NTIRE 2021 Learning the Super-Resolution Space Challenge

Andreas Lugmayr* Martin Danelljan* Radu Timofte* Christoph Busch Yang Chen
Jian Cheng Vishal Chudasama Ruipeng Gang Shangqi Gao Kun Gao
Laiyun Gong Zhejiang University Qingrui Han NetEaseYunXin Chao Huang Zhi Jin
Younghyun Jo Seon Joo Kim Younggeun Kim Seungjun Lee Yuchen Lei
Chu-Tak Li Chenghua Li Ke Li Zhi-Song Liu Youming Liu Nan Nan
Seung-Ho Park Heena Patel Shichong Peng Kalpesh Prajapati Haoran Qi
Kiran Raja Raghavendra Ramachandra Wan-Chi Siu Donghee Son Ruixia Song
Kishor Upla Li-Wen Wang Yatian Wang Junwei Wang Qianyu Wu Xinhua Xu
Sejong Yang Zhen Yuan NetEaseYunXin Liting Zhang Huanrong Zhang
Junkai Zhang Yifan Zhang Zhenzhou Zhang Hangqi Zhou Aichun Zhu
Xiahai Zhuang Jiaxin Zou

Abstract

This paper reviews the NTIRE 2021 challenge on learning the super-Resolution space. It focuses on the participating methods and final results. The challenge addresses the problem of learning a model capable of predicting the space of plausible super-resolution (SR) images, from a single low-resolution image. The model must thus be capable of sampling diverse outputs, rather than just generating a single SR image. The goal of the challenge is to spur research into developing learning formulations and models better suited for the highly ill-posed SR problem. And thereby advance the state-of-the-art in the broader SR field. In order to evaluate the quality of the predicted SR space, we propose a new evaluation metric and perform a comprehensive analysis of the participating methods. The challenge contains two tracks: $4\times$ and $8\times$ scale factor. In total, 11 teams competed in the final testing phase.

1. Introduction

Single image Super-Resolution (SR) is the task of increasing the resolution of a given image by filling in additional high-frequency content. It has been a popular research topic for decades [27, 19, 44, 54, 52, 57, 58, 59, 51, 12, 24, 53, 15, 16, 32, 34, 36, 18, 4, 5, 22, 26, 21] due to its many applications. The current trend addresses the ill-posed SR problem using deep Convolutional Neural Networks (CNNs). While initial methods focused on achieving high fidelity in terms of PSNR [15, 16, 32, 34, 36]. Recent work has put further emphasis on generating perceptually more appealing predictions using for instance adversarial losses [61, 35, 56].

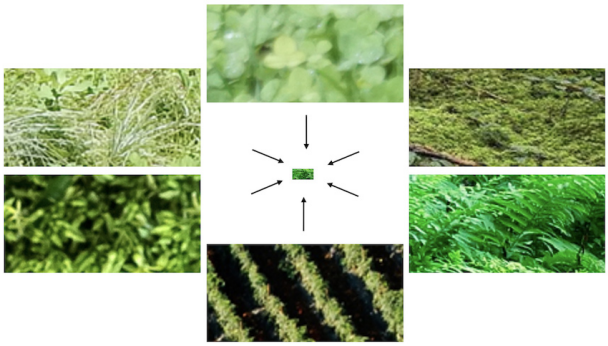

Figure 1. Many High Resolution Images can be downsampled to a single low-resolution image. Super-resolution is thus an ill-posed problem. In this challenge the goal is to take this property into account by promoting methods with a stochastic output.

Usually, super-resolution (SR) is trained using pairs of high- and low-resolution images. Infinitely many high-resolution images can be downsampled to the same low-resolution image. That means that the problem is ill-posed and cannot be inverted with a deterministic mapping. Instead, one can frame the SR problem as learning a stochastic mapping, capable of sampling from the space of plausible high-resolution images given a low-resolution image. This problem has been addressed in recent works [40, 8, 11]. The one-to-many stochastic formulation of the SR problem al-

*Andreas Lugmayr (andreas.lugmayr@vision.ee.ethz.ch), Martin Danelljan, and Radu Timofte at ETH Zürich are the NTIRE 2021 challenge organizers. The other authors participated in the challenge. Appendix A contains the authors' team names and affiliations.
https://data.vision.ee.ethz.ch/cvl/ntire21/

lows for a few potential advantages. First, it can be used to develop more robust learning formulations that better accounts for the ill-posed nature of the SR problem. Second, multiple predictions can be sampled and compared. Third, it opens the potential for controllable exploration and editing in the space of SR predictions.

The goal of the NTIRE 2021 Learning the Super-resolution Space challenge is to spur new research in the direction of stochastic super-resolution and to improve the state-of-the-art of SR in general. The participants are evaluated in terms of three criteria: photo-realism, consistency with the LR image, and how well the SR space is spanned. For the latter, we develop a new metric, based on the relative improvement of a given distance metric when using additional samples.

This challenge is one of the NTIRE 2021 associated challenges: nonhomogeneous dehazing [6], defocus deblurring using dual-pixel [1], depth guided image relighting [17], image deblurring [42], multi-modal aerial view imagery classification [37], learning the super-resolution space [39], quality enhancement of heavily compressed videos [60], video super-resolution [50], perceptual image quality assessment [20], burst super-resolution [9], high dynamic range [45].

2. NTIRE 2021 Challenge

The goals of the NTIRE 2021 Learning the Super-Resolution Space Challenge is to (i) stimulate research into learning the full space of plausible super-resolutions; (ii) develop benchmark protocols and metrics; (iii) probe the state-of-the-art in super-resolution in general. The aim of the challenge is to develop an SR method, capable of sampling diverse predictions. Each individual prediction should achieve the highest possible photo-realism, as perceived by humans. The predictions should also be consistent with the underlying LR image. Hence, content that cannot be explained from the observed LR image should not be hallucinated.

2.1. Overview

The challenge contains two tracks, targeting $4\times$ and $8\times$ super-resolution respectively. Evaluation code and information about the challenge were provided at a public GitHub page `http://git.io/SRSpace`. The challenge employs the DIV2k [2] splits for validation and testing. As the final result, the participants in the challenge were asked to submit 10 random SR predictions for each given LR image.

2.2. Rules

To guide the research towards useful and generalizable techniques, submissions needed to adhere to the following rules.

Team	Generative formulation				Additional Data
	Flow	GAN	VAE	IMLE	
BeWater	✓				
CIPLAB	✓				✓
Deepest	✓				✓
FudanZmic21			✓		✓
FutureReference				✓	✓
SR_DL		✓	✓		
SSS		✓			✓
SYSU-FVL		✓			✓
nanbeihuishi	✓				✓
njtech& seu	✓				
svnit_ntnu		✓			

Table 1. Information about the participating teams in the challenge.

- The method must be able to generate an arbitrary number of diverse samples. That is, the method cannot be limited to a maximum number of different SR samples (corresponding to *e.g.* a certain number of different output network heads).
- All SR samples must be generated by a single model. That is, no ensembles are allowed.
- No self-ensembles or test-time data augmentation (flipping, rotation, etc.).
- All SR samples must be generated using the same hyper-parameters. That is, the generated SR samples shall not be the result of different choices of hyper-parameters during inference.
- Submissions of deterministic methods were allowed. However, they will naturally score zero in the diversity measure and therefore not be able to win the challenge.
- Other than the validation and test split of the DIV2k dataset, any training data or pre-training is allowed.

Furthermore, all participants were asked to submit the code of their solution along with the final results.

2.3. Challenge phases

The challenge had two phases: (1) Development phase: the participants got training and validation images as well as the tools to evaluate the results. (2) Test phase: the participants got access to the LR test images and had to submit their super-resolved images along with the description, code and model weights for their methods.

2.4. Data

We provide the standard DIV2K dataset for $4\times$ and $8\times$ for training and validation. For testing, we only provide the LR images of the test set for both Tracks.

3. Evaluation Protocol

A method is evaluated by first predicting a set of 10 randomly sampled SR images for each low-resolution image in the dataset. From this set of images, evaluation metrics corresponding to the three criteria above will be considered. The participating methods will be ranked according to each metric. These ranks will then be combined into a final score. The three evaluation metrics are described next.

3.1. Photo-realism

Automatically assessing photo-realism and image quality is an extremely difficult task. All existing methods have severe shortcomings. As a very rough guide, the participants were asked to use the LPIPS distance [62]. However, the participants were notified that a human study will be conducted to finally evaluate photo-realism on the test set, and thus beware of overfitting to the LPIPS metric, as that can lead to worse results.

User Study To assess the photo-realism, a human study is performed on the test set for the final submission. The user is asked to rank crops according to how photo-realistic they seem for them. As a reference, the user is shown the region around this crop. To obtain an unbiased opinion, we sample the crop coordinates uniformly within the images. In total, we evaluate three crops of size 80×80 per image of the 100 DIV2K test set images. Every task is done by five different users, resulting in 1500 completed tasks in total. We report the Mean Opinion Rank (MOR) for the user study,

3.2. The spanning of the SR Space

The goal is to generate SR samples that provide meaningful diversity. While, for instance, the pixel-wise standard deviation within the set of generated SR samples measures variations, this variation is not necessarily meaningful. For example, an SR method should be able to easily super-resolve a uniform patch of sky with high accuracy. Since all surrounding pixels in the LR image have very similar color, the SR method can confidently predict the corresponding pixels of the underlying HR image. Hence, the SR model *should* generate low diversity in this case. On the other hand, such confidence cannot be achieved when super-resolving *e.g.* the fine structures in a patch of foliage. The LR image does not contain all information for reconstructing the exact arrangement of leaves and branches. Even when leveraging learned priors, there are thus multiple plausible predictions of the foliage texture. In this case, we want the network to span the space of possibilities.

