



# Auxiliary Features-Guided Super Resolution for Monte Carlo Rendering

Qiqi Hou and Feng Liu

Portland State University, Portland, OR, USA  
qiqi.hou2012@gmail.com, fliu@pdx.edu

## Abstract

*This paper investigates super-resolution to reduce the number of pixels to render and thus speed up Monte Carlo rendering algorithms. While great progress has been made to super-resolution technologies, it is essentially an ill-posed problem and cannot recover high-frequency details in renderings. To address this problem, we exploit high-resolution auxiliary features to guide super-resolution of low-resolution renderings. These high-resolution auxiliary features can be quickly rendered by a rendering engine and at the same time provide valuable high-frequency details to assist super-resolution. To this end, we develop a cross-modality transformer network that consists of an auxiliary feature branch and a low-resolution rendering branch. These two branches are designed to fuse high-resolution auxiliary features with the corresponding low-resolution rendering. Furthermore, we design Residual Densely Connected Swin Transformer groups to learn to extract representative features to enable high-quality super-resolution. Our experiments show that our auxiliary features-guided super-resolution method outperforms both super-resolution methods and Monte Carlo denoising methods in producing high-quality renderings.*

**Keywords:** super-resolution, fast-to-compute auxiliary features, transformer, Monte Carlo rendering

**CCS Concepts:** • Computing methodologies → Ray tracing

## 1. Introduction

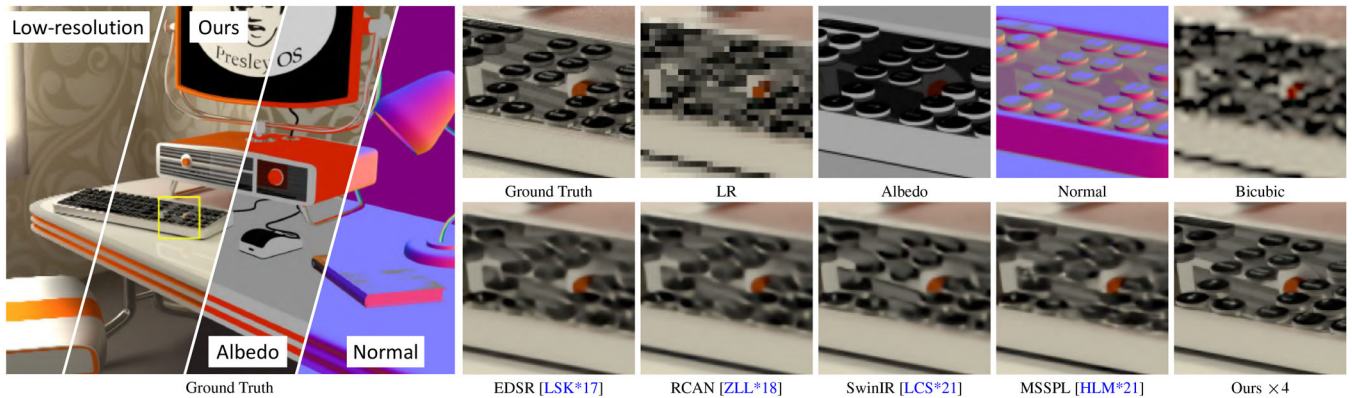
Monte Carlo rendering algorithms are now widely used to generate photo realistic computer graphics images for applications such as visual effects, video games, and computer animations. These algorithms generate a pixel's colour by integrating over all the light paths arriving at a single point [CPC84]. To rendering a high-quality image, a large number of rays need to be cast for each pixel, which makes Monte Carlo rendering a slow process (see Figure 1).

A great amount of effort has been devoted to speeding up Monte Carlo rendering. The core idea is to reduce the number of rays for each pixel. For instance, numerous denoising algorithms are now available to reconstruct a high-quality image from a rendering produced at a low sampling rate. Such Monte Carlo denoising algorithms often use auxiliary features generated by a rendering algorithm to help denoise the noisy rendering result. The recent deep neural network-based denoising algorithms can now generate very high-quality images at a fairly low sampling rate [BVM\*17, KKS\*17, KKR18, GLA\*19].

Monte Carlo rendering can also be sped up by reducing the number of pixels to render. For example, pixels from the frames that have

already been rendered can be warped to generate frames in-between existing frames to increase the frame rate [BDM\*21] or to generate future frames to reduce latency [GFL\*21]. Another approach is to only render one pixel for a block of neighbouring pixels to further reduce the total number of pixels to render. This can be implemented by first rendering a low-resolution image and then applying super-resolution to increase its resolution [XNC\*20, HLM\*21]. As super-resolution is a fundamentally ill-posed problem, it alone often cannot recover high-frequency details from only the low-resolution rendering. To address this problem, Hou *et al.* render a high-resolution rendering with a low sampling rate and use that together with the high-resolution auxiliary features to help super resolve the low-resolution rendering rendered at a high sampling rate. While this method produces a high-quality result, it needs to render the high-resolution image at a low sampling rate, which still takes a considerable amount of time [HLM\*21].

*Can we only use the fast-to-obtain high-resolution auxiliary features without the high-resolution-low-sample rendering to effectively assist super-resolution of the corresponding low-resolution rendering? If so, we can further speed up Monte Carlo rendering. We are encouraged by the recent work on neural frame synthesis that showed fast-to-obtain auxiliary features of the target frames can*



**Figure 1:** Our method uses fast-to-compute high-resolution auxiliary features to support super-resolution of Monte Carlo rendering results. Our method generates high-quality visual details than both off-the-shelf super-resolution methods and MSSPL, a dedicated super-resolution method for Monte Carlo renderings that uses both more high-resolution auxiliary features and the corresponding high resolution low sample rendering result.

greatly help interpolate or extrapolate the target frames [BDM\*21, GFL\*21]. On the other hand, Hou *et al.* showed that using a wide range of auxiliary features and the high-resolution-low-sample rendering help super-resolution more than only using a subset of auxiliary features within their own deep neural network-based super-resolution framework [HLM\*21]. Therefore, if we only use a small number of fast-to-compute auxiliary features, we need to have a better super-resolution method.

This paper presents a Cross-modality Residual Densely Connected Swin Transformer (XRDS) for super-resolution of a Monte Carlo rendering guided by its auxiliary features. For the seek of speed, we only use two auxiliary features: albedo and normal. To effectively use these features, we design a super-resolution network based on Swin Transformer that recently has been shown powerful for a wide variety of computer vision tasks. Our transformer network has two branches, one for the low resolution rendering and the other for the auxiliary features. Such two branches are designed to perform cross-modality fusion to effectively use auxiliary features to assist super-resolution of the low-resolution rendering. While the auxiliary feature branch consists of convolutional blocks, the branch for the low-resolution rendering consists of a sequence of Residual Densely Connected Swin Transformer blocks to extract effective features. The features from the two branches are combined together using a cross-modality fusion module and are finally used to generate the high-resolution high-quality rendering.

This paper contributes to Monte Carlo rendering as follows. First, we present the first super-resolution approach to Monte Carlo rendering that only uses fast-to-compute high-resolution auxiliary features to enable high-quality upsampling of a low-resolution rendering. Second, we design a dedicated Cross-modality Swin Transformer-based super-resolution network that can learn to effectively combine high-resolution auxiliary features with the corresponding low-resolution rendering to generate the final high-resolution high-quality image. Third, our experiments show that our method outperforms super-resolution and denoising methods in producing high-quality renderings.

## 2. Related Work

This section briefly discusses relevant work to our paper, including Monte Carlo denoising, super-resolution and vision transformers.

**Monte Carlo denoising.** Monte Carlo rendering algorithms need numerous samples per pixel to generate a high-quality rendering [CPC84, Kaj86]. With insufficient samples, the rendering results suffer from noise. To address this problem, many Monte Carlo denoising methods have been developed to re-construct high-quality renderings from only a small number of samples. Traditional methods re-construct renderings in a similar way to general image denoising methods by designing specific denoising kernels based on image variance or geometric features or directly regress the final result [DSHL10, JC95, LSK\*07, LR90, McC99, RW94, SIMP06, SD12, XP05]. Zwicker *et al.* provided a good survey on this [ZJL\*15]. Alternatively, adaptive sampling algorithms can be used to reduce the overall sample numbers for the whole image [BM98, ETH\*09, Jen01, LWC12, MA06, MMMG16, ODR09, RKZ11, WABG06, WRC88].

