Geometry-Aware Reference Synthesis for Multi-View Image Super-Resolution

Ri Cheng, Yuqi Sun, Bo Yan*, Weimin Tan, and Chenxi Ma.

1 INTRODUCTION

Multi-view Multimedia (MVMM) uses numerous cameras to capture multi-view images of the same scene from various angles. It is widely used in auto-driving, street-view navigation, free-viewpoint video[17, 36], and virtual reality[1, 3]. Recent MVMM applications call for high-resolution (HR) images to give a pleasing visual experience and assist downstream tasks such as detection and tracking. Nevertheless, HR multi-view images face substantial storage costs or bandwidth constraints. To address this problem, we propose the Multi-View Image Super-Resolution (MVISR) task, which increases the resolution of multi-view images by using all low-resolution (LR) multi-view images in the same scene.

A naive approach to MVISR is to perform single image SR (SISR) [6, 21] on each view. They take a single view as input and cannot utilize the relationship across all LR views. Other alternatives include video SR (VSR) methods [4, 18]. They regard multi-views as video frames with little object motion, and so cannot deal with the large-angle transformations between input views. Some reference-based...
Figure 1: Multi-view image super-resolution. We propose GARS module for extracting the HR details from all LR Multi-View. Then, our GARS module synthesizes high-frequency textural details into HR Reference Image that is aligned with the bicubic of LR Input View. The word SQUARE is more textured in HR Reference Image than in the LR Input View.

SR (Ref-SR) methods [23, 45, 52] can overcome angle constraints and extend the range of reference views. However, Ref-SR is ineffective due to the difficulties involved with reference image selection in all LR multi-views and reference alignment when the scale and rotation differences between the Ref and LR input are large. In the end, none of the approaches discussed previously can explicitly leverage all LR multi-views and their associated geometry information.

Therefore, in this paper, we propose MVSRnet, which utilizes the geometry information to exploit high-frequency (HF) reference information from all LR views for each view SR. A camera that is closer to the objects can capture sharper images than a distant one. It inspires us that a close-up view contains more textural details that can provide sufficient reference for the distant view. To realize this idea, we design Geometry-Aware Reference Synthesis (GARS) module in MVSRnet. As illustrated in Figure 1, the word SQUARE is blurred in the LR input view but sharp in other LR views. Our GARS module takes LR multi-views as input to generate reference images with rich textural details and well-aligned with the LR input view. To be specific, we use structure-from-motion [31] and multi-view stereo [32, 46] to extract scene geometry information such as the camera pose and depth map. After that, GARS conducts 3D image warping to align all LR multi-views into the target view. Then, we introduce a depth-guided patch-selection strategy to select close views on the patch level and fuse them into HR multi-view reference images (MVRs).

On the other hand, MVRs synthesized from all LR views in GARS have artifacts and some pixels that are not properly aligned with the LR input view due to inaccuracies in the geometry information. Nearby views have a small difference in viewing angle and hence suffer less from geometric inaccuracies. As a result, we also synthesize near-view reference images (NVRs) with fewer artifacts but fewer HF textural details using nearby views. Then, in order to properly leverage MVRs and NVRs, we design the Raw Selection Module (RSM) and Adaptation Selection Module (ASM) in our Dynamic High-Frequency Search (DHFS) network to exploit the HF textural details by reducing the feature difference dynamically. Finally, we perform experiments to demonstrate the proposed MVSRnet SR capability, and some comparative experiments are shown in Figure 2. Additionally, our work contributes to the Ref-SR methodology by proposing a way for performing SR using multi-view LR images rather than HR images [52] as reference.

This paper mainly has the following contributions:

- For the first time, this paper proposes the MVSRnet, a learning-based network that makes explicit use of geometric information to address the MVISR problem. In addition, this work enriches the Reference-Based SR research.
- The Geometry-Aware Reference Synthesis module synthesizes the HR reference image from all LR multi-view images using geometry information. The reference supports SR by containing rich high-frequency textural details and aligning with the input view.
- The Dynamic High-Frequency Search network is proposed for reconstructing the high-resolution view by dynamically reducing the feature difference to exploit high-frequency textural details from synthesized reference images.