From the aforementioned discussion, it is clear that diversity is not a quantity that should be simply maximized (or minimized). Instead, the model should learn meaningful diversity, corresponding to the uncertainty in the SR prediction. Simple metrics, such as pixel-wise standard deviation, are therefore not suitable. Instead, we propose a new metric, aiming to measure how well the network spans the space of possibilities.

The challenge in measuring the aforementioned ability lies in that we only have access to a single ground-truth HR sample for every LR image. However, this single sample should lie inside the solution space spanned by the SR model. The proposed metric aims at measuring how well the ground-truth SR image is represented in the predicted space. When following this strategy, the main challenge arises from the high dimensionality of the HR image space. Our key observation is that this can be mitigated by performing the analysis on smaller patches. That is, a single HR image is decomposed into multiple smaller (potentially overlapping) patches. This effectively reduces the dimensionality of the output space, allowing us to evaluate the quality of the predicted SR space from a very limited number of random samples.

Let $y_k \in \mathbb{R}^{N\times N\times 3}$ be the k-th patch in the original HR ground-truth image y. We denote the M number of predictions generated by the SR model as $\{\hat{y}^i\}_{i=1}^M$ and let $\hat{y}_k^i \in \mathbb{R}^{N\times N\times 3}$ be the corresponding decomposition into patches. We measure the similarity between two image patches with a distance metric d. To obtain the meaningful diversity that the samples represent, we calculate how much the minimum distance to the ground-truth patch decreases when using M samples,

$$S_M = \frac{1}{\bar{d}_M}\left(\bar{d}_M - \frac{1}{K}\sum_{k=1}^{K}\min\left\{d(y_k,\hat{y}_k^i)\right\}_{i=1}^{M}\right). \quad (1)$$

Note that the right term evaluates the average distance to the closest of the M patches. To obtain a relative improvement measure, we normalize it w.r.t. to a base distance $\bar{d}_M$ computed over the M samples. One alternative is to set the base distance to simply the average $\bar{d}_M = \frac{1}{KM}\sum_{k,i} d(y_k,\hat{y}_k^i)$. However, such a reference distance is sensitive to outliers. We therefore compute $\bar{d}_M$ by finding the minimum distance on a global sample level,

$$\bar{d}_M = \min\left\{\frac{1}{K}\sum_{k=1}^{K} d(y_k,\hat{y}_k^i)\right\}_{i=1}^{M}. \quad (2)$$

This choice still yields a score in the range $S_M \in [0,1]$, where $S_M = 0$ means no diversity and $S_M = 1$ means that the ground-truth HR image was exactly captured by one of the generated samples. In the tables, we report S_M in percent.

To compute the final diversity score, we average the relative score (1) over all images in the dataset. For the distance metric d, we experimented with both L_2 (*i.e.* mean squared error) and LPIPS [62]. We found the latter to be a more well suited metric for image patches, and therefore use it for our

Figure 2. Qualitative comparison between the participating approaches for $4\times$ super-resolution

final score. In particular, we compute the LPIPS in a fully convolutional manner over the full images y and $\hat{y}^i$. Instead of performing the final spatial averaging of the metric, as done for the standard case, we directly use the resulting distance map as our patch-wise distances $d(y_k, \hat{y}^i_k)$.

3.3. Low Resolution Consistency

To measure how much information is preserved in the super-resolved image from the low-resolution image, we measure the LR-PSNR. It is computed as the PSNR between the input LR image and the predicted sample downsampled with the given bicubic kernel. The goal of this challenge is to obtain an LR-PSNR of at least 45dB.

4. Challenge Results

Before the end of the final test phase, participating teams were required to submit results, code/executables, and factsheets for their approaches. From 112 registered participants, 11 valid methods were submitted. The methods of the teams that entered the final phase are described in Section 5 and the teams' members and affiliations are shown in Section Appendix A.

4.1. Baselines

We compare methods participating in the challenge with the following baseline approaches.

ESRGAN A common baseline for photo-realistic super-resolution is the ESRGAN [56]. Since it is not a stochastic method, the diversity is zero.

SRFlow The method SRFlow [40] uses image conditional normalizing flow to super-resolve images. This method inherently provides stochastic, photo-realistic and

Team	LPIPS	LR-PSNR	Div. Score S_{10} [%]	MOR	Final Rank
svnit_ntnu	0.355	27.52	$1.871_{(11)}$	-	-
SYSU-FVL	0.244	49.33	$8.735_{(10)}$	-	-
nanbeihuishi	0.161	50.46	$12.447_{(9)}$	-	-
FudanZmic21	0.273	47.20	$16.450_{(7)}$	-	-
FutureReference	0.165	37.51	$19.636_{(6)}$	-	-
SR_DL	0.234	39.80	$20.508_{(5)}$	-	-
SSS	0.110	44.70	$13.285_{(8)}$	$4.530_{(3)}$	5.5
BeWater	0.137	49.59	$23.948_{(3)}$	$4.720_{(4)}$	3.5
CIPLAB	0.121	50.70	$23.091_{(4)}$	$4.478_{(2)}$	3.0
njtech&seu	0.149	46.74	$\mathbf{26.924}_{(1)}$	$4.977_{(5)}$	3.0
Deepest	0.117	50.54	$26.041_{(2)}$	$\mathbf{4.372}_{(1)}$	**1.5**
SRFlow	0.122	49.86	25.008	4.410	-
ESRGAN	0.124	38.74	0.000	4.467	-
GT	0	∞	-	3.728	-

Table 2. Quantitative comparison of participating teams. ($4\times$)

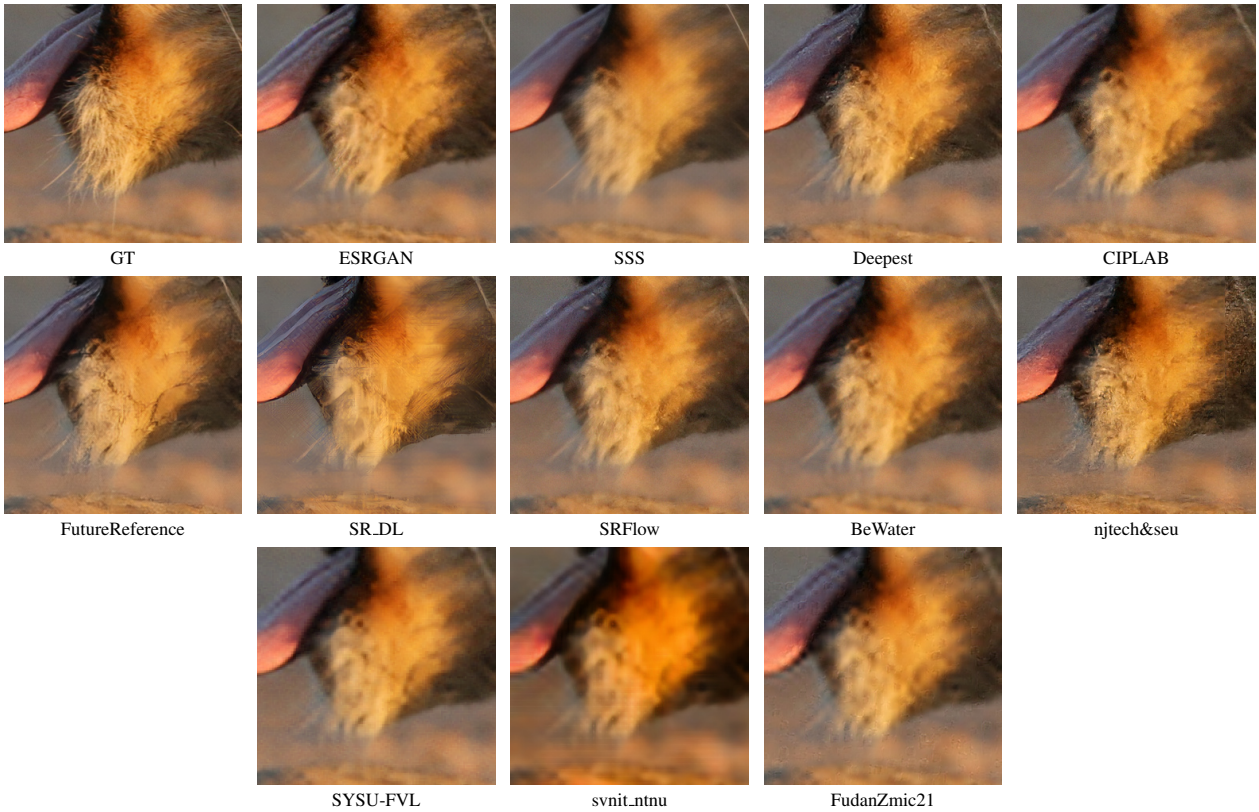

Figure 3. Qualitative comparison between the participating approaches for $8\times$ super-resolution

low-resolution consistent super-resolutions.

4.2. Architectures and Main Ideas

In this section, we discuss the four directions that methods submitted to this challenge are based on. An overview of the participating teams is shown in Table 1.

