Deep Generative Model based Rate-Distortion for Image Downscaling Assessment

Yuanbang Liang[1]   Bhavesh Garg[2*]   Paul Rosin[1]   Yipeng Qin[1†]

[1]School of Computer Science and Informatics, Cardiff University   [2]IIT Bombay & WadhwaniAI

{liangy32, rosinpl, qiny16}@cardiff.ac.uk, bh05avesh@gmail.com

Abstract

In this paper, we propose Image Downscaling Assessment by Rate-Distortion (IDA-RD), a novel measure to quantitatively evaluate image downscaling algorithms. In contrast to image-based methods that measure the quality of downscaled images, ours is process-based that draws ideas from rate-distortion theory to measure the distortion incurred during downscaling. Our main idea is that downscaling and super-resolution (SR) can be viewed as the encoding and decoding processes in the rate-distortion model, respectively, and that a downscaling algorithm that preserves more details in the resulting low-resolution (LR) images should lead to less distorted high-resolution (HR) images in SR. In other words, the distortion should increase as the downscaling algorithm deteriorates. However, it is non-trivial to measure this distortion as it requires the SR algorithm to be blind and stochastic. Our key insight is that such requirements can be met by recent SR algorithms based on deep generative models that can find all matching HR images for a given LR image on their learned manifolds. Extensive experimental results show the effectiveness of our IDA-RD measure. Our code is available at: https://github.com/Byronliang8/IDA-RD

1. Introduction

Image downscaling is a fundamental problem in image processing and computer vision. To address the diverse application scenarios, various digital devices with different resolutions, such as smartphones, iPads, and desktop monitors, co-exist, which makes this problem even more important. In contrast to image super-resolution (SR), which aims to "add" information to low-resolution (LR) images, image downscaling algorithms focus on "preserving" information present in the high-resolution (HR) images, which is especially important for applications and devices with limited screen spaces.

Traditional image downscaling algorithms low-pass filter an image before resampling it. While this prevents aliasing in the downscaled LR image, important high-frequency details

of the HR image are removed simultaneously, resulting in a blurred or overly-smooth LR image. To improve the quality of downscaled images, several sophisticated approaches have been proposed recently, including remapping of high-frequency information [12], optimization of perceptual image quality metrics [29], using $L0$-regularized priors [23], and pixelizing the HR image [13, 15, 20, 37]. Nevertheless, research in image downscaling algorithms has significantly slowed down due to the lack of a quantitative measure to evaluate them. Specifically, standard distance measures (*e.g.*, $L1$, $L2$ norms) and full-reference image quality assessment (IQA) methods are not applicable here due to the absence of ground truth LR images; existing No-Reference IQA (NR-IQA) metrics [7, 27, 28] cannot be applied either as they rely on the "naturalness" of HR images, which is not present in LR images (we will verify this in our experiments).

In this paper, we propose a new quantitative measure for image downscaling based on Claude Shannon's rate-distortion theory [5], namely Image Downscaling Assessment by Rate-Distortion (IDA-RD). The main idea of our IDA-RD measure is that a superior image downscaling algorithm would try to retain as much information as possible in the LR image, thereby reducing the distortion when being upscaled (*a.k.a.* super-resolved) to the size of the original HR image. However, such an upscaling method is non-trivial as, for our purpose, it must satisfy two challenging requirements: i) *blindness*, *i.e.*, it must apply to all kinds of downscaling algorithms without knowing them in advance; ii) *stochasticity*, *i.e.*, it must be able to generate a manifold of HR images that captures the conditional distribution of the super-resolution process. Our key insight is that both such requirements can be satisfied by the recent success of deep generative models in blind and stochastic super-resolution. To demonstrate the flexibility of our IDA-RD measure, we show that it can be successfully implemented with two mainstream generative models: Generative Adversarial Networks [26] and Normalizing Flows [24]. Extensive experiments demonstrate the effectiveness of our IDA-RD measure in evaluating image downscaling algorithms. Our contributions include:

- Drawing on Shannon's rate-distortion theory [5], we propose the Image Downscaling Assessment by Rate-
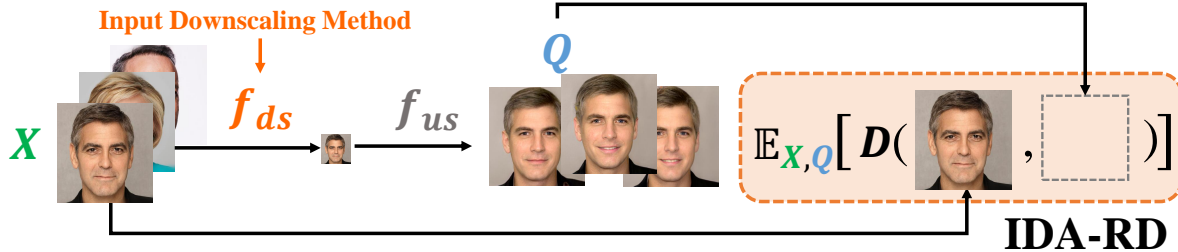

Figure 1. Illustration of the proposed IDA-RD measure. Given a downscaling method f_{ds} to be evaluated, i) we first use it to downscale several HR images; ii) then, we upscale them back to the original resolution with f_{us} and measure the distortion from the corresponding HR images. Such an upscaling method leverages the recent success in deep generative models and thus can i) apply to arbitrarily down-scaled images and ii) output a manifold of HR images that captures the conditional distribution given a downscaled image.

Distortion (IDA-RD) measure to quantitatively evaluate image downscaling algorithms, which fills a gap in image downscaling research.

- We demonstrate the effectiveness of our IDA-RD measure with extensive experiments on both synthetic and real-world image downscaling algorithms.