2 RELATED WORK

2.1 Single Image Super-Resolution

Single Image Super-Resolution (SISR) [6, 16, 21, 25, 41, 51] aims to learn the mapping function between LR and HR images. However, SISR is unable to recover HF details due to the destruction of HR textural clues in the downsampling LR image.
2.2 Reference-Based Super-Resolution

Reference-based super-resolution (Ref-SR) approaches transfer the HR details of the reference image to the LR input image. Several works [34, 53] leverage deformable convolution [5, 54] or estimated flow to align and extract reference features. However, it will seriously affect the performance of these algorithms when finding long-distance correspondence. The patch-based Ref-SR approaches [13, 23, 44, 45, 52] perform patch matching in the feature space by calculating the cosine similarity between the LR patch and Ref patch. [23, 45] formulate patch matching based on the transformer [39]. However, the difference in scale and rotation between the HR Ref and LR inputs can cause matching errors [13], which is serious in multi-view settings. In addition, for the super-resolution of multi-view LR images, there is no available view reference selection method. The SIFT-based [22] selection method used in constructing the CUFED5 [52] dataset is only applicable to selecting HR references whose content is similar to the LR input view. In general, it is difficult to apply the methods discussed above to the MVISR task directly.

3 PROPOSED METHOD

This section describes the proposed network called MVSNet, and the overview of the MVSNet is shown in Figure 3(a). We begin with a data preprocessing step that estimates depth maps \( D_{S \rightarrow T}^{hr} \) and poses \( \{ P_{S}^{lr, i} \} \) of all S LR multi-views in Section 3.1. Then, in Section 3.2, we introduce Geometry-Aware Reference Synthesis (GARS) that uses geometry priors \( \{ D_{S}^{lr} \}_{i=1}^{S} \) and \( \{ P_{S}^{lr, i} \}_{i=1}^{S} \) to synthesize V HR multi-view reference (MVR) \( \{ R_{M}^{N} \}_{i=1}^{V} \) and HR near-view reference (NVR) \( \{ R_{N}^{i} \}_{i=1}^{V} \). MVR stores more high-frequency (HF) textural details than NVR since MVR is synthesised from all S LR multi-view images \( \{ I_{S}^{lr} \}_{i=1}^{S} \), not from L nearby views like NVR. However, there are fewer artifacts caused by erroneous warping in the NVR. Therefore, in Section 3.3, we design Dynamic High-Frequency Search (DHFS) network that dynamically selects and extracts the correct HF details in MVRs and NVRs to super-resolve the LR input view \( I_{T}^{lr} \) and get HR output view \( I_{T}^{hr} \). At last, loss functions are shown in Section 3.4.

3.1 Preprocessing

To obtain the geometry information of the multi-view scene, we utilize COLMAP [31] to compute camera poses by inputting all LR multi-view images and depth maps. COLMAP computes camera poses by inputting all LR multi-view images and depth maps.

2.3 Multi-View Super-Resolution

Several traditional approaches [7, 38] and learning-based approaches [20, 28] study the multi-view texture SR problem. However, they only super-resolve the texture of the object instead of the entire image. Stereo image SR [40, 48] utilizes additional information provided from a second viewpoint, but it uses a much smaller number of viewpoint than MVISR. Light field image SR [2, 42, 43, 50] has also been extensively studied in recent years, but the view span are limited compared to the unstructured multi-views. Our MVSNet differs from multi-view video SR [19, 24] in that they utilize HR reference frames, whereas MVSNet exclusively uses LR frames. However, obtaining HR images is difficult due to substantial storage costs and bandwidth constraints in reality. Sun et al. [37] addresses the space-angle SR problem, but only two adjacent views are used to synthesize virtual view and SR. In general, it is difficult to apply the methods discussed above to the MVISR task directly.
S LR multi-views. This is known as structure-from-motion (SfM), and we can get the camera pose \( \{ P_i^r \}_{i=1}^S \) of each viewpoint. The pose is represented by the extrinsic parameters (rotation matrices and translation vectors) and intrinsic parameters. We use the depth map of the LR input view to align all LR multi-views to the LR input view. Therefore, we conduct the multi-view stereo (MVS) in COLMAP to obtain the depth map for each LR view \( \{ D_i^l \}_{i=1}^S \) by inputting all LR multi-views and the SfM output.

### 3.2 Geometry-Aware Reference Synthesis

The GARS module synthesizes references (Refs) from all multi-view images, and we display the synthesis process of a Ref in Figure 3(b). The reference image is aligned with the input view and contains high-frequency details.

