Multi–Scale Recursive and Perception–Distortion Controllable Image Super–Resolution

Pablo Navarrete Michelini, Dan Zhu, and Hanwen Liu

BOE Technology Group, Co., Ltd.
{pnavarre,zhudan,liuhanwen}@boe.com.cn

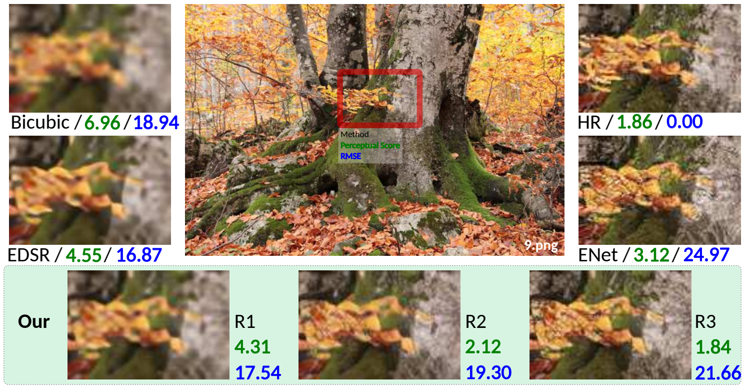


Fig. 1. Our G–MGBP super–resolution improves the perceptual quality of low distortion systems like EDSR[16] (with slightly higher RMSE), as well as baseline systems like EnhanceNet[26] (with significantly lower RMSE). Its perceptual scores are similar to the original images showing its effectiveness for ECCV PIRM–SR Challenge 2018[2]. Code and models are available at https://github.com/pnavarre/pirm-sr-2018.

Abstract. We describe our solution for the PIRM Super–Resolution Challenge 2018 where we achieved the $\mathbf{2^{nd}}$ **best perceptual quality** for average $RMSE \leqslant 16$, 5^{th} best for $RMSE \leqslant 12.5$, and 7^{th} best for $RMSE \leqslant 11.5$. We modify a recently proposed Multi–Grid Back–Projection (MGBP) architecture to work as a generative system with an input parameter that can control the amount of artificial details in the output. We propose a discriminator for adversarial training with the following novel properties: it is multi–scale that resembles a progressive–GAN; it is recursive that balances the architecture of the generator; and it includes a new layer to capture significant statistics of natural images. Finally, we propose a training strategy that avoids conflicts between reconstruction and perceptual losses. Our configuration uses only $281k$ parameters and upscales each image of the competition in $0.2s$ in average.

Keywords: backprojection · multigrid · perceptual quality

1 Introduction

We are interested in the problem of single image super–resolution (SR), which is to improve the quality of upscaled images by large factors (e.g. 4×) based on examples of pristine high–resolution images. Questions such as the objective meaning of quality, and what characterizes a pristine image, leads us towards different targets. The traditional approach is to focus on the reconstruction of high–resolution images from their downscale versions. We will refer to this target as *distortion* optimization. Alternatively, we can focus on creating upscale images that look as real as natural images to human eyes. We refer to the latter as *perception* optimization. In [3], Blau and Michaeli studied the conflicting roles of distortion and perceptual targets for image enhancements problems such as SR. Both targets cannot be achieved at the same time, one must compromise perceptual quality to reduce distortion and vice versa. Here, we are interested in the optimal balance between these two targets.

Our work follows the line of research started by SRCNN[4,5], which designed SR architectures using convolutional networks. SRCNN focused on a distortion target and it was later improved most notably by EDSR[16] and DBPN[8] in NTIRE–SR Challenges[29,30]. The work on SR architectures with a focus on perceptual targets has been possible thanks to the great progress in Generative Adversarial Networks (GAN)[7] and style transfer[6]. It began with SRGAN[15], which proposed the use of GANs, followed by Johnson[11], who proposed a real–time style transfer architecture, and later improved by EnhanceNet[26], which combined both approaches. Most recently, the Contextual (CX) loss[19] has been used in SR architectures to improve the similarity of feature distributions between artificial and natural images[18]. This latest method provides the best benchmark for perceptual quality according to non–reference metrics used in PIRM–SR 2018[2]: Ma[17] and NIQE[20].

Our system architecture was inspired by the multi–scale structure of MSLapSR[14], which we adapted to use Iterative Back–Projections (IBP) in feature space to enforce a downscaling model. In [23] we extended the classic IBP method to multiple scales by using a recursion analogous to the Full Multi–Grid algorithm, which is commonly used as PDE solver[31]. The system in [23] focused exclusively on a distortion target and now we extend it to perceptual targets.

Our main contributions are:

- We propose a novel **strategy to control the perception–distortion trade-off** in Section 2, which we adopt to design our system.
- We introduce **multi–scale diversity** into our SR architecture design, through random inputs at each upscaling level. These inputs are manipulated by the network in a recursive fashion to generate artificial details at different scales. See Section 3.
- We propose a novel **variance–normalization and shift–correlator** (VN+SC) layer that provides meaningful features to the discriminator based upon previous research on the statistics of natural images. See Section 4.1.
- We propose, to the best of our knowledge, the **first multi–scale and recursive discriminator** for adversarial training. It is a configuration symmetric

to the multi–scale upscaler, therefore it is more effective for adversarial training. It can simultaneously evaluate several upscaling factors, resembling a Progressive GAN[13] in the sense that the optimizer can focus on smaller factors first and then work on larger factors. See Section 4.2.

– We propose a novel **noise–adaptive training strategy** that can avoid conflicts between reconstruction and perceptual losses, combining loss functions with different random inputs into the system. See Section 5.

2 Controlling Distortion vs Perceptual Quality

To better illustrate our target, we present a diagram of image sets in Figure 2. Here, $\mathcal{H}$ is the set of all high–resolution images, $\mathcal{H}^{real} \subset \mathcal{H}$ is the subset of high–resolution images that correspond to natural images, and $\mathcal{L}$ is the set of all low–resolution images. Given an image $X \in \mathcal{H}^{real}$, we are interested in the set of *aliased* images:

$$\mathcal{A}(X) = \{Y \in \mathcal{H} \quad s.t. \quad S_{down}(Y) = S_{down}(X)\} \ , \tag{1}$$

where $S_{down} : \mathcal{H} \to \mathcal{L}$ is a *downscale* operator. We are particularly interested in the set $\mathcal{A}(X) \cap \mathcal{H}^{real}$ of alias images that correspond to real content.

