This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# FS-NCSR: Increasing Diversity of the Super-Resolution Space via Frequency Separation and Noise-Conditioned Normalizing Flow

Ki-Ung Song [1, 3*] Dongseok Shim[1, 3*] Kang-wook Kim[1, 3*]
Jae-young Lee[1, 3] Younggeun Kim[1, 2, 3]
[1] Seoul National University, Seoul, Republic of Korea
[2]MINDsLab Inc., Gyeonggi, Republic of Korea
[3]Deepest, Seoul, Republic of Korea
{sk851, tlaehdtjr01, full324, jerry96, eyfydsyd97}@snu.ac.kr

## Abstract

*Super-resolution suffers from an innate ill-posed problem that a single low-resolution (LR) image can be from multiple high-resolution (HR) images. Recent studies on the flow-based algorithm solve this ill-posedness by learning the super-resolution space and predicting diverse HR outputs. Unfortunately, the diversity of the super-resolution outputs is still unsatisfactory, and the outputs from the flow-based model usually suffer from undesired artifacts which causes low-quality outputs. In this paper, we propose FS-NCSR which produces diverse and high-quality super-resolution outputs using frequency separation and noise conditioning compared to the existing flow-based approaches. As the sharpness and high-quality detail of the image rely on its high-frequency information, FS-NCSR only estimates the high-frequency information of the high-resolution outputs without redundant low-frequency components. Through this, FS-NCSR significantly improves the diversity score without significant image quality degradation compared to the NCSR, the winner of the previous NTIRE 2021 challenge.*

## 1. Introduction

Single image super-resolution (SISR), the task that restores low-resolution (LR) images to high-resolution (HR) images, is an active research topic that can be utilized in several applications such as surveillance [41], medical and astronomical image processing [2, 18, 26].

Early SISR approaches [7, 12, 20, 39, 40] focus on generating a single high-quality output for a given input LR image by improving Peak Signal-to-Noise Ratio (PSNR) ratio between the input LR images and predicted HR outputs. Since those studies utilize $L_1$ or $L_2$ loss between the generated and ground-truth HR images, they suffer from an over-smoothing problem. Alternative to PSNR-oriented models, GAN-based methods [17, 33] are proposed to generate photo-realistic super-resolved images.

Unfortunately, multiple possible HR images exist for a single LR image and the aforementioned deterministic models which improve the image quality of a single output cannot solve this ill-posed nature of the super resolution. SRFlow [23] learns the distribution of the HR image consistent for the given LR images and predicts diverse HR images to improve the high photo-realism, diversity, and the LR consistency at once. Following, NCSR [13] adopts noise-conditioned layers suggested in SoftFlow [11] and HCFlow [19] proposes hierarchical conditional flow for the diversity and the higher image quality. However, flow-based models usually generate undesired artifacts in HR outputs which leads to lower image quality and the diversity of the outputs are not improved significantly compared to SRFlow.

We observe that the super-resolution models predict the missing high-frequency information of the HR images from the given LR image which takes part in generating the diverse details of the HR images such as the shape of the foliage and the direction of the fur. Previous super-resolution models [13, 23] predict not only high-frequency information, but also low-frequency information of the HR images. It leads to inefficient training and these models have difficulty in increasing the diversity and the image quality of the super-resolution outputs.

In this paper, we propose **FS-NCSR** (Frequency-Separated Noise-Conditioned Normalizing Flow for Super-Resolution) which applies frequency separation to NCSR. We reconstruct the low-frequency information of the HR outputs by upsampling LR images in bicubic without any learnable parameters and predict the high-frequency infor-

[*] Equal contribution
This work was done as a project of Deepest

mation by training flow-based model. By doing so, we increase the diversity of learned super-resolution space in both ×4 and ×8 settings and improve the super-resolution quality by reducing the number of the artifact. Our contributions can be summarized as follows:

- We propose a flow-based algorithm for high-quality diverse super-resolution output using noise-conditioned affine coupling and frequency separation.
- By filtering low-resolution information of the target image, the generative model focuses on producing high-frequency outputs and improves super-resolution quality.
- We expand the filtered input data distribution by adding noise to the sparse high-frequency image for the output diversity.

## 2. Related Works

### 2.1. Single Image Super Resolution

Super-resolution has been studied long in computer vision fields. Before deep learning-based methods have been applied, sparsed coding [4, 29, 36, 37] and local linear regression [31, 32, 35] have been highly applied. Many deep learning-based methods have been approached for SISR, since SRCNN [7] which exploited CNN layers and L1 Loss. After SRCNN was proposed, many variations have been suggested including [33]. But as CNN-based methods have relied on L1 or L2 loss, they have generated blurry images. GAN-based methods, which were first suggested by SRGAN [17], have shown improvements by employing adversarial loss and perceptual loss. Although GAN-based methods have generated images with good quality [17, 33], their diversities were so limited, thereby generating only one image.