#### Alignment Using Scene Geometry.

When the reference image has a large deformation in comparison to the input image, the alignment capabilities of recent Ref-SR approaches are reduced. For instance, Figure 3 illustrates a great differences in scale and rotation of the truck between the LR input view \( I_i^l \) and view \( I_i^h \). Therefore, we use the multi-view scene geometry to explicitly align all multi-view source images \( \{ I_i^h \}_{i=1}^S \) with the input view \( I_i^l \). Specifically, we up-sample the depth map to the target output resolution (SR factor) \( x4 \) in our experiment setting based on the bicubic mode and modify the camera intrinsic matrices according to the up-sampling factor. Then we apply the 3D warping operation [26] to obtain warped views \( \{ I_i^{hr} \}_{i=1}^S \) as follows:

\[
I_i^{hr} = W(p_i^{hr}, D_i^h, p_i^{hr}, D_i^l, t_i),
\]

where \( i \) and \( T \) refer to \( i \)-th multi-view images and input view. \( D_i^h \) and \( p_i^{hr} \) denote the depth maps and the camera pose at the same resolution as the target output solution. Specifically, the pixel coordinate correspondence between the multi-view images and the input view in the 3D warping operation is as the following formula:

\[
D_i(p_i) \ast p_i = K_i(R_iK_i^{-1}D_T + p_T + t_i), \tag{2}
\]

where \( p \) denotes pixel homogeneous coordinates. Camera intrinsic matrices \( K \), rotation matrices \( R \) and translation matrices \( t \) are provided by camera posed \( P \) as presented in Eq.(1). \( R_T = R_lR_T^l \), \( t_T = t_l - R_lT \) are relative rotation and translation matrices.

It is noted that we can directly apply 3D warping operation on LR depth maps without up-sampling and obtain LR aligned views \( \{ I_i^{hr} \}_{i=1}^S \). However, the general sampling modes in warping operation, such as nearest and bilinear, blurs and destructs high-frequency details using the weighted average method when dealing with overlapping pixels. One possible solution is to use a pre-trained network as a feature extractor to conduct 3D warping operations in the LR feature space. However, this leads to both spatially and temporally inefficient as hundreds of views need to be extracted features in each training iteration.

#### HF Details Extraction Using Depth-Guided Patch-Selection Strategy.

We should extract useful HF details from warped views \( \{ I_i^{hr} \}_{i=1}^S \) that contain redundant details as the warped view contains the LR details area as well. We assume that the view taken at close range has more high-frequency textural details than the distant view. Therefore, we apply the 3D warping operation on source depth maps \( W(D_i^l, D_i^h, p_i^{hr}, p_i^{hr}, D_i^l) \) to obtain HR warped depth maps \( \{ D_i^{hr} \}_{i=1}^S \). Warped depth maps present distance geometry information, i.e., the distance between the object and the camera. The smaller the depth value, the closer the object relative to the camera. As a result, we can leverage the HR warped depth maps to locate the HR details area in the warped views \( \{ I_i^{hr} \}_{i=1}^S \). For example, Figure 3(b) shows that the truck door has smaller depth values (darker blue) in the warped depth map \( D_i^{hr} \) than \( D_i^l \).

Hence, the HR warped view \( I_i^{hr} \) contains more truck door HR details than \( I_i^{hr} \), corresponding to more truck door high-frequency details in view \( I_i^{lr} \) than \( I_i^{lr} \).

We unfold each warped depth map into patches \( p_i^{rref}, i \in \{1,S\}, j \in \{1,(hr_h \times hr_w) / (ps^2)\} \), \( hr_h \), \( hr_w \), and \( ps \) denote the target HR height, target HR width, and patch size. \( j \) denotes the \( j \)-th patch index. We also unfold each warped view, and denote it as \( I_i^{hr} \). Then, we assign the depth value in each patch to its patch mean \( p_i^{mean} = \text{mean}(p_i^{rref}) \). This alleviates the effect of the depth value averaged in the area of each view by the up-sampling and warping operation. Then, we obtain the high-frequency index map \( H \) from the index with the smallest depth value in \( p_i^{mean} \) as follows:

\[
H_j = \arg \min_{i} p_i^{mean}, \tag{3}
\]

\( H_j \) refers to the high-frequency patch index of warped views, and we extract the high-frequency patch as the following formula:

\[
R_j = I_i^{hr} \quad i = H_j. \tag{4}
\]

Now we can obtain the multi-view reference image \( R \) by folding the multi-view reference patches \( R_j \).