A *distortion* function $\Delta(X, y)$ measures the dissimilarity between a reconstructed image y and the original image X. Popular and basic distortion metrics such as L1, L2, PSNR, etc., are sensitive to changes (any minor difference in pixel values would increase the amount of distortion) and are known to have low correlation with human perception[27]. Several distortion metrics have been proposed to approach perceptual quality by emphasizing some differences more than others, either through normalization, feature extraction or other approaches. These include metrics like SSIM[32], VIF[28] and the VGG content loss[12]. By doing so, correlation with human perception improves according to [27], but experiments in [3] show that these metrics still focus more on distortion. More recently, the contextual loss has been proposed to focus more on perceptual quality while maintaining a reasonable level of distortion[19].

The optimal solution of distortion optimization is obtained by:

$$X^* = \text{argmin}_y \mathbb{E}\left[\Delta(X, y)\right] \ . \tag{2}$$

The original image X is fixed, and the expected value in (2) removes any visible randomness in the search variable y. But, according to research on the statistics of natural images, randomness plays an essential role in what makes images look real[25]. This is well known for non–reference image quality metrics such as NIQE[20] or BRISQUE[21], and led to a definition of perceptual quality as a distance between probability distributions in [3]. It is also known that distortion optimization solutions tend to look unreal, as seen in state–of–the–art results from NTIRE–SR Challenges[29,30]. Common distortion metrics in these challenges (L1 and L2) make the image X^* lose all randomness. We argue that this removal of randomness in X^* is what moves it out of set $\mathcal{H}^{real}$, as we show in Figure 2.

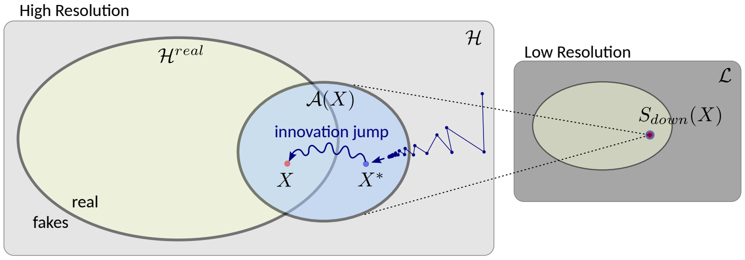


Fig. 2. Given a high–resolution image X that looks real, distortion optimization approaches an optimal solution X^* that does not look real because it lacks the random nature of natural images. We can still use X^* as a reference point to move through an *innovation jump* into the set of realistic images.

We know that $X \neq X^*$ because $X \in \mathcal{H}^{real}$ and $X^* \notin \mathcal{H}^{real}$ according to our previous discussion. However, distortion optimization can still be useful to generate realistic images. By approaching X^* we are getting closer to X. As shown in Figure 2, both X and X^* can be in $\mathcal{A}(X)$. Using a signal processing terminology, the *innovation*[22] is the difference between X and the optimal forecast of that image based on prior information, X^*. Most SR architectures take the randomness for the innovation process from the low–resolution input image, which is a valid approach but loses the ability to expose and control it.

In our proposed architecture we add randomness explicitly as noise inputs, so that we can control the amount of innovation in the output. Independent and identically distributed noise will enter the network architecture at different scales, so that each of them can target artificial details of different sizes. Generally speaking, our training strategy will be to approach X^* with zero input noise and any image in $\mathcal{A}(X) \cap \mathcal{H}^{real}$ with unit input noise. By using noise to target perceptual quality, and remove it for the distortion target, we teach the network to *jump* from X^* into $\mathcal{H}^{real}$. With probability one the network cannot hit X, but the perceptual target is any image in $\mathcal{A}(X) \cap \mathcal{H}^{real}$.

3 Generator Architecture

Our proposed architecture is shown in Figure 3 and is based on the Multi–Grid Back–Projection (MGBP) algorithm from [23], which improves a similar system used in NTIRE–SR Challenge 2018[30]. This is a multi–scale super–resolution system that updates a progressive classic upscaler (like bicubic) with the output of a convolutional network system. At each level MGBP shares the parameters of all networks. The first upcale image at each level is obtained by a Laplacian pyramid approach[14] and later improved by Iterative Back–Projections (IBP)[10] computed in latent space (e.g. features within a network). Iterative Back–projections introduces a downscaler system to recover the low–resolution image from upscale images, and thus captures information from the acquisition model of the input image. By using back–projections in latent space,

Algorithm 1 Generative Multi–Grid Back–Projection (G–MGBP)

```
$\boldsymbol{G-MGBP}(X, W, \mu, L)$:
Input: Input image $X$.
Input: Noise amplitude $W$.
Input: Numbers $\mu$ and $L$.
Output: $Y_k$, $k = 2, \ldots, L$.
$Y_1 = X$
$noise_1 = W \cdot \mathcal{N}(0, 1)$
for $k = 2, \ldots, L$ do
    $Y_k = \text{ClassicUpscale}(Y_{k-1})$
    $d = \text{Downscale}(\text{Analysis}(Y_k))$
    $u = \text{Upscale}([Y_{k-1},\ d,\ noise_{k-1}])$
    $u = BP_k^{\mu}(u, Y_1, \ldots, Y_{k-1},$
                $noise_1, \ldots, noise_{k-1})$
    $noise_k = W \cdot \mathcal{N}(0, 1)$
    $Y_k = Y_k + \text{Synthesis}(u)$
end for
```

```
$\boldsymbol{BP_k^{\mu}}(u, Y_1, \ldots, Y_{k-1}, noise_1, \ldots, noise_{k-1})$:
Input: Image $u$, level index $k$, steps $\mu$.
Input: Images $Y_1, \ldots, Y_{k-1}$ (only for $k > 1$).
Input: Images $noise_1, \ldots, noise_{k-1}$ (only for
    $k > 1$).
Output: Image $u$ (inplace)
if $k > 1$ then
    for $step = 1, \ldots, \mu$ do
        $d = BP_{k-1}^{\mu}($
            $\text{Downscale}(u), Y_1, \ldots, Y_{k-2},$
            $noise_1, \ldots, noise_{k-2}$
        $)$
        $u = u + \text{Upscale}([Y_{k-1},\ d,\ noise_{k-1}])$
    end for
end if
```

the downscaling model can be learned from training data and the iterations will enforce this model. For a multi–scale solution, MGBP uses a recursion based on multigrid algorithms[31] so that, at each upscaling level, an image is updated recursively using all previous level outputs.

