NAFSSR: Stereo Image Super-Resolution Using NAFNet

Xiaojie Chu* Liangyu Chen* Wenqing Yu
Background

- Stereo Image Super-resolution

- Low-resolution
  - Left-view
    - intra-view
  - Right-view
    - intra-view

- Cross-view

- High-resolution
Overview

NAFSSR

NAFBlock: Blocks from NAFNet
SCAM: Stereo Cross-Attention Modules
Architectures

- NAFBlock [1]

\( \text{SimpleGate}(x) = x_1 \times x_2, \text{ where } [x_1, x_2] = \text{split}(x, \text{dim}=\text{channel}) \)

[1] Chen, Liangyu, et al. "Simple baselines for image restoration." arXiv preprint arXiv:2204.04676 (2022).
Architectures

- Stereo Cross Attention Module (SCAM)
  - Scaled dot-Product Attention
    - Attention($Q, K, V$) = $\text{softmax}(QK^T/\sqrt{C})V$
  - Bidirectional Cross Attention
    - $F_{R \rightarrow L} = \text{Attention}(W_1^{L}X_L, W_1^{R}X_R, W_2^{R}X_R)$
    - $F_{L \rightarrow R} = \text{Attention}(W_1^{R}X_R, W_1^{L}X_L, W_2^{L}X_L)$
  - Fusion
    - $F_L = \gamma_L F_{R \rightarrow L} + X_L$
    - $F_R = \gamma_R F_{L \rightarrow R} + X_R$

![Diagram of Architectures](image-url)
Architectures

- Stereo Cross Attention Module (SCAM)
  - Scaled dot-Product Attention
    - $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{C})V$

- Bidirectional Cross Attention
  - $F_{R \rightarrow L} = \text{Attention}(W^L_1X_L, W^R_1X_R, W^R_2X_R)$
  - $F_{L \rightarrow R} = \text{Attention}(W^R_1X_R, W^L_1X_L, W^L_2X_L)$

- Attends to corresponding features along the horizontal epipolar line
- since image pairs has horizontal disparities only
Architectures

- NAFSSR Family

| Models     | #Channels | #Blocks | #Params  |
|------------|-----------|---------|----------|
| NAFSSR-T   | $C = 48$  | $N = 16$| 0.46M    |
| NAFSSR-S   | $C = 64$  | $N = 32$| 1.56M    |
| NAFSSR-B   | $C = 96$  | $N = 64$| 6.80M    |
## Tricks

- Data Augmentation

|          | hflip | vflip | channel shuffle | PSNR | ΔPSNR |
|----------|-------|-------|-----------------|------|-------|
|          | ✗     | ✗     | ✗               | 23.43| -     |
| ✗        | ✓     | ✗     | ✗               | 23.64| +0.21 |
| ✗        | ✓     | ✓     | ✗               | 23.63| +0.20 |
| ✗        | ✗     | ✓     | ✓               | 23.62| +0.19 |
| ✓        | ✓     | ✓     | ✗               | 23.73| +0.30 |
| ✓        | ✓     | ✓     | ✓               | 23.82| +0.39 |
Tricks

- Stochastic depth [2] for better generality.

| Model     | Training Stoch. Depth | Test TLSC | In-distribution Flickr1024 [32] | Out-distribution KITTI 2012 [9] | KITTI 2015 [25] | Middlebury [27] | Average |
|-----------|-----------------------|-----------|---------------------------------|---------------------------------|-----------------|-----------------|---------|
| NAFSSR-S  | ✓                     | ✓         | 23.85                           | 26.91                           | 26.74           | 29.63           | 27.76   |
|           | ✓                     | ✓         | 23.82 (−0.03)                   | 26.88 (−0.03)                   | 26.71 (−0.03)   | 29.61 (−0.02)   | 27.73 (−0.03) |
|           | ✓                     | ✓         | 23.78 (−0.07)                   | 26.86 (−0.05)                   | 26.67 (−0.07)   | 29.54 (−0.09)   | 27.69 (−0.07) |
| NAFSSR-B  | ✓                     | ✓         | 24.10                           | 27.05                           | 26.89           | 29.93           | 27.96   |
|           | ✓                     | ✓         | 23.98 (−0.11)                   | 26.92 (−0.13)                   | 26.70 (−0.19)   | 29.78 (−0.15)   | 27.80 (−0.16) |
|           | ✓                     | ✓         | 24.01 (−0.09)                   | 27.00 (−0.05)                   | 26.80 (−0.09)   | 29.81 (−0.12)   | 27.87 (−0.09) |