### 2.2. Normalizing Flow

Flow-based models have been first proposed by [5] for modeling complex high dimensional density. As flow-based models learn the whole distribution, they have been widely used for mapping complex distributions given simple distribution, including Gaussian distribution. Invertible neural networks have been adopted to map complex distributions from simple distributions [5, 6, 15]. Flow-based models in the early days have not shown great improvements relative to GAN-based models. However, SRFlow [23], which adopts negative log-likelihood loss, showed improvements in image quality and diversities simultaneously. As SRFlow used negative log-likelihood loss, it could learn the whole distribution, which leads to generating much more diverse images than GAN-based methods. NCSR [13] has shown further improvements in terms of image quality and diversity, by providing networks with noises. [13] has proposed adding a conditional noise layer, which essentially resolves distribution discrepancy between simple data and complex data.

### 2.3. Frequency Separation

The study of frequency domain based on Fast Fourier Transform (FFT) algorithm [3] played a crucial role in traditional signal processing. In this perspective, before the era of deep learning, studying the frequency information was important in image restoration research. In this light, it is readily known that high-frequency information of the given image contributes greatly to its sharpness and high-quality detail. Therefore, we can say that a recent huge success of deep learning-based approaches in realistic images generation is due to the success of synthesizing high-frequency information of the desired images.

Therefore, in recent image restoration research including super-resolution, there exist approaches [34] in which low-frequency and high-frequency are separated and treated by a separate neural network, and approaches [30] in which an FFT-based layer is designed to better process information of the frequency domain. We observed that when the former approaches were combined with NCSR, instability of the NLL training of the flow-based model occurred. And in the case of the latter approaches, the existing FFT-based layers are not suitable for the flow-based approach due to their non-invertible nature.

## 3. Methods

Given a LR image, our goal is to learn a diverse super-resolution space corresponding to that image. From the perspective of the frequency domain, we propose a more efficient method to increase the diversity of learned space. In this section, we introduce our point of view and proposed method. We begin with a brief background related to our work.

### 3.1. Background

Various model frameworks (*e.g.* Generative Adversarial Networks [8], Normalizing Flow [25], and Diffusion probabilistic models [10]) have been proposed in recent deep learning-based generative model research. And they show their respective strengths and weaknesses along with excellent performance. Among them, the flow-based model configures a mapping $f_\theta : X \to Z$ between the desired data distribution $X$ and latent space distribution $Z$ (*e.g.* Gaussian) through a series of invertible transformations. Such an invertible mapping architecture enables an explicit computation of negative log-likelihood (NLL) by the change of variable formula as:

$$-\log p_X(x) = -\log p_Z(f_\theta(x)) - \log|\det \frac{\partial f_\theta}{\partial x}(x)|. \quad (1)$$

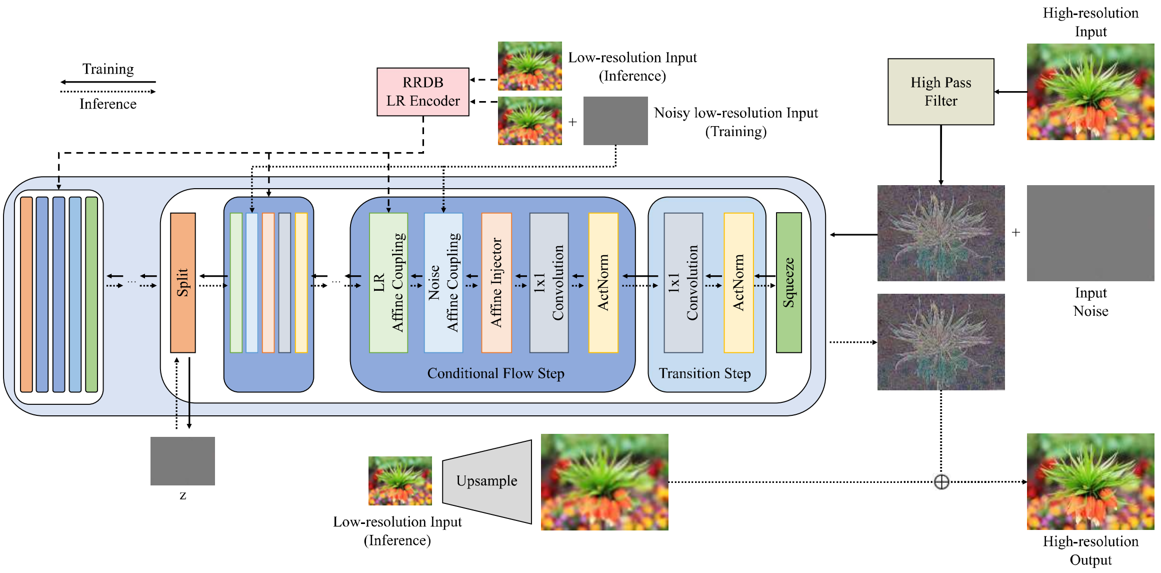


Figure 1. Algorithm overview. We propose a frequency separation on the target image and applies noise on high-frequency input with noise-conditioned coupling layers for diverse super-resolution outputs.