2. Related Work

Image Downscaling has a long history and its traditional methods (*e.g.*, bicubic) have now become the standard for image processing and computer vision software, making it difficult to trace their origins. To this end, we only review recent attempts in developing better image downscaling algorithms. For example, Gastal and Oliveira [12] conducted a discrete Gabor frequency analysis and propose to remap the high-frequency information of HR images to the representable range of the downsampled spectrum, thereby preserving high frequency details in image downscaling. Oeztireli and Gross [29] model image downscaling as an optimization problem and minimize a perceptual metric (SSIM) between the input and downscaled image. However, the limitations of SSIM are also carried over to their approach. DPID [44] preserves small details by assigning higher weights to the input pixels whose color deviates from their local neighborhood within the convolutional filter. Liu et al. [23] propose an optimization framework using two $L0$ regularized priors that addresses two issues of image downscaling, *i.e.*, salient feature preservation and downscaled image construction. Image thumbnailing, a special case of image downscaling, has been studied by Sun and Ling [38]. Their two-component thumbnailing framework, named as Scale and Object Aware Thumbnailing (SOAT) focuses on saliency measure and thumbnail cropping. Li et al. [21] term image downscaling as image Compact Resolution (CR) and address it with a Convolutional Neural Network (CNN). Inspired by the success of CNNs in image super-resolution (SR), they introduce the CNN-CR model for image downscaling that can be jointly trained with any CNN-SR model. Although their CNN-CR model results in better reconstruction quality than other downscaling algorithms, they only demonstrate results

for small downscaling factors ($\times 2$). However, the majority of both image downscaling and super-resolution algorithms tend to focus on larger scaling factors (*e.g.*, $\times 8$). Despite the aforementioned works, there does not exist a good quantitative measure for the evaluation of image downscaling methods, which impedes the research on them.

Image Quality Assessment (IQA) can be subjective or objective. Subjective methods rely on the visual inspection by human assessors while objective methods resort to quantitative measures, *e.g.*, image statistics. Examples of the most commonly used objective IQA metrics include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Multi-Scale SSIM (MS-SSIM) [42] and Learned Perceptual Image Patch Similarity (LPIPS) [49]. However, such IQA metrics are not applicable in the evaluation of image downscaling algorithms as there are no ground truth LR images for comparison. Please note that we do not consider the LR images captured by cameras to be ground truth, as they rely on the particular camera used and can thus be viewed as being captured by "hardware" downscaling methods that can also be assessed by our IDA-RD measure. Thus, most researchers rely on subjective evaluation of downscaled images, which is costly and time-consuming.

No-Reference Image Quality Assessment (NR-IQA) addresses IQA in the absence of a reference (*i.e.*, ground truth) image. For example, Mittal et al. [27] propose BRISQUE, an NR-IQA metric that uses the natural scene statistics (NSS) to quantify loss of "naturalness" in distorted images. Using locally normalized luminances, BRISQUE models a regressor which maps the feature space to image quality scores. Based on their NSS, Mittal et al. [28] further devised an Opinion Unaware (OU) and Distortion Unaware (DU) model for blind NR-IQA, which is named as NIQE. Bosse et al. [7] follow a data-driven approach for NR-IQA. Inspired by Siamese networks, they train a deep neural network for feature extraction and regression in an end-to-end manner. However, due to the lack of a large enough training dataset, their model does not generalize well. However, such NQ-IQA metrics are also not applicable, as the "naturalness" they rely on exists only in HR but not LR images. To this end, we bor-

row ideas from Claude Shannon's rate-distortion theory and propose a new measure called Image Downscaling Assessment by Rate-Distortion (IDA-RD). Our IDA-RD measure leverages the recent success in deep generative models and shows promising results in the quantitative evaluation of image downscaling methods.

Deep Generative Models. We refer interested readers to [6] for a detailed survey on deep generative modeling. Here, we review the two deep generative models used in our work, *i.e.*, Generative Adversarial Networks (GANs) and normalizing flows. Since the pioneering work by Goodfellow et al. [14], GANs have experienced significant improvements. For example, Radford et al. [33] proposed DCGAN, which incorporates convolutional neural networks for better image synthesis. Arjovsky et al. [4] addressed the notorious instability of GAN training by employing a novel loss function, *i.e.*, the Wasserstein distance loss. To date, the StyleGAN series [16–18] developed by Nvidia has shown impressive results in high-resolution and high-quality image synthesis, leading to various applications in image processing and manipulation [1, 2, 50]. In this paper, we follow [26] and implement our measure with a StyleGAN generator pre-trained on portrait images. Nevertheless, normalizing flows [19, 31, 34] that construct complex distributions by transforming a probability density function through a series of invertible mappings have attracted increasing attention in the past several years. In this paper, we employ the SRFlow [24] model to implement our measure, which directly learns the conditional distribution of the HR output given the LR input.

3. Our Approach

In this section, we first introduce the definition of our metric derived from Shannon's rate-distortion theory [5], and then detail how *deep generative models* help to sidestep the data scarcity challenge that impedes the application of the proposed metric.

3.1. Metric Definition

We create a proxy task, namely the *lossy compression problem* underpinned by Claude Shannon's rate-distortion theory [5], and formulate image downscaling as its encoding process:

$$\inf_{Q_f(\hat{x}|x)} \mathbb{E}[D_Q(X, \hat{X})] \ s.t. \ I_Q(X; \hat{X}) \leq R \quad (1)$$

where X is the set of input high-resolution images, $\hat{X}$ is the set of output reconstructed images, R is a rate constraint determined by the downscaling process[1], $Q_f(\hat{x}|x)$ or Q for short is the probability density function (PDF) of reconstructed HR images $\hat{x}$ conditioned on an input HR image x

---

[1]Note that in image downscaling, this constraint on R is always satisfied as the downscaled images are of a fixed resolution defined by users.

with respect to a given lossy image reconstruction function f that $\hat{x} = f(x) = f_{us}(f_{ds}(x))$, where f_{us} and f_{ds} denote image upscaling and downscaling functions respectively, D_Q is a distortion metric between two image sets where the image correspondence is determined by Q. Thus, we propose to use the expectation of the distortion as an evaluation metric for image downscaling:

$$S(f_{ds}) = \mathbb{E}[D_Q(X, \hat{X})] = \mathbb{E}_x\{\mathbb{E}_{\hat{x}|x}[D(x, \hat{x})]\}, \quad (2)$$

where $x \in X$, $\hat{x} \in \hat{X}$, D is a distortion metric between two images, *e.g.*, LPIPS [49]. The lower S, the better the downscaling algorithm f_{ds}. Although straightforward, the application of such a metric remained a challenge as it requires a strong upscaling function f_{us} that can:

- Reconstruct the input image x regardless of the input downscaling algorithm f_{ds}.
- Generate a conditional distribution of reconstructed images $\hat{x}|x$ for each x.