#### Multiple References Synthesis.

It is impossible to integrate all of the high-frequency details into a reference image. As a result, we synthesize \( V \) multi-view reference images (MVRs) denoted as \( \{ R_j^V \}_{j=1}^V \) using those the top \( V \) high-frequency index maps. However, the depth errors generated by MVS estimation and 3D warping causes inaccurate warped depth values, leading to incorrect MVR selection and alignment. In addition, the large angle difference between views magnifies alignment errors, resulting artifacts in MVR. Therefore, we synthesize \( V \) near-view reference images (NVRs) \( \{ R_j^V \}_{j=1}^V \) from \( L \) nearby views that have a small viewing angle gap and similar lighting conditions as the LR input view. NVRs suffer less from geometric inaccuracies, and the selection criteria for nearby views are based on their order in the dataset. NVRs includes less high-frequency details than MVR, but it also contains fewer alignment errors, and artifacts. As a result, both MVRs and NVRs can be used to cross-reference to correct selection errors and accurately supply HF details to the SR model.

### 3.3 Dynamic High-Frequency Search

MVR has both more HF textural details and more artifacts than NVR, and that NVR is the inverse. Therefore, we propose the DHFS network to maximize the use of references (MVR and NVR). As shown in Figure 4(a), we first concatenate the MVRs and NVRs to the bicubic input view at the channel level, and use a learnable feature extractor to extract their multi-scale (x1, x2, and x4) textural feature \( f_{M}^{ref} \) and \( f_{N}^{ref} \). On the other hand, we obtain the textural features \( F_T \) from LR input view using several residual layers. Then, we use three Reference Extraction Module (REM) that do
Adaptation Selection Module. As shown in Figure 4(d), ASM first adds residues, calculated from \( F_{M}^{ref} \) and \( F_{N}^{ref} \), to \( F_{M}^{asm} \) and \( F_{N}^{asm} \). Then, utilizing numerous residual layers, ASM adapts reference features to \( F_{rs}^{asm} \), reducing the feature difference and artifact features in reference features. Finally, it takes the adapted reference feature \( (F_{M}^{a}, F_{N}^{a}) \) and \( F_{rs}^{asm} \) as the input of another RSM to extract high-frequency feature details. We define the process in ASM as follows:

\[
F_{M}^{a} = \text{Resnet} \left( \text{Conv}_{2} \left( F_{M}^{ref} + \text{Conv}_{1} \left( F_{M}^{ref}, F_{rs}^{asm} \right) \right) \right)
\]

\[
F_{N}^{a} = \text{Resnet} \left( \text{Conv}_{2} \left( F_{N}^{ref} + \text{Conv}_{1} \left( F_{N}^{ref}, F_{rs}^{asm} \right) \right) \right)
\]

\[
F_{rs}^{asm} = \text{RSM} \left( F_{M}^{a}, F_{N}^{a}, F_{rs}^{asm} \right)
\]

As a result, dynamic extraction via RSM and ASM makes effective use of the HF textural details contained in the Refs.

### 3.4 Loss Functions

Similar to most SR models, we adopt the reconstruction loss \( L_{rec} \), adversarial loss \( L_{adv} \), and perceptual loss \( L_{per} \) in our training stage. The \( L_{adv} \) and \( L_{per} \) are used to improve the visual quality, and we define the overall loss as follows:

\[
L_{\text{overall}} = L_{rec} + \lambda_{adv} L_{adv} + \lambda_{per} L_{per}
\]

Specifically, we employ \( l_{1} \)-norm for reconstruction loss and adopt the WGAN-GP [9] structure to obtain adversarial loss. For perceptual loss [14], it is calculated on the VGG19 [35] features of conv3-2, conv4-2, and conv5-2.

### 4 EXPERIMENTS

#### 4.1 Datasets and Training Details

Datasets. As the multi-view image super-resolution problem focuses on multi-view settings, we compare our MVSNet to the state of the art on three challenging multi-view datasets: Tanks and

