Hierarchical Generative Adversarial Networks for Single Image Super-Resolution (HSRGAN)

김정현

Vision & Display Systems Lab.

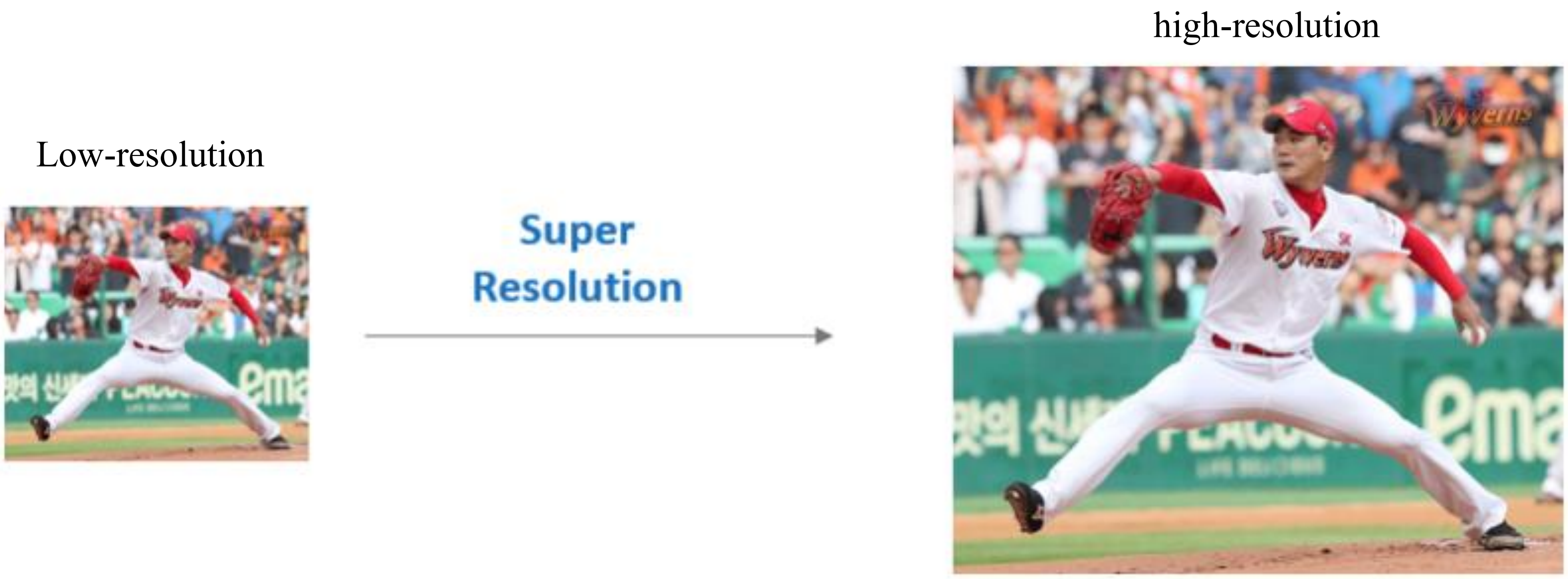
Dept. of Electronic Engineering, Sogang University

Outline

- Background
  - SISR
  - MSE
  - GAN
- HSRGAN
  - HFEM
  - HGRM
  - Formulation
- Experiment
  - Metric
  - Ablation study
  - Comparison with the state-of-the-art
- Conclusion

Background

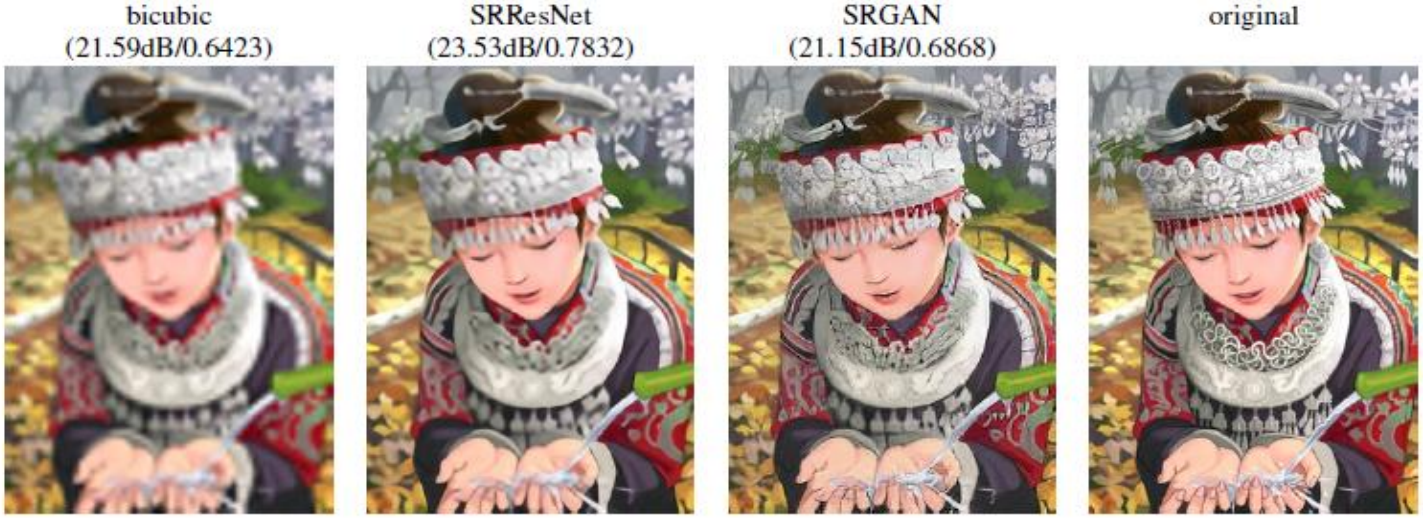
- Single Image Super-Resolution (SISR) :
  - Super-Resolution: Converting LR (low resolution) image to HR (high resolution) image



