tasks [13]. To utilize the information of relevant textures from reference images, the attention mechanism is an efficiency solution to transfer interesting regions from the reference images. *Wang et al.* [14] introduce a non-local operation, which is the first adaptation of the dot-product attention mechanism for long-range dependency modeling in computer vision. Different from dot-product attention and its variants, *Shen et al.* [15] propose a novel efficient attention mechanism equivalent to dot-product attention. This dot-product attention aggregates the values by the template attention maps to form global context information. In [16], *Yang et al.* propose a novel texture transformer network, where attention mechanism is used to transfer HR textures from reference images by LR images as quires and HR reference images as keys in a transformer. Motivated by these, we are encouraged to bridge between depth maps and texture images to guide the SR of depth maps.

## 3. PROPOSED METHOD

### 3.1. Network Architecture

Given a LR depth map $D_{LR}$, it is firstly up-sampled to the same spatial scale $\tilde{D}_{HR}$ corresponding to the HR texture image $T_{HR}$. Firstly, a pre-trained VGG model based on ImageNet is used to construct texture-depth semantic feature representation. Secondly, we design a texture-depth transformer module(TDTM), using the texture-depth transformer mechanism to convert the structural information of texture features onto the corresponding depth features. Furthermore, a multi-scale feature fusion strategy(MSF) is adopt to fuse the fusion features at multiple scales. Finally, the depth feature reconstruction module is used to obtain the reconstruction result of the HR depth map. Fig. 2 gives the diagram of our designed network.

### 3.2. Texture-Depth Transformer

Texture-Depth Transformer aims to explore what texture information matters with depth SR. As we all known, the general texture-guided depth SR methods assume that the structure is spatial uniform between depth maps and texture images. However, the spatial discontinuities between texture and depth information are always unavoidable due to the acquisition environment. Therefore, it is necessary to find uniform reference features from texture images to adapt the depth maps, not limited in spatial coordinate.

Based on the extracted texture features and depth features, we first use 3 convolutional layers with $1 \times 1$ to map features into three standard transformer paradigms $Q(query) \in \mathbb{R}^{(HW)\times C_i}$, $K(key) \in \mathbb{R}^{(HW)\times C_i}$ and $V(value) \in \mathbb{R}^{(HW)\times C_o}$[1]. Targetting to enhance the depth features by using the corresponding texture features, we construct a depth feature template based on $K$ and $V$ which is largely different from the common transformer. In the feature extraction module, we exploit two separate convolution filters to generate different representations of depth features. Then a self-attention mechanism is applied to generate the feature template. On the other hand, $Q$ is the corresponding texture features by stacking cascade convolutional layers. Therefore, the texture transformer can be formulated as,

$$F_T = \boldsymbol{Q}\left(\textit{softmax}(\boldsymbol{K}^\top)\boldsymbol{V}\right), \tag{1}$$

where $\textit{softmax}(\boldsymbol{K}^\top)\boldsymbol{V}$ is a self-attention mechanism. And the self-attention matrix with size of $C_i \times C_o$ represents $C_i$ attention maps, which can be viewed as a global semantic information description. In order to transform texture features, the module choose reliable features from the $\boldsymbol{Q}$ according to the depth semantic template. Then, the transformed texture feature $F_T \in \mathbb{R}^{HW\times C_o}$ is reshaped as $F_T \in \mathbb{R}^{H\times W\times C_o}$ to

[1] $H$ and $W$ denote the size of input features, $C_i$ and $C_o$ indicates the input and output channel number, respectively.

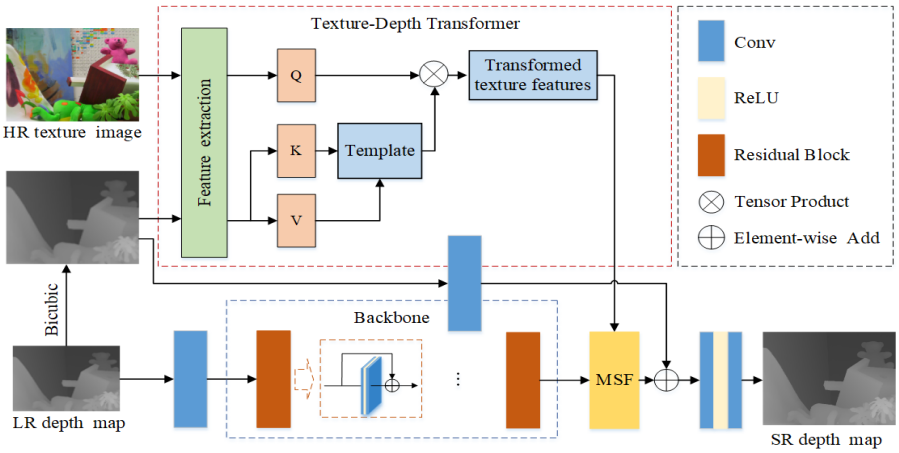



**Fig. 2**. The architecture of the proposed Texture-Depth Transformer Network (TDTN).

concatenate with LR depth features from the backbone network to generate texture-depth fused features.