Recently, deep neural network-based denoising methods have shown impressive performance. These methods learn to re-construct high-quality renderings from small number of samples. In their seminal work, Kalantari *et al.* estimated optimal filter parameters using a multi-layer perceptron neural network [KBS15]. Bako *et al.* estimated spatially adaptive kernels for denoising in a convolutional manner [BVM\*17]. Vogels *et al.* extended the concept of kernel prediction methods to temporal denoising [VRM\*18]. With asymmetric loss functions, their method could produce high-quality results for a sequence of frames. Chaitanya *et al.* developed a recurrent autoencoder to denoise a sequence of frames while maintaining temporal stability [CKS\*17]. Xu *et al.* developed an adversarial approach to Monte Carlo rendering denoising that can greatly reduce artifacts such as blurs and unfaithful details from denoising results [XZW\*19]. Gharbi *et al.* developed a kernel splatting network that re-constructs the final image by splatting samples to pixels according to the estimated splatting kernels [GLA\*19]. Munkberg *et al.* proposed to filter auxiliary layers of individual

samples [MH20]. Their method works well on outliers and complex visibility. Hasselgren *et al.* proposed a neural spatial-temporal sampling method for Monte Carlo video denoising [HMS\*20]. Their method first estimates the sampling map from the temporal reprojection and auxiliary features and then denoises the resulting image generated using the sampling map to produce high-quality results. Zheng *et al.* proposed an ensemble denoising technique that learns to combine multiple denoisers together [ZZXY21]. Yu *et al.* designed a transformer-based neural network for Monte Carlo denoising [YNL\*21]. Their network consists multi-scale feature extractor and a self-attention module and achieved promising results. Unlike these denoising methods, our method explores an orthogonal approach that speeds up Monte Carlo rendering by reducing the number of pixels to render via super-resolution.

**Super-resolution.** Super-resolution is a classic problem in computer vision. It aims to re-construct a high-resolution image from the low-resolution input. Recently, the state of the art of super-resolution research has been advanced significantly due to the use of deep neural networks [DLHT14, AKS18, HSU19, HWG18, KKLML16, LSK\*17, LTT\*19, LWLS19, XMS19, ZZZ18, ZWLQ19]. Specifically, in their seminal work, Dong *et al.* trained a three-layer fully convolutional neural network for single image super-resolution [DLHT14]. Since that, a large variety of super-resolution deep neural networks have been invented by leveraging carefully designed network architectures, including residual blocks (RBs) [HZRS16], densely connected blocks [HLVDMW17, ZTK\*18], channel attention blocks [HSS18, ZLL\*18], transformers [LCS\*21] and others.

Super-resolution has recently been explored to speed up rendering. Xiao *et al.* designed a super-resolution network that takes both the low-resolution rendering and the corresponding low-resolution auxiliary features as input and outputs a high-resolution frame. They leveraged neighbouring frames to further improve super-resolution quality [XNC\*20]. While their method was designed for a rasterization-based rendering engine, in principal, it can be applied to Monte Carlo rendering. Thomas *et al.* combined super-resolution and Monte Carlo denoising for videos. Their network takes a low-resolution rendering as well as a warped previous frame as input and produces a high-resolution frame [TLP\*22]. However, super-resolution is essentially an ill-posed problem and cannot fully recover missing high-frequency visual details from only the low-resolution input. To address this problem, Hou *et al.* developed a super-resolution approach based on multiple-resolution sampling. Their method first renders a low-resolution image at a high-sampling rate and a high-resolution image at a low sampling rate. Their method then exploits the high-resolution noisy image to recover high frequency visual details [HLM\*21]. While their method generates high-quality renderings, it needs to render the high resolution noisy image and auxiliary features, which takes a considerable amount of time. Different from the above methods, our method obtains high-frequency information from only fast-to-compute high-resolution auxiliary features as inspired by recent interpolation and extrapolation methods that use fast-to-compute auxiliary features of the target frames to help generate the target frames [GFL\*21, BDM\*21].

**Vision transformer.** Transformer was initially designed for natural language process tasks [VSP\*17]. Due to its self-attention mech-

anism, it can efficiently capture the long-term information from the input. Recently, transformer networks have attracted considerable attention in the computer vision community and achieved great success in various computer vision tasks, including image recognition [LLC\*21], object detection [CMS\*20], semantic segmentation [ZLZ\*21] and image restoration [LCS\*21]. Dosovitskiy *et al.* developed the first transformer network for image recognition [DBK\*20]. They split the input image into image patches and then feed these image patches as tokens to the transformer network. Chen *et al.* presented an image processing transformer for various restoration problems and demonstrated that pre-training on large datasets could greatly improve the capability of a transformer network for low-level computer vision tasks [CWG\*21]. Liang *et al.* developed SwinIR for image restoration. Their network adapted the Swin Transformer [LLC\*21] as their backbone and achieved promising results [LCS\*21]. However, transformer for the Monte-Carlo denoising is less explored. Inspired by the success of these vision transformer networks, we are the first to design a dedicated cross-modality transformer network for super-resolution of Monte Carlo renderings that can effectively leverage fast-to-compute high-resolution auxiliary features to recover high-frequency visual details when upsampling a low-resolution rendering.

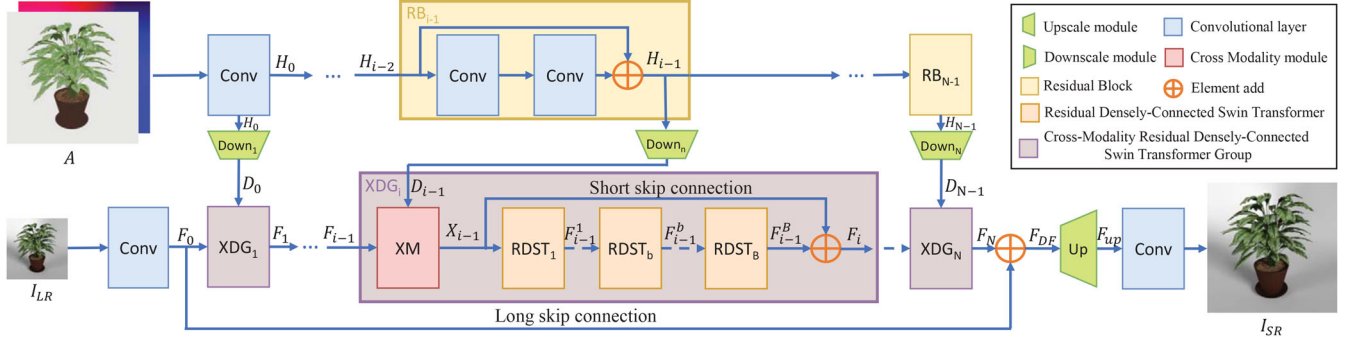
### 3. Our Method

This paper proposes a super-resolution method guided by the fast-to-compute auxiliary features to speed up the Monte Carlo rendering. Our method takes a low resolution rendering  $I_{LR}$  and its high-resolution fast-to-compute auxiliary features  $A$  as input, and outputs the corresponding high-quality high-resolution result  $I_{SR}$ . The high-resolution auxiliary features provide the essential high-frequency information for the super-resolution.

Different from the previous work [HLM\*21], which leverages a wide range of auxiliary features, our method *only* employs the auxiliary features that can be computed very fast [BDM\*21], including albedo and normal. On the one hand, although our method does not leverage the shading layers, albedo and normal could provide a lot of high-frequency information, *e.g.* the texture of the material, which is essential for super-resolution. As we will show, it can help improve the super-resolution results. On the other hand, albedo and normal can be computed fast [BDM\*21]. It not only reduces the rendering time but also enables us to render these high-resolution layers at a relatively higher sampling rate, which typically contains fewer artifacts, such as aliasing.

We design a cross-modality transformer network to effectively fuse two categories of visual input, namely, the low-resolution rendering and its corresponding high-resolution auxiliary features, to recover visual details. Figure 2 shows the architecture of our network. It contains two parallel branches, one for the low-resolution rendering and the other for the corresponding high-resolution auxiliary features.

**Auxiliary feature branch.** The auxiliary feature branch takes auxiliary features as inputs, which provide essential high frequency visual details. As discussed above, we select albedo and normal, which are relatively fast to acquire. Since this branch processes high-resolution input, we design a shallow architecture for the sake



**Figure 2:** The architecture of our network. Our network takes a low-resolution rendering and its corresponding fast-to-compute high-resolution auxiliary features as input and predicts the final high-resolution-high-quality image.

of memory and speed. As shown in Figure 2, we employ a convolutional layer and  $N = 3$  RB [HZRS16] in a sequence to get the features  $\{H_i\}_{i=0}^{N-1}$

$$\begin{aligned} H_0 &= f_{conv}^A(A), \\ H_i &= f_{RB}^i(H_{i-1}), \quad i = 1, \dots, N-1, \end{aligned} \quad (1)$$

where  $f_{conv}(\cdot)$  indicates the convolution operation.  $f_{RB}(\cdot)$  indicates the operation of a RB. In our experiments, we set the channels as 32 for the auxiliary feature branch.

We then obtain the downsampled features  $\{D_i\}_{i=0}^{N-1}$  with a group of deshuffle layers [HLM\*21], which is able to downscale the feature while keeping the high-frequency information.

$$D_i = f_{DSF}^{i+1}(H_i), \quad i = 0, \dots, N-1 \quad (2)$$

where  $f_{DSF}(\cdot)$  indicates the deshuffle layer.

**Low-resolution rendering branch.** Following the recent works on image super-resolution [ZLL\*18, ZZZ18, LCS\*21], we first adopt a  $3 \times 3$  convolutional layer with 64 channels to get the shallow feature from the low resolution rendering  $I_{LR}$ .