Example of Single Image Super-Resolution

Background

- Mean Squared Error (MSE) based
  - The optimization target of supervised SR algorithms is commonly the minimization of the MSE between the recovered HR image and the ground truth.
  - Nevertheless the ability of MSE to capture perceptually relevant differences, such as high texture detail, is very limited as they are defined based on pixel-wise image differences.
    - Highest PSNR does not necessarily reflect the perceptually better SR result



bicubic (21.59dB/0.6423)

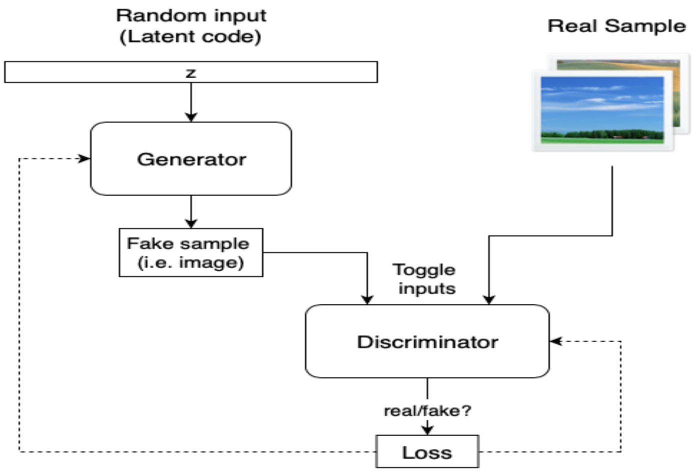
SRResNet (23.53dB/0.7832)

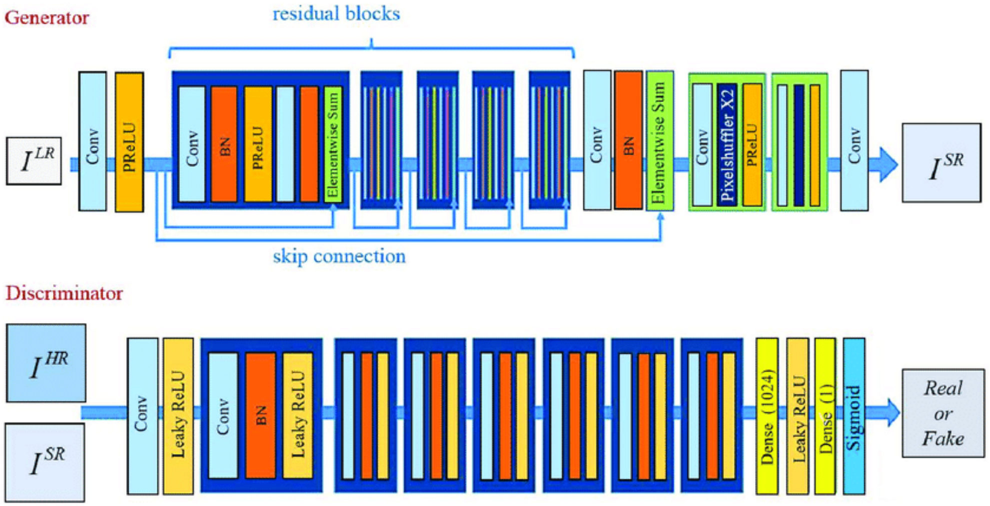
SRGAN (21.15dB/0.6868)

original

Background

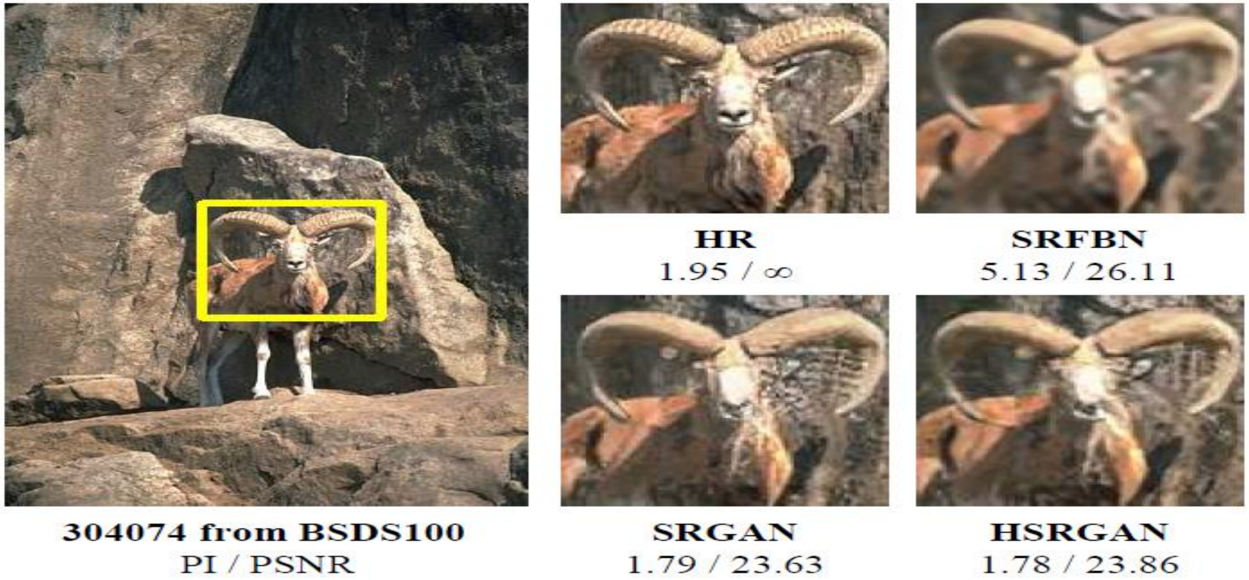
- Generative Adversarial Network (SRGAN)[1]
  - Generator can learn to create solutions that are highly similar to real images and thus difficult to classify
  - To discriminate real HR images from generated SR samples we train a discriminator network





