Task Decoupled Framework for Reference-based Super-Resolution

Yixuan Huang1*, Xiaoyun Zhang1*, Yu Fu1, Siheng Chen1,2, Ya Zhang1,2, Yanfeng Wang1,2†, Dazhi He1
1Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, 2Shanghai AI Laboratory
{huangyixuan, xiaoyun.zhang, fyuu11, sihengc, ya.zhang, wangyanfeng, hedazhi}@sjtu.edu.cn

C2-Matching Ours
(a) Reference-underuse issue
reference & input
input & reference
C2-Matching Ours
(b) Reference-misuse issue
RRDB

Figure 1. Current state-of-the-art RefSR methods suffer from (a) reference-underuse issue and (b) reference-misuse issue due to the improper coupled framework. Specifically, for the reference-underuse issue, RefSR methods transfer the detail textures from reference insufficiently, even when the reference has highly similar content to input. As for the reference-misuse issue, current methods introduce blur and artifacts, which is even worse than SISR methods, when the reference has no relevant content. Our novel task decoupled framework for RefSR mitigates these issues.

1. Introduction

Reference-based super-resolution(RefSR) has achieved impressive progress on the recovery of high-frequency details thanks to an additional reference high-resolution(HR) image input. Although the superiority compared with Single-Image Super-Resolution(SISR), existing RefSR methods easily result in the reference-underuse issue and the reference-misuse as shown in Fig.1. In this work, we deeply investigate the cause of the two issues and further propose a novel framework to mitigate them. Our studies find that the issues are mostly due to the improper coupled framework design of current methods. Those methods conduct the super-resolution task of the input low-resolution(LR) image and the texture transfer task from the reference image together in one module, easily introducing the interference between LR and reference features. Inspired by this finding, we propose a novel framework, which decouples the two tasks of RefSR, eliminating the interference between the LR image and the reference image. The super-resolution task upsamples the LR image leveraging only the LR image itself. The texture transfer task extracts and transfers abundant textures from the reference image to the coarsely upsampled result of the super-resolution task. Extensive experiments demonstrate clear improvements in both quantitative and qualitative evaluations over state-of-the-art methods.

Abstract

Reference-based super-resolution(RefSR) has achieved impressive progress on the recovery of high-frequency details thanks to an additional reference high-resolution(HR) image input. Although the superiority compared with Single-Image Super-Resolution(SISR), existing RefSR methods easily result in the reference-underuse issue and the reference-misuse as shown in Fig.1. In this work, we deeply investigate the cause of the two issues and further propose a novel framework to mitigate them. Our studies find that the issues are mostly due to the improper coupled framework design of current methods. Those methods conduct the super-resolution task of the input low-resolution(LR) image and the texture transfer task from the reference image together in one module, easily introducing the interference between LR and reference features. Inspired by this finding, we propose a novel framework, which decouples the two tasks of RefSR, eliminating the interference between the LR image and the reference image. The super-resolution task upsamples the LR image leveraging only the LR image itself. The texture transfer task extracts and transfers abundant textures from the reference image to the coarsely upsampled result of the super-resolution task. Extensive experiments demonstrate clear improvements in both quantitative and qualitative evaluations over state-of-the-art methods.

1. Introduction

The goal of Single Image Super-Resolution(SISR) [10, 16, 26, 38] is to recover the high-resolution(HR) details of the image from its low-resolution(LR) counterpart. It has been widely used in image enhancement, video surveillance and remote sensing imaging. The ill-posed essence of SISR makes it easily resulting in visual artifacts as the scale factor is large(×4). To tackle the problem, Reference-based Super-Resolution(RefSR) has attracted much attention recently. Compared with SISR that only uses one single LR input, RefSR [13, 22, 27, 36, 40, 41] super-resolves the LR image with an additional HR reference image, which can provide abundant real textures for the super-resolved image. RefSR can be applied to plenty of scenarios, such as dual-cameras [31] and video streaming [25].

Current RefSR methods mostly transfer the textures based on the alignment between reference and LR images, such as spatial alignment [27, 41] and patch matching [13, 22, 36, 40]. After alignment, RefSR methods concatenate the LR and aligned reference features as the input and transfer the detailed textures to the output through some well-designed architectures, such as attention mechanism [36] and spatial adaptation module [22].