By minimizing NLL directly, it is widely known that the flow-based models show decent performance in mode coverage of the desired data distribution.

Based on this advantage of the flow-based approach, SRFlow [23] first showed that the flow-based modeling of the conditional distribution of the HR image can successfully learn super-resolution space corresponding to the given LR input. And one of its variants model, NCSR [13], proposed an additional noise-conditional layer to SRFlow to generate more diverse super-resolution outputs. Results of the previous works show that the ill-posedness of super-resolution can be solved from the perspective of super-resolution space learning. To take advantage of the flow-based model's good mode coverage performance, we propose a method to learn more diverse super-resolution space with NCSR architecture.

## 3.2. High-Frequency Information

There are various ways to configure a High-pass filter and Low-pass filter to separate high-frequency and low-frequency information. Without affecting the stability of NLL training of the flow-based model, we utilize the bicubic downsampling-upsampling process as the Low-pass filter, $L_s$, with a specific scale factor $s$. And the corresponding High-pass filter, $H_s$, computes the high-frequency information $x_{hf}$ of the given input by subtracting low-frequency information from the HR target $x$:

$$L_s(x) = ((x)_{s\downarrow})_{s\uparrow}, \quad H_s(x) = x_{hf} = x - L_s(x). \quad (2)$$

There are also other frequency separation methods. Some can configure $L_s$ and $H_s$ based on FFT and others can utilize the known 3x3 (or 5x5) kernel. In the former case, the filtering threshold level is an additional hyperparameter that is heavily dependent on an individual image. And in both cases, to match the low-frequency information of the LR input $y$ and $L_s(x)$, additional process such as the usage of a neural network is required leading to instability of NLL training.

By using this simple kind of High-pass filter, sparse high-frequency information can be efficiently obtained since we have the LR input as $y = (x)_{s\downarrow}$. And it leads to our proposed method which achieves efficient training without the need for additional memory or network compared to the previous flow-based approaches.

## 3.3. Overall Method

We propose FS-NCSR (Frequency Separating Noise-Conditioned Normalizing Flow for Super-Resolution), the generative model for super-resolution only produces the high-frequency information of the target HR image $x$ without redundant low-frequency information readily available from $y = (x)_{s\downarrow}$. Our overall model architecture is shown in Figure 1.

In the training process of the flow-based models, dequantization processes exist [9, 15] for better performance. As can be readily checked in Figure 2 and Table 3, the high-frequency information is relatively sparse compared to HR images. And training the model with this kind of information is difficult. In the previous work of NCSR, the idea of Softflow [11] was used by adding a different level of noise to the input instead of the dequantization process. This can be interpreted as an attempt to expand the modality of the desired data distribution's sparse region in the perspective of score matching [27, 28] which is in the spotlight of the generative model today. Therefore, we applied the same idea of Softflow [11] to deal with sparse information, and it was crucial in the training stability of the proposed method.

Now, with the same analog to the work of [13, 23], we can formulate the training process of our method as follows:

$$\begin{aligned} x_{hf}^{+} &= x_{hf} + v, \quad v \sim \mathcal{N}(0, \Sigma), \\ y^{+} &= y + w, \quad w \sim \mathcal{N}(0, \hat{\Sigma}), \\ z &= f_\theta(x_{hf}^{+} | y^{+}, v). \end{aligned} \tag{3}$$

where $w$ indicates noise resized to the same size as the LR input $y$. And also similar to [13, 23], we formulate the loss function only NLL $\mathcal{L}_{nll}$ as below,

$$\begin{aligned} \mathcal{L}_{nll} &= -\log p_X(x|y^{+}, v) \\ &= -\log p_Z(f_\theta(x; y^{+}, v)) - \log|\det \frac{\partial f_\theta}{\partial y}(x; y^{+}, v)|. \end{aligned} \tag{4}$$

The model trained in proposed method does not require additional cost in the inference stage compared to the previous approaches. Since the low-frequency information $L_s(x)$ is readily given by the LR input $y = (x)_{s\downarrow}$. The super-resolution output $\hat{x}$ is obtained by:

$$\begin{aligned} \hat{x} &= f_\theta^{-1}(z; y, v) + (y)_{s\uparrow} \\ &= f_\theta^{-1}(z; (x)_{s\downarrow}, v) + L_s(x). \end{aligned} \tag{5}$$

where $v$ is the random noise from the latent space $Z$.

In this perspective of frequency domain, super-resolution is the process of generating the corresponding high-frequency information since we have $f_\theta^{-1}(\cdot; (x)_{s\downarrow}) \approx H_s(x)$.

# 4. Experiments

## 4.1. Datasets

We utilize DF2K dataset, a merged dataset of DIV2K [1] and Flickr2K[1], for training and evaluation. DIV2K dataset consists of 800, 100, and 100 high-resolution images of train, validation, and test split, respectively. Flickr2K dataset comprises 2560 high-resolution images. The train split from the DIV2K dataset and the whole Flickr2K dataset are merged and used for training. We evaluate our model with the validation split of DIV2K dataset.