Method

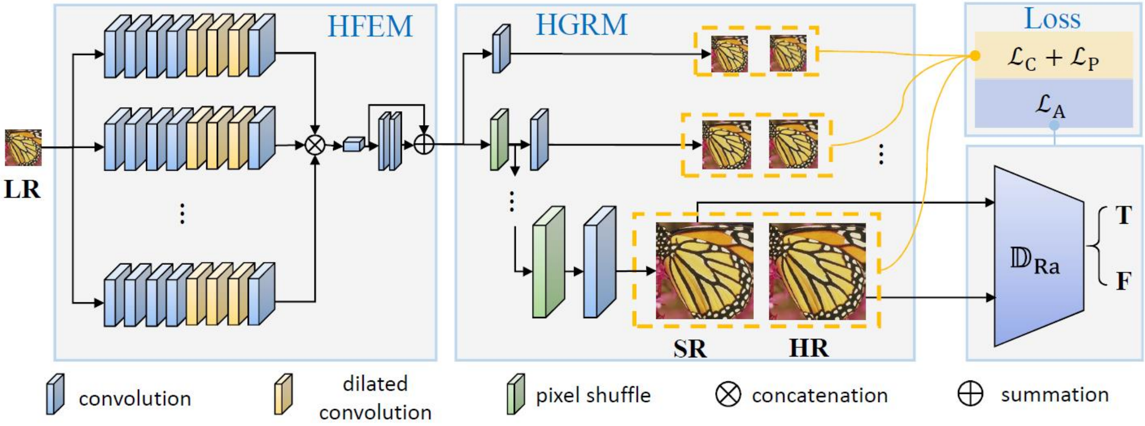
- Generative Adversarial Network (SRGAN)[1]
  - However, GAN usually extract features on a single scale and lack sufficient supervision information, leading to undesired artifacts and unpleasant noise in super-resolution (SR) images.
  - HSRGAN can reduce unpleasant artifacts and produce more convincing textures.



Comparison of state-of-the-art methods

Method

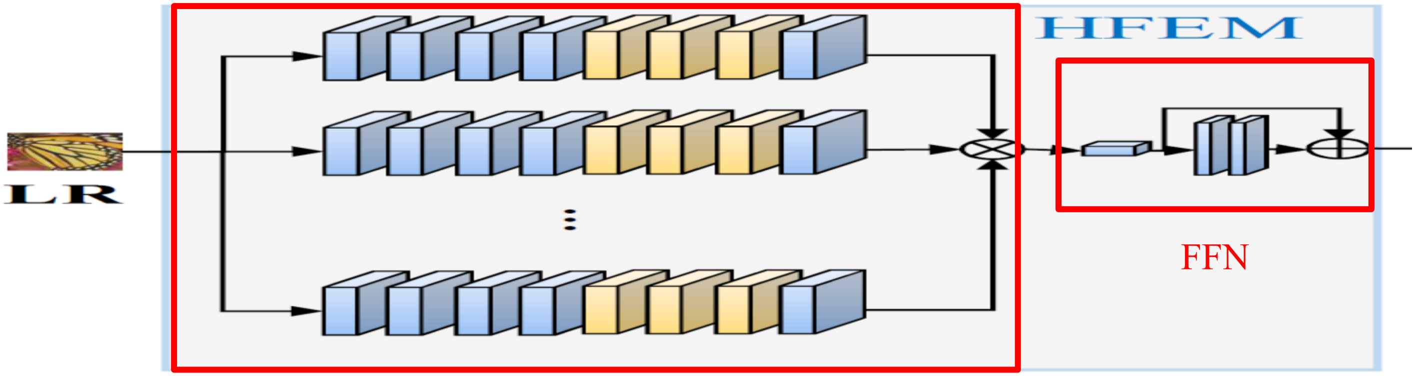
- Hierarchical Feature Extraction Module (HFEM)
  - Hierarchical Feature Extraction Module (HFEM) to extract the features of multiple scales using a multi-branch architecture, which helps our network concentrate on both local textures and global semantics
- Hierarchical Guided Reconstruction Module (HGRM)
  - Hierarchical Guided Reconstruction Module (HGRM), where we divide the SR task of a large upscale factor into a sequence of easier sub-tasks with small upscale factors



Block diagram of proposed HSRGAN

Method

- Hierarchical Feature Extraction Module (HFEM)
  - In human visual systems, when recognizing an object, we need to pay attention to the global information as well as the local details.
    - In the same way, suitable feature representations are crucial for networks to understand an image.
  - To address these problems, inspired by image inpainting[2] , we utilize the HFEM to capture different levels of features from input LR images, which considers features in various scales.
    - HFEM contains two components : multi-branch network(MBN) and future fusion network (FFN).