Although superior of the performance compared with SISR, current methods are easily plagued by the reference-underuse issue(i.e., reference textures are transferred insuf-
The two issues are mainly due to the inadequate framework design of current RefSR methods, which couple the super-resolution task of the LR image and the texture transfer task from the reference image in one single module. As shown in Fig.2(a), most RefSR methods design a single module to execute the super-resolution task and the texture transfer task, using the concatenation of LR features and aligned reference features as the input. Processing the LR and reference images simultaneously with convolution operation will introduce disturbance to each other. Specifically, on the one hand, if the reference is highly similar to the input LR image, the model should concentrate more on transferring the aligned reference features to the output. However, the texture transfer process can be interfered by the input LR image and result in the reference-underuse issue, in which the model outputs the average of the aligned reference image and the LR image, as shown in Fig.1(a). On the other hand, as the reference content is irrelevant to the LR input image, the incongruous reference will disturb the super-resolution task of the LR image, when processing the LR and reference images at the same time, which results in the reference-misuse issue as shown in Fig.1(b).

To handle the two issues, as shown in Fig.2(b), we propose a novel framework for RefSR based on decoupling the super-resolution task and the texture transfer task of RefSR. The decoupled task of the super-resolution aims to super-resolve the LR image coarsely without the disturbance of the reference image. The texture transfer task is to extract the textures from the reference image and further transfer the detailed textures to the result of the super-resolution task. In addition, in order to extract and transfer abundant textures stably from the reference, we propose the texture extraction module and the adaptive texture transfer module.

In this work, we propose a novel framework for RefSR based on decoupling the super-resolution task and the texture transfer task. The super-resolution task upsamples the LR image independently, and the texture transfer task extracts and transfers textures from the reference image. Experimental results demonstrate that our proposed method significantly outperforms the existing RefSR methods both quantitatively and qualitatively.

To summarize, our contributions include:

- We deeply investigate the cause of the reference-underuse issue and the reference-misuse issue for RefSR. The two issues are mainly due to the improper coupled framework of current RefSR methods.
- To tackle the two issues, we propose a novel framework for RefSR based on decoupling the super-resolution task and the texture transfer task. The super-resolution task upsamples the LR image independently, and the texture transfer task extracts and transfers textures from the reference image.
- Experimental results demonstrate that our proposed method significantly outperforms the existing RefSR methods both quantitatively and qualitatively.

2. Related Work

Single Image Super-Resolution. Single Image Super-Resolution (SISR) aims to recover the HR image with just a single LR image as input. In recent years, deep learning based methods [9, 10, 15–17, 20, 21, 26, 29, 38, 39] have achieved excellent performance on the SISR task. However, most of them obtain over-smooth images because of the mean square (MSE) loss. To improve the visual quality, perceptual loss [14] and adversarial loss [18, 33, 37] are proposed. Although perceptual loss and adversarial loss improve the visual quality, the methods can easily result in hallucinations and artifacts.

Reference-based Image Super-Resolution. The goal of Reference-based Image Super-Resolution (RefSR) [22,
are proposed. SRNTT [40] operated the patch matching on introduced into RefSR in TTSR [36], and the hard attention into LR images. Transformer architecture has been multi-scale features and then transferred the texture implicitly, between LR and reference images. To match the images example to solve the potential large disparity in distributions properly. MASA [22] proposed a spatial adaptation module and soft attention can extract reference features more efficiently. The alignment between the LR and reference images. Cross-ods transfer the textures from the reference image through image retrieval methods [6]. The RefSR methods transfer the textures from the reference image through the alignment between the LR and reference images. CrossNet [41] aligned the reference and LR images by the flow estimation between two images. SSEN [27] further used deformable convolution to extract the reference features instead. However, the fatal limitation of the above spatial alignment methods is that they all lack the ability to build the long-distance correlation of two different images. Therefore, patch matching based methods [13, 22, 36, 40] are proposed. SRNTT [40] operated the patch matching on multi-scale features and then transferred the texture information into LR images. Transformer architecture has been introduced into RefSR in TTSR [36], and the hard attention and soft attention can extract reference features more properly. MASA [22] proposed a spatial adaptation module to solve the potential large disparity in distributions between LR and reference images. To match the images explicitly, $C^2$-Matching [13] introduced contrastive learning and knowledge distillation into patch matching steps, overcoming the transformation gap between LR and Ref images.