For the PIRM–SR Challenge 2018[2] we extended MGBP to work as a generative system. For this purpose we added noise inputs that provide the *innovation process* as explained in Section 2. Previous work has shown the strong ability of convolutional networks to interpolate in feature space[24]. Inspired by this, we concatenate one channel of $\mathcal{N}(0, 1)$ noise to the input features of the Upscaler module in Figure 3, and we use a parameter W to control the amplitude of the noise. This parameter will later allow us to interpolate between distortion and perception optimizations (see Section 6.2). In our experiments we use 48 features, which increases to 49 features with the noise input. The new recursive method is specified in Algorithm 1.

The same noise channel is used during different IBP iterations at one scale ($\mu = 2$ times in our experiments) and i.i.d. noise is used for different scales. Figure 3 shows the unrolling recursion for $\mu = 2$ number of back–projections.

4 Discriminator Architecture

4.1 Variance Normalization and Shift Correlator

The task of the discriminator is to measure how realistic is an image. A straightforward approach is to input the color image to a convolutional network architecture. Then, we hope that the discriminator learns from adversarial training using real and fake image examples. In practice, we find that this approach works well to identify which areas of upscale images need more textures but the artificial details look noisy and have limited structure.

So what makes an image look natural? Extensive research has been carried to address this question. Here, we follow the seminal work of Ruderman[25]

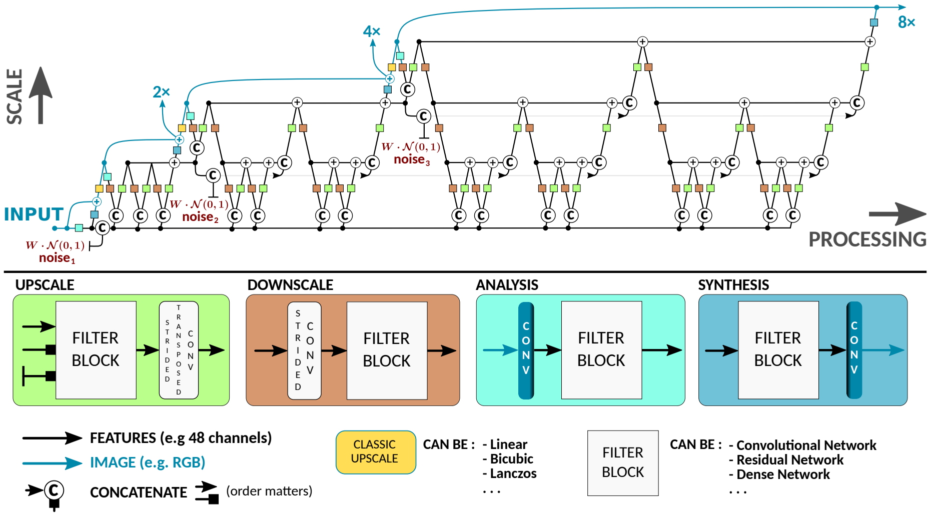

SCALE
4×
2×
8×
INPUT
$W \cdot \mathcal{N}(0,1)$ noise$_1$
$W \cdot \mathcal{N}(0,1)$ noise$_2$
$W \cdot \mathcal{N}(0,1)$ noise$_3$
PROCESSING
UPSCALE
DOWNSCALE
ANALYSIS
SYNTHESIS
FILTER BLOCK
STRIDED TRANSPOSED CONV
STRIDED CONV
CONV
FEATURES (e.g 48 channels)
IMAGE (e.g. RGB)
CONCATENATE (order matters)
CLASSIC UPSCALE
CAN BE : - Linear - Bicubic - Lanczos ...
FILTER BLOCK
CAN BE : - Convolutional Network - Residual Network - Dense Network ...

Fig. 3. Generative Multi–Grid Back–Projection (G–MGBP) workflow, obtained from the recursion in Algorithm 1 with $\mu = 2$ and $L = 3$, to output 2×, 4× and 8× upscale images. One channel of $\mathcal{N}(0,1)$ noise enters each scale in feature space, and it is reused several times within each level.

who found regular statistical properties in natural images that are modified by distortions. In particular, Ruderman observed that applying the so–called *variance normalization* operation:

$$\hat{I}_{i,j} = \frac{I_{i,j} - \mu_{i,j}(I)}{\sigma_{i,j}(I) + 1} \ , \tag{3}$$

has a decorrelating effect on natural images. Here, $I_{i,j}$ is the luminance channel of an image with values in $[0, 255]$ at pixel (i,j), $\mu(I)$ is the local mean of I (e.g. output of a Gaussian filter), and $\sigma(I)^2 = \mu(I^2) - \mu^2(I)$ is the local variance of I. Ruderman also observed that these normalized values strongly tend towards a Gaussian characteristic for natural images. These findings are used in the NIQE perceptual quality metric considered for the PIRM–SR Challenge 2018[20]. NIQE also models the statistical relationships between neighboring pixels by considering horizontal and vertical neighbor products: $\hat{I}_{i,j}\hat{I}_{i,j+1}$, $\hat{I}_{i,j}\hat{I}_{i+1,j}$, $\hat{I}_{i,j}\hat{I}_{i,j-1}$ and $\hat{I}_{i,j}\hat{I}_{i-1,j}$. Previously, the BRISQUE non–reference metric also used diagonal products[21].