MBN

Method

- Hierarchical Feature Extraction Module (HFEM)
  - Firstly, the input LR image is fed into the MBN to parallelly extract features F_M in various scales:

$$\begin{aligned}\mathbf{F}_{\mathrm{M}} &= \mathbb{E}_{\mathrm{M}}(\mathbf{I}_{\mathrm{LR}}) \\ &= \left[\mathbb{E}_{\mathrm{M}}^{(1)}(\mathbf{I}_{\mathrm{LR}}), \mathbb{E}_{\mathrm{M}}^{(2)}(\mathbf{I}_{\mathrm{LR}}), \cdots, \mathbb{E}_{\mathrm{M}}^{(B)}(\mathbf{I}_{\mathrm{LR}})\right]\end{aligned}$$

    – where E_M represents the MBN of our HFEM, B denotes the number of branches and $E_M^{(b)}$ means the b - th branch, $1 \leq b \leq B$.

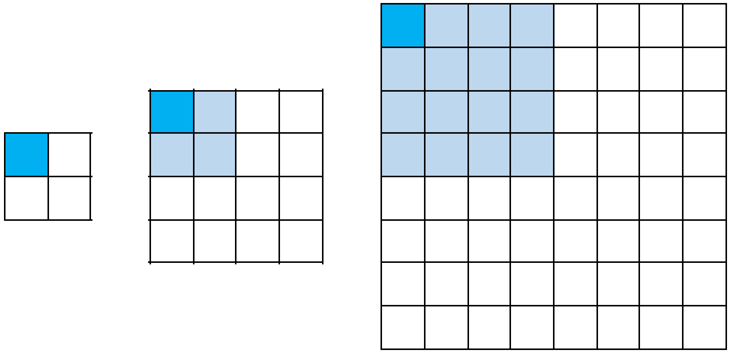
  - Secondly, feeding the concatenation of the hierarchical features FM into the FFN to jointly learn the final feature representation of the input I_{LR} :

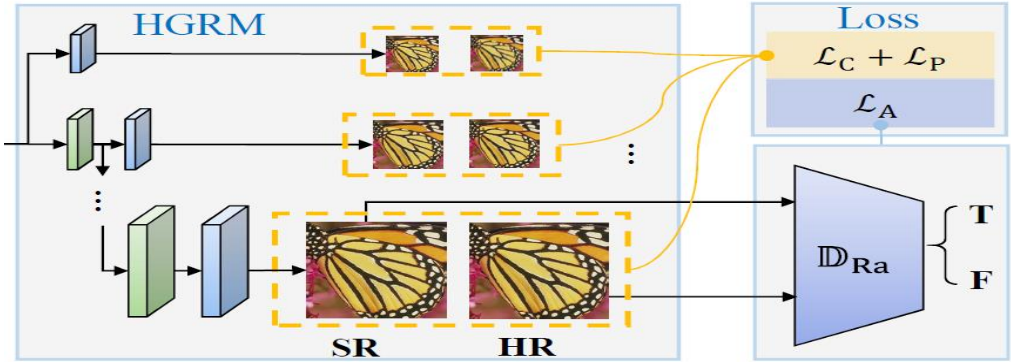
$$\mathbf{F}_{\mathrm{H}} = \mathbb{E}_{\mathrm{F}}\left(\mathbf{F}_{\mathrm{M}}\right)$$

    – where E_F represents the FFN of our HFEM.

Method

- Hierarchical Guided Reconstruction Module (HGRM)
  - Compared with the problem with a small upscale factor, the network will create more pixels based on one pixel, which may force the generator to produce a lot of unreal image details.
    - To solve this drawback, we propose a hierarchical guided reconstruction module (HGRM) to recover the final SR image in an easy-to-hard way.





Method

- Hierarchical Guided Reconstruction Module (HGRM)
  - Different from conventional SISR methods, our HGRM introduces more supervision information into the model.
  - The main branch reconstructs the LR image to the target resolution and the rest branches produce the intermediate SR images with corresponding upscale factors to provide more supervision information:

$$\mathbf{I}_{\mathrm{SR}}^{(t)} = \mathbb{U}^{(t)}\left(\mathbf{F}_{\mathrm{H}}\right)$$

  - where $I_{SR}^{(\mathrm{t})}$ represents the *t*-th output image and $U^{(t)}$ denotes the *t*-th branch, $1 \leq t \leq \mathrm{T}$
  - The main branch generates the final SR image I_{SR} by

$$\mathbf{I}_{\mathrm{SR}} = \mathbf{I}_{\mathrm{SR}}^{(T)} = \mathbb{U}^{(T)}\left(\mathbf{F}_{\mathrm{H}}\right).$$

Method

- Hierarchical Guided Reconstruction Module (HGRM)
  - The main purpose to introduce hierarchical branches into the reconstruction module is to provide more supervision information about the image content
    - By penalizing the network with losses between the outputs of intermediate branches and intermediate HR images generated from the ultimate HR image.
  - Due to the fact that employing adversarial loss may force the generator to produce sharp but incorrect details, we disable the adversarial loss to intermediate branches to avoid wrong supervision information.
    - Thus the intermediate branches mainly concentrate on the content information, while the final branch preserves the structure information and generate realistic results.