However, most current RefSR methods are designed to upsample the LR image and transfer textures from the reference image through a single module simultaneously, which easily results in the reference-underuse issue and the reference-misuse issue as shown in Fig.1. In contrast, our framework decouples the super-resolution task and the texture transfer task of RefSR, excluding the interference between the LR image and the reference image, thus solving the two issues to a large extent.

3. Analysis of the RefSR framework

3.1. Coupled framework for RefSR

Given a LR image $I_{LR}$ as the input, Reference-based Super-Resolution(RefSR) aims to produce the high-resolution image $I_{RefSR}$ with the guidance of an additional HR reference image $I_{Ref}$. As shown in Fig.2(a), the framework of most existing methods [13, 22, 36, 40] can be summarized as the coupled framework. The coupled framework firstly makes the alignment between $I_{LR}$ and $I_{Ref}$, obtaining the LR features $F_{LR}$ and aligned reference features $F_{Ref}^{align}$, and then reconstruct the HR image:

$$I_{RefSR} = G\left(F_{LR}, F_{Ref}^{align}\right),$$

where $G$ denotes the RefSR network, and $[.]$ denotes the concatenation operation.

3.2. Analysis of the reference-underuse issue and the reference-misuse issue

The coupled framework easily results in the reference-underuse issue and the reference-misuse issue, due to the interference between the LR image and the reference image. Therefore, we’re curious about whether the two issues can be tackled by separating the process of the LR and reference images. To further investigate the two issues, we retrain the $C^2$-Matching [13] model, which is the state-of-the-art method recently, for the following settings:

- **Model-Ref**: Only the aligned reference image is input to reconstruct the final HR image without the LR image. So the model only processes the texture transfer task from the reference image.
- **Model-LR**: Only the LR image is input to reconstruct the SR image without reference image. So the model turns into a SISR model with single image input.
- **Model-Both**: Both the LR image and aligned reference image are input to reconstruct the final HR image.

**Reference-underuse issue.** As shown in Fig.3(a), when similar reference images are given, the results of Model-Ref, which only use the reference image, have more similar detailed textures to the ground-truth images than Model-Both, which input both the LR image and reference image. The experiments demonstrate that the model can transfer the textures more sufficiently from the reference image.
when removing the interference of the input LR image. To a large extent, the removal of the LR image mitigates the reference-underuse issue.

Reference-misuse issue. As shown in Fig.3(b), when irrelevant reference images are given, Model-LR eliminates the blur and artifacts of the Model-Both results. Therefore, by removing the influence of the incongruous reference content, the RefSR model can upsample the LR image better. The removal of the irrelevant reference mitigates the reference-misuse issue.

In conclusion, as we separate the super-resolution and the texture transfer process of RefSR, the reference-underuse issue and the reference-misuse issue are both mitigated by a large margin. Therefore, inspired by this finding, we design a novel framework based on decoupling the super-resolution task and the texture transfer task for RefSR.

4. Our Approach

4.1. Overview

In this work, to tackle the reference-underuse issue and the reference-misuse issue, we propose a novel framework based on decoupling the super-resolution task and texture transfer task of RefSR. We set two different missions for the input and reference images. Specifically, The input LR image is expected to provide the structure and content information for the final output, while the reference HR image should supply additional detailed textures. Consequently, we utilize the input LR image $I_{LR}$ and the reference image $I_{Ref}$ separately and integrate them subsequently. The overview of our proposed framework is shown in Fig.4, which consists of two major parts:

For the Super-Resolution task, we upsample the input LR image by itself without any interference from the reference image, obtaining an initial high-resolution image $I_{SISR}$ and its features $F_{SISR}$:

$$F_{SISR} = F_{SR}(I_{LR}),$$  \hfill (2)

where $F_{SR}$ denotes the SISR model, in which we use RRDB [33] as the basic block. As for Texture Transfer task, we extract and transfer the fine texture feature from the reference image to the SISR feature $F_{SISR}$. We first extract the reference texture feature $F_{Tex}$ through the texture extraction module:

$$F_{Tex} = F_{TE}(I_{Ref}, I_{SISR}),$$  \hfill (3)

where $F_{TE}$ denotes the texture extraction module. It’s worth noting that $I_{SISR}$ is only used for calculating the spatial correlation between the reference image and the input LR image. Then the textures $F_{Tex}$ are transferred to $F_{SISR}$ by the adaptive texture transfer module according to the similarity between the reference image and the LR input image.
image, obtaining the final RefSR output $I_{RefSR}$:

$$I_{RefSR} = F_{ATT} (F_{Tex}, F_{SISR}),$$  \hspace{1cm} (4)

where $F_{ATT}$ denotes the adaptive texture transfer module.

### 4.2. Texture Extraction with Alignment

The key task of texture extraction is how to process the spatial alignment of the reference image $I_{Ref} \in \mathbb{R}^{H \times W \times 3}$ with the SISR image $I_{SISR} \in \mathbb{R}^{H \times W \times 3}$. In the alignment module, we use two feature extractors to map $I_{SISR}$ and $I_{Ref}$ into the same feature space. The architecture of the two feature extractors are shared. After that, we use the operation of patch matching [40] in the feature space to calculate the spatial correlation map $M \in \mathbb{R}^{H \times W \times W'}$ between $I_{SISR}$ and $I_{Ref}$. Then we calculate the index map $P$ and confidence map $C$ from the correlation matrix $M$:

$$P_i = \arg \max_j M_{i,j}, \quad C_i = \max_j M_{i,j}. \hspace{1cm} (5)$$

The index map $P$ can be regarded as flow information between $I_{SISR}$ and $I_{Ref}$. Inspired by the video super-resolution task [3, 5, 30, 32] and the video frame interpolation task [1, 2, 7, 19], we use two different methods for alignment. The first method is flow-based warping. Regarded $P$ as the flow map, we process the backwarp operation on the $F_{Ref}$ to get the aligned features. The second method is deformable [8] alignment. Previous studies [4] have shown that deformable alignment has significant improvements over flow-based alignment because of the offset diversity. However, the training of deformable alignment is unstable, which can cause offset overflow [4] thus limiting the performance. To overcome the instability of deformable alignment, we combine deformable alignment and flow-based alignment. The combination of two methods utilizes the offset diversity in deformable alignment and the stability of flow-based alignment.

**Flow-based Alignment.** Given the flow information $P$, we directly warp the feature maps $F_{Ref}$ to get the $F_{Tex}$ following [22, 36, 40]:

$$F_{Tex} = W(F_{Ref}, P), \hspace{1cm}(6)$$

Where $W$ denotes the operation of spatial warping.

**Deformable Alignment.** The difficulty to train the deformable convolution (DCN), illuminated by the usage of deformable convolution [5, 32] in the video super-resolution task, we employ optical flow to guide deformable alignment like [5]. The result of spatial warping $F_{Tex}$ is utilized to estimate the offsets $o$ and modulation masks $m$ for DCN:

$$o = P + \mathcal{E}_o (F_{Tex}, F_{SISR}), \hspace{1cm} (7)$$

$$m = \sigma(\mathcal{E}_m (F_{Tex}, F_{SISR})), \hspace{1cm} (8)$$

where $\mathcal{E}_o$ and $\mathcal{E}_m$ denote the stacks of convolution layers, and $\sigma$ denotes the sigmoid function. Then we employ DCN on $F_{Ref}$ to get the aligned feature $F_{Tex}$:

$$F_{Tex} = D(F_{Ref}, o, m), \hspace{1cm} (9)$$

where $D$ denotes the deformable convolution. We fuse the $F_{Tex}$ and $F_{SISR}$ through a convolution layer, obtaining the texture features $F_{Tex}$. Through a decoder, we obtain the texture image $I_{Tex}$. To extract the high-frequency detailed textures, the $I_{Tex}$ is supervised by the residual of the HR image $I_{HR}$ and the SISR image $I_{SISR}$.