### 3.3. Multi-scale feature fusion

In order to further integrate the texture-depth fusion features, We adopt a multi-scale feature fusion strategy. As mentioned in Sec. 3.2, we can obtain three-scale texture-depth fusion features, which not only contains the required structural information corresponding to depth features, but includes the feature transform cross different scales. Thereby, we also conduct a cross-scale feature fusion as shown in Fig. 3. Firstly, a residual block is used to further implement fused feature the transform on feature space. Then, to exchange structural information with cross-scale, we use Bicubic method to scale the size of different fused features for matching the spatial size. At last, we apply a $1 \times 1$ convolution operation to fuse these features, as

$$F_{out} = W^T * \left[F_r^1, F_r^{\frac{1}{2}} \uparrow_{2\times}, F_r^{\frac{1}{4}} \uparrow_{4\times}\right] + b \qquad (2)$$

where $\uparrow$ is up-sampling operation, $F_r^1$, $F_r^{1/2}$ and $F_r^{1/4}$ are the fused texture-depth features by a CNN block at different scales, respectively.

### 3.4. Implementation details

As shown in Fig. 2, some details of the designed network are described as follows. The feature extraction module in the texture-depth transformer is composed of the first 12 layers of VGG-19 [17], including 5 convolutional layers with ReLU and 2 pooling layers. Therefore, the transformed texture features and the fused features are scaled to different spatial resolutions, which are $1$, $1/2$ and $1/4$, respectively.

Except for the convolutional layer with labeled parameters, the size of the convolution kernel is 3, the number of

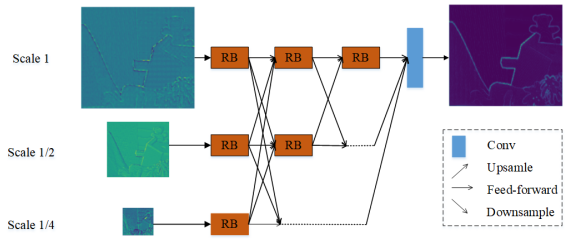



**Fig. 3**. The architecture of multi-scale feature fusion.

channels is 64 and the batch size is 16. $L1$ Loss is used to train of TDTN, with 200 epochs and initial learning rate is $10^{-4}$. The network is optimized by the ADAM with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\mathcal{E} = 10^{-8}$.

## 4. EXPERIMENTS

In this section, we design several experiments to verify the efficiency of our proposed method. Just like as [10], for training, we select 34 pairs of texture-depth images from the "Middlebury Datasets" [18] and 58 pairs of texture-depth images from the "MPI Sintel depth" dataset [19], repectively. For evaluation, we choose 10 pairs of texture-depth images from "Middlebury Datasets" as the testing set, including *Art, Books, Moebius, Dolls, Laundry, Reindeer, Tsukuba, Venus, Teddy and Cones*.

We pre-process the original HR depth maps by using Bicubic algorithm to obtain the LR depth maps. During training, random rotation and flip operations are employed for data augmentation. The root mean square error (RMSE) and the peak signal to noise ratio (PSNR) are used to evaluate the performance.