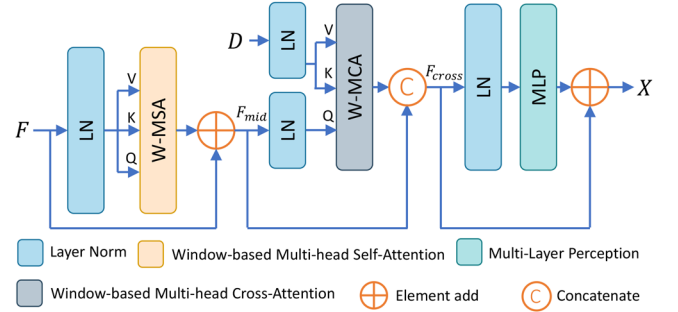
$$F_0 = f_{conv}^{LR}(I_{LR}) \quad (3)$$

We feed the resulting feature  $F_0$  to a sequence of cross-modality Residual Densely Connected Swin Transformer groups (XDG).

$$F_i = f_{XDG}^i(F_{i-1}, D_{i-1}), \quad i = 1, \dots, N \quad (4)$$

where  $f_{XDG}(\cdot)$  indicates the XDG module.  $N$  indicates the number of XDG. We choose  $N = 3$  in our experiments. XDG is designed to fuse the auxiliary features  $D_{i-1}$  and the low-resolution rendering features  $F_{i-1}$ . It consists of a cross-modality module (XM) and a sequence of Residual Densely Connected Swin Transformer blocks (RDST). Specifically, XM is designed to fuse the local information from the low-resolution rendering and the high-frequency information from the auxiliary features, while the RDST sequence learns more dedicated representations for super-resolution from them.

**Cross-modality module (XM).** Inspired by the success of Swin Transformer [LLC\*21, LCS\*21] and Transformer Decoder [GLDZ22], we design XM based on Swin Transformer, which can efficiently model the long-range dependency. Figure 3 shows the



**Figure 3:** The cross-modality module. It takes feature  $F$  from the low-resolution rendering branch and  $D$  from the auxiliary feature branch, and outputs the fused feature  $X$ .

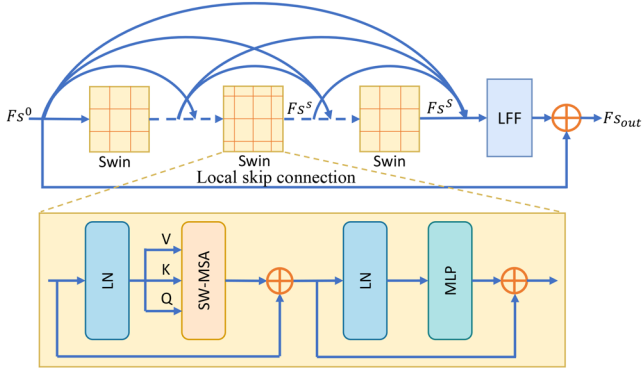
architecture of XM. It takes features  $F$  from the low-resolution rendering branch and features  $D$  from the auxiliary feature branch as input and outputs the fused feature  $X$ . It consists of Layer Norm layers (LN), a Window-based Multi-head Self-Attention layer (W-MSA), a Window-based Multi-head Cross Attention layer (W-MCA), and a Multi-Layer Perception layer (MLP). The key idea behind XM is to combine the features  $F$  from the low-resolution rendering branch with the features  $D$  from the high-resolution auxiliary branch using cross-attention, creating a more comprehensive representation for super-resolution. The process starts by extracting intermediate features  $F_{mid}$  from  $F$ , which serve as the ‘query’  $Q$ . From  $D$ , which holds high-resolution information, the ‘key’  $K$  and ‘value’  $V$  are extracted. Then, the cross-attention is calculated following [VSP\*17] and combined with  $F_{mid}$  to generate  $F_{cross}$ . Finally, an MLP layer is used to integrate the features from the low-resolution branch and the cross-attention.

**Residual Densely Connected Swin Transformer block (RDST).** As shown in Figure 2, we feed the fused feature  $X$  from XM to a sequence of  $B = 5$  Residual Densely Connected Swin Transformer blocks (RDST),

$$F_{i-1}^b = f_{RDST}(F_{i-1}^{b-1}), \quad (5)$$

where  $f_{RDST}$  indicates the RDST block. We also use a short skip connection to combine the shallow feature  $X_{i-1}$  with the deep





**Figure 4:** The Residual Densely Connected Swin Transformer block (RDST). Red lines indicate the window partitions.

feature  $F_{i-1}^B$

$$F_i = F_{i-1}^B + X_{i-1}. \quad (6)$$

We design RDST by combining the ideas of the Residual Densely Connected Network (RDN) [ZTK\*18] and Swin Transformer [LLC\*21]. We are specifically inspired by SwinIR [LCS\*21] that explores Swin Transformers for image restoration tasks. It replaces traditional convolutional layers with Swin layers in RBs, allowing for the learning of more descriptive features and delivering impressive results. Taking inspiration from RDN [ZTK\*18], we introduce RDST, where the convolution layers in densely connected blocks are replaced with Swin layers. As shown in Figure 4, RDST consists of a sequence of Densely Connected Swin Transformer blocks and a local feature fusion block. For the Densely Connected Swin Transformer blocks, we shift the windows. We also use a local skip connection to fuse the features from the shallow layer.

**Upscale.** We adopt the pixel shuffle layer [SCH\*16] to upscale the dense feature  $F_{DF}$  to a high-resolution feature. We also use a  $3 \times 3$  convolutional layer with three channels to predict the final high resolution image  $I_{SR}$ .

$$I_{SR} = f_{conv}(f_{UP}(F_{DF})), \quad (7)$$

where  $f_{UP}$  indicates the operation of the pixel shuffle layer.

**Training details.** We adopt the robust loss to handle the prediction with a high dynamic range image [HLM\*21].

$$\ell_r = \frac{1}{M} \sum_{p \in I_{HR}} \frac{|I_{HR}^p - I_{SR}^p|}{\beta + |I_{HR}^p - I_{SR}^p|}, \quad (8)$$

where  $I_{HR}$  indicates the ground truth image.  $M$  indicates the number of pixels.  $\beta$  indicates the robust factor, which is set to 0.1.

We implement our network in PyTorch. We train our super-resolution network on examples of size  $256 \times 256$ . We select Adam [KB14] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  as the optimizer. The learning rate is set to 0.0001. We train the network for 400 epochs with a mini-batch size of 16 for our  $4 \times$  super-resolution models, and we fine-tune our other models using the  $4 \times$  pre-trained weights. It takes about 1 week to train a single model using 4 Nvidia A40 GPUs.

**Table 1:** Comparison with super-resolution methods with different upsampling scales on the BCR dataset [HLM\*21].

Methods	$\times 2$		$\times 4$		$\times 8$	
	PSNR	RelMSE	PSNR	RelMSE	PSNR	RelMSE
Bicubic	30.57	0.0141	25.39	0.0858	22.36	0.2473
EDSR	32.01	0.0079	30.70	0.0119	27.97	0.0241
RCAN	32.03	0.0084	30.73	0.0117	27.92	0.0253
SwinIR	31.05	0.0118	30.78	0.0152	28.04	0.0364
MSSPL	38.40	0.0015	34.27	0.0039	31.08	0.0079
Ours	<b>42.48</b>	<b>0.0007</b>	<b>37.45</b>	<b>0.0021</b>	<b>31.94</b>	<b>0.0076</b>

We adopt the BCR dataset [HLM\*21] as the training dataset. BCR dataset contains 2449 images from 1463 scenes rendered by Blender Cycles. Following MSSPL [HLM\*21], we use 2126 images from 1283 scenes for training, 193 images from 76 scenes for validation and 130 images from 104 scenes for testing.

## 4. Experiments

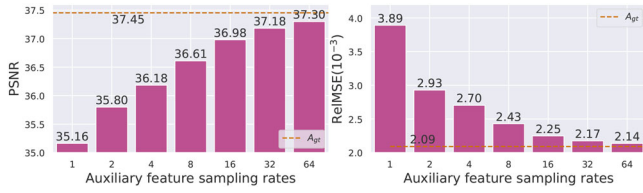
We evaluate our network by quantitatively and qualitatively comparing them with state-of-the-art image super-resolution methods and the Monte Carlo denoising methods on the BCR dataset [HLM\*21] and the Gharbi dataset [GLA\*19]. We also conduct the ablation study to examine our method. Following Hou *et al.* [HLM\*21], we adopt Relative Mean Square Error (RelMSE) and PSNR to evaluate our methods in the scene linear colour space and the sRGB space, respectively. Please refer to the supplementary material for an interactive demo that provides more results.

### 4.1. Comparison with super-resolution methods

We compare our method with state-of-the-art super-resolution methods, including EDSR [LSK\*17], RCAN [ZLL\*18], and SwinIR [LCS\*21], a recent transformer-based approach, as well as the multiple sampling-based super-resolution method MSSPL [HLM\*21]. We obtained the results of compared methods either from the authors [HLM\*21] or from finetuning the official models [LSK\*17, ZLL\*18, LCS\*21] on the BCR dataset.