Method

- Formulation
  - Content loss can be expressed as :

$$\mathcal{L}_{\mathrm{C}} = \frac{1}{T} \sum_{i=1}^{T} \left\| \mathbf{I}_{\mathrm{SR}}^{(i)} - \mathbf{I}_{\mathrm{HR}}^{(i)} \right\|_1$$

  - Perceptual loss is expressed as :

$$\mathcal{L}_{\mathrm{P}} = \frac{1}{T} \sum_{i=1}^{T} \left\| \phi\left(\mathbf{I}_{\mathrm{SR}}^{(i)}\right) - \phi\left(\mathbf{I}_{\mathrm{HR}}^{(i)}\right) \right\|_1$$

  - Adversarial loss for generator is expressed as:

$$\begin{aligned}\mathcal{L}_{\mathrm{A}} = & -\log\left(1 - \mathbb{D}_{\mathrm{Ra}}\left(\mathbf{I}_{\mathrm{HR}}, \mathbf{I}_{\mathrm{SR}}\right)\right) \\ & -\log\left(\mathbb{D}_{\mathrm{Ra}}\left(\mathbf{I}_{\mathrm{SR}}, \mathbf{I}_{\mathrm{HR}}\right)\right).\end{aligned}$$

  - Total loss of generator is defined as :

$$\mathcal{L}_{\mathrm{G}} = \mathcal{L}_{\mathrm{C}} + \lambda \mathcal{L}_{\mathrm{P}} + \eta \mathcal{L}_{\mathrm{A}}$$

Experiment

- Metric
  - SR algorithms have gradually developed into two directions:
  - One is to obtain higher restoration accuracy measured by PSNR [3, 4, 5]
    - PSNR calculates the pixel-wise difference between SR image and ground-truth PSNR where the higher is the better.

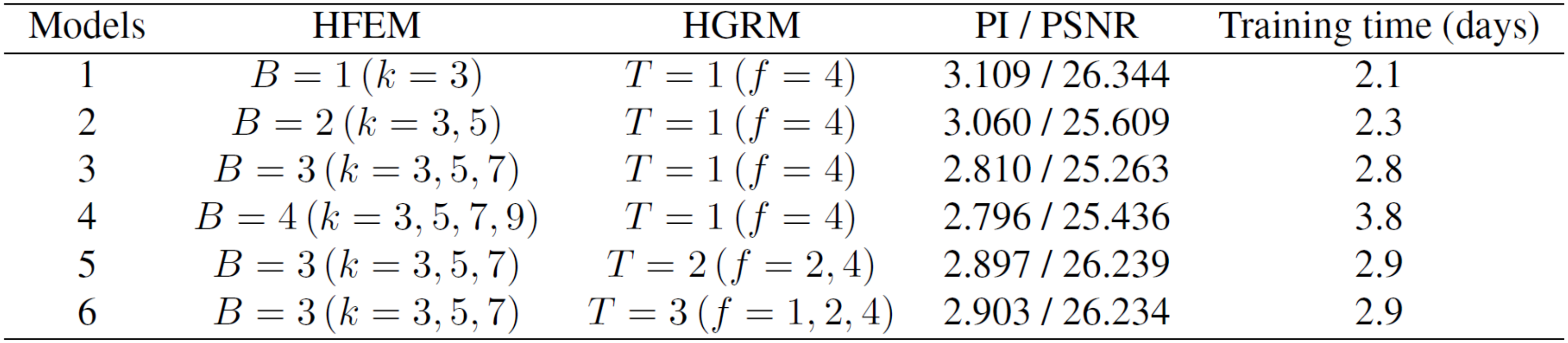
$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{\mathrm{MAX_I}^2}{\mathrm{MSE}}$$

  - The other is to measure perceptual quality of reconstructed images, among which the perceptual index [6] is the most commonly used metric PI
    - PI combines the no-reference image quality measures of Ma score [7] and NIQE [8], and the lower is the better.

$$\mathrm{PI} = \frac{1}{2}((10 - \mathrm{Ma}) + \mathrm{NIQE})$$

Experiment

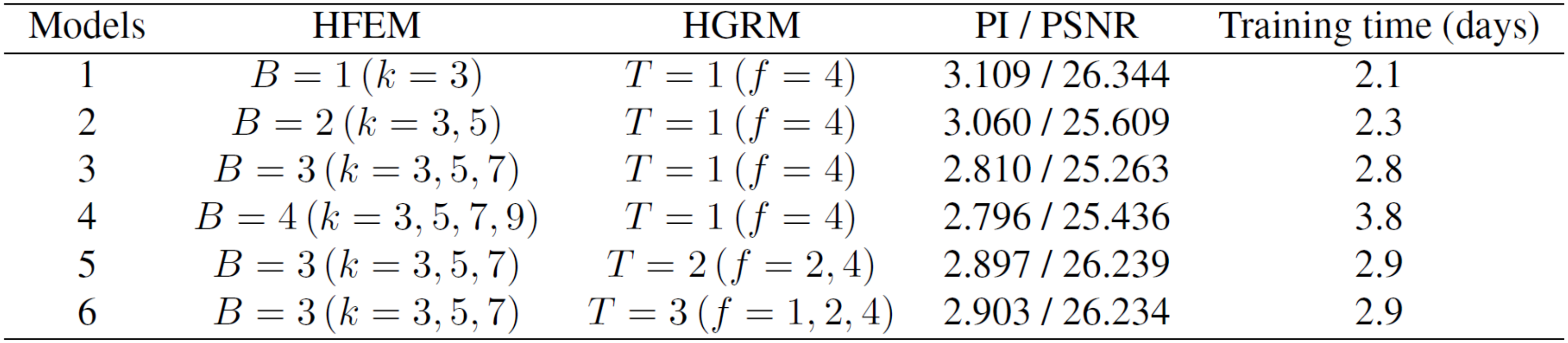
- Ablation Study
  - We can see that model 1 achieves PI = 3.109 and PSNR = 26.344. we employ our HFEM and add branches as model 2 to model 4.
    - Results from model 2 to model 4 verify the effectiveness of the hierarchical features extracted from our HFEM.
  - Specifically, model 3 significantly optimizes PI to 2.810 and model 4 further reduces PI to 2.796 while costs more than 1 day in training, since the computing complexity caused by kernel size of 9.
    - The pixel-wise accuracy reflected by PSNR deteriorates about 1dB.