Between them, the first is commonly known as *blind image super-resolution* that is essentially a many-to-one mapping problem that aims to map different distorted downscaled images to the same high-resolution image; the second is commonly known as *one-to-many super-resolution* due to its ill-posed nature caused by the information loss during downscaling [24].

Data Scarcity Challenge. Combining the above two requirements makes the desired f_{us} an extremely challenging many-to-many mapping problem that has remained unsolved for decades. Specifically, the numerous kinds of distorted downscaled images and the corresponding countless high-resolution images for each of them makes it infeasible to collect sufficient data for supervised learning methods:

$$f_{us} = \arg\min_{f_\theta} \mathbb{E}_{I_{LR}}(\mathbb{E}_{I_{HR}}||I_{HR} - f_\theta(I_{LR})||) \quad (3)$$

where I_{HR} and I_{LR} denote the high-resolution (HR) and low-resolution (LR) training images respectively, $\mathbb{E}_{I_{HR}}$ indicates that there are many I_{HR} corresponding to the same I_{LR}, $\mathbb{E}_{I_{LR}}$ indicates that there are many I_{LR} obtained by different image downscaling methods f_{ds}.
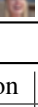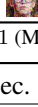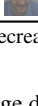
3.2. Evaluation with Deep Generative Models

Our key insight is that the above-mentioned data scarcity challenge (Eq. 3) can be overcome by the recent successes in deep generative modeling [4, 14, 16–19, 31, 33, 34]. In deep generative modeling, a neural network model is trained to learn a manifold of natural and high-resolution (HR) images from samples in the training dataset. This has been successfully applied to various image processing tasks [1, 2, 50]. To demonstrate the flexibility of our metric, we show its two implementations using two mainstream deep generative models: i) Generative Adversarial Networks (GANs) and ii) Normalizing Flows respectively as follows.

Table 1. IDA-RD scores for synthetic image downscaling with different types and levels of degradations (a), (b); with mixed degradations (c). The numbers in parentheses denote degradation parameters. As a reference, the IDA-RD score for the bicubic-downscaled image without degradation is 0.11±0.145. It is best to **Zoom In** to view the examples of downscaled images with different types and levels of degradations. ρ: Spearman' rank coefficient between our IDA-RD metric and levels of degradations, where 1/-1 means that they are monotonically correlated (positive or negative); Gauss. : Gaussian; Contrast Inc.: Contrast increase; Contrast Dec.: Contrast decrease. Please see Sec. 3 of the supplementary material for results on more types of degradation.

	Gauss. Blur		Gauss. Noise		Contrast Inc.
(1.0)	0.320±0.048	(0.05)	0.482±0.051	(1.5)	0.231±0.042
(2.0)	0.434±0.057	(0.10)	0.640±0.052	(2.0)	0.317±0.041
(4.0)	0.579±0.065	(0.20)	0.659±0.052	(2.5)	0.462±0.043

(a) $\rho = 1$ (Monotonic Increasing).

	Quantization		Contrast Dec.
(15)	0.164±0.002	(0.75)	0.330±0.047
(10)	0.205±0.003	(0.50)	0.644±0.074
(5)	0.463±0.064	(0.25)	0.669±0.034

(b) $\rho = -1$ (Monotonic Decreasing).

Gauss. Blur (1)	0.320±0.048
+ Gauss. Noise (0.05)	0.585±0.062
+ Contrast Dec. (0.75)	0.664±0.046
+ Quantization (10)	0.795±0.063

(c) Mixed Degradations.

Table 2. IDA-RD scores for synthetic image downscaling methods with different scaling factors. $(\cdot)$: the resolution of downscaled images. Bicubic: bicubic-downscaled image without degradation. G.B.: Gaussian Blur. The $32\times$ super-resolution is achieved by a concatenation of a $8\times$ and a $4\times$ upscaling implemented by pretrained SRFlow models.

S.F.	Bicubic	G.B. ($\sigma = 1.0$)	G.B. ($\sigma = 2.0$)	G.B. ($\sigma = 4.0$)
$4\times$ (256×256)	0.058±0.142	0.146±0.032	0.269±0.043	0.412±0.055
$8\times$ (128×128)	0.110±0.145	0.320±0.048	0.434±0.057	0.579±0.065
$32\times$ (32×32)	0.228±0.056	0.614±0.068	0.680±0.066	0.741±0.065

Implementation with a Pre-trained GAN generator. Similar to [26], we implement the upsampling function f_{us} in our metric using an optimization-based GAN inversion method [1, 2]. Leveraging the power of a pre-trained Style-GAN [16] generator G, we define our GAN-based f_{us} (Eq. 2) as locating the optimized StyleGAN latent code $\mathbf{z_i^*}$ so that its corresponding HR image $G(\mathbf{z_i^*})$ synthesized by G shares the same downscaled image as an input LR image $I_{LR} = f_{ds}(x)$:

$$f_{us}(I_{LR}, i) = G(\mathbf{z_i^*}) = \arg\min_{G(\mathbf{z_i})} ||I_{LR} - f_{ds}(G(\mathbf{z_i}))|| \quad (4)$$

where $I_{LR} = f_{ds}(x)$ denotes the input LR image downscaled by f_{ds}, $\mathbf{z_i}$ denotes the i-th randomly initialized latent code to be optimized to get the i-th sample from $\hat{x}|I_{LR}$ (i.e., $G(\mathbf{z_i^*})$), $i = 1, 2, 3, ...$ is the index. It can be observed that i) our f_{us} sidesteps the data scarcity challenge (Eq. 3) by using a StyleGAN generator that is trained with HR images only (i.e., without any many-to-many LR-HR training pairs); ii) it relocates the supervision to downscaling (i.e., enforcing different HR images to be downscaled to the same LR image) and thus outputs high quality HR images $G(\mathbf{z_i^*})$ that applies to an arbitrary choice of f_{ds}; iii) it is inherently stochastic given the random choices of $\mathbf{z_i}$.