---

Figure 4: Overview of the proposed Dynamic High-Frequency Search (DHFS) network, which contains three Reference Extraction Modules (REM). Each REM includes a Raw Selection Module (RSM) and an Adaptation Selection Module (ASM) to exploit the high-frequency textural details in reference images dynamically.
| Model  | Train | Playground | M60 | Truck | Total |
|--------|-------|------------|-----|-------|-------|
| EDSR [21] | 28.05 / 0.8294 / 0.226 | 28.53 / 0.8070 / 0.318 | 31.38 / 0.9072 / 0.165 | 28.19 / 0.8535 / 0.209 | 28.99 / 0.8493 / 0.230 |
| RCAN [51] | 28.56 / 0.8351 / 0.228 | 29.68 / 0.8117 / 0.320 | 32.41 / 0.9126 / 0.163 | 28.93 / 0.8583 / 0.213 | 29.89 / 0.8544 / 0.231 |
| HAN [27] | 28.43 / 0.8321 / 0.222 | 29.59 / 0.8098 / 0.315 | 32.25 / 0.9103 / 0.167 | 28.77 / 0.8551 / 0.207 | 29.76 / 0.8518 / 0.228 |
| ESRGAN [41] | 25.53 / 0.7512 / 0.109 | 25.48 / 0.6810 / 0.151 | 29.52 / 0.8666 / 0.091 | 26.00 / 0.7891 / 0.097 | 26.63 / 0.7729 / 0.112 |
| SPSR [25] | 25.68 / 0.7495 / 0.106 | 26.61 / 0.6900 / 0.133 | 29.85 / 0.8631 / 0.087 | 26.49 / 0.7920 / 0.090 | 27.16 / 0.7736 / 0.104 |
| VSR | | | | |
| RBPN [10] | 28.35 / 0.8314 / 0.236 | 29.66 / 0.8131 / 0.521 | 32.25 / 0.9117 / 0.171 | 28.77 / 0.8563 / 0.217 | 29.75 / 0.8531 / 0.236 |
| RSDN [12] | 27.68 / 0.8148 / 0.240 | 29.05 / 0.7954 / 0.314 | 31.05 / 0.8886 / 0.189 | 28.23 / 0.8402 / 0.218 | 29.00 / 0.8347 / 0.240 |
| MuCAN [18] | 28.46 / 0.8363 / 0.215 | 29.66 / 0.8148 / 0.296 | 32.56 / 0.9175 / 0.150 | 28.90 / 0.8607 / 0.200 | 29.90 / 0.8573 / 0.215 |
| IconVSR [4] | 28.31 / 0.8319 / 0.227 | 29.69 / 0.8172 / 0.310 | 32.36 / 0.9157 / 0.151 | 28.62 / 0.8538 / 0.217 | 29.75 / 0.8547 / 0.227 |

| Model  | BlendedMVS | GTAV |
|--------|------------|-----|
| EDSR [21] | 27.80 / 0.7198 / 0.393 | 30.93 / 0.8478 / 0.211 |
| RCAN [51] | 27.90 / 0.7272 / 0.393 | 31.23 / 0.8530 / 0.212 |
| HAN [27] | 27.86 / 0.7211 / 0.396 | 31.13 / 0.8515 / 0.206 |
| ESRGAN [41] | 25.50 / 0.6315 / 0.173 | 28.26 / 0.7476 / 0.105 |
| SPSR [25] | 25.72 / 0.6406 / 0.168 | 28.37 / 0.7739 / 0.102 |
| VSR | | |
| RBPN [10] | 27.79 / 0.7179 / 0.408 | 31.27 / 0.8575 / 0.218 |
| RSDN [12] | 26.38 / 0.6412 / 0.480 | 30.26 / 0.8365 / 0.223 |
| MuCAN [18] | 27.72 / 0.7162 / 0.400 | 31.24 / 0.8585 / 0.196 |
| IconVSR [4] | 27.53 / 0.7089 / 0.408 | 31.41 / 0.8635 / 0.205 |
| Ref-SR | | |
| TTSR [45] | 26.09 / 0.6567 / 0.201 | 28.80 / 0.7914 / 0.121 |
| TTSR-rec [45] | 27.68 / 0.7219 / 0.396 | 30.79 / 0.8485 / 0.216 |
| MASA [23] | 25.84 / 0.6468 / 0.220 | 28.05 / 0.7690 / 0.131 |
| MASA-rec [23] | 27.82 / 0.7335 / 0.375 | 31.12 / 0.8631 / 0.196 |
| C²-Matching [13] | 27.24 / 0.7039 / 0.281 | 30.57 / 0.8396 / 0.153 |
| C²-Matching-rec [13] | 27.99 / 0.7377 / 0.370 | 31.21 / 0.8628 / 0.194 |
| MVSNet | 26.90 / 0.7080 / 0.182 | 30.27 / 0.8462 / 0.085 |
| MVSNet-rec | 28.03 / 0.7457 / 0.360 | 31.60 / 0.8775 / 0.179 |

We compare our MVSRnet with recent SISR, VSR, and Ref-SR methods are grouped separately. ESRGAN, SPSR, TTSR, MASA, C²-Matching, and MVSNet are GAN-based methods and the best results are in bold.
model with overall loss (Eq.(7)) and solely reconstruction loss and refer to them as MVSRnet and MVSRnet-rec, respectively.