We try to increase the amount of training dataset by including crawled images from Unsplash website[2], but there are no performance improvements in the diversity and visual quality of super-resolved images. Thus, we do not include our crawled dataset in this research.

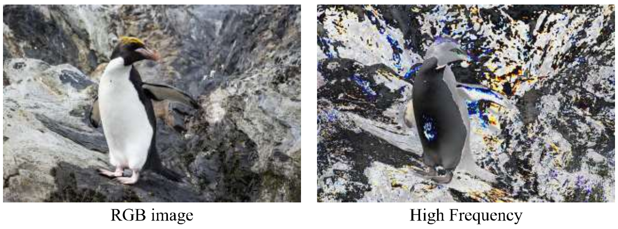


Figure 2. Full RGB vs. High-frequency information. High-frequency information is relatively sparse compared to its original RGB images. The high-frequency information was obtained based on ×128 scale for clear visual confirmation. For ×4 and ×8 cases, the high-frequency information is much sparser, making it difficult to see.

During training, we randomly crop 160×160 patches from original HR images and use them as HR samples. We obtain LR samples by downsampling HR patches and utilizing HR and LR patches as HR-LR pairs for training. The LR samples are downsampled via bicubic kernel. We train our model in RGB channels, and randomly apply horizontal flips and 90-degree rotation for data augmentation.

## 4.2. Training

We use the Adam optimizer [14] with $\beta_1$ = 0.9, $\beta_2$ = 0.99, $\epsilon$ = $10^{-8}$, and set the initial learning rate as $2 \times 10^{-4}$. Following [13], the learning rate is halved at 50%, 75%, 90%, and 95% of the total training steps. We train our network with a batch size of 16 on a V100 GPU. The ×4 network was trained at 180k steps and the ×8 network at 220k steps.

## 4.3. Evaluation

We evaluate our model and other baselines based on three criteria: photo-realism, diversity of super-resolution space, and image consistency on LR. We adopt LPIPS [38] to evaluate photo-realism, diversity score to evaluate diversity, and LR PSNR to evaluate LR consistency.

**LPIPS.** LPIPS is the distance between the super-resolved and the ground-truth HR image. The distance is measured on the feature space of AlexNet [16].

**Diversity Score.** To obtain meaningful diversity of models, Lugmayr *et al.* [21] proposed the diversity score. Let the ground-truth HR image $y$ and $y_k$ be the $k$-th patch of $y$. Generating $M$ samples from the super-resolution models, the $i$-th super-resolved images from the model is $\hat{y^i}$, and its $k$-th patch is $\hat{y_k^i}$, where $i \in \{1, 2, ..., M\}$. Than the diversity

[1] https://github.com/limbee/NTIRE2017

[2] https://unsplash.com

| Model | Diversity↑ | LPIPS↓ | LR PSNR↑ |
|---|---|---|---|
| RRDB [33] | 0 | 0.253 | 49.20 |
| ESRGAN [33] | 0 | 0.124 | 39.03 |
| ESRGAN+ [24] | 22.13 | 0.279 | 35.45 |
| SRFlow [23] | 25.26 | 0.120 | 49.97 |
| HCFlow [19] | 22.73 | **0.116** | 49.46 |
| NCSR [13] | 26.72 | 0.119 | **50.75** |
| **FS-NCSR (Ours)** | **29.44** | 0.127 | 49.31 |

Table 1. General image ×4 super-resolution results on the DIV2K validation set. We measure all the metrics with $M = 10$ samples for each HR image.

score $S_M$ can be computed as follows:

$$S_M = \frac{1}{\bar{d_M}} \left( \bar{d_M} - \frac{1}{K} \sum_{k=1}^{K} \min \left\{ d\left( y_k, \hat{y_k^i} \right) \right\}_{i=1}^{M} \right), \quad (6)$$

where minimum distance on a global sample, $\bar{d_M}$, defined as follows:

$$\bar{d_M} = \min \left\{ \frac{1}{K} \sum_{k=1}^{K} d\left( y_k, \hat{y_k^i} \right) \right\}_{i=1}^{M}. \quad (7)$$

We use LPIPS as distance function $d$, and set $M = 10$.

**LR PSNR.** In LR, the super-resolved output of the model must be consistent with the original LR input. Thus, we measure PSNR (Peak Signal-to-Noise Ratio) between downsampled super-resolved image and given input LR image.

### 4.4. Quantitative Results

We compare our model, FS-NCSR, with diverse baseline models: RRDB [33], ESRGAN [33], ESRGAN+ [24], SRFlow [21], HCFlow [19], and NCSR [13]. RRDB is the model trained with $L_1$ loss with ground-truth HR image, consequently oriented to minimizing PSNR. ESRGAN and ESRGAN+ are GAN-based methods that are the common baselines for photo-realistic super-resolution. RRDB and ESRGAN are deterministic models, so their diversity scores are zero. SRFlow, HCFlow, and NCSR are stochastic super-resolution models that can super-resolve diverse photo-realistic images from the given input LR image. For all the flow-based super-resolution models, the temperature is set to 0.9. However, the temperature is 0.85 for NCSR ×8 model.