Flow-Based Inspired by the baseline SRFlow [40] the teams BeWater, CIPLAB, Deepest, nanbeihuishi and njtech&seu submitted Flow-Based approaches. This approach aims to learn the conditional probability distribution of HR images given an LR image. The flow network learns to map an HR-LR pair into a latent space, where the probability density can be evaluated. Since the network is invertible [14], it can be driven in the reverse direction to generate images by sampling a latent vector. Hence, this approach is an inherent stochastic method that draws samples from the space of plausible SR images. Another benefit is that the outputted SR images are highly consistent with the LR images. This was observed by measuring the PSNR of the downsampled SR image compared to the input LR image [40]. The team Deepest worked on the information content gap between the HR image and the latent space. The method submitted by njtech&seu achieved the highest Diversity Score in both $4\times$ and $8\times$ using their multi-head attention module and the normalization flow module. However, this method did not reach the quality in terms of MOR of the baseline SRFlow. The teams BeWater, CIPLAB and nanbeihuishi focused on improving parts of the original SRFlow architecture.

Team	LPIPS	LR-PSNR	Div. Score S_{10} [%]	MOR	Final Rank
svnit_ntnu	0.481	25.55	$4.516_{(10)}$	-	-
SYSU-FVL	0.415	47.27	$8.778_{(9)}$	-	-
FudanZmic21	0.496	46.78	$14.287_{(7)}$	-	-
FutureReference	0.291	36.51	$17.985_{(5)}$	-	-
njtech&seu	0.366	29.65	$\mathbf{28.193}_{(1)}$	-	-
SSS	0.237	37.43	$13.548_{(8)}$	$4.692_{(3)}$	5.5
SR_DL	0.311	42.28	$14.817_{(6)}$	$4.738_{(4)}$	5.0
BeWater	0.297	49.63	$23.700_{(3)}$	$5.133_{(5)}$	4.0
CIPLAB	0.266	50.86	$23.320_{(4)}$	$4.637_{(2)}$	3.0
Deepest	0.259	48.64	$26.941_{(2)}$	$\mathbf{4.630}_{(1)}$	**1.5**
SRFlow	0.282	47.72	25.582	4.635	-
ESRGAN	0.284	30.65	0	4.323	-
GT	0	∞	-	2.613	-

Table 3. Quantitative comparison of participating teams. ($8\times$)

GAN-Based The teams SR_DL, SSS, svnit_ntnu and SYSU-FVL submitted GAN-Based approaches. The team svnit_ntnu is based on the MUNIT [25] approach and samples the style control signal. With this approach, they did not reach the required LR PSNR or reached the baseline in diversity score. The two teams SSS and SYSU-FVL are us-

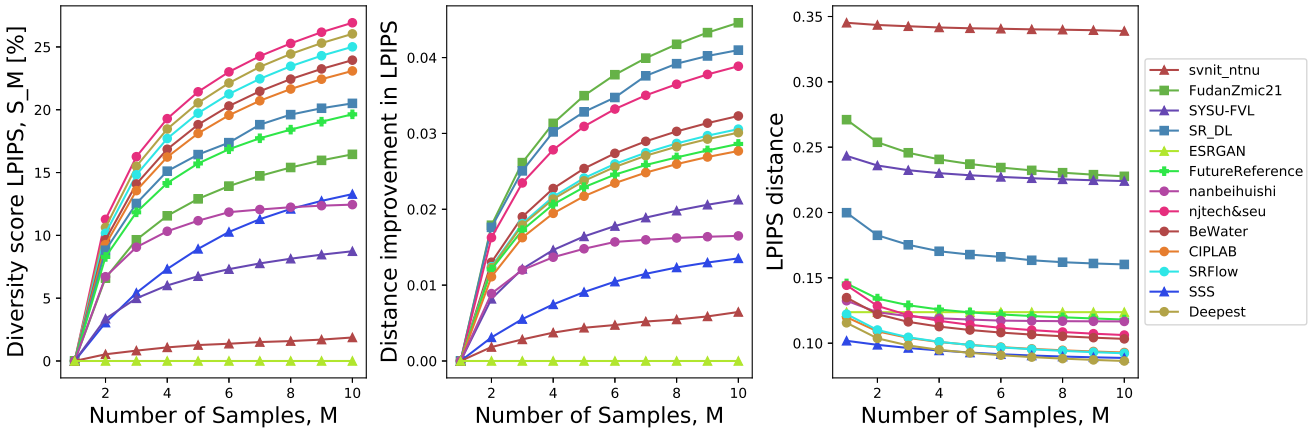

Figure 4. Visualization of improvement in LPIPS for $4\times$ by number of samples. Flow: Circle, VAE: Square, IMLE: Plus, GAN: Triangle

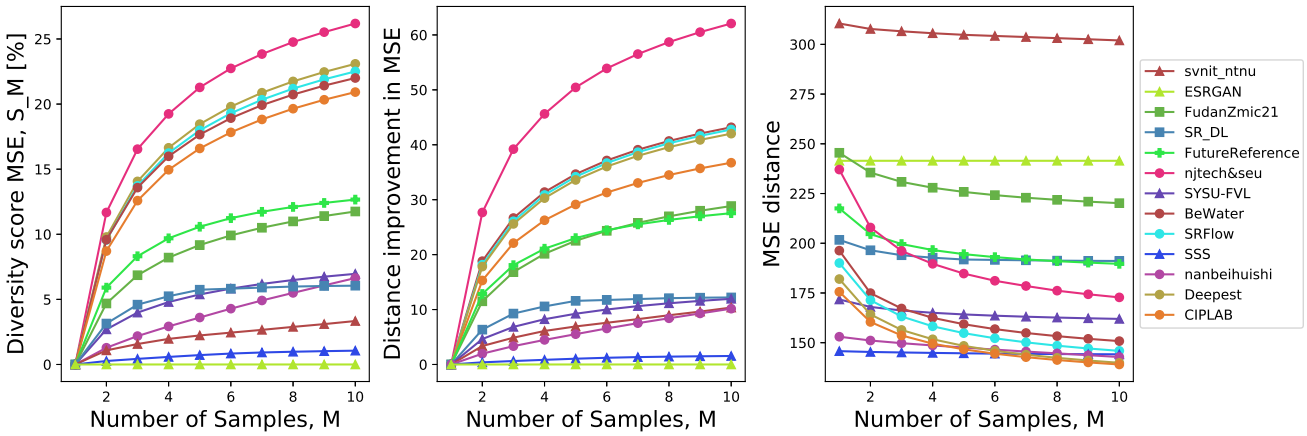

Figure 5. Visualization of improvement in MSE for $4\times$ by number of samples. Flow: Circle, VAE: Square, IMLE: Plus, GAN: Triangle

ing a purely GAN-based approach. Since such approaches are commonly deterministic and have a low LR-PSNR, they add modules to make it stochastic and LR consistent. To enable sampling for the network, they both add layers that inject randomness. The LR consistency is encouraged using the CEM module described in [8].

VAE-Based The teams FudanZmic21 and SR_DL used a VAE-Based approach. Similar to flow models, these approaches are inherently able to sample output images. Using VAE-Based method has the advantage over Flow-Based methods that the network components are not restricted to be bijective and having a tractable determinant of the Jacobian.

IMLE-based The team FutureReference is based on the implicit generative model [41] (IMLE). This method explicitly aims to cover all modes by reversing the direction in which generated samples are matched to real data.

4.3. Discussion

Here we present the results for both $4\times$ and $8\times$ super-resolution. All experiments were conducted on the DIV2k test set. The numerical results are shown in Tables 2 and 3 for $4\times$ and $8\times$ respectively. The user study is conducted for the 5 teams with the highest photorealism according to an initial analysis. The final ranking score (right column) is computed as the average of the team's rank in the diversity measure S_{10} and the MOR. For the team Deepest, which scored highest in the final ranking, we additionally show all ten submitted samples of a crop of a test image in Figure 19 and 20 for $4\times$ and $8\times$ respectively.

The team that performs best in the user study (MOR) in both tracks is Deepest. They improve SRFlow, by using the SoftFlow approach to mitigate the problems arising from the unbalanced information content in HR image and latent space. The better photo-realism is confirmed by the visual examples shown in Figures 2 and 3, where it has the highest level of details among the participating methods.