Inspired by previous research we define the Variance Normalization and Shift Correlator (VN+SC) layer as follows:

$$V_{i,j}^{7(p+3)+q+3}(I) = \hat{I}_{i,j} \cdot \hat{I}_{i+p,j+q} \ , \qquad p = -3, \ldots, 3 \quad , q = -3, \ldots, 3 \ . \tag{4}$$

Here, we transform a color image into a set of neighbor products (shift correlator) $V_{i,j}^k$ with $k = 0, \ldots, 48$, using the variance normalized image $\hat{I}$. The number of

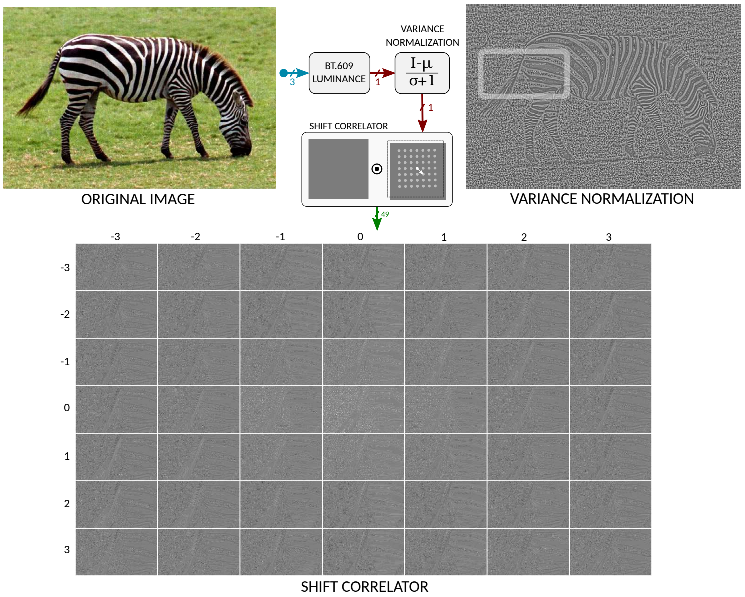


Fig. 4. We propose a Variance Normalization and Shift Correlator (VN+SC) layer that transforms the inputs of the discriminator into 49 channels that, according to research on the statistics of natural images, capture the essential information to discriminate between natural and unnatural images.

neighbor products can be any number, and we set it to 7×7 in our experiments to get a number similar to the 48 features used in our discriminator architecture. Figure 4 shows the visual effect of the the VN+SC operation. We use a VN+SC layer for each input of our discriminator, as shown in Figure 5.

4.2 Multi–Scale and Recursive Architecture

The G–MGBP upscaler designed in Section 3 is multi–scale and recursive. We can then take advantage of the multi–scale distortion optimization training strategy proposed in [14]. This strategy is difficult for adversarial training because the outputs at different levels contain different artifacts and might need an ensemble of discriminators. We simplify this problem by using a multi–scale and recursive architecture as shown in Figure 5. The system takes several upscaled images using different factors (2×, 4× and 8× in our experiments) and, based on all of them, it outputs one score to decide if the images are real or fake. The parameters of each block (pooling, analysis and synthesis) are shared at each level. Thus, the system keeps the same number of parameters, either in a small configuration with $L = 1$ level to evaluate a single 2× upscale output, or in a large configuration with $L = 3$ levels to simultaneously evaluate 2×, 4× and 8× upscale outputs. Adversarial training with this discriminator resembles a Progressive GAN[13] because it can adjust parameters to first solve the simpler

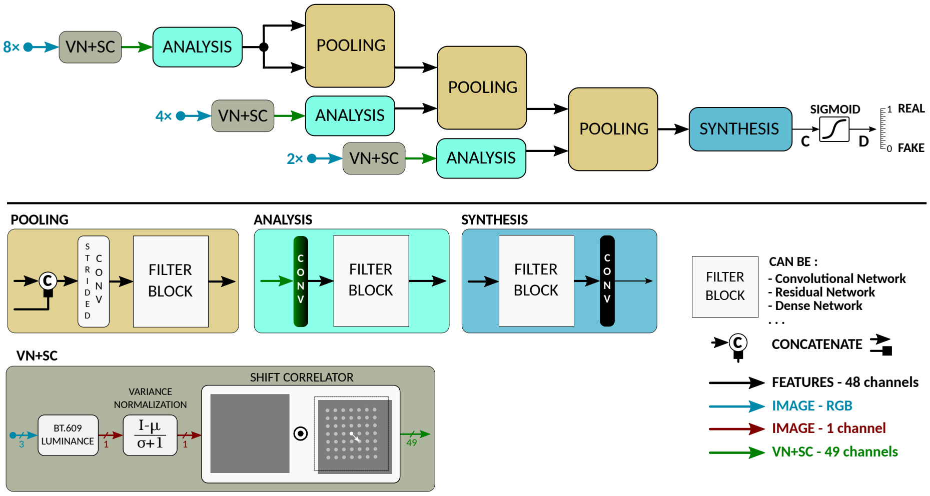


Fig. 5. Multi–level recursive discriminator used for adversarial training. The diagram shows D^3, the system unfold for 3 levels to simultaneously evaluate 2×, 4× and 8× upscale outputs. Each module shares parameters in different scales.

problem of 2× upscaling, and then follow with larger factors. But, at the same time, it is significantly different because a Progressive GAN system is neither multi–scale nor recursive.