We measure the diversity score, LPIPS, LR PSNR of our model and compare them with the reported results of other baselines. We evaluate all the models in ×4 super-resolution setting. As shown in Table 1, our proposed model, FS-NCSR, achieves the highest diversity score in ×4 setting. The diversity score of FS-NCSR is significantly higher than NCSR [13], which indicates frequency separation plays a key role to improve diversity. Although FS-NCSR achieves the lower LR PSNR and higher LPIPS than SRFlow [23], HCFlow [19] and NCSR, diversity increase is significant compared to such performance degradation so can be compensated. In addition, we observe that the number of artifacts and failure cases in the generated samples of FS-NCSR is less than that of NCSR. We will discuss this qualitative comparison in 4.5.

| Model | Diversity↑ | LPIPS↓ | LR PSNR↑ |
|---|---|---|---|
| RRDB [33] | 0 | 0.419 | 45.43 |
| ESRGAN [33] | 0 | 0.277 | 31.35 |
| SRFlow [23] | 25.31 | 0.272 | **50.00** |
| NCSR [13] | 26.8 | 0.278 | 44.55 |
| **FS-NCSR (Ours)** | **26.9** | **0.257** | 48.90 |

Table 2. General image ×8 super-resolution results on the DIV2K validation set. We measure all the metrics with $M = 10$ samples for each HR image.

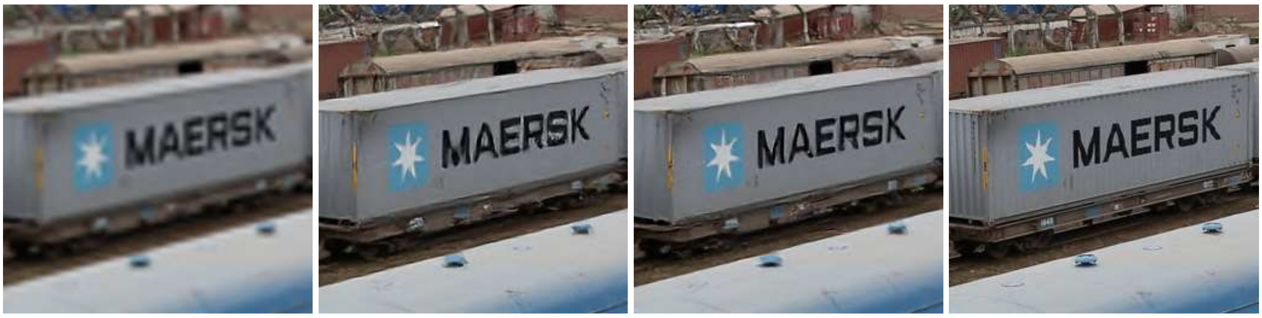
We also evaluate all the models except ESRGAN+ [24] in ×8 super-resolution setting. As presented in Table 2, FS-NCSR outperforms all the other methods in terms of diversity score and LPIPS. Also, FS-NCSR achieves comparable LR PSNR with SRFlow [23], the model which achieved the highest LR PSNR. These results show that FS-NCSR outperforms all the other methods in terms of photo-realism and diversity, and frequency separation is a decisive factor.

To clearly demonstrate the effect of frequency separation, we additionally report the metric trajectories during the training process of FS-NCSR and NCSR [13]. We measure LPIPS and diversity score in 150k, 160k, 170k, 180k steps for each model. The results of such models during the training process are presented in Figure 6. For trained weights of FS-NCSR, higher diversity and lower LPIPS than NCSR weights of the same iteration are measured. These results show that frequency separation consistently improves the diversity and photo-realism of the model output during the training process.

### 4.5. Qualitative Results

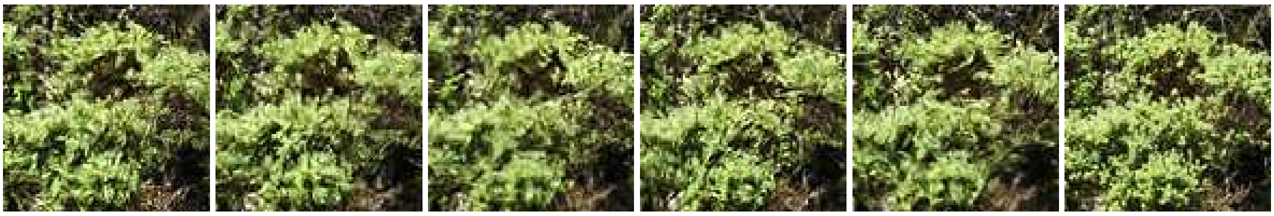
The qualitative result in Figure 4 shows that the direction and the degree of density of the leaves are slightly different for every 5 outputs. Thus, we can say that the proposed method not only shows a higher diversity score than previous approaches but also can generate outputs with diverse details that are distinguishable visually. It means that the frequency separation can enhance the high mode coverage performance of the flow-based model.