The team that performs best in Diversity Score in both

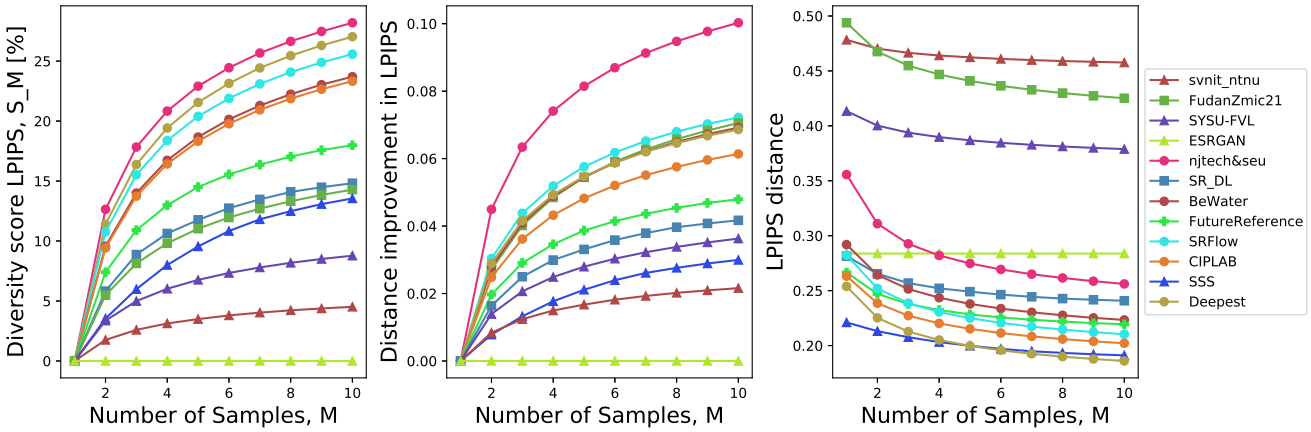

Figure 6. Visualization of improvement in LPIPS for $8\times$ by number of samples. Flow: Circle, VAE: Square, IMLE: Plus, GAN: Triangle

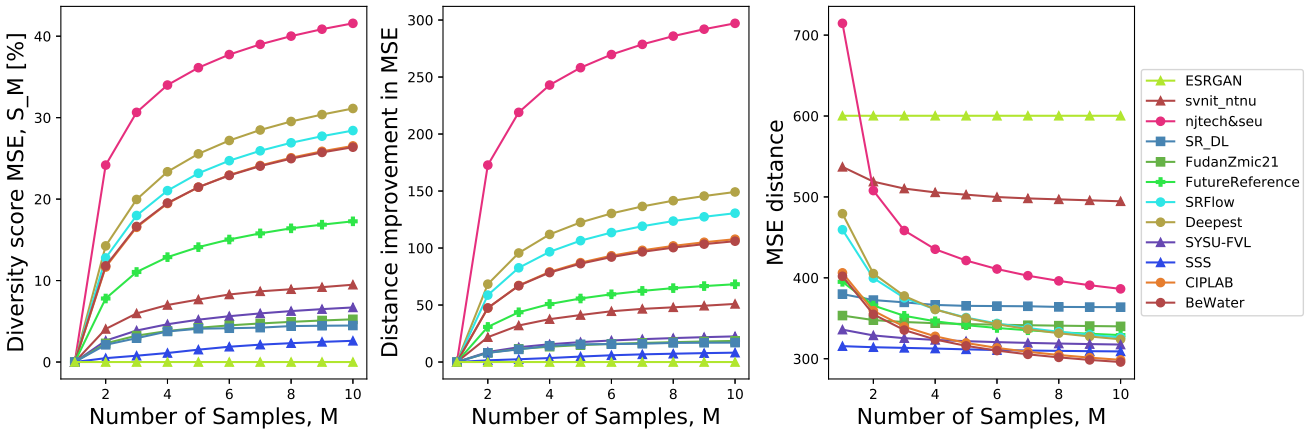

Figure 7. Visualization of improvement in MSE for $8\times$ by number of samples. Flow: Circle, VAE: Square, IMLE: Plus, GAN: Triangle

tracks is njtech&seu. This can be attributed to their multi-head attention module. However, for $8\times$ it fails to reach the LR-PSNR threshold set in the challenge description. For $4\times$, this method scores significantly worse in the user study compared to Deepest, which has the second rank in terms of Diversity Score. Notably, Deepest is the only method that outperforms the baseline SRFlow in terms of photo-realism and diversity on both scale levels. For $8\times$ SR, Deepest, CIPLAB, and SRFlow achieve very similar user scores.

While lagging behind in the $4\times$ case, ESRGAN interestingly achieves the best MOR for $8\times$ SR. On the other hand, ESRGAN only obtains an LR-PSNR of 30.65dB in this setting, which is far lower than the challenge goal of 45dB. Regarding LR-PSNR 7 of 11 methods in Track $4\times$ and 5 of 10 methods for $8\times$ reached the 45dB threshold. All methods that used the CEM [8] module or that are based on SRFlow [40] satisfied this criterion. In general, the VAE-based methods FudanZmic21 and SR_DL do not reach the SRFlow [40] baseline in terms of LPIPS and Diversity Score. Moreover, the GAN-based methods SR_DL, SSS, svnit_ntnu and SYSU-FVL obtain substantially lower diversity scores compared to the Flow-based competitors. This can indicate a higher susceptibility to mode collapse, which is a well-known problem in conditional GANs.

Under the assumption that the GT image is only one plausible HR image that corresponds to an LR image, ideal stochastic SR methods could come arbitrarily close to the GT for a sufficiently large number of samples. To visualize this effect for the participating methods, we show how close the SR images comes to the GT when increasing the number of samples. In Figures 4 and 6 we show this effect using LPIPS as the distance metric d. In Figures 5 and 7 we use MSE as the distance metric d. Each figure contains three plots to present the following aspects of the diversity. The plot on the right side depicts the locally best LPIPS or MSE, *i.e.* the right term in (1). To remove effects from the ordering of the submitted samples we first sort the samples corresponding to one GT image according to their best global LPIPS or MSE. For the LPIPS setting we calculate the local metric by using the dense pixel-wise distance and for MSE we use a patch size of $N = 16$.

To better visualize how much the method improves by

sampling more images, we show the absolute improvement compared to the reference distance (2) when using M samples in the middle figure. Since it is much more difficult to improve a method that already has a low LPIPS or MSE, they would be disadvantaged in this setting. To mitigate this unfair advantage, compute the final diversity score (1) relative to the reference distance (2) by dividing with it. The final diversity score (1) for different number of samples M are shown in the plots on the left.

The methods based on SRFlow, marked with a circle, are in a distinct group on top of the Diversity Score for both scale factors and metrics. The IMLE based method FutureReference, marked with a plus, is in the middle field for all scales and metrics. Methods that are based on VAEs are in the middle field as well, marked with a square. The GAN-Based methods are based on deterministic approaches that were made stochastic by injecting randomness. They are in the lower Diversity Score section, marked with a triangle. The baseline ESRGAN has diversity zero since it is deterministic.

5. Teams

5.1. Deepest: Noise Conditioned Flow Model for Learning the Super-Resolution Space

This method is based on SoftFlow [31] and SRFlow [40]. With the use of SoftFlow they alleviate the the problem of unbalanced information content in HR image and latent space. The key idea of SoftFlow is to add noise that is obtained from randomly selected distribution and to use these distribution parameters as conditions. [31] has shown that these methods can experimentally succeed in capturing the innate structure of manifold data. They show that in this same principle, they can increase performance on SR tasks using Flow models through adding noise and noise (distribution parameters) condition training. The difference from SRFlow [40] is that the proposed model adds the Noise Conditional Layer (NCL) to the flow step. The NCL is added to all levels in SRFlow, except the finest level, where the NCL tended to generate artifacts. They add noise to the data, *i.e.* the high-resolution image, and create a conditional layer for this noise distribution as depicted in Figure 8. They conducted noise condition training in two ways, one for noise itself and one for standard deviation for noise distribution. They proceed both methods in a similar way to the conditional affine coupling of [40]. Although standard deviation conditional training, such as those used in [31], improved diversity and LPIPS, it tended to create artifact from the generated images. In contrast, with noise conditional training, the numerical performance was slightly lower, but the number of artifacts occurring in the generated images was reduced and they finally applied noise condition training. Only the negative log-likelihood was used for loss, as in [40]. 498 additional images was used for training. Their method is as follows. Initially, random value c is obtained from uniform distribution $U(0, M)$ as [31] did. Next, set noise distribution $N(0, \Sigma)$, where $\Sigma = c^2 I$. Then, we sample noise vector v from $N(0, \Sigma)$ and add noise to the original high resolution image x to obtain perturbed data x^+. Finally, resize these vector v to get noise vector w for low-resolution images and obtain y^+ by adding w to the original low resolution image y. During inference, we add a zero vector instead of noise. Thus, the approach learns a flow network $f(z|y, v)$ that, given the noise vector v and LR image y predicts an HR image $x = f(z|y, v)$ from a random latent variable z. Details about this method can be found in [33]

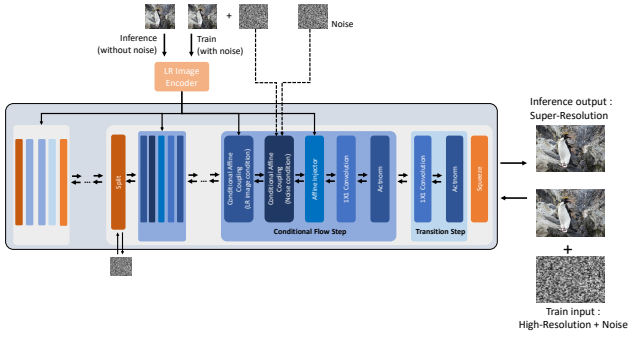

Figure 8. Method of Team Deepest.

5.2. CIPLAB: SRFlow-DA

This method is based on SRFlow [40]. To increase the receptive field, this method increases the depth of the non-invertible networks that calculate the mappings for the affine couplings as shown in Figure 9. They stack six 3×3 convolutional layers followed by ReLU activation except for the last convolutional layer, and its receptive field is 13×13.