**Table 1**. Quantitative comparisons on test for $2\times$ scaling factor in terms of RMSE values

| Method | Art | Books | Moebius | Dolls | Laundry | Reindeer | Tsukuba | Venus | Teddy | Cones | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bicubic | 2.66 | 1.08 | 0.85 | 0.94 | 1.61 | 1.97 | 5.81 | 1.32 | 1.99 | 2.45 | 2.068 |
| GF [20] | 3.14 | 1.28 | 1.09 | 1.09 | 1.87 | 2.17 | 6.82 | 1.29 | 2.21 | 2.99 | 2.395 |
| SRAM [4] | 1.67 | 1.02 | 0.72 | 0.97 | 1.43 | 1.67 | / | / | / | / | 1.247 |
| GSR-PPT [12] | 4.57 | 1.57 | 1.51 | 1.30 | 1.99 | 2.48 | 9.99 | 1.46 | 2.41 | 3.44 | 3.072 |
| SRCNN [7] | 2.48 | 1.03 | 0.81 | 0.90 | 1.52 | 1.84 | 5.47 | 1.27 | 1.88 | 2.34 | 1.954 |
| PS-DMSR [21] | 0.66 | 0.54 | 0.52 | 0.58 | 0.52 | 0.59 | 1.41 | 0.56 | 0.85 | 0.88 | 0.711 |
| MSG [10] | 0.66 | 0.37 | 0.36 | 0.35 | 0.37 | 0.42 | 1.85 | **0.14** | 0.71 | 0.90 | 0.613 |
| RDN-GDE [22] | 0.56 | 0.36 | 0.38 | 0.56 | 0.48 | 0.51 | / | / | / | / | 0.475 |
| MFR-SR [23] | 0.71 | 0.42 | 0.42 | 0.60 | 0.61 | 0.65 | / | / | / | / | 0.568 |
| PMBA [6] | 0.61 | 0.41 | 0.39 | 0.36 | 0.38 | 0.40 | / | / | / | / | 0.425 |
| DepthSR [5] | 0.53 | 0.42 | / | / | 0.44 | 0.51 | 1.33 | / | 0.83 | / | 0.677 |
| Ours(TDTN) | **0.37** | **0.28** | **0.31** | **0.34** | **0.30** | **0.34** | **1.01** | 0.16 | **0.54** | **0.59** | **0.424** |

**Table 2**. Quantitative comparisons on test for $4\times$ scaling factor in terms of RMSE values

| Method | Art | Books | Moebius | Dolls | Laundry | Reindeer | Tsukuba | Venus | Teddy | Cones | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bicubic | 3.90 | 1.63 | 1.29 | 1.33 | 2.39 | 2.86 | 8.56 | 1.91 | 2.90 | 3.60 | 3.037 |
| GF [20] | 3.82 | 1.62 | 1.32 | 1.30 | 2.34 | 2.69 | 8.44 | 1.69 | 2.73 | 3.63 | 2.958 |
| SRAM [4] | 2.57 | 1.33 | 0.85 | 1.07 | 2.00 | 2.07 | / | / | / | / | 1.648 |
| GSR-PPT [12] | 3.79 | 1.65 | 1.44 | 1.32 | 1.99 | 2.43 | 9.96 | 1.48 | 2.39 | 3.71 | 3.016 |
| SRCNN [7] | 3.71 | 1.58 | 1.23 | 1.28 | 2.31 | 2.73 | 8.11 | 1.85 | 2.77 | 3.43 | 2.900 |
| PS-DMSR [21] | 1.59 | 0.83 | 0.86 | 0.91 | 0.92 | 1.11 | 3.73 | 0.72 | 1.58 | 2.38 | 1.463 |
| MSG [10] | 1.47 | 0.67 | 0.66 | **0.69** | 0.79 | 0.98 | 4.29 | **0.35** | 1.49 | 2.60 | 1.399 |
| RDN-GDE [22] | 1.47 | 0.62 | 0.69 | 0.88 | 0.96 | 1.17 | / | / | / | / | **0.965** |
| MFR-SR [23] | 1.54 | 0.63 | 0.72 | 0.89 | 1.11 | 1.23 | / | / | / | / | 1.020 |
| PMBA [6] | 2.04 | 0.92 | 0.84 | 0.95 | 1.14 | 1.39 | / | / | / | / | 1.213 |
| DepthSR [5] | **1.20** | 0.60 | / | / | 0.78 | 0.96 | **3.26** | / | 1.37 | / | 1.362 |
| Ours(TDTN) | 1.24 | **0.48** | **0.61** | 0.76 | **0.68** | **0.95** | 3.50 | 0.36 | **1.21** | **1.56** | 1.135 |