We now qualitatively compare our result with the output of NCSR to verify the effect of the frequency separation. As discussed in 4.4, the FS-NCSR's LPIPS was lower than the existing approaches. But Figure 3 shows that FS-NCSR



(a) ×4 LR (b) NCSR [13] (c) FS-NCSR (Ours) (d) Ground Truth

Figure 3. Qualitative results with comparison to NCSR on the DIV2K validation set for SR ×4 results. The cropped part of the ground truth is from 0850 from DIV2K. Each output of NCSR and FS-NCSR was chosen randomly from 10 generated outputs respectively.



(a) Output 1 (b) Output 2 (c) Output 3 (d) Output 4 (e) Output 5 (f) Ground Truth

Figure 4. Qualitative result to check the FS-NCSR's diversity of generated details on the DIV2K validation set for ×4 super-resolved results. The ground truth is a cropped part of 0875 from DIV2K. 5 outputs were chosen from 10 generated FS-NCSR outputs.

can reproduce the characters more clearly than NCSR. This qualitatively confirmed that although the training focused on high-frequency information performs slightly lower on LPIPS, actual outputs do not suffer a degradation of image quality than the existing methodologies.

The existing SRFlow and NCSR models show repeated failure cases where artifacts appear in a specific image (*e.g.* 0807, 0828 from DIV2K validation set). In the case of 0807 from DIV2K, for instance, when both SRFlow and NCSR generated the corresponding ×4 super-resolved outputs, all outputs were failure cases since some artifacts appeared. On the other hand, when FS-NCSR generate the ×4 super-resolved outputs of the given image, 4 out of 10 outputs were made without any artifact, and even for 6 failure cases, the degree of the artifact was relatively less than that of NCSR. Figure 5 presents the degree of the artifact differs between NCSR and FS-NCSR output and the FS-NCSR's artifact-free results compared to the ground truth image.

### 4.6. Ablation: Comparison of Generated High-Frequency Information

So far, we have discussed the results both quantitatively and qualitatively with the super-resolved outputs only. But we tried to compare the results from the perspective of frequency information additionally. Since the sparse high-frequency information plays a key role in the proposed method, we investigated how the proposed method affects the sparsity of the generated high-frequency information. For this purpose, the generated high-frequency information in $[0, 1]$ range is first quantized to the uint8 $[0, 255]$ range. And then **Sparsity** and **Relative Sparsity (RS)** is computed as follows:

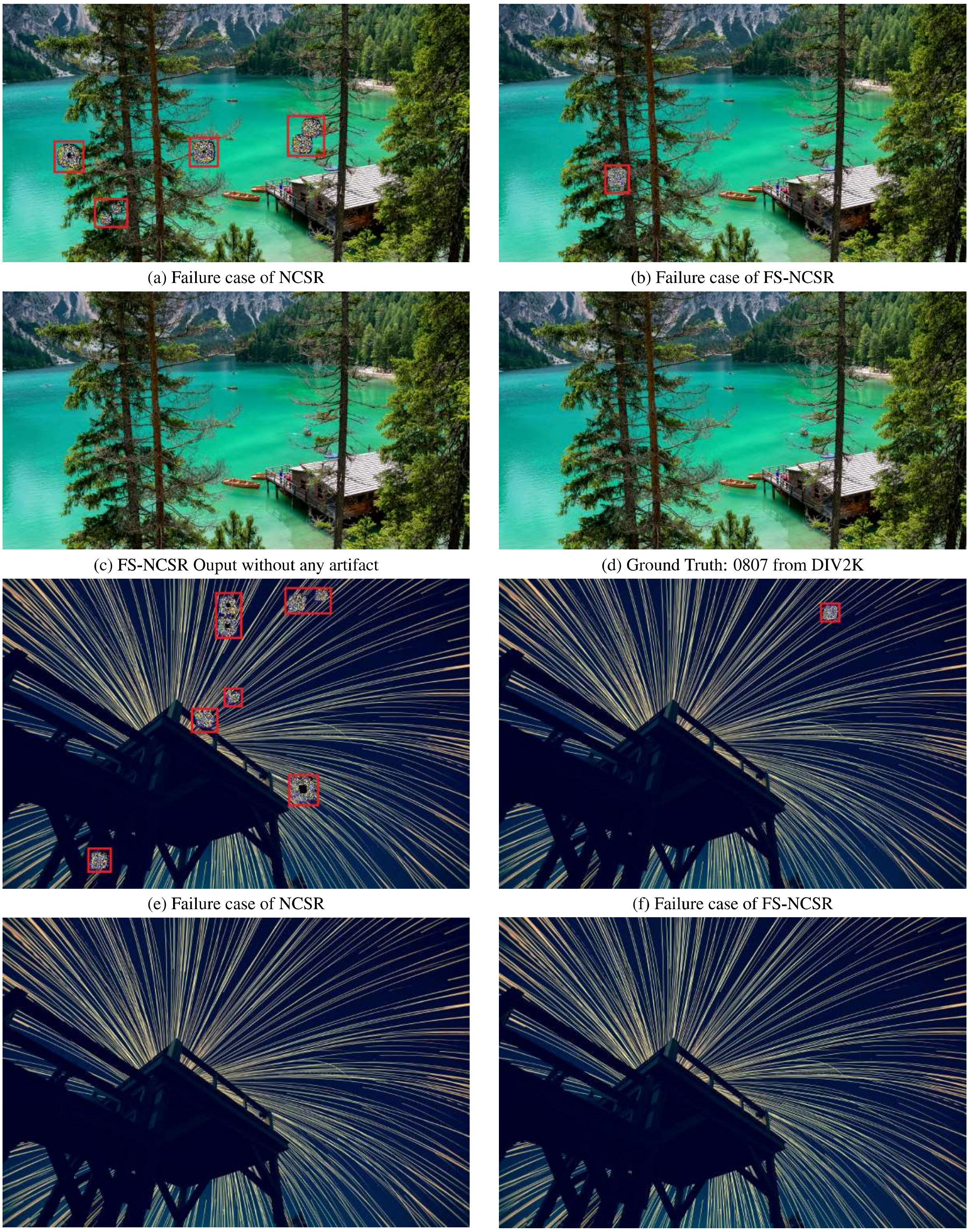
$$\begin{aligned}\textbf{Sparsity} &= 1 - \frac{\textbf{Number of non-zero pixels}}{H \times W \times C} \\ \textbf{RS} &= 1 - \frac{\textbf{Number of non-zero pixels}}{\textbf{Number of non-zero gt pixels}}\end{aligned} \tag{8}$$

