PARSING MAP GUIDED MULTI-SCALE ATTENTION NETWORK FOR FACE HALLUCINATION

Chenyang Wang, Zhiwei Zhong, Junjun Jiang, Deming Zhai, and Xianming Liu
School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
Introduction—Super-resolution

Image super-resolution (SR):
transforms low-resolution (LR) images to high-resolution (HR) images
Introduction—Face hallucination

Face SR VS General image SR:
Face images have their special structural information and prior knowledge such as:
1. parsing map
2. landmark
3. heatmap

a) landmark  b) parsing map  c) heatmaps
Our contributions:

1. We design a FishSRNet to generate features in a variety of resolution
2. We propose a multi-scale channel and spatial attention block (MSAB)
3. We get prior knowledge directly from input LR faces.
FishSRNet

**Existing methods:** pre or post-upampling model, but features in low-resolution or high-resolution don’t work well

**Our method:** designs a FishSRNet to generate features in a variety of resolution
Our method: designs a FishSRNet to generate features in a variety of resolution

FishSRNet first up-samples the input then down-samples and up-samples again.
- up-sampling module (UM)
- down-sampling module (DM)
FishSRNet

Feature extraction layer: extracts features from the input

\[ F_0 = Feature\ \text{extraction}(I_{LR}), \]

where \( F_0 \) is the output of the feature extraction layer.
Fish Head: up-samples features to increase the receptive field and the resolution of the features

\[
F_1, F_2, F_3, F_4 = \text{Fish Head}(F_0),
\]

where \(F_1, F_2, F_3\) are the features from every UM for much richer variety of the features, \(F_4\) is for deep layer.
**Fish Body**: down-samples features to improve the diversity of resolution

\[ F_5, F_7, F_9, F_{10} = Fish\ Body(F_4, F_1, F_2, F_3), \]  

where \( F_5, F_7, F_9 \) are the features from every DM for much richer variety of the features, \( F_{10} \) is for deep layer.
Fish Tail: up-samples the feature maps to the same resolution as HR

\[ F_t = Fish\ Tail(F_{10}, F_9, F_7, F_5), \]

where \( F_t \) is for deep layer.
FishSRNet

**Reconstruction layer:** generates the final output

\[ I_{SR} = \text{Reconstruction}(F_t), \]

where \( I_{SR} \) is the result of our network.
FishSRNet—Experiment

The effectiveness of FishSRNet

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>post-upsampling</td>
<td>25.12</td>
<td>0.8705</td>
</tr>
<tr>
<td>FishSRNet</td>
<td>25.26</td>
<td>0.8745</td>
</tr>
</tbody>
</table>
**Existing methods**: ignore the attention mechanism which is proved useful in general image SR.

**Our method**: introduces attention mechanism to face SR and constructs a multi-scale channel and spatial attention block (MSAB).
Multi-scale convolution: extracts multi-scale information
MSAB

- Channel attention: generates channel mask
- Spatial attention: generates spatial mask
The effectiveness of MSAB

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FishSRNet + Resblock</td>
<td>25.26</td>
<td>0.8745</td>
</tr>
<tr>
<td>FishSRNet + MSAB</td>
<td>25.39</td>
<td>0.8773</td>
</tr>
</tbody>
</table>
Existing methods: the prior knowledge derived from the intermediate results is directly affected by the quality of intermediate results.

Our method: gets prior knowledge directly from input LR faces.
Overall Framework

- LR denotes the input LR face
- SR denotes the output of our network
ParsingNet

Overall framework

- common residual network
- parsing map: mask matrix with 0 in skin region and 255 in other components
- other components have much richer textual and structural information
Overall Framework—ParsingNet

Overall framework
Overall Framework–FishSRNet

FishSRNet with parsing map

P denotes the parsing map. We concat the paring map at the front of the FishSRNet and before the Fish Tail.
The effectiveness of ParsingNet

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FishSRNet + MSAB</td>
<td>25.39</td>
<td>0.8773</td>
</tr>
<tr>
<td>FishSRNet + MSAB + ParsingNet</td>
<td>25.34</td>
<td>0.8758</td>
</tr>
</tbody>
</table>

ParsingNet can’t improve PSNR and SSIM.
Illustrations of influences of our different components: (a) LR. (b) The results of FishSRNet. (c) The results of FishSRNet + MSAB. (d) The results of FishSRNet + MSAB + ParsingNet. (e) Ground truth.

ParsingNet contributes to visual quality.
Quantitative evaluation of various face hallucination methods

<table>
<thead>
<tr>
<th></th>
<th>Bicubic</th>
<th>SRCNN</th>
<th>VDSR</th>
<th>URDGN</th>
<th>[26]</th>
<th>FSRNet</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>22.60</td>
<td>23.18</td>
<td>22.60</td>
<td>23.42</td>
<td>24.71</td>
<td>25.08</td>
<td><strong>25.34</strong></td>
</tr>
<tr>
<td>SSIM</td>
<td>0.8104</td>
<td>0.8301</td>
<td>0.8164</td>
<td>0.8375</td>
<td>0.8587</td>
<td>0.8670</td>
<td><strong>0.8758</strong></td>
</tr>
</tbody>
</table>
Experiment

Qualitative comparison of various face hallucination methods
Experiment

Failure cases

Our method exhibits poor performances when encountering special faces.
Thanks for your attention!

Any questions?