5 Adversarial Training Strategy

We follow the design of multi–scale loss from MSLapSR[14] with 3 scales: 2×, 4× and 8×. For each scale $L \in \{1, 2, 3\}$ we take X^L, as patches from the HR dataset images. High–resolution references X^k with $k = 1, \ldots, L-1$ are obtained by downscaling the dataset HR images with factor $L - k$. This is:

$$X^L = \text{HR image from dataset}\ , \quad L = 1\,,2\,,3\ , \tag{5}$$

$$X^k = S^{L-k}_{down}(X^L)\ , \quad k = 1, \ldots, L-1\ . \tag{6}$$

We denote $Y_{W=0}$ and $Y_{W=1}$ the outputs of our generator architecture using noise amplitudes $W = 0$ and $W = 1$, respectively. Then, we combine the multi–scale loss from [14] and the perceptual loss from [18] with different noise inputs. Our total loss is given by:

$$\mathcal{L}(Y, X; \theta) = \sum_{L=1,2,3} \Big\{ \ 0.001 \cdot \mathcal{L}^{GAN-G}_L(Y_{W=1}) + 0.1 \cdot \mathcal{L}^{context}_L(Y_{W=1}, X) + 10 \cdot \mathcal{L}^{rec}_L(Y_{W=0}, X) + 10 \cdot \mathcal{L}^{cycle}_L(Y_{W=0}, Y_{W=1}, X) \Big\}\ . \tag{7}$$

Here, colors represent the target of each loss term according to Figure 6. First,

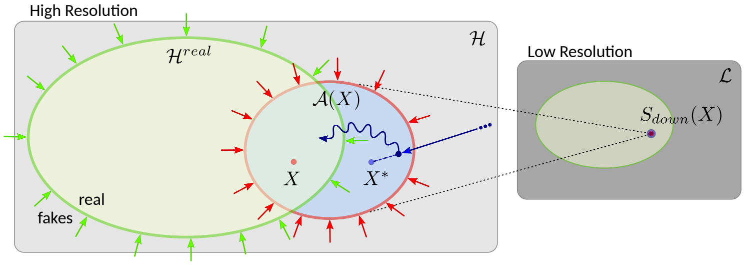


Fig. 6. Our loss function tries to: look real by moving into $\mathcal{H}^{real}$ (GAN and CX loss), enforce a downscaling model by moving into $\mathcal{A}(X)$ (cycle loss), and be reachable by latent space interpolation from the optimal distortion solution X^* (reconstruction loss).

$$\mathcal{L}_L^{GAN-G}(Y_{W=1}) = \mathbb{E}\left[\log(D^L(Y_{W=1}^k|k=1,\ldots,L)\right] \ , \tag{8}$$

$$\begin{aligned}\mathcal{L}_L^{GAN-D}(Y_{W=1}) = \mathbb{E}\left[\log(D^L(X^k|k=1,\ldots,L))\right] + \\ \mathbb{E}\left[\log(1-D^L(Y_{W=1}^k|k=1,\ldots,L))\right]\end{aligned} \tag{9}$$

follows a standard adversarial loss[7], where D^L is our L–level recursive discriminator evaluating L output images, as shown in Figure 5. Then,

$$\mathcal{L}_L^{context}(Y_{W=1}, X) = -\mathbb{E}\left[\sum_{k=1}^{L}\log\left(CX(\Phi(Y_{W=1}^k), \Phi(X^k)\right)\right] \tag{10}$$

uses the *contextual similarity* CX as defined in [19] and Φ are features from *conv3–4* of VGG–19 network as suggested in [18]. The contextual loss is designed to give higher importance to the perceptual quality[19]. Next,

$$\mathcal{L}_L^{rec}(Y_{W=0}, X) = \mathbb{E}\left[\sum_{k=1}^{L}||Y_{W=0}^k - X^k||_1\right] \tag{11}$$

is a standard $L1$ distortion loss, equivalent to the multi–scale loss in [14]. We note that here the noise input is set to zero, which prevents this term to interfere with the generation of details as it does not see randomness in the outputs. Finally, the *cycle* regularization loss enforces the downscaling model by moving the outputs back to low–resolution, analogous to the cycle–loss in CycleGAN[33]. This is,

$$\mathcal{L}_L^{cycle}(Y_{W=0}, Y_{W=1}, X) = \mathbb{E}\left[\sum_{k=1}^{L}\sum_{f=1}^{k}||S_{down}^f(Y_{W=0}^k) - S_{down}^f(X^k)||_1\right] + \tag{12}$$

$$\mathbb{E}\left[\sum_{k=1}^{L}\sum_{f=1}^{k}||S_{down}^f(Y_{W=1}^k) - S_{down}^f(X^k)||_1\right] \ , \tag{13}$$

where we use the $L1$ distance between downscaled outputs and low–resolution inputs. The first term, with noise amplitude zero, forces $Y_{W=0}$ to stay in $\mathcal{A}(X)$

as it approaches the image X^*. The second term, with unit noise, forces $Y_{W=1}$ to stay in $\mathcal{A}(X)$ as it approaches the set $\mathcal{H}^{real}$.

6 Experiments

6.1 Configuration

For training and validation data we resized images to the average mega–pixels of PIRM–SR dataset (0.29 Mpx), taking all images from: DIV2K[1], FLICKR–2K, CLIC (professional sets), and PIRM–SR self–validation[2]. We selected $4,271$ images for training and 14 images for validation during training.

We used one single configuration to test our system. We configure the *Analysis*, *Synthesis*, *Upscale*, *Downscale* and *Pooling* modules in Figure 3 and 5 using 4–layer dense networks[9] as filter–blocks. We use 48 features and growth rate 16 within dense networks. For classic upscaler we started with Bicubic and we set the upscaling filters as parameters to learn as proposed in [14].

We trained our system with Adam optimizer and a learning rate initialized as 10^{-3} and square root decay, both for generator and discriminator systems. We use 128×128 patches with batch size 16. We pre–trained the network with $W = 0$ and only L1 loss, and used as initial setting for our overall loss (7).

6.2 Moving on the Perception–Distortion plane

An essential part of our generative SR architecture is the noise inputs. The training strategy introduced in Section 5 teaches the system to optimize distortion when the noise is set to zero, and maximize perceptual quality when the noise is enabled. Thus, noise provides the randomness needed for natural images and represents the *innovation jump* according to Figure 2.

After training, we are free to control the noise inputs. In particular, we can move the noise amplitude smoothly between $W = 0$ and $W = 1$ to inspect the path to jump from distortion to perception optimization. Figure 7 shows an example of this transition. Here, it is important to note that our training strategy does not optimize the trajectory in the perception–distortion plane, but only the corner cases of best distortion ($W = 0$) and best perception ($W = 1$). The corner cases are clearly verified in Figure 7. At this point, it is unkown which trajectory will the the network take to move from one case to the other.