where $(H, W, C)$ is the shape of a given image. Since the ground truth high-frequency information is already sparse, the RS reflects the sparsity of each ground truth image for a more fair comparison.

| Model | Average Sparsity | Average RS |
|---|---|---|
| NCSR [13] | 66.2% | 1.123 |
| **FS-NCSR (Ours)** | **66.0%** | **1.120** |

Table 3. As a ×4 super-resolution results on DIV2K validation set, **Average Sparsity of GT high frequency information is 59.0%**.

See Table 3. Although our proposed method shows less average sparsity and average RS slightly, the average sparsity of the ground truth high-frequency information and the



(a) Failure case of NCSR

(b) Failure case of FS-NCSR

(c) FS-NCSR Ouput without any artifact

(d) Ground Truth: 0807 from DIV2K

(e) Failure case of NCSR

(f) Failure case of FS-NCSR

(g) FS-NCSR Ouput without any artifact

(h) Ground Truth: 0828 from DIV2K

Figure 5. Visual comparison of failure cases on SR ×4 results from NCSR and FS-NCSR. The degree of the artifact is relatively less in the result of FS-NCSR than that of NCSR. And FS-NCSR could generate clear SR output without any artifact while NCSR couldn't. Each output was chosen randomly.

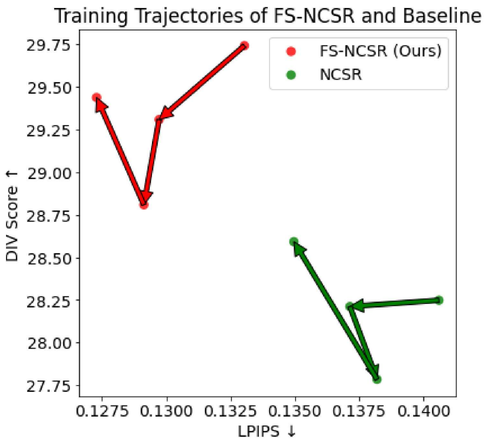


Figure 6. LPIPS and diversity scores of multiple checkpoints of FS-NCSR and NCSR [13]. The super-resolution ratio is ×4 setting. We measure LPIPS and diversity at 150k, 160k, 170k, 180k steps of training procedures. The arrows in the figure indicate the change in both metrics as the iteration increases by 10k.

| Team | LPIPS | LR PSNR | Div. Score | MOR |
|---|---|---|---|---|
| IMAG_ZW | 0.171 | 48.14 | 21.938 | 3.57 |
| **FS-NCSR (Ours)** | 0.126 | 50.13 | **28.853** | 3.67 |
| IMAG_WZ | 0.169 | 45.20 | 27.320 | **3.34** |
| SSS | 0.110 | 44.70 | 13.285 | _ |
| NCSR | 0.117 | 50.54 | 26.041 | _ |
| SRFlow | 0.122 | 49.86 | 25.008 | 3.62 |
| ESRGAN | 0.124 | 38.74 | 0.000 | 3.52 |