**Table 3**. Quantitative comparisons on test for $8\times$ scaling factor in terms of RMSE values

| Method | Art | Books | Moebius | Dolls | Laundry | Reindeer | Tsukuba | Venus | Teddy | Cones | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bicubic | 5.50 | 2.36 | 1.89 | 1.87 | 3.43 | 4.05 | 12.3 | 2.76 | 4.07 | 5.30 | 4.353 |
| GF [20] | 5.34 | 2.30 | 1.86 | 1.80 | 3.32 | 3.87 | 12.1 | 2.58 | 3.91 | 5.22 | 4.230 |
| SRAM [4] | 3.20 | 1.46 | 1.10 | 1.19 | 2.11 | 2.47 | / | / | / | / | 1.922 |
| GSR-PPT [12] | 3.98 | 1.82 | 1.51 | 1.36 | 2.03 | 2.57 | 10.0 | 1.51 | 2.46 | 3.95 | 3.119 |
| SRCNN [7] | 5.28 | 2.30 | 1.84 | 1.82 | 3.32 | 3.92 | 11.8 | 2.67 | 3.95 | 5.15 | 4.205 |
| PS-DMSR [21] | 2.57 | 1.19 | 1.21 | 1.31 | 1.52 | 1.80 | 7.79 | 1.09 | 2.88 | 4.66 | 2.602 |
| MSG [10] | 2.46 | 1.03 | 1.02 | **1.05** | 1.51 | 1.76 | 8.42 | 1.04 | 2.76 | 4.23 | 2.528 |
| RDN-GDE [22] | 2.60 | 1.00 | 1.05 | 1.21 | 1.63 | 2.05 | / | / | / | / | **1.590** |
| MFR-SR [23] | 2.71 | 1.05 | 1.10 | 1.22 | 1.75 | 2.06 | / | / | / | / | 1.648 |
| PMBA [6] | 3.63 | 1.68 | 1.41 | 1.47 | 2.19 | 2.74 | / | / | / | / | 2.187 |
| DepthSR [5] | **2.22** | 0.89 | / | / | 1.31 | **1.57** | **6.89** | / | **1.85** | / | 2.455 |
| Ours(TDTN) | 2.45 | **0.86** | **0.91** | 1.15 | **1.29** | 1.75 | 8.86 | **0.80** | 2.20 | **3.09** | 2.336 |

### 4.1. Comparison Experiments

We choose Bicubic, GF [20], SRAM [4] as representatives of the classical traditional methods, and some CNN-based methods including GSR-PPT [12], SRCNN [7], PS-DMSR [21], MSG [10], RDN-GDE [22], MFR-SR [23], PMBA [6], DepthSR [5]. When the scale factor is $2\times$, $4\times$, and $8\times$, the objective quality comparison results of the depth SR reconstruction are shown in Table. 1 to Table. 3. The best and the second best are indicated in bold and underline, respectively.

As shown in Table. 1 to Table. 3, for different scale factors, our model almost achieves the best results in the most of testing depth maps. Specially, when scale factor is $4\times$ and $8\times$, respectively, the average RMSE performance of our model slightly inferior to MFR-SR [23] and RDN-GDE [22]. However, in each given testing result, our performance is better than these both methods. Moreover, when the scale factor is $8\times$, the RMSE result of our model on the "Art" is also $0.23$ higher than that of the DepthSR [5], but in the subjective evaluation (as shown in Fig. 5), our model indeed retains more edge detail information compared to DepthSR [5].

Some visual results are shown in Fig. 4 and Fig. 5. It can prove that the texture features are indeed helpful to depth SR. Especially for the sharp edges, all of DepthSR [5], PMBA [6] and our model can clearly recover the edges, compared to Bicubic and SRCNN [7]. Nevertheless, DepthSR [5] and PMBA [6] justly use texture images of different scales according to the coordinate, and both the extracted depth features and texture features are directly concatenated to implement the feature fusion, which results in the blending on the edge regions. Partly because the consistence between depth maps and texture images cannot be guaranteed in practice. However, our model utilizes the TDTM to select the reliable texture features to adapt with depth features, and uses a multi-scale feature fusion mechanism to implement the features fusion at different scales, thereby, the edge detail information of depth maps is better restored.

### 4.2. Ablation experiments

To explore what matters with depth SR in our model, we construct different ablation experiments. Experimental setup set that the scale factor is $2\times$.