It is interesting to see in Figure 7 that the transition from best perception to best distortion happens within a narrow margin of $\Delta W = 0.02$ amplitude values and much closer to $W = 0$ than $W = 1$ (around $W \sim 0.01$). Similar transitions were observed in other images of the PIRM dataset, for both test and validation.

We also observe that the parametric curve in the perception–distortion plane looks like a monotonically non–increasing and convex function, similar to the optimal solution studied in [3]. But, it is important to emphasize that the curve in Figure 7 is not optimal as we are not enforcing optimality and, as a matter of fact, for the PIRM–SR Challenge we ended up using different training results for R1, R2 and R3, each one performing better than the others in its own region.

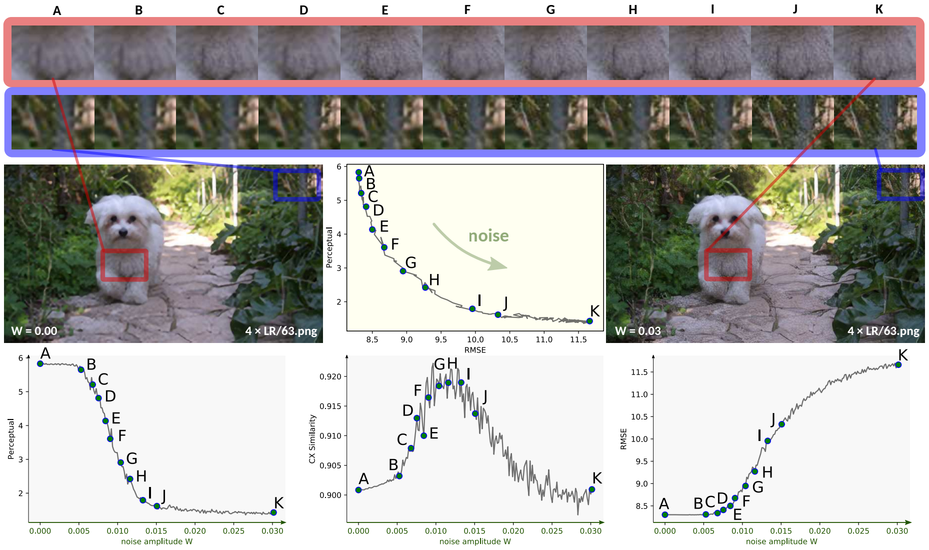
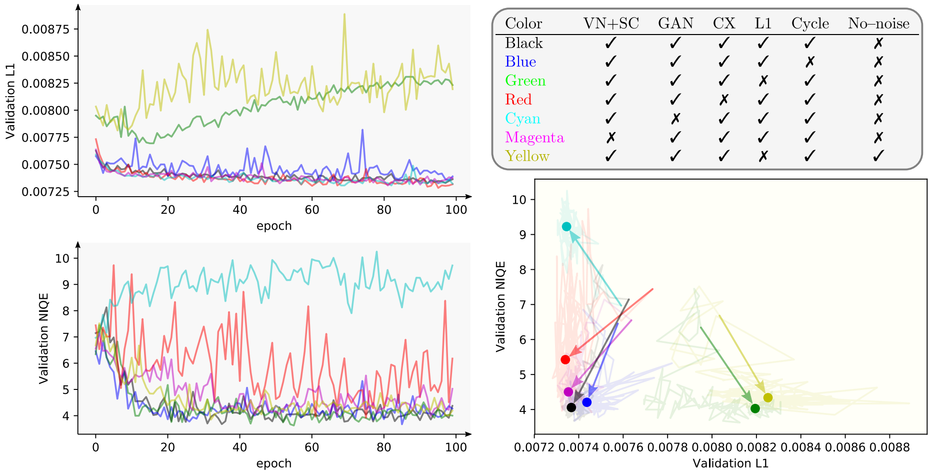


Fig. 7. Our SR architecture uses noise inputs that can be use to move from distortion to perception optimization, without retraining the system. The plot shows how perceptual quality improves as we increase the noise amplitude in our R3 model. Output images show how artificial details appear in different areas of the image as noise amplifies.

Regarding image quality metrics, we see with no surprise that the *Perceptual* index proposed for the PIRM–SR Challenge[2] improves as noise increases, while the distortion measured by RMSE increases. We observed very similar results for the perceptual metrics NIQE and Ma, as well as the L1 distortion metric. More interesting is the transition observed in the *contextual similarity* index. First, it behaves as a perceptual score with the CX similarity improving consistently as noise increases. Then, when the *Perceptual* score seems to stall, but RMSE keeps increasing, the CX similarity changes to a distortion metric pattern, reducing as noise increases. This is consistent with the design target of *CX similarity* to focus more on perceptual quality while maintaining a reasonable level of distortion[19].

6.3 Ablation Tests

Our overal loss combines terms focused on different targets (e.g. low distortion, perceptual quality). In Section 5 we explained the purpose of each term using the diagram in Figure 2. It remains to verify this design and to quantify the relevance of each term. We also want to quantify the contribution of our novel VN+SC layer. For this purpose we trained our network architecture for 100 epochs according to the configuration in section 6.1. In Figure 8 we show our measurements of L1 (distortion) and NIQE (perceptual) in a small validation set of 14 images after each epoch. We display the evolution through the number of epochs as well as the trajectories on the perception–distortion plane.



Color	VN+SC	GAN	CX	L1	Cycle	No–noise
Black	✓	✓	✓	✓	✓	✗
Blue	✓	✓	✓	✓	✗	✗
Green	✓	✓	✓	✗	✓	✗
Red	✓	✓	✗	✓	✓	✗
Cyan	✓	✗	✓	✓	✓	✗
Magenta	✗	✓	✓	✓	✓	✗
Yellow	✓	✓	✓	✗	✓	✓

Fig. 8. Ablation tests show the validation scores when training our network for 100 epochs. We consider removal of the loss terms: GAN, CX, L1 and Cycle in (7), as well as VN+SC layers in the discriminator, and training the system without noise inputs.