As shown in Table 1, our method outperforms super-resolution methods. This improvement can be largely attributed to the use of high-resolution auxiliary features to capture high-frequency visual details. For this experiment, we use the ground truth auxiliary features as they are fast to acquire. We also vary the number of samples used to generate these features in order to examine their effect on our method. As shown in Figure 5, while having more samples to generate these auxiliary features benefits our method, the features generated with only one sample per pixel allow our method to outperform the standard super-resolution methods.

MSSPL takes both the low-resolution rendering and the high-resolution noisy rendering as well as a wide variety of high resolution auxiliary features as input [HLM\*21]. In this test, the high-resolution rendering and features are rendered with one sample per pixel. As shown in Figure 5 and Table 1, our method,



**Figure 5:** The effects of the sample rates used to generate fast-to-compute auxiliary features on the performance of our method.  $A_{gr}$  indicates the ground truth auxiliary features (4000spp).

**Table 2:** Comparison of runtime cost and peak memory with super-resolution methods to produce a  $1024 \times 1024$  image on an Nvidia Titan XP.

	Scale	EDSR	RCAN	SwinIR	MSSPL	Ours
Runtime (ms)	$\times 4$	503.96	280.51	1149.25	125.24	1009.08
Peak memory (MB)	$\times 4$	2493.9	672.1	806.0	739.70	941.3
Peak memory (MB)	$\times 8$	2375.6	621.3	659.0	1010.1	803.8
Peak memory (MB)	$\times 16$	2359.7	615.4	608.4	1008.0	783.4

when only using albedo and normal as auxiliary features obtained with one sample per pixel, can achieve 35.16 dB which is higher than MSSPL (34.27 dB) for the  $\times 4$  task, even though our method takes much less input information from the high-resolution input.

**Speed and memory.** Table 2 reports the speeds and the peak memory of the above methods. As our method is based on the transformer, our method is slower than CNN-based methods, including EDSR [LSK\*17], RCAN [ZLL\*18] and MSSPL [HLM\*21]. This is consistent with many other works that transformer tends to be slower than CNN [DBK\*20, TCD\*21, LCS\*21]. Compared to transformer-based method SwinIR, our method is slightly faster. We also compare the peak memory to produce a  $1024 \times 1024$  image in Table 2. Our method uses less peak memory than EDSR and MSSPL but more memory than RCAN and SwinIR.

#### 4.2. Comparison with denoising methods

We compare our methods to the state-of-the-art Monte Carlo denoising methods, including Sen [SD12], Rousselle [RKZ11], Kalantari [KBS15], Bitterli [BRM\*16], KPCN[BVM\*17], Gharbi [GLA\*19], MSSPL [HLM\*21], AdvMC [XZW\*19] and MCSA [YNL\*21]. Tables 3 and 4 report results on the BCR dataset [HLM\*21] and the Gharbi dataset [GLA\*19], respectively. We obtain the results of the comparing methods either from their authors [HLM\*21] or from their project websites [GLA\*19]. MSSPL [HLM\*21] was trained on the BCR dataset. For KPCN [BVM\*17], AdvMC [XZW\*19] and MCSA [YNL\*21], we finetuned their official models on the BCR dataset using their official training scripts. For our model, We trained a distinct model for each scale and sampling count.

As most denoising methods do not take high-resolution auxiliary features as input, we follow MSSPL [HLM\*21] to compute the average spp for our method and MSSPL as  $app_{avg} = spp_{LR}/s^2 + spp_{HR}$ , where  $s$  indicates the scale.  $spp_{LR}$  and  $spp_{HR}$  indicate the sampling

**Table 3:** Comparison on the BCR dataset [HLM\*21]. (4 - 1) indicates that our method takes a 4-spp low-resolution rendering and 1-spp fast-to-compute auxiliary feature as input. MSSPL also takes the 1-spp high-resolution rendering as input, including the diffusion and specular layers. Our method takes less shading information compared to MSSPL [HLM\*21].

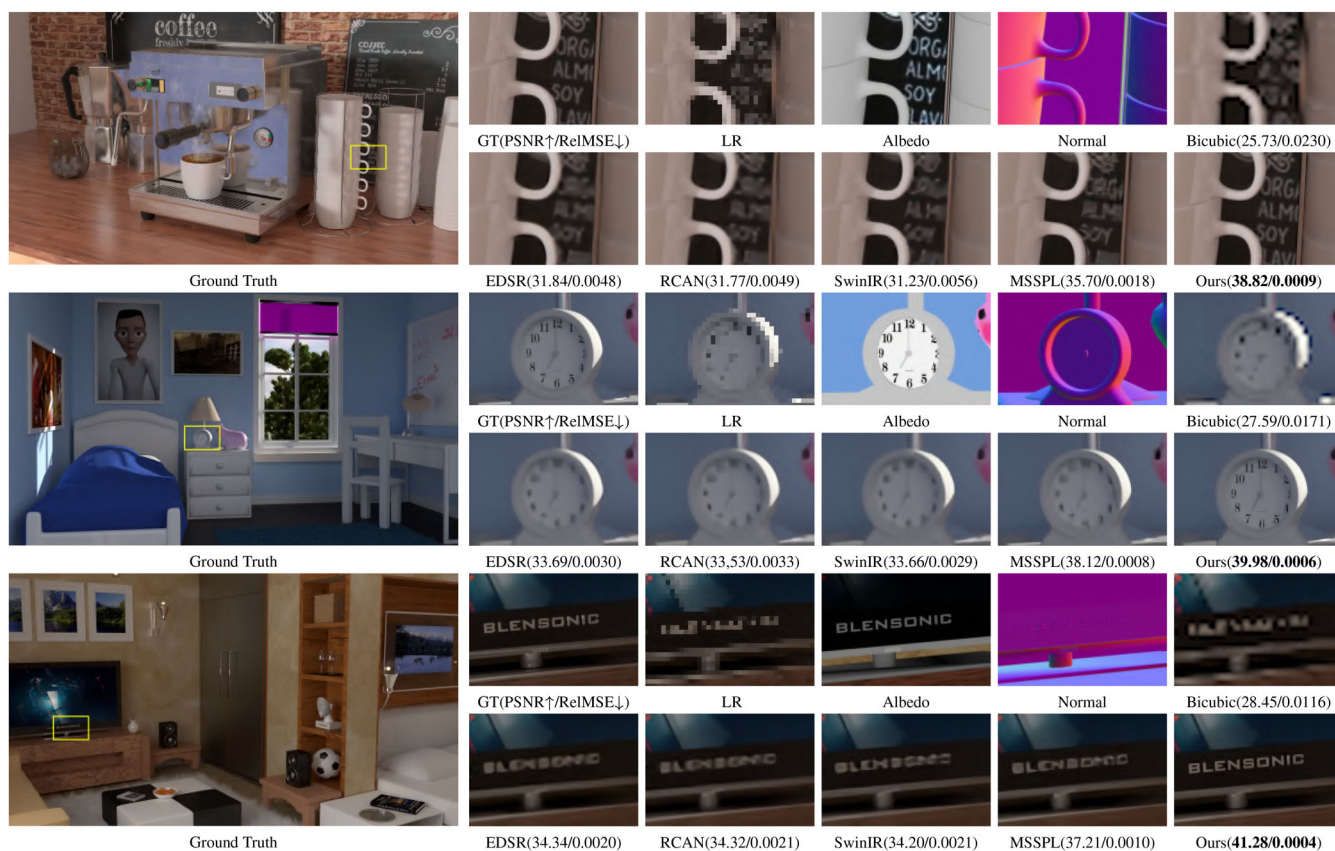
Method	2 spp		4 spp		8 spp	
	PSNR	ReIMSE	PSNR	ReIMSE	PSNR	ReIMSE
Input	18.12	0.2953	21.51	0.1400	24.75	0.0646
KPCN	25.87	0.0390	27.31	0.0299	28.11	0.0276
KPCN-ft	31.03	0.0078	33.69	0.0043	35.83	0.0026
Bitterli	26.67	0.0293	27.22	0.0252	27.45	0.0226
Gharbi	30.73	0.0068	31.61	0.0057	32.29	0.0050
MSSPL $\times 2$	33.27	0.0044	35.15	0.0027	36.74	0.0019
MSSPL $\times 4$	33.94	0.0039	35.21	0.0028	36.31	0.0022
MSSPL $\times 8$	31.37	0.0075	32.35	0.0057	33.14	0.0049
AdvMC-ft	30.33	-	32.30	-	33.69	-
MCSA-ft	32.68	0.0049	34.81	0.0031	36.61	0.0021
Ours $\times 1$	(1 - 1)		(2 - 2)		(4 - 4)	
	31.04	0.0078	34.67	0.0030	36.62	0.0020
Ours $\times 2$	(4 - 1)		(8 - 2)		(16 - 4)	
	<b>34.12</b>	<b>0.0035</b>	<b>35.49</b>	<b>0.0026</b>	<b>37.09</b>	<b>0.0018</b>
Ours $\times 4$	(16 - 1)		(32 - 2)		(64 - 4)	
	34.08	0.0046	35.06	0.0034	35.77	0.0029
Ours $\times 8$	(64 - 1)		(128 - 2)		(250 - 4)	
	31.10	0.0095	31.36	0.0093	31.62	0.0093

rates for the low- and high-resolution inputs, respectively. In our case, we take the sampling rates for the auxiliary features as  $spp_{HR}$ . We would like to note that this measurement of spp is **unfair** to our method, as our method only uses high-resolution albedo and normal features which takes much less time than rendering all the shading layers to obtain the high-resolution rendering as done in MSSPL (see Figure 6).