Implementation with a Pre-trained Flow model. We use a pre-trained SRFlow model [24] that implements the f_{us}

in our metric with a conditional invertible neural network. Leveraging its invertible nature, f_{us} is trained to explicitly learn the conditional distribution $\hat{x}|I_{LR}$ by minimizing the negative log-likelihood:

$$f_{us} = \arg\min_{f_\theta} -\log p_{\mathbf{z}}(f_\theta(x|I_{LR})) \quad (5)$$

where $I_{LR} = f_{ds}^{\text{bicubic}}(x)$ is a bicubic downscaled image of HR input x, $\mathbf{z}$ denotes a random latent variable whose distribution encodes $\hat{x}|I_{LR}$ with a 'reparameterization trick'. Although trained with only bicubic downscaling, surprisingly, we observed that the resulting f_{us} can also be applied to evaluate other downscaling methods.

We use SRFlow in the final version of our metric as it shares similar performance as the GAN-based implementation but has a much lower time cost.

4. Experiments

To validate the effectiveness of our IDA-RD measure, we first test it with synthetic image downscaling methods whose performance are known beforehand (Sec. 4.2). Specifically, we simulate different types and levels of downscaling distortions by adding controllable degradations (e.g., Gaussian Blur, Contrast Change) to bicubic-downscaled images. In principle, the heavier the degradation, the worse the results

Table 3. (a) IDA-RD scores for real-world image downscaling methods (4×) on DIV2K [3], Flickr30k [46] and RealSR [8] datasets. N.N.: Nearest Neighbour. $L0$-reg.: L0-regularized. UD: "unknown downscaled" images provided by DIV2K. Camera: LR images "downscaled" by a camera provided by RealSR. (b) IDA-RD scores for real-world image downscaling methods with different scaling factors. S.F.: Scaling Factor, the resolutions of downscaled images (*e.g.*, 512×512 for 2×, 64×64 for 16×), are omitted for simplicity. Note that the relatively large standard deviations in some cases (especially when the scaling factors are small) indicate the algorithmic biases of image downscaling methods against individual images, *e.g.*, flat images with large color blocks may suffer less from information loss. The 32× super-resolution is achieved by a concatenation of a 8× and a 4× upscaling implemented by pretrained SRFlow models.

	Bicubic	Bilinear	N.N.	DPID	Perceptual	$L0$-reg.	Camera	UD
DIV2K	0.157±0.073	0.129±0.089	0.374±0.079	0.216±0.057	0.336±0.068	0.226±0.072	—	0.355±0.128
Flickr30K	0.263±0.102	0.239±0.112	0.452±0.105	0.357±0.097	0.367±0.080	0.364±0.103	—	—
RealSR	0.116±0.052	0.114±0.055	0.389±0.102	0.224±0.079	0.341±0.083	0.264±0.075	0.047±0.125	—

(a) IDA-RD scores for real-world image downscaling methods (4×).

S.F.	Bicubic	Bilinear	N.N.	DPID	Perceptual	$L0$-reg.
4×	0.058±0.142	0.031±0.053	0.335±0.310	0.122±0.234	0.388±0.321	0.136±0.251
8×	0.110±0.145	0.090±0.067	0.512±0.340	0.127±0.294	0.398±0.337	0.213±0.301
32×	0.228±0.056	0.272±0.056	0.601±0.163	0.291±0.076	0.514±0.152	0.307±0.050

(b) IDA-RD scores for real-world image downscaling methods with different scaling factors.

of downscaling, and the higher our measure should be. We also validate the effectiveness of our IDA-RD measure across different scaling factors. Then, we show that our measure can also be used to evaluate real-world image downscaling methods like Bicubic, Bilinear, Nearest Neighbour, and state-of-the-art downscaling methods like L0-regularized [23], Perceptual [29] and DPID [44] (Sec. 4.3). Please see Sec. 2 of the supplement for examples of downscaled images.

4.1. Experimental Setup

Dataset Unless specified, we use a balanced subset of 900 images from the FFHQ dataset [16], including face images at 1024×1024 resolution, as the set of input high-resolution images X in Eq. 2 for our IDA-RD measure. Please see Sec. 4 of the supplementary materials for more details on how we construct balanced subsets of images from FFHQ. We also use real-world datasets that contain images for all domains, including DIV2K [3], Flickr2K[2] and RealSR [8], for the evaluation. However, observing that SRFlow is unstable on them (Sec. 8 in supplementary material), we only use real-world datasets for the 4× downscaling assessment in Sec. 4.3 and use domain-specific datasets for other experiments.

Image Upscaling Algorithms We use SRFlow [24] as the f_{us} in Eq. 2. Specifically, we used the models provided by the authors for 4× and 8× super resolution that are pre-trained on DIV2K [3] and Flickr2K datasets. Unless specified, we use the 8× model for all experiments. For PULSE [26], we use the same StyleGAN generator pre-trained with FFHQ [16]. This model generates face images of size 1024×1024. We use a learning rate of 0.4, and stop the optimization for each image after 200 steps of spherical gradient descent. The noise signals of the StyleGAN generator were kept fixed.

Hyperparameters Unless specified, we use i) $N_Q = 5$ as the number of images upscaled from a single downscaled image for the estimation of Q in Eq. 2; ii) LPIPS [49] as the distortion measure D in Eq. 2; iii) $N_X = 900$ as the number of images in the set of high-resolution image X in Eq. 2.

4.2. Test with Synthetic Downscaling Methods

In this section, we demonstrate the effectiveness of our IDA-RD measure by testing its performance on synthetic downscaling methods. Without loss of generality, we simulate the effects of different downscaling methods by adding controllable degradations after bicubic downscaling, whose rationale is justified in Sec. 9 of the supplementary materials where we show that applying degradations before and after downscaling yield similar results.

4.2.1 Effectiveness across Degradation Types

As detailed below, we test our IDA-RD measure with four sets of synthetic downscaling methods that apply different types and levels of degradations to bicubic-downscaled images respectively and compute the Spearman coefficients ρ between levels of degradations and our IDA-RD metrics to assess their correlations.

Gaussian Blur. We apply Gaussian blur to the bicubic-downscaled images. The standard deviation of the blur kernel σ is chosen from $\{1.0, 2.0, 4.0\}$. The kernel size was set as 3. The results are shown in Table 1 (a).