### 4.3. Adaptive Texture Transfer

To accurately combine the content of the SISR feature $F_{SISR}$ and the textures of the reference feature $F_{Tex}$, we design the adaptive texture transfer module to blend $F_{Tex}$ and $F_{SISR}$ according to the similarity between the two images.

To adaptively select the texture from $F_{Tex}$, preventing the misaligned texture from deteriorating the output, we use the confidence map $C$ calculated in Eq. (5) to suppress the weight of the misaligned texture region:

$$F_{RefSR} = \text{Conv}(F_{Tex}) \cdot C + F_{SISR}, \hspace{1cm} (10)$$

where $\text{Conv}$ denotes a convolution layer. And then after eight residual blocks for restoration, we get the final output $I_{RefSR}$. More details of confidence map $C$ are discussed in the supplementary material.

### 4.4. Implementation Details

The overview network of the proposed framework is trained by two steps: 1) training of the singe image super-resolution network. 2) training of the texture extraction and transfer network.

**Training of SISR Network.** To ensure the content information of SISR results, we only use reconstruction loss, which is calculated by $L_1$ loss, to train the SISR network. After training, we fix the SISR network.

**Training of Texture Transfer Network.** The super-resolution image $I_{SISR}$ obtained from the SISR network is used for the feature alignment step. The texture transfer network has two outputs for supervision, the result of texture extraction $I_{Tex}$ and the final result $I_{RefSR}$. To let the network extract the high-frequency detailed texture information, we adopt the residual of the ground-truth image $I_{HR}$ and $I_{SISR}$ as supervision (More discussions of $L_{Tex}$ are in the supplementary material.):

$$L_{Tex} = ||I_{HR} - (I_{SISR} + I_{Tex})||_1, \hspace{1cm} (11)$$

As for the final output, we adopt reconstruction loss $L_{rec}$, perceptual loss $L_{per}$ [14] and adversarial loss $L_{adv}$ [11]. The weight for $L_{rec}, L_{per}$ and $L_{adv}$ are $1, 10^{-2}$ and $10^{-4}$, respectively. The initial training rate is set as $10^{-4}$. 

5935
Table 1. Quantitative Comparisons with the state-of-the-art methods. We use PSNR/SSIM metrics for evaluation. The best and second-best performances are marked by red and blue colors, respectively. The RefSR models with ‘-rec’ suffix are trained only by the reconstruction loss. Our proposed method outperforms significantly against current SISR and RefSR state-of-the-art methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>CUFED5</th>
<th>Sun80</th>
<th>Urban100</th>
<th>Manga109</th>
<th>WR-SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SISR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRCNN [10]</td>
<td>25.33/0.745</td>
<td>28.26/0.781</td>
<td>24.41/0.738</td>
<td>27.12/0.850</td>
<td>27.27/0.767</td>
</tr>
<tr>
<td>EDSR [21]</td>
<td>25.93/0.777</td>
<td>28.52/0.792</td>
<td>25.51/0.783</td>
<td>28.93/0.890</td>
<td>28.07/0.793</td>
</tr>
<tr>
<td>RRDB [33]</td>
<td>26.41/0.783</td>
<td>29.99/0.814</td>
<td>25.98/0.788</td>
<td>29.87/0.907</td>
<td>27.96/0.793</td>
</tr>
<tr>
<td>RCAN [38]</td>
<td>26.33/0.781</td>
<td>29.97/0.814</td>
<td>25.99/0.787</td>
<td>30.11/0.907</td>
<td>27.91/0.793</td>
</tr>
<tr>
<td>SRGAN [18]</td>
<td>24.40/0.702</td>
<td>26.76/0.725</td>
<td>24.07/0.729</td>
<td>25.12/0.802</td>
<td>26.21/0.728</td>
</tr>
<tr>
<td>ENet [24]</td>
<td>24.24/0.695</td>
<td>26.24/0.702</td>
<td>23.63/0.711</td>
<td>25.25/0.802</td>
<td>25.47/0.699</td>
</tr>
<tr>
<td>ESRGAN [33]</td>
<td>21.90/0.633</td>
<td>24.18/0.651</td>
<td>20.91/0.765</td>
<td>23.53/0.797</td>
<td>26.07/0.726</td>
</tr>
<tr>
<td>RankSRGAN [37]</td>
<td>22.32/0.692</td>
<td>24.20/0.702</td>
<td>22.07/0.708</td>
<td>25.25/0.802</td>
<td>25.57/0.726</td>
</tr>
</tbody>
</table>