SRFlow uses 3 and 4 levels multi-scale architecture with 16 flow steps for each scale, for ×4 and ×8 SR respectively. From the default SRFlow setting, they reduce the multi-scale levels to 2 and 3, for ×4 and ×8 SR respectively. In addition, they reduce the number of flow steps from 16 to 6. The proposed method SRFlow-DA (Deep convolutional block in the Affine layers) reduces the total number of parameters of the original SRFlow model and can be trained on a single GPU (<11GB). Details about this method can be found in [28]

5.3. BeWater: SRFlow with Respective Field Block

This method is based on SRFlow [40] and improves the LR encoding and the affine couplings. First, they replaces the RRDB LR encoding network with the RRFDB encoder [47]. The overall structure is shown in Figure 10. Secondly, in SRFlow, the scale and shift used in Affine In-

jector and Affine Coupling are predicted in one network. By contrast, they use two separate networks for more precise predictions. This method uses the additional 2650 images from Flickr2K [3].

5.4. njtech&seu: Learning Spatial Attention with Normalization Flow for Image Super-Resolution

This method proposes a Flow-based Pixel Attention Network to establish the spatial relationships between pixels, thereby increasing the realism of super-resolution images. As shown in Figure 11, the proposed network consists of three parts: the RRDB block, the multi-head attention module and the normalization flow module.

First, they employ a CNN-based architecture named Residual-in-Residual Dense Blocks (RRDB) [56] to extract the rich information in the low-resolution image. The introduced RRDB block has a series of convolutions with the same kernel size, and residual connections are adopt to fuse the features of different convolutional layers.

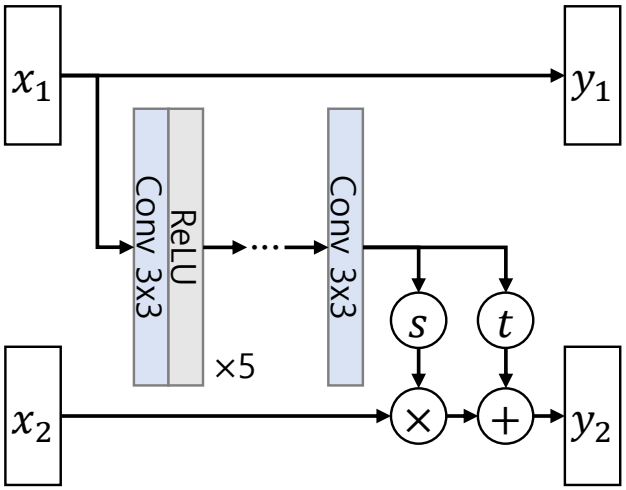

Figure 9. Method of Team CIPLAB.

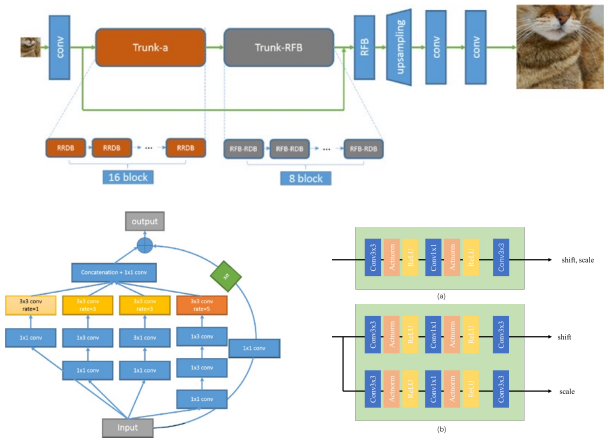

Figure 10. Method of Team BeWater.

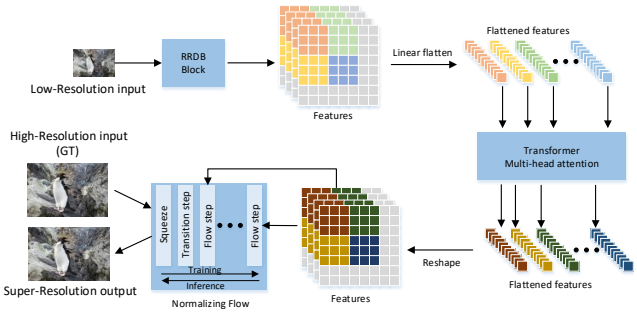

Figure 11. Method of Team njtech&seu.

Second, a multi-head attention module is proposed to learn the spatial pixel-level relations of the low-resolution image. Since the real-world images have many areas with rich texture details, the deep network may lose subtle clues when extracting features. Therefore, some super-resolution images tend to be blurred, distorted, etc. To generate more realistic image, they establish the spatial relationships between pixels. Specifically, each $width \times height \times channel$ patch is compressed into $1 \times (width*height) \times channel$, and the module learns the relation between pixels across channels.

Third, to tackle the ill-posed problem of super-resolution, they adopt the SRFlow network [40] as the normalization flow module in Figure 11. It can learn to predict diverse photo-realistic high-resolution images.

5.5. SSS: Flexible SR using Conditional Objective

The generator of this method consists of two streams, an SR branch and a condition branch as shown in Figure 12. The SR branch is built with basic blocks consisting of Residual in Residual Dense Block (RRDB) [56] equipped with the SFT layers [55]. Since most of the existing methods calculate perceptual losses on an entire image in the same feature space, the results tend to be monotonous and unnatural. For this reason, they define a style control map that is fed to the SR network at the inference phase to explore various pixel-wise HR solutions. During training, they optimize an SR model with a conditional objective, which is a weighted sum of multiple perceptual losses at different feature levels. During inference the style control map is used to generate a stochastic output.

5.6. SR_DL: Variational AutoEncoder for Image Super-Resolution

This method proposes a reference based image super-resolution model. As shown in Figure 13, the approach takes arbitrary references $\mathbf{R}$ and LR images $\mathbf{X}$ for training and testing. It consists of three components the VGG Encoder, the CVAE, and the image decoder. The VGG

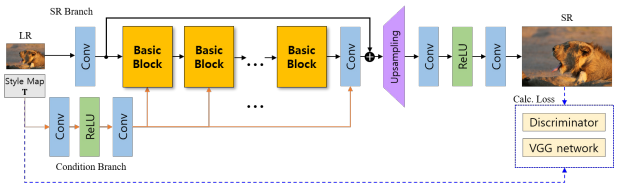

Figure 12. Method of Team SSS.

Enocder is based on the fully convolutional part of the VGG-16 network. They directly use pre-trained VGG-16 to extract feature maps for references (F_R) and bicubic upsampled LR images (F_X). A Conditional Variational AutoEncoder (CVAE) then encodes the reference feature maps to a latent space to learn the hidden distribution. The Feature Decoder learns to transfer the reference features as conditions C_R for LR feature maps. In order to have a flexible control over the LR feature maps, they use a convolution block to learn the mean and variance for the LR feature maps as F_μ and F_σ. They then have the conditioned feature maps as $F_{X|R} = C_R \cdot (1 + F_\sigma) + F_\mu$. Finally, the Image Decoder learns to reconstruct the conditioned feature maps to the SR image $\mathbf{Y}'$. The image decoder is similar to the VGG Encoder which followed by 3 layers of convolution with simple bilinear interpolation.

During training, they encourage the model to use reference features for super-resolution. They adopt the style and content losses from style transfer [29, 43] to align the statistics of feature maps between SR $\mathbf{Y}'$ and HR $\mathbf{Y}$ images. They use pretrained VGG-19 to extract intermediate feature maps for content loss as,

$$L_{\text{content}} = \left\|\phi^{4\text{-}1}_{\mathbf{Y}} - \phi^{4\text{-}1}_{\mathbf{Y}'}\right\|^1 + \|\mathbf{X} - D(\mathbf{Y}', \alpha)\|^1 + \|\mathbf{Y} - \mathbf{Y}'\|^1 + \|\mathbf{Lap}(\mathbf{Y}) - \mathbf{Lap}(\mathbf{Y}')\|^1 \quad . \tag{1}$$

where $\phi(\cdot)^{4\text{-}1}$ is the feature map on *relu4_1* layer. They also include L_1 loss between SR and HR image pairs. For $\alpha\times$ super-resolution, after upsampling, they also include the downsampling loss $D(\cdot, \alpha)$ to calculate the loss between original and estimated LR images. Meanwhile, they also use Laplacian loss [10] to calculate the structural loss between HR and SR image to pursue structural similarity.

The style loss is calculated by using *relu1_2, relu2_2, relu3_4, relu4_1*-th feature maps from VGG-19 network. Similarly to [29, 43], they align the statistics between SR and HR feature maps using mean and variance as,

$$L_{\text{style}} = \sum_i \left\|\mu(\phi^i(\mathbf{R})) - \mu(\phi^i(\mathbf{Y}'))\right\|^1 + \left\|\sigma(\phi^i(\mathbf{R})) - \sigma(\phi^i(\mathbf{Y}'))\right\|^1 \quad . \tag{2}$$

For the KL divergence, they learn the lower bound of the hidden distribution $N(\mu, \sigma))$ as

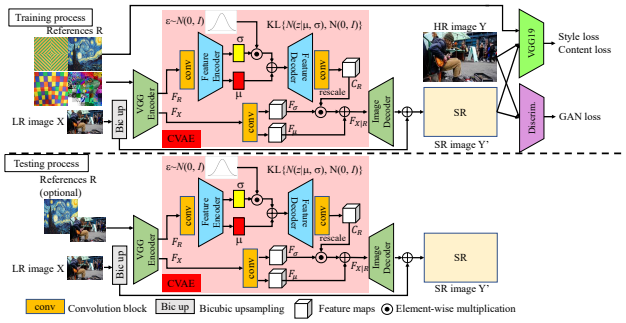

Figure 13. Method of Team SR_DL.