Gaussian Noise. We add Gaussian noise to the bicubic-downscaled images. The standard deviation σ of the noise is chosen from $\{0.05, 0.1, 0.2\}$ (for reference, the mean intensity range of bicubic-downscaled images is $[0.022, 0.964]$). The results are shown in Table 1 (a).

---
[2]https://github.com/andreas128/SRFlow

Table 4. Ablation study of N_X for IDA-RD implemented with PULSE. Synthetic image downscaling methods with Contrast Decrease with $\sigma = 0.75$ (DG1); Gaussian Noise with $\sigma = 0.05$ (DG2); mixed noise consisting of Gaussian Blur with $\sigma = 1.0$, Contrast Decrease with $\sigma = 0.75$, and Gaussian Noise with $\sigma = 0.05$ (DG3); are used in the experiments.

N_X	30	300	600	900	1200	1500
DG1	0.351±0.014	0.342±0.019	0.343±0.012	0.339±0.022	0.339±0.021	0.339±0.023
DG2	0.361±0.011	0.383±0.011	0.374±0.012	0.351±0.023	0.353±0.022	0.352±0.021
DG3	0.471±0.011	0.483±0.012	0.391±0.013	0.293±0.019	0.289±0.022	0.291±0.021

Table 5. Ablation study of f_{us}, the image upscaling algorithms. PULSE [26] and SRFlow [24] have similar results but those of SRFlow are more distinguishable. Please see Sec. 5 of the supplementary materials for the results when using f_{us} based on stable diffusion.

	Bicubic	Bilinear	N.N.	DPID	Perceptual	$L0$-reg.
PULSE	0.171±0.015	0.164±0.015	0.254±0.018	0.179±0.016	0.223±0.017	0.205±0.016
SRFlow	0.110±0.145	0.090±0.067	0.512±0.340	0.127±0.294	0.398±0.337	0.213±0.301

Contrast Change. We apply contrast change to bicubic-downscaled images. To increase the contrast, we select the scale factor from $\{1.5, 2.0, 2.5\}$ in Table 1 (a). Note that such scaling can cause degradation due to the clipping of extreme intensity values. Similarly, to decrease the contrast, we select the contrast parameter from $\{0.25, 0.50, 0.75\}$ in Table 1 (b).

Quantization. We apply pixel quantization to bicubic-downscaled images and select the number of color thresholds from $\{5, 10, 15\}$. Specifically, we apply Otsu's multilevel thresholding algorithm [30] to the graylevel histogram which is derived from the color image, and then apply these thresholds uniformly to each of the RGB color channels. The results are shown in Table 1 (b).

Mixed Degradations. In addition to single degradations mentioned above, we also demonstrate the effectiveness of our IDA-RD measure on their mixtures. The results are shown in Table 1 (c).

It can be observed that our IDA-RD measure works as expected (*i.e.*, the stronger the degradation, the worse the downscaling algorithm, and the higher the IDA-RD) for all synthetic image downscaling methods, which demonstrates its effectiveness. In addition, we investigate the minimum degradation that causes differences in IDA-RD values in Sec. 10 of the supplementary materials, which justifies the effectiveness of IDA-RD in assessing small degradations.

4.2.2 Effectiveness across Scale Factors

We further demonstrate the effectiveness of our IDA-RD measure on synthetic downscaling algorithms across different scale factors. As Table 2 shows, we test our IDA-RD on synthetic downscaling algorithms of different levels of Gaussian Blur degradation as mentioned above. It can be observed that: i) the larger the scaling factor, the more the

information loss, and the higher the IDA-RD; ii) the stronger the degradation, the worse the downscaling algorithm, and the higher the IDA-RD; which justifies the validity of our IDA-RD measure.

4.3. Evaluating Existing Downscaling Methods

We apply our method to compare six existing downscaling algorithms, consisting of three traditional methods: Bicubic, Bilinear, Nearest Neighbor (N.N.), and three state-of-the-art methods: DPID [44], L0-regularized downscaling [23], and Perceptual [29] downscaling. Please see Sec. 12 of the supplementary materials for a visualization of the six downscaling methods. We conduct experiments on both real-world datasets, *i.e.*, DIV2K, Flickr30k and RealSR, which contain images for all domains, and FFHQ. As mentioned above in Sec. 4.1, we use FFHQ for the evaluation against different scaling factors as it is more stable. The results are shown in Table 3. For Table 3a, it can be observed that: i) when applied to classical downscaling algorithms (*i.e.*, Bicubic, Bilinear, and N.N.), our IDA-RD measure identifies the quality of these algorithms in the correct order (Bilinear > Bicubic > N.N.), although the difference between the results of Bicubic and Bilinear downscaling is not significant as expected; ii) our method can also evaluate the "unknown downscaling" in DIV2K and camera-captured LR images, which shows that camera-captured LR images do lose less information; iii) when applied to SOTA ones, the common belief is that these algorithms should perform better than Bilinear downscaling. However, none of these methods achieve a better IDA-RD, suggesting that although SOTA image downscaling methods excel in perceptual quality, they actually lose more information than Bilinear downscaling[3]. Nevertheless, it

---

[3]Note that our results do *not* contradict previous perception-based evaluations, but rather provide a new, objective and orthogonal dimension, *i.e.*, the extent to which they retain the information of their corresponding HR images.

Table 6. Ablation study of N_X, the number of images in test dataset X in Eq. 2 in the main paper. Synthetic image downscaling methods with Contrast Decrease with $\sigma = 0.75$ (DG1); Gaussian Noise with $\sigma = 0.05$ (DG2); mixed noise consisting of Gaussian Blur with $\sigma = 1.0$, Contrast Decrease with $\sigma = 0.75$, and Gaussian Noise with $\sigma = 0.05$ (DG3); are used in the experiments.

N_X	30	300	600	900	1200	1500
DG1	0.320±0.026	0.321±0.047	0.321±0.046	0.330±0.047	0.325±0.047	0.329±0.047
DG2	0.501±0.055	0.473±0.051	0.481±0.050	0.482±0.051	0.483±0.051	0.484±0.051
DG3	0.483±0.088	0.312±0.048	0.321±0.045	0.320±0.048	0.321±0.047	0.322±0.048

can be observed that DPID and L0-regularized methods are slightly better than Perceptual downscaling on our IDA-RD measure, which is consistent with previous understanding. These indicate that our IDA-RD measure is a useful complement to visual inspection, *i.e.*, a good image downscaling algorithm should be both visually satisfying and achieve a low IDA-RD score, which further validates the role of our measure in providing new insights into image downscaling algorithms. For Table 3b, it can be observed that the larger the scaling factor, the more the information loss, and the higher the IDA-RD, which is consistent with the observation of synthetic results. Please see Sec. 13 of the supplementary materials for a qualitative comparison and Sec. 6 of the supplementary materials for validation of our IDA-RD using "camera" images.