Overall, we see that our strategy adding all the losses (in black color) gives the best perception–distortion balance. In the extremes we see that removing the L1 and GAN losses have catastrophic effects on distortion and perception, respectively. Still, these cases do not diverge to infinity because of other loss terms. Next, it is clear that the contextual loss helps improving the perceptual quality, and regarding distortion the amount of improvement is not conclusive. Then, the addition of the cycle loss shows a clear improvement over distortion, with unconclusive improvements on perceptual quality. And finally, we observe that the addition of the VN+SC layer in the discriminator clearly improves perceptual quality, although not as much as CX and GAN losses.

Figure 8 also shows a test in which we avoid the use of noise imputs by setting $W = 0$ in all losses. In this case we remove the L1 loss that would otherwise interfere with the GAN loss, causing a catastrophic effect. In this case distortion is controlled by the cycle loss, equivalent to how it is done in [18]. In this configuration the network performs slightly worse in perceptual quality and clearly worse on distortion, similar to only removing the L1 loss. In this case, we believe that the network uses the randomness in the input as innovation process, which cannot be controlled and limits the diversity of the generator.

6.4 Challenge Results

Table 1 shows our best average scores in the PIRM–SR Challenge 2018[2] for Region 1 ($RMSE \leqslant 11.5$), Region 2 ($11.5 < RMSE \leqslant 12.5$) and Region 3 ($12.5 < RMSE \leqslant 16$), compared to baseline methods: EDSR[16], CX[19] and EnhanceNet[26]. We achieved better perceptual scores compared to all baselines.

Table 1. Quantitative comparison between our solutions for R1, R2 and R3 and baseline methods in the test set. Best numbers in each row are shown in bold.

		EDSR	Our			CX	ENet
	Unit/Metric		R1	R2	R3		
Parameters	[k]	43, 100	**281**	**281**	**281**	853	853
Perceptual	$\frac{1}{2}((10 - \text{Ma}) + \text{NIQE})$	4.904	3.817	2.484	**2.019**	2.113	2.723
Distortion	$RMSE$	**10.73**	11.50	12.50	14.24	15.07	15.92

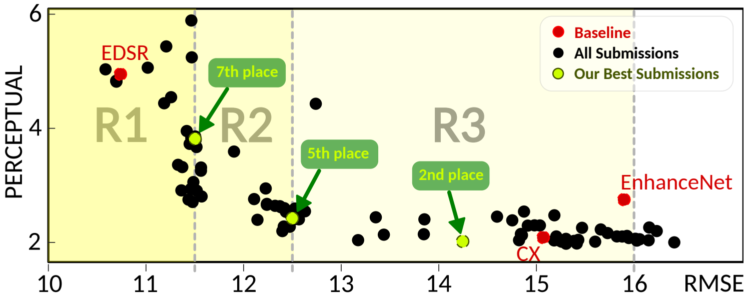


Fig. 9. Perception–distortion plane with average scores in the test set showing all submissions, from all teams in PIRM–SR 2018[2]. Our best scores are shown in green color together with the final ranking in PIRM–SR Challenge 2018.

Beyond the target of the competition, we also observe that we use significantly less parameters. This shows the advantage of the recursive structure of our system, which successfully works across multiple scales to achieve the target. Our system can upscale one image of the self–validation set in $0.2s$ in average.

Compared to other submissions, we observe in Figure 9 that our system performs better in Region 3. Here, we achieve the 2^{nd} place within very small differences in perceptual scores but with significantly lower distortion. This shows the advantage of our training strategy to optimize the perception–distortion trade–off. In Regions 1 and 2 we were one among only two teams that reached the exact distortion limit (11.5 in Region 1 and 12.5 in Region 2). We were able to achieve this by controlling the noise amplitude, without retraining the system. Our ranking lowers as the distortion target gets more difficult. We believe that this is caused by the small size of our system that becomes more important for low distortion targets, since we use only $281k$ parameters compared to $43M$ of the EDSR baseline in Region 1.

Finally, Figure 1 and 10 show comparisons of our results with the baselines, using images from our validation set. We observe that in Region 3 we achieve better perceptual scores even compared to the original HR images. While we subjectively confirm this in some patches, we do not make the same conclusion after observing the whole images. Somehow, we believe that our design for adversarial training and validation strategy managed to overfit the perceptual scores. Nevertheless, we observe clear advantages to the baselines, showing better structure in textures and more consistent geometry in edges and shapes.