As shown in Table 3, our method generates better results than the state-of-the-art methods on the BCR dataset [HLM\*21]. Ours  $\times 2$  model wins 0.18, 0.28 and 0.35 dB in terms of PSNR on 2 spp, 4 spp and 8 spp, respectively.

We also conduct our experiments on  $\times 16$  scale. On the one hand, with  $\times 16$ , our method produces worse results than MSSPL because MSSPL uses the high-resolution RGB image as input that is not available to our method. While the high-resolution RGB input to MSSPL is rendered at a low sampling rate, it still provides useful information. As shown in the existing literature on Monte Carlo denoising, even the rendering result at 1 spp can be denoised to a reasonable quality. At such a high upsampling rate of  $\times 16$ , super-resolution is very difficult. On the other hand, in practice, given a target overall spp rate, our method can select an optimal combination of (spp rate, super-resolution scale) that outperforms MSSPL and other methods, as shown in Table 3. In practice,  $\times 16$  will not be used for rendering by either MSSPL or our method to achieve an overall target spp as it produces the worst results among alternative combinations of spp rate and super-resolution scale.

Figure 7 shows the visual comparisons. Our results are more visually plausible. Briefly, instead of working in the pixel colour space



**Figure 6:** Visual comparison with super-resolution methods on the BCR dataset [HLM\*21].

**Table 4:** Comparison on the Gharbi dataset [GLA\*19]. We directly test our models pre-trained on the BCR dataset without finetuning.

Method	4 spp		8 spp		16 spp	
	PSNR	ReIMSE	PSNR	ReIMSE	PSNR	ReIMSE
Input	19.58	17.5358	21.91	7.5682	24.17	11.2189
Sen	28.23	1.0484	28.00	0.5744	27.64	0.3396
Rousselle	30.01	1.9407	32.32	1.9660	34.36	1.9446
Kalantari	31.33	1.5573	33.00	1.6635	34.43	1.8021
Bitterli	28.98	1.1024	30.92	0.9297	32.40	0.9640
KPCN	29.75	1.0616	30.56	7.0774	31.00	20.2309
KPCN-ft	29.86	0.5004	31.66	0.8616	33.39	0.2981
Gharbi	33.11	<b>0.0486</b>	34.45	<b>0.0385</b>	35.36	<b>0.0318</b>
MSSPL×2	34.02	1.5025	35.30	1.4902	<b>36.43</b>	1.4748
MSSPL×4	33.94	5.5586	35.22	5.6781	35.97	5.7436
MSSPL×8	31.56	3.7228	32.60	4.2300	33.22	4.5045
Ours×1	(2 - 2)		(4 - 4)		(8 - 8)	
	27.41	0.3438	30.39	0.3092	32.88	0.3062
Ours×2	(8 - 2)		(16 - 4)		(32 - 8)	
	<b>34.29</b>	2.2587	<b>35.47</b>	1.5480	36.37	1.5417
Ours×4	(32 - 2)		(64 - 4)		(128 - 8)	
	34.26	20.7861	35.12	29.0364	35.52	28.1264
Ours×8	(128 - 2)		(16 - 8)		(32 - 16)	
	31.57	1.3474	31.26	1.1718	31.51	1.0940

that can potentially cause the colour fidelity problem, our method fuses the low-resolution RGB and high-resolution feature maps in the feature space and learns to fuse them into correct colours, thus alleviating the colour ambiguities/artifacts at fine details. For example, in Figure 7, the wall of our results is less noisy and more accurate than the results from other methods that are either blurred or inconsistent with the ground truth. In the second example, our method produces high-frequency geometric details in the wine basket area that well differentiates the mesh colour and the background colour.

Table 4 reports the comparison on the Gharbi dataset [GLA\*19]. Following MSSPL [GLA\*19], we directly test our models pre-trained on the BCR dataset without fine-tuning as the training set of the Gharbi dataset is not available. Our ×2 model wins 0.27 and 0.17 dB in terms of PSNR on 4 and 8 spp, respectively. When the spp is 16, our PSNR is slightly lower than MSSPL [HLM\*21]. We would like point out our method takes less high-resolution information than MSSPL. Our input high-resolution auxiliary features only include the albedo and normal, while MSSPL also takes all the shading layers as inputs. When the high-resolution input is rendered at a high spp, the shading layers can contribute a lot of high-frequency information. Similar to the findings in MSSPL [HLM\*21], our results on ReIMSE are heavily affected by a small number of pixels with abnormal large errors. Excluding these abnormal pixels can greatly improve our scores on ReIMSE. As shown in Figure 8, our method



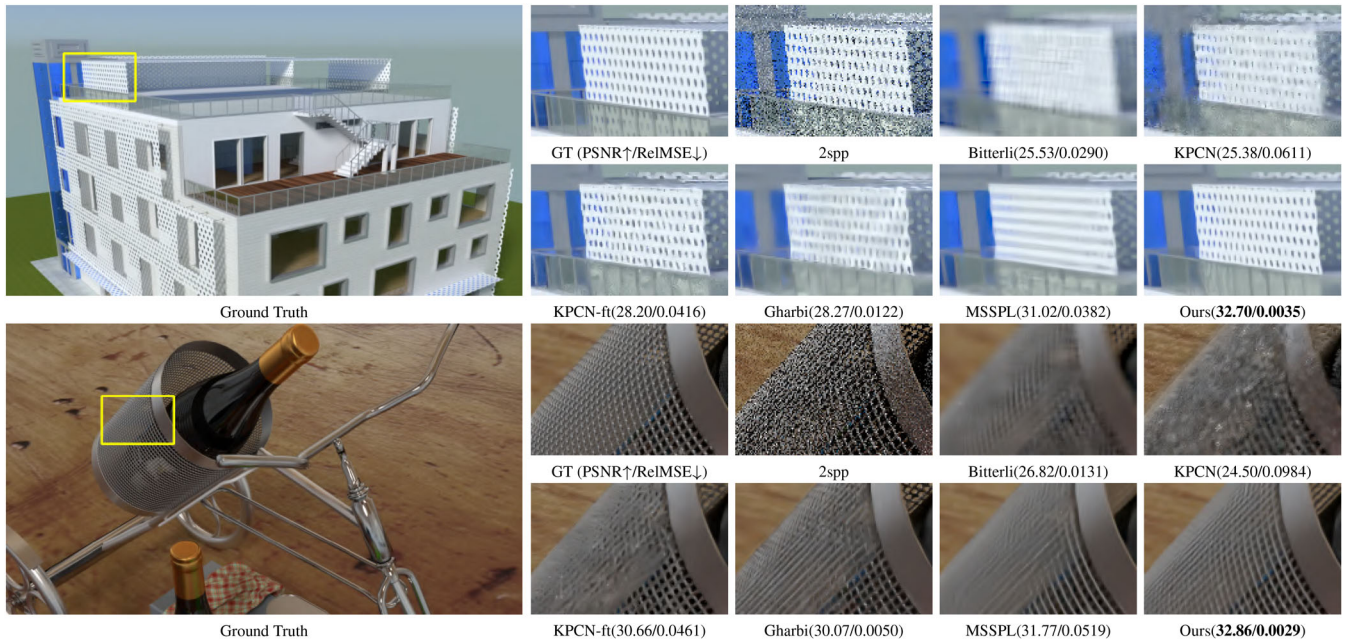


Figure 7: Visual comparison with denoising methods on the BCR dataset [HLM\*21].