4.4. Time Complexity

Sec. 1 of the supplementary materials shows the running times of our IDA-RD measure using PULSE and SRFlow as f_{us} (Eq. 2 in the main paper) on an Nvidia RTX3090 GPU, respectively. It can be observed that the SRFlow implementation runs much faster, which justifies our choice of using it in our IDA-RD measure.

4.5. Ablation Study

In this experiment, we justify the algorithmic choices of our IDA-RD measure, *i.e.*, f_{us}, D, the number of images used to estimate Q and in X, and the content of X in Eq. 2, by performing a thorough ablation study on them.

Table 7. Ablation study of the contents of dataset X in Eq. 2 in the main paper. (1) Bicubic (2) Bilinear (3) Nearest Neighbor (N.N.) (4) DPID (5) Perceptual (6) $L0$-regularized.

	FFHQ	NPRportrait 1.0	AFHQ-Cat
(1)	0.110±0.145	0.119±0.166	0.107±0.029
(2)	0.090±0.067	0.100±0.101	0.091±0.033
(3)	0.512±0.340	0.329±0.292	0.277±0.103
(4)	0.127±0.294	0.119±0.099	0.152±0.047
(5)	0.398±0.337	0.391±0.231	0.289±0.067
(6)	0.213±0.301	0.166±0.234	0.211±0.025

Choice of f_{us}. As Table 5 shows, both PULSE [26] and SRFlow [24] have similar results when used as f_{us} in our

IDA-RD measure, *i.e.*, N.N. > Perceptual > L0-regularized > DPID > Bicubic > Bilinear. However, since SRFlow yields more distinguishable results and runs much faster (Table 1 in Sec. 1 of the supplementary materials), we use it in our IDA-RD measure. Nevertheless, our IDA-RD is very flexible (*i.e.*, not restricted to PULSE or SRFlow) and will benefit from future progresses of blind and stochastic super-resolution methods. The invalidity of non-blind or non-stochastic SR methods is discussed in Sec. 5.

Number of Images in X. As Table 6 shows, we investigate how many images are required in the test dataset X consisting of high-resolution images to achieve a robust estimation of IDA-RD, namely N_X. It can be observed that the results become stable when $N_X \geq 900$, so we choose $N_X = 900$ for our IDA-RD measure. We also justify this choice on the PULSE version of our measure. As Table 4 shows, we also investigate how many images are required in the test dataset X consisting of high-resolution images to achieve a robust estimation of IDA-RD implemented with PULSE [26]. Similarly, it can be observed that the results become stable when $N_X \geq 900$, which further justifies our choice of $N_X = 900$ for IDA-RD.

The Content of X. As Table 7 shows, in addition to FFHQ [16], we test our IDA-RD measure on another two datasets: the NPRportrait 1.0 benchmark set [35] and AFHQ-Cat [10]. Between them, we use all 60 images at around 800×1024 resolution from the NPRportrait 1.0 benchmark set as X, which was carefully constructed so as to include a controlled diversity of gender, age and ethnicity; we use a random sample of 900 images at 512×512 resolution from the AFHQ-Cat dataset as X. We test them with $4 \times$ image downscaling. It can be observed that our conclusions hold for all datasets, which further verifies the flexibility of our method against the content of X. Without loss of generality, we use FFHQ in our IDA-RD measure.

Number of Images used to Estimate Q. As Table 8 shows, for a downscaled image, we investigate how many images are required to be upscaled from it (by f_{us}) to achieve a robust estimation of the conditional distribution Q and thus our IDA-RD, namely N_Q. It can be observed that the results become stable when $N_Q \geq 5$, so we choose $N_Q = 5$ for our IDA-RD measure.

Table 8. Ablation study of N_Q, the number of images required for a robust estimation of Q in Eq. 2 in the main paper.

N_Q	1	3	5	10	15
Bicubic	0.103±0.141	0.109±0.142	0.110 ±0.145	0.111±0.145	0.110±0.145
Bilinear	0.090±0.069	0.090±0.067	0.090 ±0.067	0.090±0.067	0.090±0.067
N.N.	0.513±0.341	0.512±0.340	0.512 ±0.340	0.512±0.340	0.511±0.340

Table 9. Ablation study of D, the distortion measure in Eq. 2 of the main paper. Dec.: Decrease. Param.: Parameter.

	Param.	PSNR	SSIM	MS-SSIM	LPIPS
Contrast Dec.	0.75	22.137±4.020	0.834±0.159	0.881±0.101	0.330±0.047
	0.50	17.814±2.148	0.714±0.087	0.819±0.080	0.644±0.074
	0.25	14.790±1.461	0.578±0.072	0.579±0.028	0.669±0.034
Contrast Inc.	1.50	16.641±4.019	0.603±0.223	0.772±0.150	0.231±0.042
	2.00	13.450±3.539	0.482±0.192	0.693±0.131	0.317±0.041
	2.50	11.032±2.893	0.357±0.159	0.602±0.120	0.462±0.043
Gaussian Noise	0.05	20.784±0.160	0.597±0.004	0.648±0.071	0.482±0.051
	0.10	18.121±1.713	0.563±0.029	0.576±0.066	0.640±0.052
	0.20	16.120±1.751	0.520±0.029	0.376±0.066	0.659±0.052
Gaussian Blur	1.00	25.159±1.999	0.744±0.059	0.929±0.017	0.320±0.048
	2.00	22.365±1.875	0.646±0.073	0.849±0.033	0.434±0.057
	4.00	19.738±1.739	0.558±0.080	0.715±0.051	0.579±0.065

Choice of D. As Table 9 shows, we test different choices of D including multiple image distortion metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [43], MS-SSIM (Multi-Scale SSIM), and LPIPS [49]. Experimental results demonstrate a similar trend across all of them, indicating the flexibility of our IDA-RD measure. Nevertheless, since LPIPS is a more advanced metric that has been shown to be more consistent with human perception, we use it in the final version of our IDA-RD measure.