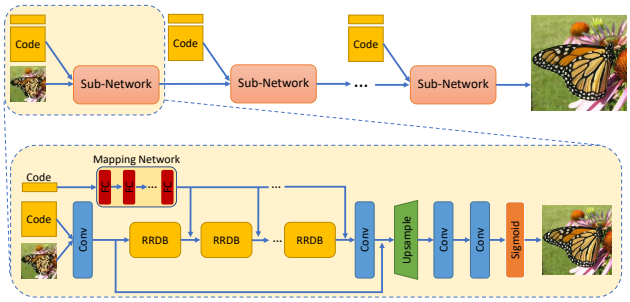

Figure 14. Method of Team FutureReference.

$L_{\text{KL}} = KL(N(0, I)||N(\mu, \sigma))$. They also has a discriminator to supervise the spatial correlation between HR and SR images. The GAN loss is defined as $\log(1 - D(\mathbf{Y}'))$ During inference, the reference image is optional. It can be any external images or the bicubic upsampled LR itself. If no reference is used, a random map $R \sim N(0, I)$ will be computed for super-resolution. Details about this method can be found in [38]

5.7. FutureReference: Generating Unobserved Alternatives

The FutureReference team formulate the one-to-many SR problem as training an implicit generative model [41]. More precisely, the predicted SR image is given by $\mathbf{y} = T_\theta(\mathbf{x}, \mathbf{z})$, where $\mathbf{x}$ is the input LR image and $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ is a random latent variable. Such a model can be trained as a conditional GAN, where $T_\theta(\cdot, \cdot)$ is interpreted as the generator. In practice, due to mode collapse, some valid predictions cannot be produced by the generator. This problem is exacerbated in the presently considered setting with one-to-one supervision, which leads to all samples of the generator conditioned on the same input $\mathbf{x}$ being identical and the random variable $\mathbf{z}$ is effectively ignored. To obtain non-deterministic predictions $\mathbf{y}$ despite the availability of only a single observation, we propose training the model using Implicit Maximum Likelihood Estimation (IMLE), which avoids mode collapse.

Compared to GANs, IMLE explicitly aims to cover all

modes by reversing the direction in which generated samples are matched to real data. Rather than making each generated sample similar to some real data point, it makes sure each real data point has a similar generated sample. IMLE can be further extended to model conditional distributions by separately applying IMLE to each member of a family of distributions $\{p(\mathbf{y}|\mathbf{x}_i)\}_{i=1}^n$. The denote the generator as $T_\theta(\cdot,\cdot)$, which takes in an input $\mathbf{x}_i$ and a random code $\mathbf{z}_{i,j}$ and outputs a sample from $p(\cdot|\mathbf{x}_i)$, the method optimizes the following objective:

$$\min_\theta \mathbb{E}_{\mathbf{z}_{1,1},\ldots,\mathbf{z}_{n,m}\sim\mathcal{N}(0,\mathbf{I})}\left[\sum_{i=1}^{n}\min_{j\in\{1,\ldots,m\}} d(T_\theta(\mathbf{x}_i,\mathbf{z}_{i,j}),\mathbf{y}_i)\right],$$

where $\mathbf{y}_i$ is the observed output that corresponds to $\mathbf{x}_i$, $d(\cdot,\cdot)$ is a distance metric and m is a hyperparameter. They use LPIPS perceptual distance [62] as the distance metric.

The proposed architecture relies on a backbone consisting of two branches. The first branch mainly consists of a sequence of residual-in-residual dense blocks (RRDB) [56], which is a sequence of three dense blocks connected by residual connections. The number of RRDB blocks are reduced by a factor of 4 and substantially expanded the number of channels compared to ESRGAN [56]. The second branch consists of a mapping network [30] produces a scaling factor and an offset for each of the feature channels after each RRDB in the first branch. Additionally they added weight normalization [46] to all convolution layers.

They adopt an approach of progressive upscaling, where they upscale the image by 2 times at a time. They chain together several backbone architectures which become subnetworks in a larger architecture, as shown in Figure 14. Each sub-network takes a latent code and the output of the previous sub-network, or if there is no previous subnetwork, the input image. They add intermediate supervision to the output of each sub-network, so that the distance metric in IMLE is chosen to be the sum over LPIPS distances between the output of each sub-network and the original image downsampled to the same resolution.

5.8. FudanZmic21: VSpSR: Explorable Super-Resolution via Variational Sparse Representation

This method combines a deterministic and a stochastic model inspired by Conditional Variational AutoEncoder (CVAE) [49]. Their stochastic model, called variational sparse representation guided explorable module VSPM, has a basis and a coefficient branch as shown in Figure 15. To improve the LR-consistency they employ a Consistency Enforcing Module (CEM) similar to [7]. Details about this method can be found in [48]

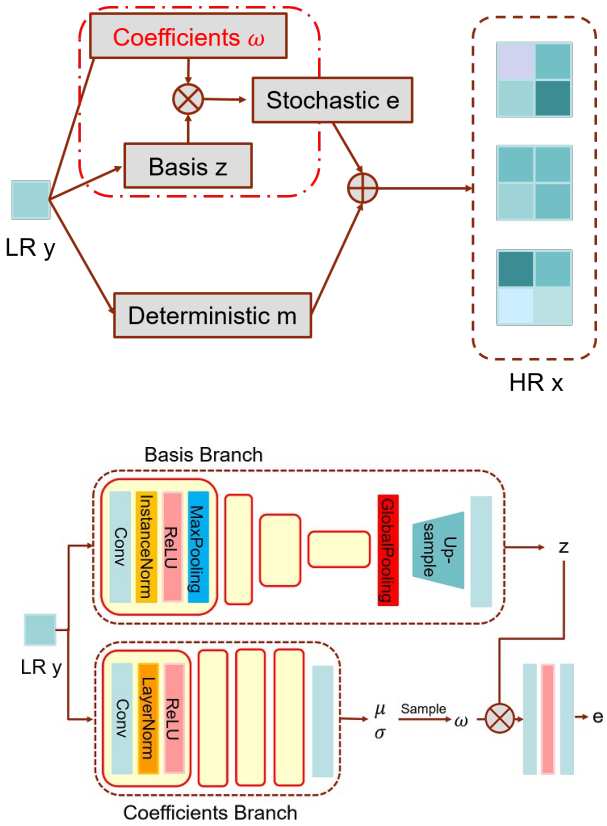

Figure 15. Method of Team FudanZmic21.

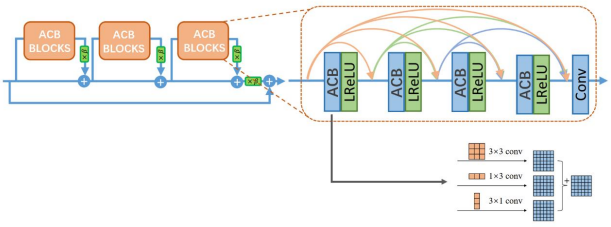

Figure 16. Method of Team nanbeihuishi.

5.9. nanbeihuishi: Modified Encoder in SRFlow via Asymmetric Convolution Blocks

This method is based on SRFlow [40] and replaces the filters in the RRDB network with Asymmetric Convolution Block (ACB) [13]. Their method is depicted in Figure 16 This method only took part in the $4\times$ Track.

5.10. SYSU-FVL

This method uses the enforcing module (CEM) [8] with LPIPS [62] loss and Quality Network loss that estimates the MOS during training. The generative network is based on the hierarchical ResNet structure [23].

The proposed generator, as shown in Figure 17, consists

of a series of residual blocks with upsampling layers. In the residual blocks of the generator, they inject noise and information through multiple normalization layers. The generator is then wrapped by CEM module to enforce low-resolution consistency. The adversarial loss follows the patch GAN approach with Hinge loss. Architecturally, there are 4 convolutional and spectral instance normalization layers process the input, with leaky ReLU as the activation function, which is the same as in the DeepSEE [11].

5.11. svnit_ntnu: Learning Multiple Solutions for Super-Resolution based on Auto-Encoder and Generative Adversarial Network

The proposed AEGAN method, is a modification of the MUNIT [25] approach for generating stochastic SR solutions. This method first extracts a content and style encoding of the LR image using two separate encoders. Those generated features are then decoded to generate SR images. The network architecture of the encoder and decoder are depicted in Figure 18. To encourage the decoder to invert the encoder network, the generated SR image is passed through the encoder by applying down-scaling operator. To make the method stochastic, they sample randomly drawn style features.

Acknowledgements

We thank the NTIRE 2021 sponsors: Huawei, Facebook Reality Labs, Wright Brothers Institute, MediaTek, and ETH Zurich (Computer Vision Lab).