Table 4. Quantitative results for NTIRE 2022 Challenge on ”Learning Super Resolution Space” on ×4 track. The results were taken from [22]. The top block of the table is this year’s result.

| Team | LPIPS | LR PSNR | Div. Score | MOR |
|---|---|---|---|---|
| **FS-NCSR (Ours)** | 0.257 | 50.37 | 26.539 | 4.510 |
| SSS | 0.237 | 37.43 | 13.548 | 4.850 |
| NCSR | 0.259 | 48.64 | 26.941 | 4.503 |
| SRFlow | 0.282 | 47.72 | 25.582 | 4.775 |
| ESRGAN | 0.284 | 30.65 | 0.000 | 4.452 |

Table 5. Quantitative results for NTIRE 2022 Challenge on ”Learning Super Resolution Space” on ×8 track. The results were taken from [22]. The top block of the table is this year’s result.

that of generated output from both NCSR and FS-NCSR was about 10% more sparse, resulting in a lack of information compared to the ground truth. This margin of difference with the ground truth verifies that a significant loss of information still exists from the perspective of the frequency domain. Therefore, it seems that it needs to be addressed in future studies.

## 5. NTIRE 2022 Challenge

Our proposed method, FS-NCSR, achieved competitive results in both tracks of NTIRE 2022 ”Learning Super Resolution Space Challenge” [22]. See table 4 and 5 for the challenge result of ×4 and ×8 tracks respectively. In the ×4 track, FS-NCSR obtained the highest diversity score among the existing and newly proposed methods by a relatively large margin. Also, it obtained the best LPIPS and LR-PSNR results among this year’s participants, although it did not lead to the best MOR. In the ×8 track, FS-NCSR was this year’s only method that achieved comparable results compared to the last year’s approaches. Through the improvement of LR-PSNR, it seems that the frequency separation affected improving the consistency with low-resolution.

## 6. Conclusion

We propose a flow-based algorithm, FC-NCSR, to learn high-frequency information of super-resolution space. Based on the relation between the high-frequency information and the high-quality details of the given image, we train the generative model for super-resolution to produce the high-frequency information corresponding to the low-resolution input. With a simple high-pass filter using the low-frequency information of the low-resolution input, we successfully increase the super-resolution diversity without any influence on the stability of the flow-based NLL training and visual quality degradation. We also confirm that the frequency separation of FS-NCSR reduces the failure cases due to artifacts, and therefore, significantly improves the quality of the super-resolution output.

## 7. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]

## References

[1] 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

[2] Yuhua Chen, Feng Shi, Anthony G Christodoulou, Yibin Xie, Zhengwei Zhou, and Debiao Li. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 91–99. Springer, 2018. 1

[3] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965. 2

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

[5] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2

[6] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *ICLR*, 2017. 2

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. In *ECCV*, pages 0–0, 2014. 1, 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[9] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pages 2722–2730. PMLR, 2019. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[11] Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. Softflow: Probabilistic framework for normalizing flow on manifolds. *Advances in Neural Information Processing Systems*, 33:16388–16397, 2020. 1, 3, 4

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 1

[13] Younggeun Kim and Donghee Son. Noise conditional flow model for learning the super-resolution space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 424–432, 2021. 1, 2, 3, 4, 5, 6, 8

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[15] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 2, 3

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 4

[17] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 0–0, 2017. 1, 2

[18] Zhan Li, Qingyu Peng, Bir Bhanu, Qingfeng Zhang, and Haifeng He. Super resolution for astronomical observations. *Astrophysics and Space Science*, 363(5):1–15, 2018. 1

[19] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *IEEE International Conference on Computer Vision*, 2021. 1, 5

[20] 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. 1

[21] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 596–612, June 2021. 4, 5

[22] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2022 challenge on learning the super-resolution space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 8

[23] 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. 1, 2, 3, 4, 5

[24] Nathanael Carraz Rakotonirina and Andry Rasoanaivo. Esrgan+: Further improving enhanced super-resolution generative adversarial network. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020. 5

[25] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2

[26] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, Kanwal Bhatia, Antonio M Simoes Monteiro de Marvao, Tim Dawes, Declan O'Regan, and Daniel Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In *International conference on medical image computing and computer-assisted intervention*, pages 9–16. Springer, 2013. 1

[27] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[28] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based

generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[29] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. In *ICCP*, 2012. 2

[30] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 2

[31] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, pages 1920–1927, 2013. 2

[32] 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. 2

[33] 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) Workshops*, pages 0–0, 2018. 1, 2, 5

[34] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. *arXiv preprint arXiv:2112.02475*, 2021. 2

[35] Chih-Yuan Yang and Ming-Hsuan Yang. Fast direct super-resolution by simple functions. In *ICCV*, pages 561–568, 2013. 2

[36] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *CVPR*, 2008. 2

[37] 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. 2

[38] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 4

[39] 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. 1

[40] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 1

[41] Wilman WW Zou and Pong C Yuen. Very low resolution face recognition problem. *IEEE Transactions on image processing*, 21(1):327–340, 2011. 1