| RefSR           |         |         |          |          |         |
| CrossNet [41]   | 25.48/0.764 | 28.52/0.793 | 25.11/0.764 | 23.36/0.741 | -       |
| SRNTT [40]      | 25.61/0.764 | 27.59/0.756 | 25.09/0.774 | 23.54/0.826 | 26.53/0.745 |
| SRNTT-rec [40]  | 26.24/0.784 | 28.54/0.793 | 25.50/0.783 | 28.95/0.885 | 27.59/0.780 |
| TTSR [36]       | 25.30/0.765 | 28.59/0.774 | 24.62/0.747 | 28.70/0.886 | 26.83/0.762 |
| TTSR-rec [36]   | 27.09/0.804 | 30.02/0.814 | 25.87/0.784 | 30.09/0.907 | 27.97/0.792 |
| MASA [22]       | 24.92/0.729 | 27.12/0.708 | 23.78/0.712 | 27.34/0.848 | -       |
| MASA-rec [22]   | 27.54/0.814 | 30.15/0.815 | 25.50/0.783 | 30.47/0.911 | 28.32/0.801 |
| C²-Matching [13] | 27.16/0.805 | 29.75/0.799 | 25.52/0.764 | 29.73/0.893 | 27.80/0.780 |
| C²-Matching-rec [13] | 28.24/0.841 | 30.18/0.817 | 26.03/0.785 | 30.47/0.911 | 28.32/0.801 |
| Ours            | 27.37/0.816 | 28.85/0.768 | 25.80/0.776 | 30.12/0.889 | 27.40/0.769 |
| Ours-rec        | 28.64/0.850 | 30.31/0.820 | 26.71/0.807 | 31.23/0.917 | 28.52/0.807 |

Table 2. Quantitative comparison at different levels on the CUFED5 testing set. ‘L1’ is the most relevant reference and ‘L4’ is the least. ‘LR’ means directly using the LR image itself as the reference. For different reference levels, our method all obtains the best performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CrossNet [41]</td>
<td>25.48/0.764</td>
<td>25.48/0.764</td>
<td>25.47/0.763</td>
<td>25.46/0.763</td>
<td>25.46/0.763</td>
</tr>
<tr>
<td>SRNTT-rec [40]</td>
<td>26.15/0.781</td>
<td>26.04/0.776</td>
<td>25.98/0.775</td>
<td>25.95/0.774</td>
<td>25.91/0.776</td>
</tr>
<tr>
<td>SSEN-rec [27]</td>
<td>26.78/0.791</td>
<td>26.52/0.783</td>
<td>26.48/0.782</td>
<td>26.42/0.781</td>
<td>-</td>
</tr>
<tr>
<td>CIMR-rec [35]</td>
<td>27.32/0.805</td>
<td>27.05/0.799</td>
<td>26.92/0.796</td>
<td>26.86/0.794</td>
<td>-</td>
</tr>
<tr>
<td>TTSR-rec [36]</td>
<td>26.99/0.800</td>
<td>26.74/0.791</td>
<td>26.64/0.788</td>
<td>26.58/0.787</td>
<td>26.43/0.782</td>
</tr>
<tr>
<td>MASA-rec [22]</td>
<td>27.35/0.814</td>
<td>26.92/0.796</td>
<td>26.82/0.793</td>
<td>26.74/0.790</td>
<td>26.59/0.784</td>
</tr>
<tr>
<td>C²-Matching-rec [13]</td>
<td>28.24/0.841</td>
<td>27.39/0.813</td>
<td>27.17/0.806</td>
<td>26.94/0.799</td>
<td>26.53/0.784</td>
</tr>
<tr>
<td>Ours-rec</td>
<td>28.64/0.850</td>
<td>27.77/0.821</td>
<td>27.46/0.815</td>
<td>27.23/0.807</td>
<td>26.83/0.794</td>
</tr>
</tbody>
</table>

5. Experiments

5.1. Datasets and Metrics

Training datasets. The CUFED5 [40] training set consists of 11,871 image pairs, and each image pair has one HR image and one reference image both with the resolution of 160 × 160. We train our model with the scale factor x4.