Models	HFEM	HGRM	PI / PSNR	Training time (days)
1	$B = 1\,(k = 3)$	$T = 1\,(f = 4)$	3.109 / 26.344	2.1
2	$B = 2\,(k = 3, 5)$	$T = 1\,(f = 4)$	3.060 / 25.609	2.3
3	$B = 3\,(k = 3, 5, 7)$	$T = 1\,(f = 4)$	2.810 / 25.263	2.8
4	$B = 4\,(k = 3, 5, 7, 9)$	$T = 1\,(f = 4)$	2.796 / 25.436	3.8
5	$B = 3\,(k = 3, 5, 7)$	$T = 2\,(f = 2, 4)$	2.897 / 26.239	2.9
6	$B = 3\,(k = 3, 5, 7)$	$T = 3\,(f = 1, 2, 4)$	2.903 / 26.234	2.9

Experiment

- Ablation Study
  - Specifically, model 5 contains an intermediate branch of upscale factor of 2 (T = 2) and model 6 contains one more branch of upscale factor of 1 (T = 3), which guides the output of HFEM to estimate the original LR image.
    - Experiments indicate that with the help of intermediate guided supervision, our model can generate higher quality images with a raise of PSNR by nearly 1dB.
  - Nevertheless, model 6 achieves almost the same quantitative results as model 5 in PI and PSNR
    - It implies that the adding the intermediate branch of original resolution does not further improve the performance.

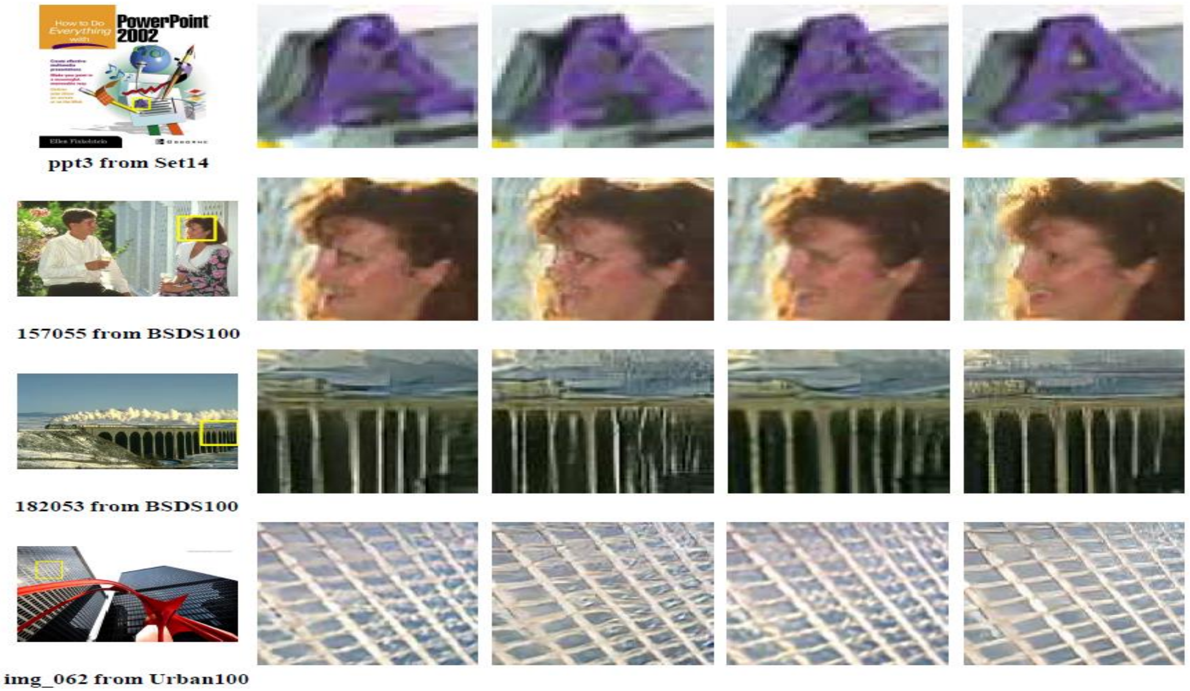


Models	HFEM	HGRM	PI / PSNR	Training time (days)
1	$B = 1\,(k = 3)$	$T = 1\,(f = 4)$	3.109 / 26.344	2.1
2	$B = 2\,(k = 3, 5)$	$T = 1\,(f = 4)$	3.060 / 25.609	2.3
3	$B = 3\,(k = 3, 5, 7)$	$T = 1\,(f = 4)$	2.810 / 25.263	2.8
4	$B = 4\,(k = 3, 5, 7, 9)$	$T = 1\,(f = 4)$	2.796 / 25.436	3.8
5	$B = 3\,(k = 3, 5, 7)$	$T = 2\,(f = 2, 4)$	2.897 / 26.239	2.9
6	$B = 3\,(k = 3, 5, 7)$	$T = 3\,(f = 1, 2, 4)$	2.903 / 26.234	2.9

Experiment

- Ablation Study
  - To compare the visual quality of different models, we test model 1, model 2, model 3 and model 5 on BSDS100 and Urban100.