Quantitative Comparison. Table 1 presents the quantitative results of the Tanks and Temples dataset. In Table 2, we also verify the generalization capacity of our MVSRnet using BlendedMVS and GTAV, and our approach continues to perform the best. The PSNR of GAN-based $C^2$-Matching in Table 1 and Table 2 is higher than our MVSRnet, while MVSRnet reduces the LPIPS [49] error by around 35%-46% when compared to $C^2$-Matching. LPIPS is validated to human perception and is commonly used as a metric for evaluating GAN-based models. While our LPIPS is comparable to that of SPSR in BlendedMVS, our PSNR and SSIM are 1.18dB and 0.0674 higher, respectively. In summary, our proposed method outperforms the state of the art by a significant margin.

Qualitative Evaluation. The visual results shown in Figure 5 reveal that our method is capable of producing finer details, and they lead to two conclusions. (a) Our reference image generated from all LR multi-view images is more beneficial for SR than the image self-reference (SISR) and adjacent views (VSR). For example, the reference image of the fourth example contains more realistic horse details than the bicubic input view. Thus, using our reference image, MVSRnet generates more fine textural details of the horse when compared to SISR methods (ESRGAN and SPSR) and VSR methods (MuCAN and IconVSR). (b) Our DHFS network, including RSM and ASM, shows its effectiveness and robustness in identifying and extracting useful features from the reference images. Specifically, because our synthesized reference image is used as an input, RefSR methods such as TTSR, MASA, and $C^2$-Matching can produce textural results. They are, however, not robust to artifacts in the synthesized reference. For example, in Figure 5, the reference image of the third example contains blur artifacts, resulting in more blur areas in the $C^2$-Matching image than in our MVSRnet image.

4.3 Ablation Study

Multi-View Reference Image (MVR) and Near-View Reference Image (NVR). We analyze the MVR and NVR that have complementary effects on SR tasks. The ablation Model 1 does not input both MVR and NVR, and Model 2 and Model 3 do not input NVR.
and MVR, respectively. As illustrated in Figure 6(a), MVR contains high-frequency details (sharp number ‘713’) as well as noticeable artifacts (the incorrect structure of letter ‘PAC’). In contrast, NVR has fewer artifacts and details of ‘PAC’ and ‘713’ than MVR, but there are still more high-frequency details than the bicubic input view. We discover that MVSRnet recovers the exact number ‘713’ than Model 3 and produces fewer ‘PAC’ artifacts than Model 2. Additionally, we conduct quantitative ablation study, and Table 3 reveals that MVSRnet has superior PSNR performance than Model 2 and Model 3 by 0.38dB and 0.45 dB, respectively. These experiments fully demonstrate the complementary effects of MVR and NVR.

In addition, we search for the optimal number $V$ of MVR and NVR. The LR details and artifacts areas grow in relation to the number $V$ of reference images, making it challenging for the network to identify and extract relevant high frequencies details. We perform ablation studies with $V$ to 1, 2, 4, 6, and 8, and we achieve the best performance when we set $V$ to 6, as shown in Table 4. We also conduct an ablation study on the hyper-parameter of patch size $p_s$ as 4/8/16/32/64 in GARS module, and achieve the best performance when we set $p_s$ to 16, as shown in Table 4. The reason is that when the patch size is too small, there are numerous selection mistakes and block-boundary artifacts in the synthesized reference image. When the block size is too large, the selected patches are blurry and lack a significant amount of high-frequency texture details.

Raw Selection Module (RSM) and Adaptation Selection Module (ASM). The ablation Model 4 does not include both RSM and ASM, and Model 5 and Model 6 do not include ASM and RSM, respectively. Figure 6(b) shows that selected areas (denoted as white) in the selection map of ‘ASM in MVSRnet’ are more than ‘RSM in Model 5’ and ‘ASM in Model 6’, indicating that more MVR HR feature details are selected for SR. This is because a small difference between features can help the network identify and extract useful features. As mentioned in Section 3.3, ASM implicitly decreases the difference between reference features and input view features and reduces artifact features in reference features, making it easier for the network to extract useful features in MVR. RSM, on the other hand, uses a residual process to adapt input view features to reference features, as indicated in Eq.(5). The different ways ASM and RSM incorporate features makes feature extraction more sufficient. Thus, combining RSM and ASM significantly improves the power of DHFS to extract and identify HF details when compared to using only RSM or ASM. We also verify them using the quantitative evaluation displayed in Table 3. MVSRnet improves the PSNR value by 0.12dB and 0.11dB, respectively, when compared to Model 5 and 6, demonstrating the usefulness of using both RSM and ASM.