5. Motivation Justification

Invalidity of Non-blind and Non-stochastic SR method As Table 11 from Sec. 11 of supplementary materials shows, non-blind or non-stochastic SR methods i) ESRGAN [40], BSRGAN [48], and Real-ESRGAN [41] fail to distinguish among image downscaling algorithms; ii) SR3 [36] and RSR [9] are slightly better but still not comparable to SR-Flow; which justifies the choice of blind and stochastic SR methods in our IDA-RD.

Invalidity of NR-IQA Metrics As Table 12 from Sec. 11 of supplementary materials shows, existing NR-IQA metrics, such as NIQE [28] and BRISQUE [27], MANIQA[4] [45] and CONTRIQUE [25], are not suitable for the image downscaling problem, especially extreme downscaling. It can be observed that i) NIQE struggles to calculate proper scores at all resolutions below 128×128; ii) BRISQUE does not provide the correct scores at a resolution of 32×32; iii) MANIQA and CONTRIQUE also rely on the "naturalness"

of HR images that is not present in LR images, thus cannot distinguish between images with relatively high degradations (*e.g.* $\sigma = 2.0$ and $\sigma = 4.0$). Also, both MANIQA and CON-TRIQUE are biased in terms of image resolutions: MANIQA is trained with 224×224 images and thus achieves higher scores with 256×256 images; CONTRIQUE is trained with 500×500 images and achieves higher scores with 512×512 images. In contrast, our measure correctly shows that the higher the downscaling factor (*i.e.*, the lower the resolution), the greater the information loss (*i.e.*, the lower the quality).

6. Conclusion

In this paper, we presented Image Downscaling Assessment by Rate Distortion (IDA-RD), a quantitative measure for the evaluation of image downscaling algorithms. Our measure circumvents the requirement of a ground-truth LR image by measuring the distortion in the HR space, which is enabled by the recent success of blind and stochastic super-resolution algorithms based on deep generative models. We validate our approach by testing various synthetic downscaling algorithms, simulated by adding degradations, on various datasets. We also test our measure on real-world image downscaling algorithms, which further validates the role of our measure in providing new insights into image downscaling algorithms. Please see Sec. 14 of the supplementary materials for **Limitation and Future Work**.

Acknowledgements

---

[4]Please note that MANIQA won the first place in the NTIRE2022 Perceptual Image Quality Assessment Challenge Track 2 No-Reference competition. https://github.com/IIGROUP/MANIQA

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3, 4

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 3, 4

[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. 5

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017. 3

[5] Toby Berger. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*, 2003. 1, 3

[6] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *arXiv preprint arXiv:2103.04922*, 2021. 3

[7] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1): 206–219, 2017. 1, 2

[8] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 5

[9] Angela Castillo, María Escobar, Juan C Pérez, Andrés Romero, Radu Timofte, Luc Van Gool, and Pablo Arbelaez. Generalized real-world super-resolution through adversarial robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1855–1865, 2021. 8

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 7

[11] Deng-Ping Fan, ShengChuan Zhang, Yu-Huan Wu, Yun Liu, Ming-Ming Cheng, Bo Ren, Paul L Rosin, and Rongrong Ji. Scoot: A perceptual metric for facial sketches. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5612–5622, 2019. 4

[12] Eduardo SL Gastal and Manuel M Oliveira. Spectral remapping for image downscaling. *ACM Transactions on Graphics (TOG)*, 36(4):1–16, 2017. 1, 2

[13] Timothy Gerstner, Doug DeCarlo, Marc Alexa, Adam Finkelstein, Yotam I Gingold, and Andrew Nealen. Pixelated image abstraction. In *NPAR@ Expressive*, pages 29–36, 2012. 1

[14] 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. 3

[15] Chu Han, Qiang Wen, Shengfeng He, Qianshu Zhu, Yinjie Tan, Guoqiang Han, and Tien-Tsin Wong. Deep unsupervised pixelization. *ACM Transactions on Graphics (TOG)*, 37(6): 1–11, 2018. 1

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3, 4, 5, 7, 1

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 3

[19] Thomas A Keller, Jorn WT Peters, Priyank Jaini, Emiel Hoogeboom, Patrick Forré, and Max Welling. Self normalizing flows. In *International Conference on Machine Learning*, pages 5378–5387. PMLR, 2021. 3

[20] Hailan Kuang, Nan Huang, Shuchang Xu, and Shunpeng Du. A pixel image generation algorithm based on CycleGAN. In *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pages 476–480. IEEE, 2021. 1

[21] Yue Li, Dong Liu, Houqiang Li, Li Li, Zhu Li, and Feng Wu. Learning a convolutional neural network for image compact-resolution. *IEEE Transactions on Image Processing*, 28(3): 1092–1107, 2018. 2

[22] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2

[23] Junjie Liu, Shengfeng He, and Rynson WH Lau. L_0-regularized image downscaling. *IEEE Transactions on Image Processing*, 27(3):1076–1085, 2017. 1, 2, 5, 6

[24] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. SRFlow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020. 1, 3, 4, 5, 6, 7

[25] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 8

[26] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020. 1, 3, 4, 5, 6, 7

[27] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12): 4695–4708, 2012. 1, 2, 8

[28] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 1, 2, 8

[29] A Cengiz Oeztireli and Markus Gross. Perceptually based downscaling of images. *ACM Transactions on Graphics (TOG)*, 34(4):1–10, 2015. 1, 2, 5, 6

[30] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 6

[31] George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. 3

[32] Jordi Pont-Tuset and Ferran Marques. Measures and meta-measures for the supervised evaluation of image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2131–2138, 2013. 4

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3

[34] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015. 3

[35] Paul L Rosin, Yu-Kun Lai, David Mould, Ran Yi, Itamar Berger, Lars Doyle, Seungyong Lee, Chuan Li, Yong-Jin Liu, Amir Semmo, Ariel Shamir, Minjung Son, and Holger Winnemöller. NPRportrait 1.0: A three-level benchmark for non-photorealistic rendering of portraits. *Computational Visual Media*, 8(3):445–465, 2022. 7