Testing datasets. To evaluate our method, we adopt five benchmarks: CUFED5 [40] testing set, Sun80 [28], WR-SR [13], Urban100 [12] and Manga109 [23]. CUFED5 testing set consists of 126 image pairs, and each input image has 5 reference images with different similarity levels. Sun80 has 80 images, and each image has 20 reference images. WR-SR consists of 80 images, and each image has one reference image searched through Google Image. We evaluate our method on Urban100 and Manga109 following [13, 36, 40]. Urban100 has 100 building images with highly self-similarity, so we use LR images as the reference image. For Manga109 dataset, we randomly select HR images from the dataset as the reference image.

Evaluation metrics. The RefSR results are evaluated by metrics of PSNR and SSIM on Y channel of YCrCb space.

5.2. Comparison with State-of-the-art Methods

Quantitative Comparison. We compare the proposed method with the following SISR and RefSR methods. The SISR methods include SRCNN [10], EDSR [21], RCAN [38], SRGAN [18], ENet [24], ESRGAN [33]
Figure 5. Visual comparisons on the CUFED5 testing set. We compare the proposed methods with 6 different SISR and Ref state-of-the-art methods. Among the methods, only RCAN is trained with $l_1$ loss, and others are all trained with GAN loss. Our method restores more realistic textures compared with other RefSR and SISR methods. (Zoom-in for best view)

Table 1 shows that our method outperforms existing state-of-the-art methods on all five datasets. The quantitative comparisons demonstrate that our proposed model has a significant improvement over 0.4dB compared with the state-of-the-art methods in standard CUFED5 benchmark.

Qualitative Evaluation. In Fig.5, we show the visual comparisons with state-of-the-art SISR and RefSR methods. Our method can achieve more pleasing visual quality with more detailed textures transferred from the reference HR images. As shown in the first row of the Fig.5, our method can restore the face with more realistic details. Besides, in the last row of Fig.5, our method obtains clear letters, while others generate artifacts or blurry results.

Robustness to irrelevant references. To evaluate the robustness to irrelevant references, we compare our method with other RefSR methods at different reference levels on the CUFED5 testing set. Table 2 shows the results of five relevant degrees. For different similarity levels, our method all achieves the best performance, which not only demonstrates our superiority on texture transfer, but also proves the robustness to irrelevant references of our method.

5.3. Ablation Studies

In this section, we conduct ablation studies on the two key components in our methods: (1) the task decoupled framework,(2) the texture extraction and transfer module.

Task decoupled framework. We set the model without the SISR part of our framework as the baseline, which achieves the super-resolution and texture transfer jointly through one single network. Table 3 evaluates the effectiveness of the task decoupled framework. With similar reference, we all get a gain over 0.26dB for RefSR datasets, CUFED5, Sun80 and WR-SR. Meanwhile, in Fig.6(a), the model with the decoupled framework can produce realistic images highly similar to the GT, with more textures and details transferred from the reference image. Both the quantitative evaluation and the qualitative comparison indicate that the task decoupled framework of the RefSR task can prompt the model to extract and transfer more textures from reference images. Without similar reference, as for SISR datasets, Urban100 and Manga109 datasets, Table 3 demonstrates that the decoupled framework achieves a significant boost over 1dB compared with baseline. Besides, Fig.6(b) shows that the baseline method can easily result in blur and artifacts, while our method with the decoupled framework alleviates the issue and obtains more details. The results prove that the task decoupled framework is more robust as irrelevant references are given. In conclusion, our framework effectively decouples the super-resolution task and the texture transfer task of RefSR, thus obtaining better performance.