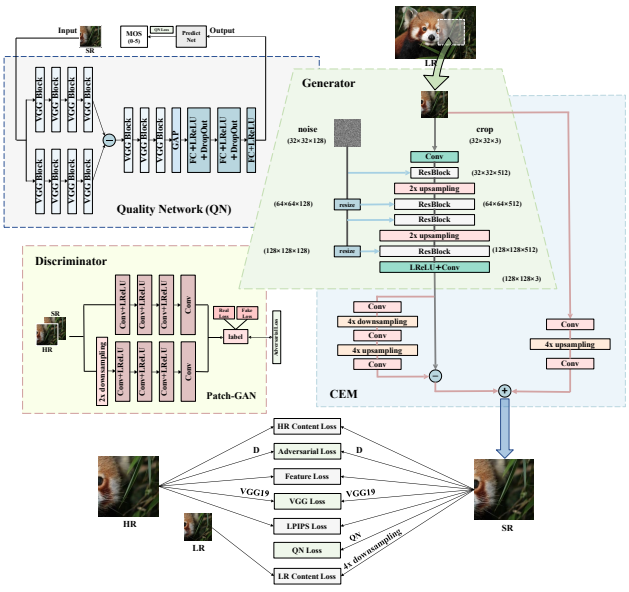

Figure 17. Method of Team SYSU-FVL.

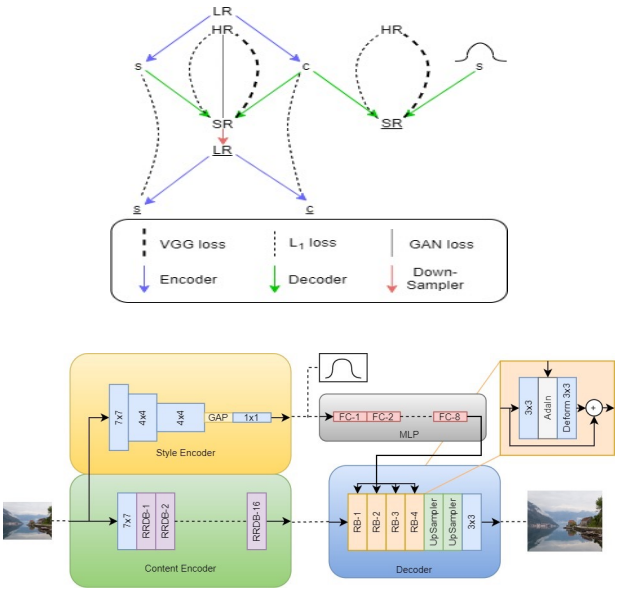

Figure 18. Method of Team svnit_ntnu.

Appendix A. Teams and affiliations

NTIRE 2021 challenge organizers

Members:
Andreas Lugmayr (andreas.lugmayr@vision.ee.ethz.ch)
Martin Danelljan (martin.danelljan@vision.ee.ethz.ch)
Radu Timofte (radu.timofte@vision.ee.ethz.ch)
Affiliation: Computer Vision Lab, ETH Zurich

BeWater

Title: SRFlow with Respective Field Block
Team Leader: Sicheng, Pan
Members:
Laiyun, Gong, Zhejiang University
Zhen, Yuan, NetEaseYunXin
Qingrui, Han, NetEaseYunXin

CIPLAB

Title: SRFlow-DA
Team Leader: Younghyun, Jo
Members:
Younghyun, Jo, Yonsei University
Sejong, Yang, Yonsei University
Seon Joo, Kim, Yonsei University

Deepest

Title: Noise Conditioned Flow Model for Learning the Super-Resolution Space

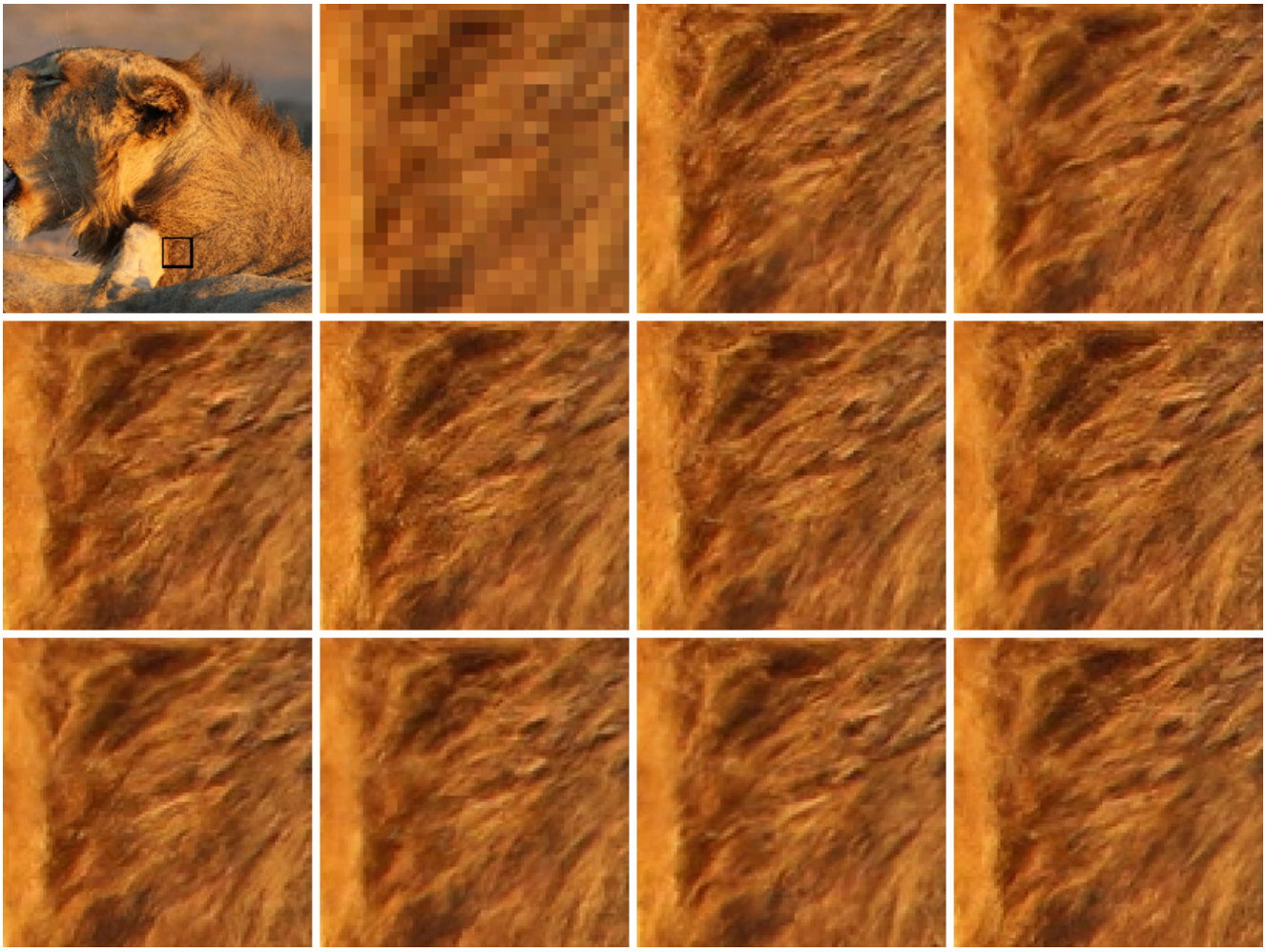

Figure 19. Visual example of diversity in super-resolution samples. The top left image is the input LR image, to the right is the ground truth and the ten remaining the samples from Deepest. ($4\times$)

Team Leader: Younggeun, Kim
Members:
Younggeun, Kim, Seoul National University
Seungjun, Lee, University of Ulsan College of Medicine, Asan Medical Center
Donghee, Son, Lomin Inc.

FudanZmic21

Title: VSpSR: Explorable Super-Resolution via Variational Sparse Representation
Team Leader: Xiahai, Zhuang
Members:
Shangqi, Gao, Fudan University
Hangqi, Zhou, Fudan University
Chao, Huang, Fudan University
Xiahai, Zhuang, Fudan University

FutureReference

Title: Generating Unobserved Alternatives
Team Leader: Shichong, Peng
Members:
Shichong, Peng, Simon Fraser University
Ke, Li, Simon Fraser University

SR_DL

Title: Variational AutoEncoder for Image Super-Resolution
Team Leader: Zhi-Song, Liu
Members:
Zhi-Song,Liu,Caritas Institute of Higher Education
Li-Wen,Wang,The Hong Kong Polytechnic University
Chu-Tak,Li,The Hong Kong Polytechnic University
Wan-Chi,Siu,The Hong Kong Polytechnic University