5 CONCLUSION

We propose the MVSRnet for multi-view image super-resolution that aims to super-resolve each view by utilizing all views as references. Our Geometry-Aware Reference Synthesis module, in particular, can synthesize reference images from all LR multi-view images by explicitly using geometry information. The reference image contains high-frequency details and is aligned with the input view. Then we implement a Dynamic High-Frequency Search network which includes the Raw Selection Module and Adaptation Selection Module, to dynamically select useful features in the synthesized reference features for super-resolution. We have achieved best experimental results compared with state-of-the-art methods and this work enriches the Reference-Based SR research.
REFERENCES

[1] Leif Bég and Judy Vance. 2017. Industry use of virtual reality in product design and manufacturing: a survey. Virtual Reality 21 (03 2017). https://doi.org/10.1007/s10055-016-0209-0

[2] Tom E. Bishop, Sara Zanetti, and Paolo Favaro. 2009. Light field superresolution. In 2009 IEEE International Conference on Computational Photography (ICCP). 1–9. https://doi.org/10.1109/ICCPHOT.2009.5559810

[3] KP Brooks. 1999. What’s real about virtual reality? IEEE Computer Graphics and Applications 19, 6 (1999), 16–27. https://doi.org/10.1109/38.799723

[4] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2021. BasicVR: The Search for Essential Components in Video Super-Resolution and Beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 4947–4956.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2016. Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 2 (2016), 295–307. https://doi.org/10.1109/TPAMI.2015.2439281

[7] Bastian Goldbecke and Daniel Cremers. 2009. Superresolution texture maps for multiview reconstruction. In 2009 IEEE 12th International Conference on Computer Vision. 1677–1684. https://doi.org/10.1109/ICCV.2009.5459378

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In NIPS. 2672–2680.

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Ian J. Goodfellow. 2017. Improved Training of Wasserstein GANs. In NIPS. 5769–5779. http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf

[10] Mohammad Haris, Gregor Schahnavazich, and Norimichi Uka. 2019. Recurrent Back-Projection Network for Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[11] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. 2018. DeepMSV: Learning Multi-View Stereopsis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[12] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. 2021. Robust Reference-Based Super-Resolution via a Holistic Attention Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[13] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 7900–7908.

[14] Tom E. Bishop, Sara Zanetti, and Paolo Favaro. 2009. Light field superresolution. In 2009 IEEE International Conference on Computational Photography (ICCP). 1–9. https://doi.org/10.1109/ICCPHOT.2009.5559810

[16] Alok Choudhary, Pratul S. Progress. In Proceedings of the European Conference on Computer Vision (ECCV). 12216–12225.

[17] Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-From-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[18] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Patchwise View Selection for Unstructured Multi-View Stereo. In European Conference on Computer Vision – ECCV 2016. Cham, 501–518.

[19] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. 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 (CVPR).

[20] Karen Simonyan, Andrew Zisserman, Karen Simonyan, and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations.

[21] Aljoscha Smolic, Karsten Mueller, Philipp Merkle, Christoph Fehn, Peter Kauff, Peter Eisert, and Thomas Wiegand. 2006. 3D Video and Free Viewpoint Video - Technologies, Applications and MPEG Standards. In 2006 IEEE International Conference on Multimedia and Expo. 2161–2164. https://doi.org/10.1109/ICME.2006.262683

[22] Yuyi Sun, Bing Cheng, Bo Yan, and Shih Zhou. 2021. Space-Angle Super-Resolution for Multi-View Images. In MM ’21: ACM Multimedia Conference, Virtual Event, China, October 20 – 24, 2021. ACM, 750–759. https://doi.org/10.1145/3474085.3475244

[23] Yugui Sun, Jingyu Yang, Jing Lu, and Zhe Zhang. 2021. Learning Back-Projection Network for Video Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[24] Yingqian Wang, Longguang Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. 2019. Learning Parallax Attention for Stereo Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[25] Xintao Wang, Ke Yu, Shixiang Gu, Tiande Wang, and Wenzhe Shi. 2017. Photo-Realistic Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2103–2112.

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Computer Vision – ECCV 2016. Cham, 694–711.