Texture Extraction Module. Table 4 evaluates the effectiveness of flow-based alignment and deformable alignment in texture extraction module. We train the following variations. (A) baseline SISR without reference. (B) with...
Table 3. **Quantitative evaluation for ablation study of the decouple framework.** The decouple framework brings an enormous improvement on the five datasets. Especially we obtain over 1dB improvement on Urban100 and Manga109.

<table>
<thead>
<tr>
<th>Decouple</th>
<th>CUFED5</th>
<th>Sun80</th>
<th>Urban100</th>
<th>Manga109</th>
<th>WR-SR</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o Decouple</td>
<td>28.36/0.842</td>
<td>30.01/0.812</td>
<td>25.66/0.769</td>
<td>29.98/0.901</td>
<td>28.16/0.797</td>
</tr>
<tr>
<td>w/ Decouple</td>
<td>28.64/0.850</td>
<td>30.31/0.820</td>
<td>26.71/0.807</td>
<td>31.23/0.917</td>
<td>28.52/0.807</td>
</tr>
</tbody>
</table>

Figure 6. **Qualitative comparison for ablation study of the decouple framework.** (a) Situation with similar reference. The model with the decoupled framework results in more real textures transferred from reference. (b) Situation without similar reference. The model with decoupled framework can recover more details. (Zoom-in for best view)

Table 4. **Quantitative evaluation for ablation study of texture extraction and transfer modules.** The PSNR/SSIM is computed on CUFED5.

<table>
<thead>
<tr>
<th>ID</th>
<th>Flow</th>
<th>DCN</th>
<th>ATT</th>
<th>PSNR/SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td></td>
<td></td>
<td></td>
<td>26.41/0.783</td>
</tr>
<tr>
<td>(B)</td>
<td>✓</td>
<td></td>
<td></td>
<td>27.91/0.832</td>
</tr>
<tr>
<td>(C)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>28.29/0.843</td>
</tr>
<tr>
<td>(D)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>28.50/0.848</td>
</tr>
<tr>
<td>(E)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>28.64/0.850</td>
</tr>
</tbody>
</table>

Figure 7. **Ablation study on texture extraction and transfer modules.** flow-based (Flow) alignment. (C) with deformable convolution (DCN) alignment. (D) with both Flow and DCN alignment. (E) further with adaptive texture transfer(ATT).

Compared with (A), the setting (B) and (C) turn the SISR task into the RefSR task, with a gain over 1dB, due to the additional reference image. (C) has better performance than (B), indicating the superiority of deformable alignment against flow-based alignment in RefSR. As shown in Fig.7, the deformable alignment and the flow-based alignment can recover clear textures of different regions. Therefore, we combine the two alignment methods in the texture extraction module. As shown in Table 4, (D) achieves significant improvement compared with (B) and (C). Fig.7 also shows that results of (D) aggregate the realistic texture in (B) and (C). In conclusion, after combining the stability of flow-based alignment and the superiority of deformable alignment, the model obtains a better performance.

**Adaptive Texture Transfer.** As show in the last row in Table 4 and the result with “Flow+DCN+ATT” in Fig.7, the adaptive texture transfer module refines textures obtained from the texture extraction module and has a further gain of 0.14dB in PSNR on the standard CUFED5 benchmark.

6. Conclusion

In this paper, we first deeply investigate the reference-underuse issue and the reference-misuse issue for RefSR, which are due to the improper coupled framework design. Therefore, we propose a novel framework, decoupling the super-resolution task and the texture transfer task for RefSR. For the LR image, the super-resolution task coarsely upsamples the LR image only by LR itself. For the reference image, the texture transfer task extracts and transfers the realistic textures from the reference image. The texture extraction and the adaptive texture transfer modules are further proposed to migrate the textures more sufficiently. Extensive experiments on five benchmarks quantitatively and qualitatively demonstrate the superior performance of our proposed method over state-of-the-art methods.

Acknowledgements

This work was supported in part by Chinese National Key R&D Program (2019YFB1804304), State Key Laboratory of UHD Video and Audio Production and Presentation, BirenTech Research, Shanghai Key Laboratory of Digital Media Processing and Transmissions(STCSM 18DZ2270700) and 111 plan(BP0719010).
References


[27] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware de-