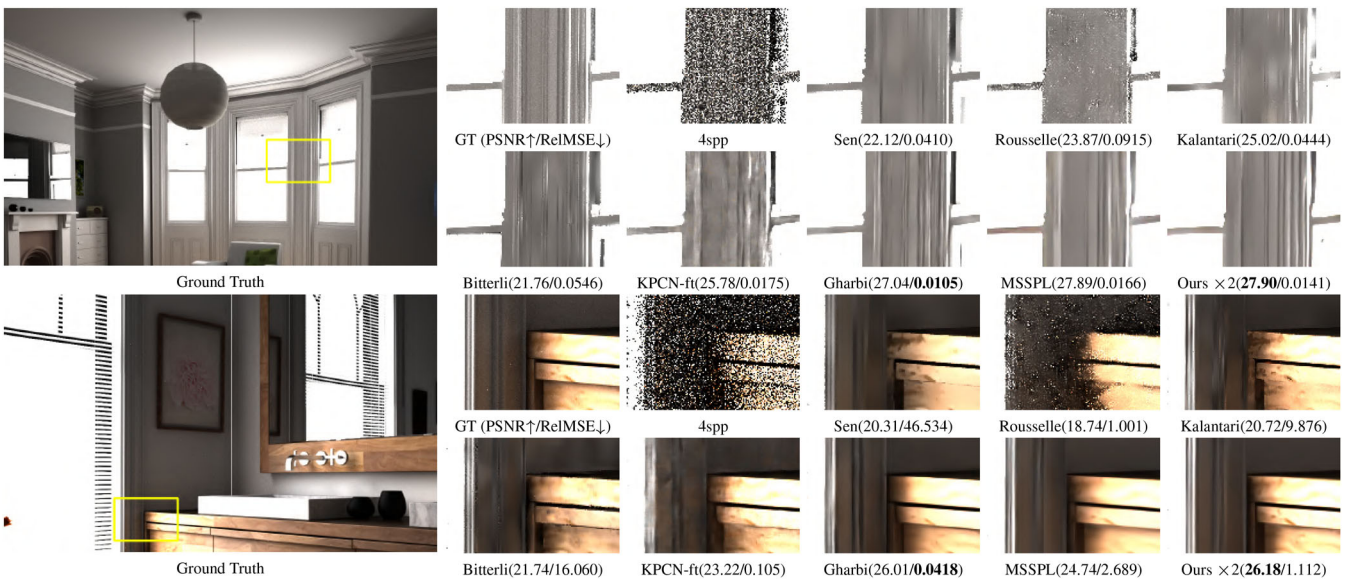


Figure 8: Visual comparison with denoising methods on the Gharbi dataset [GLA\*19].

produces high-quality results with much fewer artifacts when compared to the ground truth.

#### 4.3. Discussions

**Auxiliary features sampling rates.** As discussed above and shown in Figure 5, using more samples to generate the auxiliary features help our method generate better super-resolution results. However, even using one sample per pixel to generate the auxiliary features

can already enable our method to significantly outperform standard super-resolution methods. Moreover, when we use 16 samples to generate these features, our results are already very close to the results that use the features generated using 4000 samples per pixel denoted as  $A_{gt}$  in the figure.

**Input layers of auxiliary features.** We examine how our method works with different auxiliary feature layers. The upsampling scale is set to  $4\times$ . We use 4000 spp for  $I_{LR}$  and  $A$ . As shown in Table 5, both albedo and normal can improve the results significantly, as they



**Table 5:** The effects of input fast-to-compute auxiliary feature layers on the BCR dataset [HLM\*21].

Auxiliary Layer	None	Normal	Albedo	Normal + Albedo
PSNR	30.49	34.85	36.42	<b>37.45</b>
RelMSE	0.0141	0.0042	0.0030	<b>0.0021</b>

**Table 6:** Ablation study w.r.t. MSSPL [HLM\*21] on the BCR dataset [HLM\*21]. We compare their performance using fast-to-compute auxiliary features layers ('Fast') and full auxiliary feature layers ('Full').

Method	Buffer	2 spp		4 spp		8 spp	
		PSNR	RelMSE	PSNR	RelMSE	PSNR	RelMSE
MSSPL	Full	(16 - 1)		(32 - 2)		(64 - 4)	
		33.94	0.0039	35.21	0.0028	36.31	0.0022
MSSPL	Fast	(16 - 1)		(32 - 2)		(64 - 4)	
		32.32	0.0079	33.42	0.0060	33.96	0.0050
Ours	Full	(16 - 1)		(32 - 2)		(64 - 4)	
		34.84	0.0033	36.01	0.0024	37.22	0.0018
Ours	Fast	(16 - 1)		(32 - 2)		(64 - 4)	
		34.08	0.0046	35.06	0.0034	35.77	0.0029

can provide the essential high-frequency visual details for super-resolution. The performance of our network can be further improved if we take both of them as inputs. These findings are consistent with previous denoising methods [BVM\*17, GLA\*19] where intermediate layers can improve the final results.

**Ablation study w.r.t. MSSPL [HLM\*21].** We evaluated the performance of both our method and MSSPL [HLM\*21] using fast-to-compute auxiliary features as well as full auxiliary features. In the experiments, the upsampling scale is set to  $\times 4$ . As shown in Table 6, both our network and MSSPL benefit from using the full auxiliary features due to the richer high-resolution information they provide. However, our method with fast-to-compute layers still outperforms MSSPL with full auxiliary layers, which demonstrates the effectiveness of our network architecture.

**Network effectiveness.** We examine how our network architecture works by comparing to AdvMC [XZW\*19] and MCSA [YNL\*21]. Specifically, we feed high-resolution 1-spp RGB and 1-spp auxiliary buffers to AdvMC and MCSA and fine tune them on the BCR dataset. In this experiment, our method takes 4-spp low-resolution RGB ( $\times 2$ , effectively the same sampling rate as 1 spp at the high resolution) and 1-spp high-resolution auxiliary buffers. Table 7 shows our method outperforms these methods, which demonstrates the effectiveness of our transformer-based network architecture.

**Network architecture components.** We examine the effect of the network architecture. The upsampling scale is set to  $4\times$ . In this test, we remove XM modules and replace our RDST with the state-of-the-art blocks, including RDB from RRN [ZTK\*18] and RSTB from SwinIR [LCS\*21]. As shown in Table 8, our RDST can greatly improve the results. These improvements can be attributed to the strong generalization capability of RDST. Besides, XM modules can further improve the results.

**Table 7:** The effects of network architectures on the BCR dataset [HLM\*21]. AdvMC-ft [XZW\*19] and MCSA [YNL\*21] take 1-spp RGB and 1-spp auxiliary buffers as inputs. Our method takes 4-spp low-resolution RGB ( $\times 2$ , effectively the same sampling rate as 1 spp at the high resolution) and 1-spp high-resolution auxiliary buffers).

Method	AdvMC-ft	MCSA	Ours
PSNR	27.96	30.01	<b>34.12</b>
LPIPS	0.320	0.202	<b>0.090</b>

**Table 8:** The effects of network architecture components on the BCR dataset [HLM\*21]. We compare the proposed RDST with RDB [ZTK\*18] and RSTB [LCS\*21].

Network	RDB	RSTB	RDST	RDST + XM
PSNR	35.56	36.63	37.27	<b>37.45</b>
RelMSE	0.0034	0.0098	0.0022	<b>0.0021</b>

**Table 9:** The effects of the number of RDST blocks on the BCR dataset [HLM\*21]. We measured the flops and macs for a single  $1024 \times 1024$  image [RRRH20].

RDST Num	PSNR	RelMSE	Flops(T)	Macs(G)	Params(M)
5	34.08	0.0046	1.45	723.68	9.36
3	33.47	0.0056	1.04	519.13	6.21
1	32.60	0.0091	0.63	314.59	3.06

**Table 10:** The effects of the number of XDG blocks on the BCR dataset [HLM\*21]. We measured the flops and macs for a single  $1024 \times 1024$  image [RRRH20].

XDG Num	PSNR	RelMSE	Flops(T)	Macs(G)	Params(M)
3	34.08	0.0046	1.45	723.68	9.36
2	33.19	0.0066	1.02	507.86	6.42
1	32.30	0.1193	0.59	292.04	3.49

**Number of RDST blocks.** We examine how our network architecture works with different RDST blocks in each XDG block on the BCR dataset [HLM\*21]. In this test, the upsampling scale is set to  $\times 4$ . To check the impact of RDST, we set the XDG number as 3, and we investigated our results across different RDST numbers of each XDG block, including 1, 3 and 5. Besides, we also measure the flops, macs and parameters for a single  $1024 \times 1024$  image [RRRH20]. As shown in Table 9, decreasing the number of RDST blocks accelerates the network but compromises performance.

**Number of XDG blocks.** Similar to RDST, we investigate our results across different XDG numbers, including 1, 2 and 3. The upsampling scale is set to  $\times 4$  and the number of RDST of each XDG block is set to 5. As the results reported in Table 10, reducing the number of XDG blocks accelerates the network but also compromises performance.

**Table 11:** Comparison on the perceptual quality on the BCR dataset [HLM\*21]. We utilize the LPIPS [ZIE\*18] metric as a measure of perceptual quality.

Method	2 spp		4 spp		8 spp	
	PSNR	LPIPS	PSNR	LPIPS	PSNR	LPIPS
AdvMC-ft	30.33	0.209	32.30	0.155	33.69	0.126
MCSA-ft	32.68	0.108	34.81	0.080	36.61	0.068
Ours	<b>34.12</b>	<b>0.090</b>	<b>35.49</b>	<b>0.070</b>	<b>37.09</b>	<b>0.057</b>

**Our robust loss versus SMAPE loss** [Mea86]. Our robust loss is used based on our observations that there are a very small number of pixels with abnormally large intensity values in our dataset, mostly due to the firefly artifacts. These pixels will often incur very large errors during training and thus compromise the performance of our model. We use the robust loss to reduce these undesirable impacts of these pixels as this robust loss will limit the maximal loss value to 1 no matter how large the pixel error is. We compared these two loss functions. In our experiments, the upsampling factor is set to 4, and we set the sampling rate to (16 - 1). Models trained with the SMAPE loss showed slightly worse results: 33.96 versus 34.12 in PSNR, and 0.0046 versus 0.0035 in RelMSE.