[27] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2016. Photo-Realistic Single Image Super-Resolution with a Generative Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[28] Wilfredo Munding, Luiz Alexandre, and Konrad Schindler. 2019. Learned Multi-View Texture Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[29] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Ian J. Goodfellow. 2017. Improved Training of Wasserstein GANs. In NIPS. 5769–5779. http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf

[30] Ranganath Ananthakrishnan, Michael F. Cohen, and Chen Change Loy. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

[31] Shao-Ping Lu, Sen-Mao Li, Rong Wang, Gauthier Lafraiture, Ming-Ming Cheng, and Adrian Munteanu. 2021. Low-Rank Constrained Super-Resolution for Mixed-Resolution Multiview Videos. IEEE Transactions on Image Processing 30 (2021). https://doi.org/10.1109/TIP.2020.3042604

[32] Shao-Ping Lu, Sen-Mao Li, Rong Wang, Gauthier Lafraiture, Ming-Ming Cheng, and Adrian Munteanu. 2021. Low-Rank Constrained Super-Resolution for Mixed-Resolution Multiview Videos. IEEE Transactions on Image Processing 30 (2021). https://doi.org/10.1109/TIP.2020.3042604

[33] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. 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 (CVPR).

[34] Karen Simonyan, Andrew Zisserman, Karen Simonyan, and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations.

[35] Karen Simonyan, Andrew Zisserman, Karen Simonyan, and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations.

[36] Karen Simonyan, Andrew Zisserman, Karen Simonyan, and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations.

[37] Karen Simonyan, Andrew Zisserman, Karen Simonyan, and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations.
We use the WGAN-GP [9] structure to obtain where $I$ denotes the output of our MVSRnet and the $T$ ground truth. As a result, we utilize the features of conv3-2, conv4-2, and conv5-2 to calculate the perceptual loss.

$$L_{rec} = \left\| \frac{I_T^{hr}}{I_T^{pl}} - \frac{I_T^{pl}}{I_T^{pl}} \right\|_1,$$

where $I_T^{hr}$ and $I_T^{pl}$ represent the output of our MVSRnet and the $T$ ground truth, respectively.

### Adversarial Loss
We use the WGAN-GP [9] structure to obtain adversarial loss [8]. WGAN-GP replaces the weight clipping with a gradient norm penalty to achieve stable training and generate high-quality samples. We define adversarial loss as follows:

$$L_D = \mathbb{E}_{I_T^{hr}} [D(I_T^{hr})] - \mathbb{E}_{I_T^{pl}} [D(I_T^{pl})] + \lambda \mathbb{E}_{I_T^{hr}} \left( \left\| \nabla_D(D) \right\|_2 - 1 \right)^2,$$

where $I_T^{hr}$ is the random sample of our network output $I_T^{hr}$ and $I_T^{pl}$ through combination, $(I \leftarrow \epsilon I_T^{pl} + (1 - \epsilon) I_T^{hr})$, $\epsilon \sim U[0, 1])$.

### Perceptual Loss
We utilize perceptual loss [14] to generate the photo-realistic image, which we define as follows:

$$L_{per} = \left\| \phi_i(I_T^{hr}) - \phi_i(I_T^{pl}) \right\|_1,$$

where $\phi_i$ denotes the $i$-th layer feature map of VGG19 [35], and we use the features of conv3-2, conv4-2, and conv5-2 to calculate the perceptual loss.

**Table 5: Comparison to multi-view super-resolution method.**

| Model         | Tanks and Temple | BlendedMVS |
|---------------|------------------|------------|
| SASRnet [37]  | 28.99 / 0.8493 / 0.230 | 27.80 / 0.7198 / 0.393 |
| MVSRnet-rec   | 30.26 / 0.8748 / 0.191 | 28.03 / 0.7457 / 0.360 |
Table 6: The network structure of our Dynamic High-Frequency Search network. MVR, NVR, and BiC denote the multi-view reference image, near-view reference image, and the bicubic of the LR input view, respectively. We present the convolution layer as Conv(input channels, output channels, kernel size, stride, padding). RB × N is a residual block with N residual layers, and we adopt the residual layer with channel attention described in [51]. In addition, we use Pixelshuffle to upsample, and use ReLU in RB and LeakyReLU other places.

Table 7: Comparison of the model size. light denotes the light version of our proposed network. Red and blue indicate first and second performance, respectively. We include the parameter of discriminator in the total parameters of the GAN-based model.