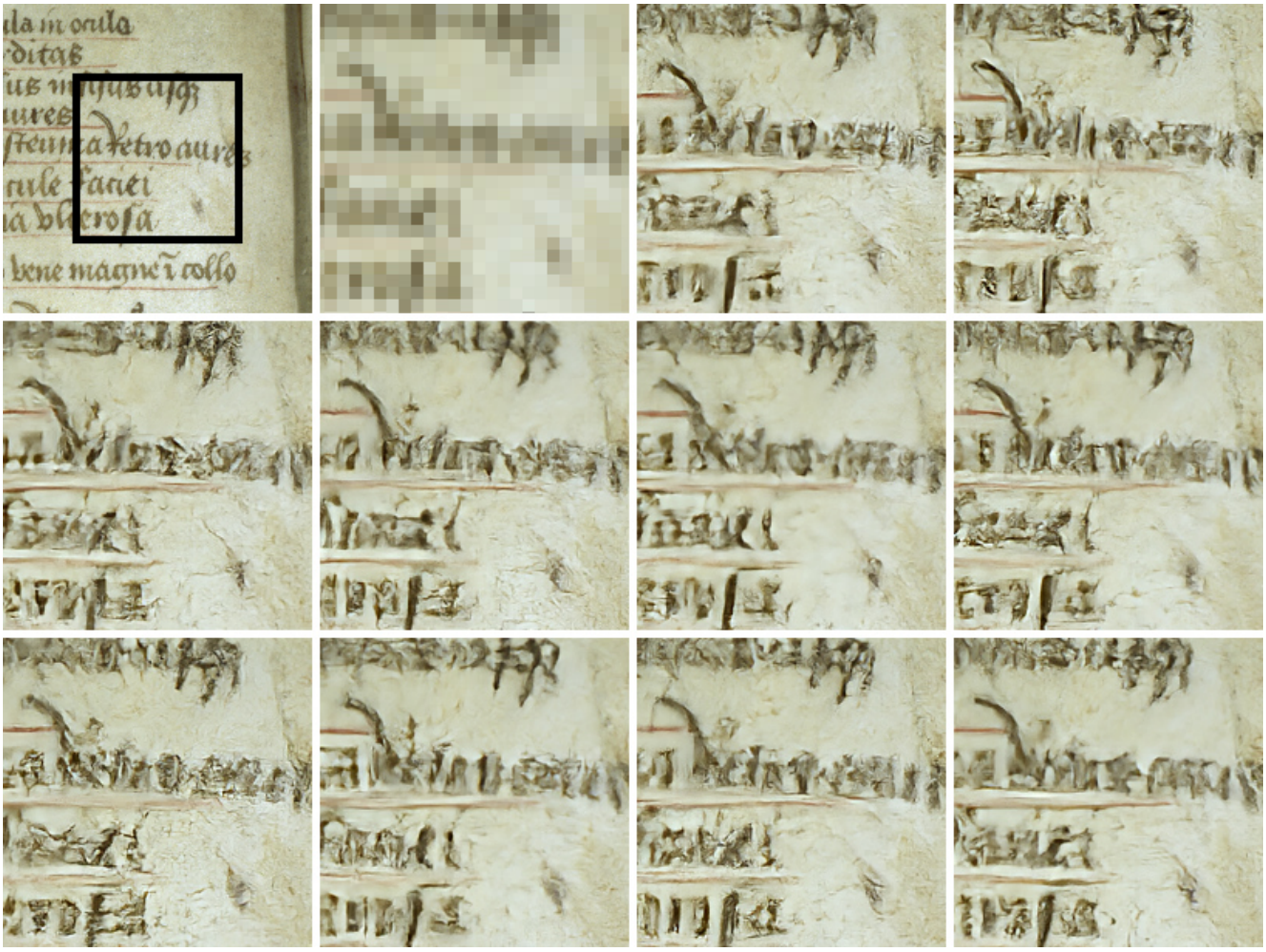

Figure 20. Visual example of diversity in super-resolution samples. The top left image is the input LR image, to the right is the ground truth and the ten remaining the samples from Deepest. ($8\times$)

SSS

Title: Flexible SR using Conditional Objective
Team Leader: Seung-Ho, Park
Members:
Seung-Ho, Park, Seoul National University

SYSU-FVL

Team Leader: Zhi, Jin
Members:
Youming, Liu, Sun Yat-sen University
Xinhua, Xu, Sun Yat-sen University
Yatian, Wang, Sun Yat-sen University
Liting, Zhang, Sun Yat-sen University
Haoran, Qi, Sun Yat-sen University
Huanrong, Zhang, Sun Yat-sen University
Zhi, Jin, Sun Yat-sen University

nanbeihuishi

Title: Modified Encoder in SRFlow via Asymmetric Convolution Blocks
Team Leader: Nan, Nan
Members:
Nan, Nan, North China University of Technology
Junkai, Zhang, University of Electronic Science and Technology of China
Chenghua, Li, CASIA
Ruipeng, Gang, NRTA
Ruixia, Song, NCUT
Yifan, Zhang, CASIA
Jian, Cheng, CASIA

njtech&seu

Title: Learning Spatial Attention with Normalization Flow for Image Super-Resolution

Team Leader: Wu, Qianyu
Members:
Qianyu, Wu, School of Computer Science and Technology, Nanjing Tech University, China
Aichun, Zhu, School of Computer Science and Technology, Nanjing Tech University, China
Yuchen, Lei, The Laboratory of Image Science and Technology, Southeast University, China
Jiaxin, Zou, The Laboratory of Image Science and Technology, Southeast University, China
Yang, Chen, The Laboratory of Image Science and Technology, Southeast University, China

svnit_ntnu

Title: Learning Multiple Solutions for Super-Resolution based on Auto-Encoder and Generative Adversarial Network
Team Leader: Kalpesh, Prajapati
Members:
Kalpesh, Prajapati, SVNIT
Vishal, Chudasama, SVNIT
Heena, Patel, SVNIT
Kishor, Upla, SVNIT
Kiran, Raja, NTNU
Raghavendra, Ramachandra, NTNU
Christoph, Busch, NTNU

References

[1] Abdullah Abuolaim, Radu Timofte, Michael S Brown, et al. NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017.

[3] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017.

[4] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, 2018.

[5] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Image super-resolution via progressive cascading residual network. In *CVPR*, 2018.

[6] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2021 nonhomogeneous dehazing challenge report. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[7] Yuval Bahat and Tomer Michaeli. Explorable super resolution. *CoRR*, abs/1912.01839, 2019.

[8] Yuval Bahat and Tomer Michaeli. Explorable super resolution. In *CVPR*, pages 2713–2722. IEEE, 2020.

[9] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 challenge on burst super-resolution: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[10] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv*, 2019.

[11] Marcel C. Bühler, Andrés Romero, and Radu Timofte. Deepsee: Deep disentangled semantic explorative extreme super-resolution. In *ACCV*, volume 12625 of *Lecture Notes in Computer Science*, pages 624–642. Springer, 2020.

[12] Dengxin Dai, Radu Timofte, and Luc Van Gool. Jointly optimized regressors for image super-resolution. *Comput. Graph. Forum*, 34(2):95–104, 2015.

[13] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. In *ICCV*, pages 1911–1920. IEEE, 2019.

[14] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. 2014.

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014.

[16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2016.

[17] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. NTIRE 2021 depth guided image relighting challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[18] Yuchen Fan, Honghui Shi, Jiahui Yu, Ding Liu, Wei Han, Haichao Yu, Zhangyang Wang, Xinchao Wang, and Thomas S Huang. Balanced two-stage residual networks for image super-resolution. In *CVPR*, 2017.

[19] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 2002.

[20] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[21] Shuhang Gu, Martin Danelljan, Radu Timofte, et al. Aim 2019 challenge on image extreme super-resolution: Methods and results. In *ICCV Workshops*, 2019.

[22] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[24] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015.

[25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[26] Yiwen Huang and Ming Qin. Densely connected high order residual network for single frame image super resolution. *arXiv preprint arXiv:1804.05902*, 2018.

[27] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP*, 1991.

[28] Younghyun Jo, Sejong Yang, and Seon Joo Kim. Srflow-da: Super-resolution using normalizing flow with deep convolutional block. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, 2016.

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[31] Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. Softflow: Probabilistic framework for normalizing flow on manifolds, 2020.

[32] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[33] Younggeun Kim and Donghee Son. Noise conditional flow model for learning the super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[34] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017.

[35] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*, 2017.

[36] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *CVPR*, 2017.

[37] Jerrick Liu, Oliver Nina, Radu Timofte, et al. NTIRE 2021 multi-modal aerial view object classification challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[38] Zhi-Song Liu, Wan-Chi Siu, and Li-Wen Wang. Variational autoencoder for reference based image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[39] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 learning the super-resolution space challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[40] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *ECCV*, pages 715–732. Springer, 2020.

[41] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[42] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on image deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[43] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[44] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 2003.

[45] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Aleš Leonardis, Radu Timofte, et al. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[46] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *ArXiv*, abs/1602.07868, 2016.

[47] Taizhang Shang, Qiuju Dai, Shengchen Zhu, Tong Yang, and Yandong Guo. Perceptual extreme super-resolution network with receptive field block. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 440–441, 2020.

[48] Gao Shangqi, Zhou Hangqi, Huang Chao, and Zhuang Xiahai. Vspsr: Explorable super-resolution via variational sparse representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[49] K. Sohn, X. Yan, H. Lee, and A. Arbor. Learning structured output representation using deep conditional generative models. In *International Conference on Neural Information Processing Systems*, 2015.

[50] Sanghyun Son, Suyoung Lee, Seungjun Nah, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on video super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[51] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. In *ICCP*, 2012.

[52] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *ACCV*, pages 111–126. Springer, 2014.

[53] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *CVPR*, pages 1865–1873. IEEE Computer Society, 2016.

[54] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, pages 1920–1927, 2013.

[55] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. *CVPR*, 2018.

[56] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks. *ECCV*, 2018.

[57] Chih-Yuan Yang and Ming-Hsuan Yang. Fast direct super-resolution by simple functions. In *ICCV*, pages 561–568, 2013.

[58] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *CVPR*, 2008.

[59] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Processing*, 19(11):2861–2873, 2010.

[60] Ren Yang, Radu Timofte, et al. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[61] Xin Yu and Fatih Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, pages 318–333, 2016.

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018.