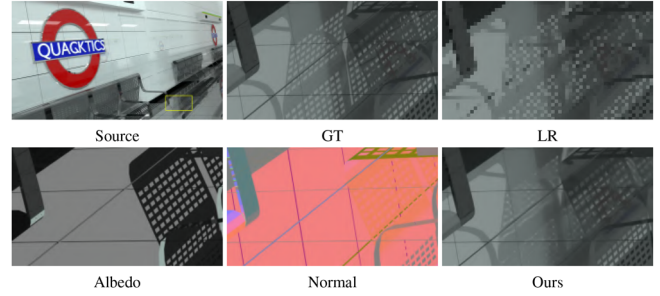
**Super-resolution scale.** We investigate our results across multiple scales, including  $\times 1$ ,  $\times 2$ ,  $\times 4$  and  $\times 8$ . Among them, scales  $\times 1$  and  $\times 8$  exhibit weaker performance compared to  $\times 2$  and  $\times 4$ . When comparing scales  $\times 4$  and  $\times 2$ ,  $\times 4$  takes less peak memory and is faster than  $\times 2$ , but  $\times 2$  leads to better quality. To make a fair comparison, we maintain a consistent average sampling rate across different scales. Consequently, the low-resolution input of our  $\times 1$  model is rendered at a much lower average sampling rate than that of our  $\times 2$  model. This makes the resulting input RGB image to our model very noisy for  $\times 1$  and thus comprises the final quality of Ours  $\times 1$ , as reported in the 2-spp column of Table 3. In the 4-spp column of the same table, the difference between Ours  $\times 1$  and Ours  $\times 2$  is less significant as in this setting, the average sampling rate of Ours  $\times 1$  is reasonably higher and provides more information for our model to synthesize higher-quality results.

In addition, we used the same training pipeline for our  $\times 1$  model as we did for other scales, keeping the number of epochs consistent across all scales. However, due to the high memory requirement to train the  $\times 1$  model, we have to set a smaller mini-batch size. This would also potentially impact the performance, but we believe that this is not as significant as the first reason we discussed above.

**Perceptual quality.** We examine the perceptual quality of our results using the LPIPS metric [ZIE\*18]. Tables 7 and 11 present the results for AdvMC [XZW\*19], MCSA [YNL\*21] and our method. Our approach outperforms the others in terms of both PSNR and LPIPS, thereby demonstrating its ability to generate images with high perceptual quality.

## 5. Limitations and Future Work

The fusion for the high-reflection parts is challenging. Our method produces high-frequency visual details by two means: (1) train a



**Figure 9:** Failure example. The performance of our method is compromised in the area where the albedo and normal could not provide high-frequency details.

neural network to learn to recover high-frequency information from low-resolution input and (2) use high-frequency information from the high-resolution albedo and normal maps. Our neural network can learn to produce visual details for many examples. However, super-resolution from a low-resolution input alone is necessarily an ill-posed problem. In the high-reflection parts of the scene, such as the example shown in Figure 9, when the high-resolution normal and albedo map cannot, by its nature, provide high-frequency details in those regions, our method may fail.

Compared to CNN-based methods, our method is slow. However, compared to another transformer-based method [YNL\*21], our method uses less peak memory (0.89 vs. 30.56Gb) and is faster (1.0 vs. 2.5 s) when producing a  $1024 \times 1024$  image using an Nvidia A40. Research on fast transformers has been advancing quickly recently. Patro *et al.* [PA23] offer an extensive review of efficient vision transformers. Through the advancement of effective token mixing strategies and efficient MLP layers, vision transformers can be significantly accelerated [LWZ\*22, GHW\*22, YPL\*22]. For example, both CMT [GHW\*22] and WaveViT [YPL\*22] outperform EfficientNet [TL19] while maintaining a lower computational complexity. Moreover, several transformer hardware accelerators have been introduced to expedite transformer networks, such as Swift-Tron [MDC\*23]. We believe that our method can benefit from the quick advance of research on transformer.

In this paper, we specifically explored albedo and normal as quick-to-compute auxiliary features. However, we acknowledge that other auxiliary features, such as a whitted ray-traced layer, could offer valuable high-frequency information and be generated fast. Incorporating such a layer can potentially improve the performance of our method. Unfortunately, the BCR dataset does not contain such layers. We plan to explore this in our future research.

## 6. Conclusion

This paper explored high-resolution fast-to-compute auxiliary features to guide super-resolution of Monte Carlo renderings. We developed a dedicated cross-modality transformer network to fuse high-resolution fast-to-compute auxiliary features with the corresponding low-resolution rendering. We designed a transformer-based cross-modality module to fuse the features from two modalities. We also developed a Residual Densely Connected Swin Transformer block to learn more representative features.

Experimental results indicate that our proposed method surpasses existing state-of-the-art super-resolution and denoising techniques in producing high-quality images.

### Acknowledgements

The authors have nothing to report.

### References

- [AKS18] AHN N., KANG B., SOHN K.-A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision* (2018), pp. 252–268.
- [BDM\*21] BRIEDIS K. M., DJELOUAH A., MEYER M., MCGONIGAL I., GROSS M., SCHROERS C.: Neural frame interpolation for rendered content. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–13.
- [BM98] BOLIN M. R., MEYER G. W.: A perceptually based adaptive sampling algorithm. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (1998), pp. 299–309.
- [BRM\*16] BITTERLI B., ROUSSELLE F., MOON B., IGLESIAS-GUITIÁN J. A., ADLER D., MITCHELL K., JAROSZ W., NOVÁK J.: Nonlinearly weighted first-order regression for denoising Monte Carlo renderings. In *Computer Graphics Forum* (2016), Wiley Online Library, vol. 35, pp. 107–117.
- [BVM\*17] BAKO S., VOGELS T., MCWILLIAMS B., MEYER M., NOVÁK J., HARVILL A., SEN P., DEROSE T., ROUSSELLE F.: Kernel-predicting convolutional networks for denoising Monte Carlo renderings. *ACM Transactions on Graphics* 36, 4 (2017), 97.
- [CKS\*17] CHAITANYA C. R. A., KAPLANYAN A. S., SCHIED C., SALVI M., LEFOHN A., NOWROUZEZAHRAI D., AILA T.: Interactive reconstruction of Monte Carlo image sequences using a recurrent denoising autoencoder. *ACM Transactions on Graphics* 36, 4 (2017), 98.
- [CMS\*20] CARION N., MASSA F., SYNNAEVE G., USUNIER N., KIRILLOV A., ZAGORUYKO S.: End-to-end object detection with transformers. In *European Conference on Computer Vision* (2020), Springer, pp. 213–229.
- [CPC84] COOK R. L., PORTER T., CARPENTER L.: Distributed ray tracing. *ACM SIGGRAPH Computer Graphics* 18 (1984), 137–145.
- [CWG\*21] CHEN H., WANG Y., GUO T., XU C., DENG Y., LIU Z., MA S., XU C., XU C., GAO W.: Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 12299–12310.
- [DBK\*20] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISENBORN D., ZHAI X., UNTERTHINER T., DEGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J., HOULSBY N.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [DLHT14] DONG C., LOY C. C., HE K., TANG X.: Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision* (2014), Springer, pp. 184–199.
- [DSHL10] DAMMERTZ H., SEWTEZ D., HANIKA J., LENSCH H.: Edge-avoiding À-Trous wavelet transform for fast global illumination filtering. In *Proceedings of the Conference on High Performance Graphics* (2010), Eurographics Association, pp. 67–75.
- [ETH\*09] EGAN K., TSENG Y.-T., HOLZSCHUCH N., DURAND F., RAMAMOORTHI R.: Frequency analysis and sheared reconstruction for rendering motion blur. *ACM Transactions on Graphics* 28 (2009), 93.
- [GFL\*21] GUO J., FU X., LIN L., MA H., GUO Y., LIU S., YAN L.-Q.: ExtraNet: Real-time extrapolated rendering for low-latency temporal supersampling. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–16.
- [GHW\*22] GUO J., HAN K., WU H., TANG Y., CHEN X., WANG Y., XU C.: CMT: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12175–12185.
- [GLA\*19] GHARBI M., LI T.-M., AITTALA M., LEHTINEN J., DURAND F.: Sample-based Monte Carlo denoising using a kernel-splating network. *ACM Transactions on Graphics* 38, 4 (2019), 1–12.
- [GLDZ22] GENG Z., LIANG L., DING T., ZHARKOV I.: RSTT: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).
- [HLM\*21] HOU Q., LI Z., MARSHALL C. S., PANNEER S., LIU F.: Fast Monte Carlo rendering via multi-resolution sampling. In *Graphics Interface Conference* (2021), pp. 25:1–9.
- [HLVDMW17] HUANG G., LIU Z., VAN DER MAATEN L., WEINBERGER K. Q.: Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4700–4708.
- [HMS\*20] HASSELGREN J., MUNKBERG J., SALVI M., PATNEY A., LEFOHN A.: Neural temporal adaptive sampling and denoising. In *Computer Graphics Forum* (2020), Wiley Online Library, vol. 39, pp. 147–155.