#### *4.2.1. Ablation study on Transformer*

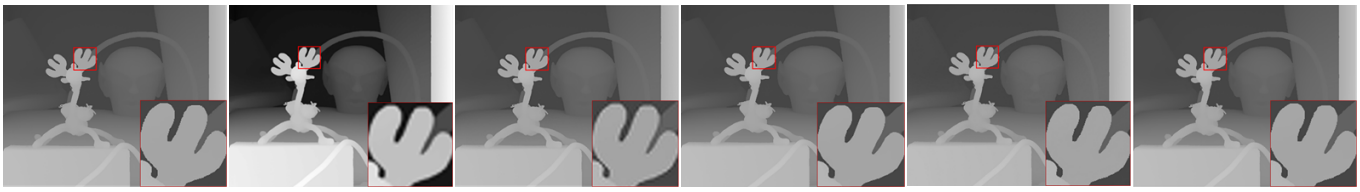
As is shown in Table. 4, where "T-Q and T-KV" denotes all $Q$, $K$ and $V$ are the features extracted from $T_{HR}$, and the degraded depth map $\tilde{D}_{HR}$ is not required in the designed network. The "D-Q and D-KV" represents the $Q$ is the extracted feature from $\tilde{D}_{HR}$ and $K$, $V$ are from $T_{HR}$. And the "Non-transformer" indicates the model directly fuses the texture features from $T_{HR}$ and depth features from $\tilde{D}_{HR}$ in the form of coordinate. In addition, the "T-Q and D-KV" is our setting as mentioned in Sec. 3.2.

**Table 4**. Ablation study on Transformer

| Method | PSNR/RSME |
|---|---|
| T-Q and T-KV | 55.79/0.47 |
| D-Q and T-KV | 55.91/0.47 |
| Non-transformer | 56.13/0.46 |
| T-Q and D-KV | **56.23/0.45** |

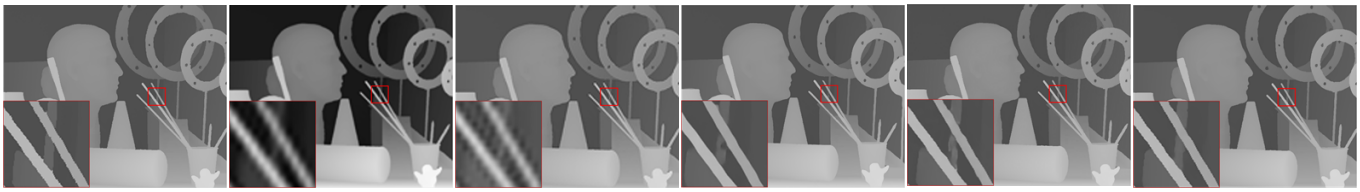
As shown in Table 4, "T-Q and "D-KV" setting achieves the best performance of SR, and the corresponding PSNR value is $0.47$dB higher than that of the "texture-Q and texture-KV", $0.32$dB higher than that of the "depth-Q and texture-KV", and $0.10$dB higher than the "Non-transformer".

The visual features are shown in Fig. 1. Fig. 1(b) is the extracted depth features, while the corresponding texture features contain a lot of texture information, such as uselessly



(a) Ground truth (b) Bicubic (c) SRCNN (d) PMBA (e) DepthSR (f) Ours

**Fig. 4**. Visual quality comparison results of the depth map "Reindeer" at scale 4×.



(a) Ground truth (b) Bicubic (c) SRCNN (d) PMBA (e) DepthSR (f) Ours

**Fig. 5**. Visual quality comparison results of the depth map "Art" at scale 8×.

edge inside the surface of the image, as shown in Fig. 1(c). The "T-Q and T-KV" tries to learn the texture features in the form of self attention, thus the extracted features focus on the semantic level, rather than edges, as shown in Fig. 1(d). On the other hand, by using the "D-Q and T-KV", the edge information can be enhanced, but the texture semantic information is discarded. It means that probably few texture features can be transferred, as shown in Fig. 1(e). The "T-Q and D-KV" uses the semantic information of the depth map as a template and select proper texture features to match with the distribution of depth features, as shown in Fig. 1(f) . Therefore, the transformed texture features not only retain the guidance information needed in the depth map reconstruction process, but also the interference information on the surface of the texture object is suppressed to a certain extent.

#### *4.2.2. Ablation study on Texture-depth transformer network*

In order to verify and analyze the role of each module in the network structure, we use the LR depth features extracted by the Backbone network composed of 5-layer residual blocks and three-scale texture features extracted by the VGG-19 network as the baseline in this experiment, and then add Residual learning (RL), Multi-Scale Fusion (MSF) and Transformer (T) to verify the role of each module. The experimental results are shown in Table. 5.

As we all known, RL can reduce the training difficulty of the network, and the reconstruction results reduces by 0.07 on RMSE compared to Baseline.When we further introduce the MSF module, its RMSE is reduced by 0.03. Lastly, the TDTM furthermore improve the performance of depth SR due to the possible advantage mentioned above.