Model 1 Model 2 Model 3 Model 5



ppt3 from Set14

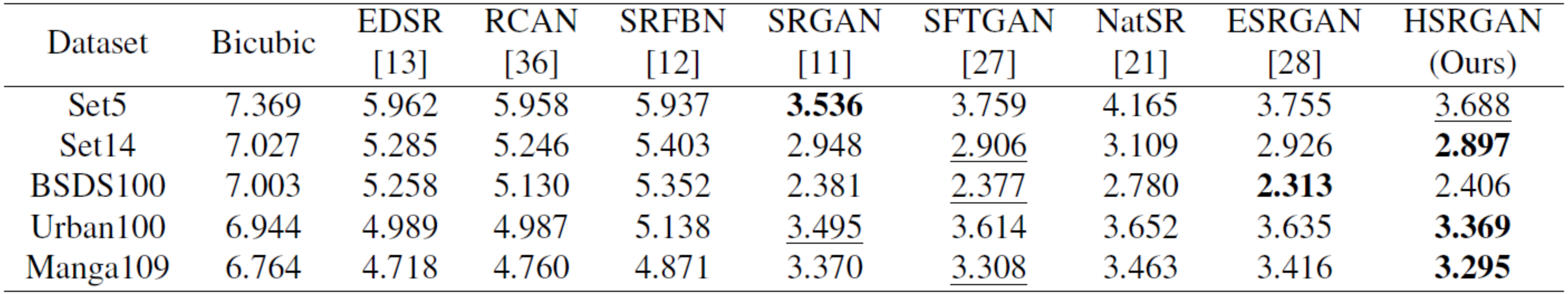
157055 from BSDS100

182053 from BSDS100

img_062 from Urban100

Experiment

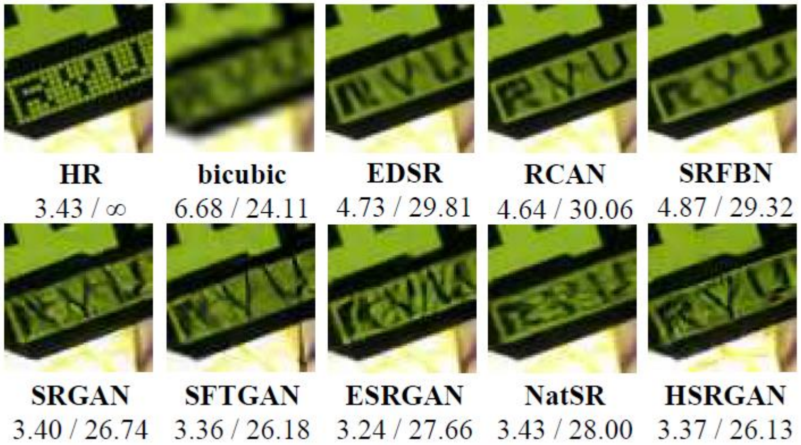
- Comparison with the state-of-the-art
  - We employ Bicubic, EDSR [3], RCAN [4], SRFBN [5], SRGAN [1], SFTGAN [10], NatSR[11], ESRGAN [6] as our comparison methods.
  - Compared to distortion oriented methods, such as EDSR, RCAN and SRFBN, GAN-based or perception-oriented methods show significant advantages in perceptual index, which indicates that the methods generate clear edges of images to some extent.
  - Among all the GAN-based methods, our HSRGAN outperforms the other methods on Set14, Urban100, Manga109 datasets and achieves the second best on Set5 dataset, which is comprehensively the best quantitative results.

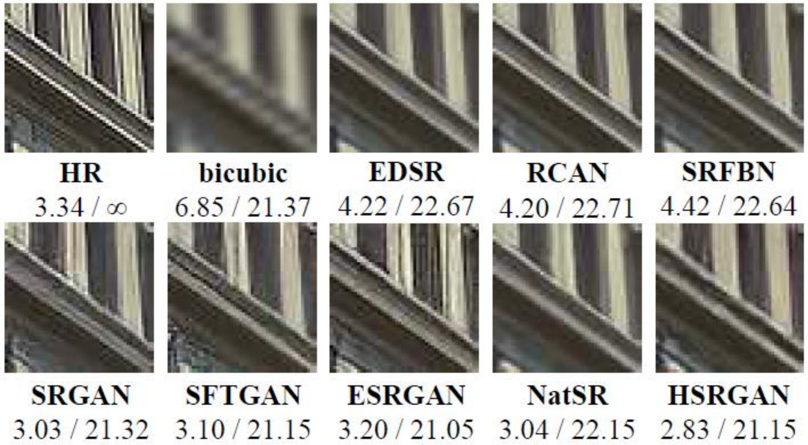


Dataset	Bicubic	EDSR [13]	RCAN [36]	SRFBN [12]	SRGAN [11]	SFTGAN [27]	NatSR [21]	ESRGAN [28]	HSRGAN (Ours)
Set5	7.369	5.962	5.958	5.937	**3.536**	3.759	4.165	3.755	3.688
Set14	7.027	5.285	5.246	5.403	2.948	2.906	3.109	2.926	**2.897**
BSDS100	7.003	5.258	5.130	5.352	2.381	2.377	2.780	**2.313**	2.406
Urban100	6.944	4.989	4.987	5.138	3.495	3.614	3.652	3.635	**3.369**
Manga109	6.764	4.718	4.760	4.871	3.370	3.308	3.463	3.416	**3.295**