[2] Huang, Gao, et al. “Deep networks with stochastic depth.” European conference on computer vision. Springer, Cham, 2016.
Tricks

- Train-test Inconsistency: Patches vs. Image
- Inference: Test-time Local Statistics Converter (TLSC) [3]

| Model       | Training Stoch. Depth | Test TLSC | In-distribution Flickr1024 [32] | Out-distribution | KITTI 2012 [9] | KITTI 2015 [25] | Middlebury [27] | Average |
|-------------|-----------------------|-----------|----------------------------------|------------------|----------------|----------------|----------------|---------|
| NAFSSR-S    | ✓                     | ✓         | 23.85                            | 26.91            | 26.74          | 29.63          | 27.76          |         |
|             | ✓                     |           | 23.82 (-0.03)                    | 26.88 (-0.03)    | 26.71 (-0.03) | 29.61 (-0.02) | 27.73 (-0.03) |         |
|             | ✓                     |           | 23.78 (-0.07)                    | 26.86 (-0.05)    | 26.67 (-0.07) | 29.54 (-0.09) | 27.69 (-0.07) |         |
| NAFSSR-B    | ✓                     | ✓         | 24.10                            | 27.05            | 26.89          | 29.93          | 27.96          |         |
|             | ✓                     |           | 23.98 (-0.11)                    | 26.92 (-0.13)    | 26.70 (-0.19) | 29.78 (-0.15) | 27.80 (-0.16) |         |
|             | ✓                     |           | 24.01 (-0.09)                    | 27.00 (-0.05)    | 26.80 (-0.09) | 29.81 (-0.12) | 27.87 (-0.09) |         |

[3] Chu, Xiaojie, et al. "Revisiting Global Statistics Aggregation for Improving Image Restoration." arXiv preprint arXiv:2112.04491 (2021).
Results

- More Stereo Cross Attention Modules (SCAM), better results

| #SCAM | 0   | 1   | 4   | 8   | 16  | 32  |
|-------|-----|-----|-----|-----|-----|-----|
| PSNR  | 23.56 | 23.74 | 23.76 | 23.79 | 23.82 | 23.85 |
| ΔPSNR | -   | +0.18 | +0.20 | +0.23 | +0.26 | +0.29 |
Results

• #Parameters vs. PSNR
Results

- Runtime speedup

| Models            | PSNR     | Time (ms) | Speedup |
|-------------------|----------|-----------|---------|
| SSRDEFNet [4]     | 23.59    | 238.5     | 1.00×   |
| NAFSSR-T (Ours)   | 23.64 (+0.05) | 46.7   | 5.11×   |
| NAFSSR-S (Ours)   | 23.88 (+0.29) | 91.8   | 2.60×   |
| NAFSSR-B (Ours)   | 24.07 (+0.48) | 224.9  | 1.06×   |
Visual Examples
NTIRE Stereo Image Super-Resolution Challenge

• Additional Tricks for challenge
  • Further enlarge model by increasing its depth and width
  • Test-time data augmentations for self-ensemble [4]
  • Ensemble multiple models trained with various hyper-parameters [5]

• Result
  • 24.239 dB PSNR on the validation set
  • 23.787 dB PSNR on the test set (First place)

[4] Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." CVPRW, 2017.
[5] Wortsman, Mitchell, et al. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time." arXiv preprint arXiv:2203.05482 (2022).
Summary

• NAFSSR
  • Single View: NAFNet Block [1]
  • Cross-view: Stereo Cross Attention Module

• Tricks:
  • Training
    • Data augmentation: flip + RGB shuffle
    • Regularization: stochastic depth [2]
  • Inference
    • Test-time Local Statistics Converter [3]

[1] Chen, Liangyu, et al. "Simple baselines for image restoration." arXiv preprint arXiv:2204.04676 (2022).
[2] Huang, Gao, et al. "Deep networks with stochastic depth." ECCV, 2016.
[3] Chu, Xiaojie, et al. "Revisiting Global Statistics Aggregation for Improving Image Restoration." arXiv preprint arXiv:2112.04491 (2021).

https://github.com/megvii-research/NAFNet