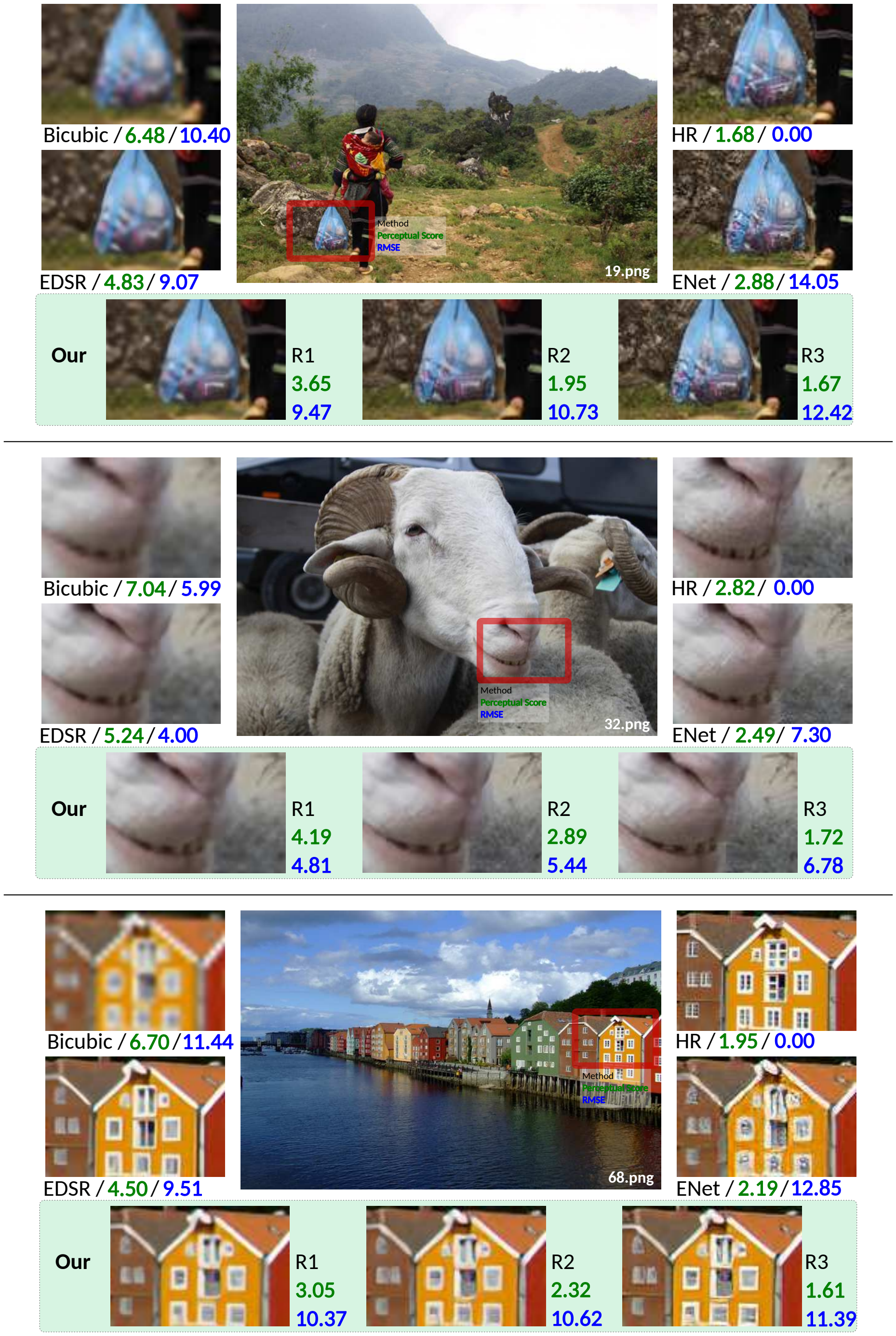


Fig. 10. Image comparisons of 4× upscaling between our solutions in R1, R2 and R3 (see Figure 9) and baseline methods in our validation set. Perceptual and distortion scores of whole images are shown in green and blue colors, respectively.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
2. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: 2018 PIRM challenge on perceptual image super-resolution (2018), http://arxiv.org/abs/1809.07517
3. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
4. Dong, C., Loy, C., He, K., Tang, X.: Learning a deep convolutional network for image super–resolution. In: et al., D.F. (ed.) Proceedings of European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 8692, pp. 184–199. Springer, Zurich (September 2014)
5. Dong, C., Loy, C., He, K., Tang, X.: Image super–resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2015, to appear)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. Journal of Vision (August 2015), http://arxiv.org/abs/1508.06576
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
8. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back–projection networks for super–resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
9. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
10. Irani, M., Peleg, S.: Improving resolution by image registration. CVGIP: Graph. Models Image Process. **53**(3), 231–239 (Apr 1991). https://doi.org/10.1016/1049-9652(91)90045-L, http://dx.doi.org/10.1016/1049-9652(91)90045-L
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (2016)
12. Johnson, J., Alahi, A., Li, F.: Perceptual losses for real-time style transfer and super–resolution. CoRR **abs/1603.08155** (2016), http://arxiv.org/abs/1603.08155
13. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
14. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Fast and accurate image super–resolution with deep laplacian pyramid networks. arXiv:1710.01992 (2017)
15. Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo–realistic single image super–resolution using a generative adversarial network. CoRR **abs/1609.04802** (2016), http://arxiv.org/abs/1609.04802
16. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super–resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)

17. Ma, C., Yang, C., Yang, X., Yang, M.: Learning a no–reference quality metric for single–image super–resolution. Computer Vision and Image Understanding **158**, 1–16 (2017), http://arxiv.org/abs/1612.05890
18. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Learning to maintain natural image statistics, [arxiv](https://arxiv.org/abs/1803.04626). arXiv preprint arXiv:1803.04626 (2018)
19. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. arXiv preprint arXiv:1803.02077 (2018)
20. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "Completely Blind" Image Quality Analyzer. IEEE Signal Processing Letters **20**, 209–212 (Mar 2013). https://doi.org/10.1109/LSP.2012.2227726
21. Mittal, A., Moorthy, A.K., Bovik, A.C.: No–reference image quality assessment in the spatial domain. IEEE Trans. Image Process pp. 4695–4708 (2012)
22. Mitter, S.K.: Nonlinear filtering of diffusion processes a guided tour. In: Fleming, W.H., Gorostiza, L.G. (eds.) Advances in Filtering and Optimal Stochastic Control. pp. 256–266. Springer Berlin Heidelberg, Berlin, Heidelberg (1982)
23. Navarrete Michelini, P., Liu, H., Zhu, D.: Mutigrid backprojection super-resolution and deep filter visualization (2018), http://arxiv.org/abs/1809.09326
24. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR **abs/1511.06434** (2015)
25. Ruderman, D.L.: The statistics of natural images. In: Network computation in neural systems. vol. 5, pp. 517–548 (1994)
26. Sajjadi, M.S.M., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
27. Seshadrinathan, K., Soundararajan, R., Bovik, A., Cormack, L.: Study of subjective and objective quality assessment of video. IEEE Transactions on Image Processing **19**(6), 1427–1441 (June 2010)
28. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. Trans. Img. Proc. **15**(2), 430–444 (Feb 2006). https://doi.org/10.1109/TIP.2005.859378, http://dx.doi.org/10.1109/TIP.2005.859378
29. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L., et al.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July)
30. Timofte, R., Gu, S., Wu, J., Van Gool, L., Zhang, L., Yang, M.H., et al.: NTIRE 2018 challenge on single image super-resolution: Methods and results. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2018)
31. Trottenberg, U., Schuller, A.: Multigrid. Academic Press, Inc., Orlando, FL, USA (2001)
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image–to–image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2017)