**Table 5**. Ablation study on texture-depth transformer network

| Method | RL | MSF | T | PSNR/RSME |
|---|---|---|---|---|
| Baseline | | | | 54.45/0.56 |
| Baseline+RL | ✓ | | | 55.41/0.49 |
| Baseline+RL+MSF | ✓ | ✓ | | 56.13/0.46 |
| Baseline+RL+MSF+T | ✓ | ✓ | ✓ | **56.23/0.45** |

## 5. CONCLUSION

In this paper, a texture-depth transformer is proposed for depth super-resolution task which can learn the corresponding structural information of the HR texture images and the corresponding interpolated depth maps. Furthermore, a multi-scale feature fusion strategy is adopt to fuse the fused features at multiple scales. Extensive experiments demonstrate the superior performance of our TDTN over state-of-the-art approaches on both quantitative and qualitative evaluations.

**Acknowledgment:** This work is supported by National Natural Science Foundation of China(No. 61972028, No. 61902022) and the Fundamental Research Funds for the Central Universities (No. 2019JBM018, No. FRF-TP-19-015A1), the computing work is partly supported by USTB MatCom of Beijing Advanced Innovation Center for Materials Genome Engineering.

## 6. REFERENCES

[1] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi, "Joint geodesic upsampling of depth images," in *Proceedings*

*of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 169–176.

[2] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang, "Color-guided depth recovery from rgb-d data using an adaptive autoregressive model," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3443–3458, 2014.

[3] Meiqin Liu, Yao Zhao, Jie Liang, Chunyu Lin, Huihui Bai, and Chao Yao, "Depth map up-sampling with fractal dimension and texture-depth boundary consistencies," *Neurocomputing*, vol. 257, pp. 185–192, 2017.

[4] Jin Wang, Wei Xu, Jian-Feng Cai, Qing Zhu, Yunhui Shi, and Baocai Yin, "Multi-direction dictionary learning based depth map super-resolution with autoregressive modeling," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1470–1484, 2019.

[5] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2545–2557, 2018.

[6] Xinchen Ye, Baoli Sun, Zhihui Wang, Jingyu Yang, Rui Xu, Haojie Li, and Baopu Li, "Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution," *IEEE Transactions on Image Processing*, vol. 29, pp. 7427–7442, 2020.

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence(TPAMI)*, vol. 38, no. 2, pp. 295–307, 2015.

[8] Xibin Song, Yuchao Dai, Dingfu Zhou, Liu Liu, Wei Li, Hongdong Li, and Ruigang Yang, "Channel attention based iterative residual learning for depth map super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 5631–5640.

[9] Zhongyu Jiang, Huanjing Yue, Yu-Kun Lai, Jingyu Yang, Yonghong Hou, and Chunping Hou, "Deep edge map guided depth super resolution," *Signal Processing: Image Communication*, vol. 90, pp. 116040, 2020.

[10] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 353–369.

[11] Lijun Zhao, Huihui Bai, Jie Liang, Bing Zeng, Anhong Wang, and Yao Zhao, "Simultaneous color-depth super-resolution with conditional generative adversarial networks," *Pattern Recognition(PR)*, vol. 88, pp. 356–369, 2019.

[12] Riccardo de Lutio, Stefano D'aronco, Jan Dirk Wegner, and Konrad Schindler, "Guided super-resolution as pixel-to-pixel transformation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 8829–8837.

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020.

[14] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 7794–7803.

[15] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li, "Efficient attention: Attention with linear complexities," *arXiv preprint arXiv:1812.01243*, 2018.

[16] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo, "Learning texture transformer network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 5791–5800.

[17] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision(IJCV)*, vol. 47, no. 1-3, pp. 7–42, 2002.

[19] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black, "A naturalistic open source movie for optical flow evaluation," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 611–625.

[20] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 1–14.

[21] Liqin Huang, Jianjia Zhang, Yifan Zuo, and Qiang Wu, "Pyramid-structured depth map super-resolution based on deep dense-residual network," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1723–1727, 2019.

[22] Yifan Zuo, Yuming Fang, Yong Yang, Xiwu Shang, and Bin Wang, "Residual dense network for intensity-guided depth map enhancement," *Information Sciences*, vol. 495, pp. 52–64, 2019.

[23] Yifan Zuo, Qiang Wu, Yuming Fang, Ping An, Liqin Huang, and Zhifeng Chen, "Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 297–306, 2019.