Experiment

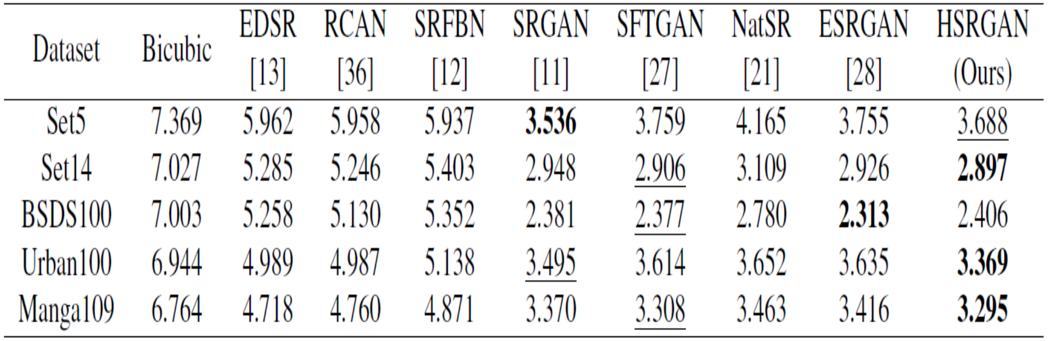
- Comparison with the state-of-the-art
  - As we can see, the distortion-oriented methods produce over-smoothing images, while recent GAN-based methods outperform in both sharpness and details.
  - Another problem of GAN based methods is that they sometimes add undesired noise into the final SR images.
  - Our HSRGAN can get rid of the unpleasant artifacts while maintaining enough details and generate clearer images.



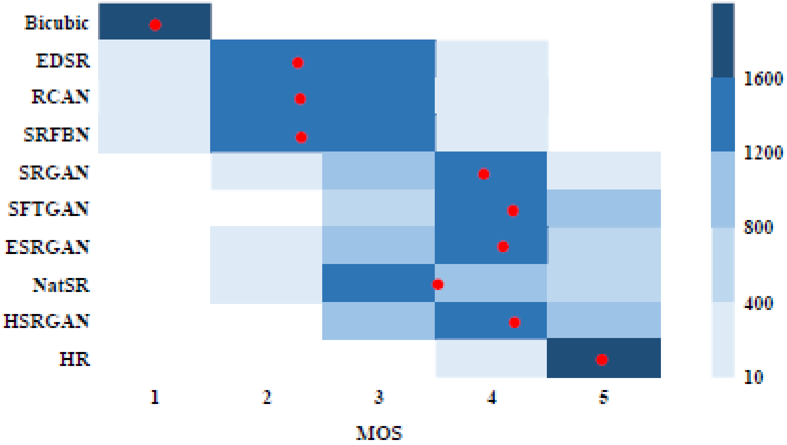


Experiment

- Comparison with the state-of-the-art
  - Perceptual index does not fully reflect the visual quality of the image.
    - The lower perceptual index does not always guarantee a better visual quality.
    - To provide a better reference standard for visual quality assessment, we use the mean opinion score (MOS) to quantify our performance.
  - HSRGAN slightly outperforms SFTGAN, NatSR, ESRGAN and SRGAN.



Dataset	Bicubic	EDSR [13]	RCAN [36]	SRFBN [12]	SRGAN [11]	SFTGAN [27]	NatSR [21]	ESRGAN [28]	HSRGAN (Ours)
Set5	7.369	5.962	5.958	5.937	**3.536**	3.759	4.165	3.755	3.688
Set14	7.027	5.285	5.246	5.403	2.948	2.906	3.109	2.926	**2.897**
BSDS100	7.003	5.258	5.130	5.352	2.381	2.377	2.780	**2.313**	2.406
Urban100	6.944	4.989	4.987	5.138	3.495	3.614	3.652	3.635	**3.369**
Manga109	6.764	4.718	4.760	4.871	3.370	3.308	3.463	3.416	**3.295**



Conclusion

- HSRGAN
  - We proposed hierarchical generative adversarial networks (HSRGAN) for the SISR problem.
  - Specifically, the hierarchical feature extraction module (HFEM) extracts the hierarchical features in multiple receptive fields, concentrating on both local texture and global semantics.
    - In addition, we proposed a hierarchical guided reconstruction module (HGRM).
  - It reconstructs the SR image by adding intermediate supervision branches in a progressive manner.
  - Extensive experiments on 5 common datasets show that our method achieves state-of-the-art performance in terms of both quantitative metrics and visual quality.

Reference

- Christian Ledig, Lucas Theis, Ferenc Husz´ar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photorealistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
- Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In Advances in Neural Information Processing Systems, pages 331–340, 2018.
- Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 136–144, 2017.
- Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 286–301, 2018.
- Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3867–3876, 2019.
- Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 0–0, 2018.
- Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding, 158:1–16, 2017.
- Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. IEEE Signal Processing Letters, 20(3):209–212, 2012.

Reference

- Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 606–615, 2018.
- Jae Woong Soh, Gu Yong Park, Junho Jo, and Nam Ik Cho. Natural and realistic single image super-resolution with explicit natural manifold discrimination. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.