- [HSS18] HU J., SHEN L., SUN G.: Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141.
- [HSU19] HARIS M., SHAKHAROVICH G., UKITA N.: Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 3897–3906.
- [HWG18] HUI Z., WANG X., GAO X.: Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 723–731.
- [HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.
- [JC95] JENSEN H. W., CHRISTENSEN N. J.: Optimizing path tracing using noise reduction filters. In *3rd International Conference in Central Europe on Computer Graphics and Visualization* (1995).
- [Jen01] JENSEN H. W.: *Realistic Image Synthesis Using Photon Mapping*. AK Peters/CRC Press, Natick, MA, 2001.
- [Kaj86] KAJIYA J. T.: The rendering equation. *ACM SIGGRAPH Computer Graphics* 20, 4 (1986), 143–150.
- [KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations* (2015).
- [KBS15] KALANTARI N. K., BAKO S., SEN P.: A machine learning approach for filtering Monte Carlo noise. *ACM Transactions on Graphics* 34, 4 (2015), 122.
- [KKLML16] KIM J., KWON LEE J., MU LEE K.: Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 1637–1645.
- [KKR18] KUZNETSOV A., KALANTARI N. K., RAMAMOORTHY R.: Deep adaptive sampling for low sample count rendering. *Computer Graphics Forum* 37 (2018), 35–44.
- [LCS\*21] LIANG J., CAO J., SUN G., ZHANG K., VAN GOOL L., TIMOFTE R.: SwinIR: Image restoration using Swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 1833–1844.
- [LLC\*21] LIU Z., LIN Y., CAO Y., HU H., WEI Y., ZHANG Z., LIN S., GUO B.: Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 10012–10022.
- [LR90] LEE M. E., REDNER R. A.: A note on the use of nonlinear filtering in computer graphics. *IEEE Computer Graphics and Applications* 10, 3 (1990), 23–29.
- [LSK\*07] LAINE S., SARANSAARI H., KONTKANEN J., LEHTINEN J., AILA T.: Incremental instant radiosity for real-time indirect illumination. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (2007), Eurographics Association, pp. 277–286.
- [LSK\*17] LIM B., SON S., KIM H., NAH S., MU LEE K.: Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2017), pp. 136–144.
- [LTT\*19] LI Y., TSIMINAKI V., TIMOFTE R., POLLEFEYS M., GOOL L. V.: 3D appearance super-resolution with deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 9671–9680.
- [LWC12] LI T.-M., WU Y.-T., CHUANG Y.-Y.: Sure-based optimization for adaptive sampling and reconstruction. *ACM Transactions on Graphics* 31, 6 (2012), 194.
- [LWLS19] LIU Z.-S., WANG L.-W., LI C.-T., SIU W.-C.: Hierarchical back projection network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019)
- [LWZ\*22] LI K., WANG Y., ZHANG J., GAO P., SONG G., LIU Y., LI H., QIAO Y.: UniFormer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [MA06] MEYER M., ANDERSON J.: Statistical acceleration for animated global illumination. *ACM Transactions on Graphics* 25, 3 (2006), 1075–1080.
- [McC99] MCCOOL M. D.: Anisotropic diffusion for Monte Carlo noise reduction. *ACM Transactions on Graphics* 18, 2 (1999), 171–194.
- [MDC\*23] MARCHISIO A., DURA D., CAPRA M., MARTINA M., MASERA G., SHAFIQUE M.: SwiftTron: An efficient hardware accelerator for quantized transformers. *International Joint Conference on Neural Networks* (2023).
- [Mea86] MEADE N.: Long range forecasting: From crystal ball to computer. *Journal of the Operational Research Society* 37, 5 (1986), 533–535.
- [MH20] MUNKBERG J., HASSELGREN J.: Neural denoising with layer embeddings. In *Computer Graphics Forum* (2020), Wiley Online Library, vol. 39, pp. 1–12.
- [MMMG16] MOON B., McDONAGH S., MITCHELL K., GROSS M.: Adaptive polynomial rendering. *ACM Transactions on Graphics* 35, 4 (2016), 40.
- [ODR09] OVERBECK R. S., DONNER C., RAMAMOORTHY R.: Adaptive wavelet rendering. *ACM Transactions on Graphics* 28, 5 (2009), 140.

- [RKZ11] ROUSSELLE F., KNAUS C., ZWICKER M.: Adaptive sampling and reconstruction using greedy error minimization. *ACM Transactions on Graphics* 30, 6 (2011), 159.
- [RRRH20] RASLEY J., RAJBHANDARI S., RUWASE O., HE Y.: DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), pp. 3505–3506.
- [RW94] RUSHMEIER H. E., WARD G. J.: Energy preserving non-linear filters. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques* (1994), pp. 131–138.
- [SCH\*16] SHI W., CABALLERO J., HUSZÁR F., TOTZ J., AITKEN A. P., BISHOP R., RUECKERT D., WANG Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 1874–1883.
- [SD12] SEN P., DARABI S.: On filtering the noise from the random parameters in Monte Carlo rendering. *ACM Transactions on Graphics* 31, 3 (2012), 18.
- [SIMP06] SEGOVIA B., IEHL J. C., MITANCHEY R., PÉROCHE B.: Non-interleaved deferred shading of interleaved sample patterns. In *Graphics Hardware* (2006), pp. 53–60.
- [TCD\*21] TOUVRON H., CORD M., DOUZE M., MASSA F., SABLAYROLLES A., JEGOU H.: Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning* (July 2021), vol. 139, pp. 10347–10357.
- [TL19] TAN M., LE Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (2019), PMLR, pp. 6105–6114.
- [TLP\*22] THOMAS M. M., LIKTOR G., PETERS C., KIM S., VAIDYANATHAN K., FORBES A. G.: Temporally stable real-time joint neural denoising and supersampling. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5, 3 (2022), 1–22.
- [VRM\*18] VOGELS T., ROUSSELLE F., MCWILLIAMS B., RÖTHLIN G., HARVILL A., ADLER D., MEYER M., NOVÁK J.: Denoising with kernel prediction and asymmetric loss functions. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–15.
- [VSP\*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), vol. 30.
- [WABG06] WALTER B., ARBREE A., BALA K., GREENBERG D. P.: Multidimensional lightcuts. *ACM Transactions on Graphics* 25, 3 (2006), 1081–1088.
- [WRC88] WARD G. J., RUBINSTEIN F. M., CLEAR R. D.: A ray tracing solution for diffuse interreflection. *ACM SIGGRAPH Computer Graphics* 22, 4 (1988), 85–92.
- [XMS19] XU X., MA Y., SUN W.: Towards real scene super-resolution with raw images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 1723–1731.
- [XNC\*20] XIAO L., NOURI S., CHAPMAN M., FIX A., LANMAN D., KAPLAYAN A.: Neural supersampling for real-time rendering. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 142.
- [XP05] XU R., PATTANAIK S. N.: A novel Monte Carlo noise reduction operator. *IEEE Computer Graphics and Applications* 25, 2 (2005), 31–35.
- [XZW\*19] XU B., ZHANG J., WANG R., XU K., YANG Y.-L., LI C., TANG R.: Adversarial Monte Carlo denoising with conditioned auxiliary feature modulation. *ACM Transactions on Graphics* 38, 6 (2019), 224.
- [YNL\*21] YU J., NIE Y., LONG C., XU W., ZHANG Q., LI G.: Monte Carlo denoising via auxiliary feature guided self-attention. *ACM Transactions on Graphics* 40, 6 (2021), 273.
- [YPL\*22] YAO T., PAN Y., LI Y., NGO C.-W., MEI T.: WaveViT: Unifying wavelet and transformers for visual representation learning. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV* (2022), Springer, pp. 328–345.
- [ZIE\*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*(2018).
- [ZJL\*15] ZWICKER M., JAROSZ W., LEHTINEN J., MOON B., RAMAMOORTHY R., ROUSSELLE F., SEN P., SOLER C., YOON S.-E.: Recent advances in adaptive sampling and reconstruction for Monte Carlo rendering. *Computer Graphics Forum* 34 (2015), 667–681.
- [ZLL\*18] ZHANG Y., LI K., LI K., WANG L., ZHONG B., FU Y.: Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 286–301.
- [ZLZ\*21] ZHENG S., LU J., ZHAO H., ZHU X., LUO Z., WANG Y., FU Y., FENG J., XIANG T., TORR P. H., ET AL.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 6881–6890.
- [ZTK\*18] ZHANG Y., TIAN Y., KONG Y., ZHONG B., FU Y.: Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 2472–2481.

[ZWLQ19] ZHANG Z., WANG Z., LIN Z., QI H.: Image super-resolution by neural texture transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 7982–7991.

[ZZXY21] ZHENG S., ZHENG F., XU K., YAN L.-Q.: Ensemble denoising for Monte Carlo renderings. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–17.

[ZZZ18] ZHANG K., ZUO W., ZHANG L.: Learning a single convolutional super-resolution network for multiple degradations. In

*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3262–3271.

### Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information

Supporting Information