[36] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 8

[37] Yunyi Shang and Hon-Cheng Wong. Automatic portrait image pixelization. *Computers & Graphics*, 95:47–59, 2021. 1

[38] Jin Sun and Haibin Ling. Scale and object aware image thumbnailing. *International Journal of Computer Vision*, 104 (2):135–153, 2013. 2

[39] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. In *arXiv preprint arXiv:2305.07015*, 2023. 2

[40] 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. 8

[41] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021. 8

[42] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. IEEE, 2003. 2

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 8

[44] Nicolas Weber, Michael Waechter, Sandra C Amend, Stefan Guthe, and Michael Goesele. Rapid, detail-preserving image downscaling. *ACM Transactions on Graphics (TOG)*, 35(6): 1–6, 2016. 2, 5, 6, 4

[45] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1191–1200, 2022. 8

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5

[47] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[48] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pages 4791–4800, 2021. 8

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 2, 3, 5, 8

[50] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. SEAN: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 3

Deep Generative Model based Rate-Distortion for Image Downscaling Assessment
Supplementary Material

Yuanbang Liang[1]   Bhavesh Garg[2]   Paul Rosin[1]   Yipeng Qin[1]

[1]School of Computer Science and Informatics, Cardiff University   [2]IIT Bombay & WadhwaniAI

{liangy32, rosinpl, qiny16}@cardiff.ac.uk, bh05avesh@gmail.com

1. Time Complexity

Table 1. Running times of our IDA-RD with PULSE and SRFlow as f_{us} (Eq. 2 in the main paper) respectively. N_X: the number of images in test dataset X in Eq. 2 in the main paper.

N_X	300	600	900
PULSE	3h08min	6h10min	9h08min
SRFlow	18min	35min	55min

Table 1 shows the running times of our IDA-RD measure using PULSE and SRFlow as f_{us} (Eq. 2 in the main paper) on an Nvidia RTX3090 GPU, respectively. It can be observed that the SRFlow implementation runs much faster, which justifies our choice of using it in our IDA-RD measure.

2. Examples of Downscaled Images used in our experiments

Table 6 and Table 7 show examples of images downscaled by synthetic and real-world image downscaling methods used in our experiments, respectively.

3. Additional Results for Different Types of Degradations

As Table 2 shows, we tested our IDA-RD using BSRGAN's more complex Type IV degradations. It can be observed that our IDA-RD remains effective across these additional degradation types.

Table 2. IDA-RD scores for synthetic image downscaling methods used in BSRGAN. The random degradation parameters for [G.N. levels, blur σ, JPEG noise] are: Random-1: [0.667, 0.026, 48]; Random-2: [0.824, 1.233, 75]; Random-3: [0.283, 1.719, 49]; Random-4: [0.404, 0.233, 35]; and Random-5: [0.771, 1.902, 50].

	Random-1	Random-2	Random-3	Random-4	Random-5
Type IV	0.537±0.002	0.820±0.004	0.410±0.001	0.0480±0.001	0.548 ±0.001

4. Balancing FFHQ into Age-, Gender-, and Race-Balanced Subsets

We balance the FFHQ dataset [16] into subsets (*i.e.*, X in Eq. 2 in the main paper) that are balanced in age, gender and ethnicity for a fair evaluation of our IDA-RD measure. For the gender and age labels of FFHQ images, we use those offered by the FFHQ-features-dataset[5]; for the ethnicity labels of FFHQ images, we use the recognition results of Deep-Face[6]. According to the above, we define i) four age groups: Minors (0-18), Youth (19-36), Middle Aged (36-54) and Seniors (54+); ii) three major ethnic groups: Asian, White and Black; iii) two gender groups: Male and Female. We apply K-means to cluster FFHQ images in 24 ($4\times3\times2$) groups and select images from them evenly to generate the subsets used in our experiments. As Table 8 shows, the subsets used in our experiments are highly-balanced in terms of age, gender and ethnicity.

5. IDA-RD Based on Stable Diffusion (SD)

As Table 3 shows, implementing our IDA-RD metric with SD models produces the same ranking as PULSE and SR-Flow, further validating the effectiveness of our method.

6. Validation Using "Camera" Images

The results in Table 4 show the same ranking of image downscaling algorithms by our IDA-RD metric, further validating the correctness of our approach. Notably, our method is superior as it does not require any reference images (*e.g.*, "camera" images).

7. IDA-RD Results on Lanczos Algorithm

As Table 4 and Table 5 show, the Lanczos algorithm loses slightly more information than the Bicubic and Bilinear algorithms, but less than the SOTA methods. This reflects a trend to sacrifice some information preservation for improved perceptual quality in image downscaling.

8. Results of SRFlow ($8\times$) on Real-world Datasets (Unstable)

As Fig. 1 shows, SRFlow becomes unstable for a scaling factor of $8\times$. For stable uses of SRFlow, we intentionally used domain-specific datasets in the main paper. Note that all state-of-the-art image downscaling methods (*i.e.*, Perceptual, L0-regularized, DPID) used in our experiments are general ones that are applicable to all domains (*i.e.*, not tuned for specific domains).

---

[5]https://github.com/DCGM/ffhq-features-dataset
[6]https://github.com/serengil/deepface

Table 3. Results of IDA-RD implementations using three SD-based methods: ResShift [47] and Diffbir [22], StableSR [39].

	Bicubic	Bilinear	N.N.	DPID	Perceptual	$L0$-reg.
ResShift($\times100$)	0.349±0.081	0.343±0.097	0.553±0.329	0.356±0.086	0.537±0.201	0.483±0.129
Diffbir($\times100$)	0.340±0.167	0.333±0.163	0.703±0.353	0.313±0.136	0.681±0.217	0.437±0.192
StableSR($\times100$)	0.680±0.243	0.650±0.226	0.773±0.341	0.697±0.252	0.739±0.274	0.698±0.187

Figure 1. SRFlow becomes unstable for a scaling factor of $8\times$ on real-world datasets, *e.g.*, DIV2K (Row 1), while such cases never happen for domain-specific datasets, *e.g.*, FFHQ (Row 2). From the left to right, the method to down scaling are N.N., DPID, Perceptual and $L0$-reg. separately.
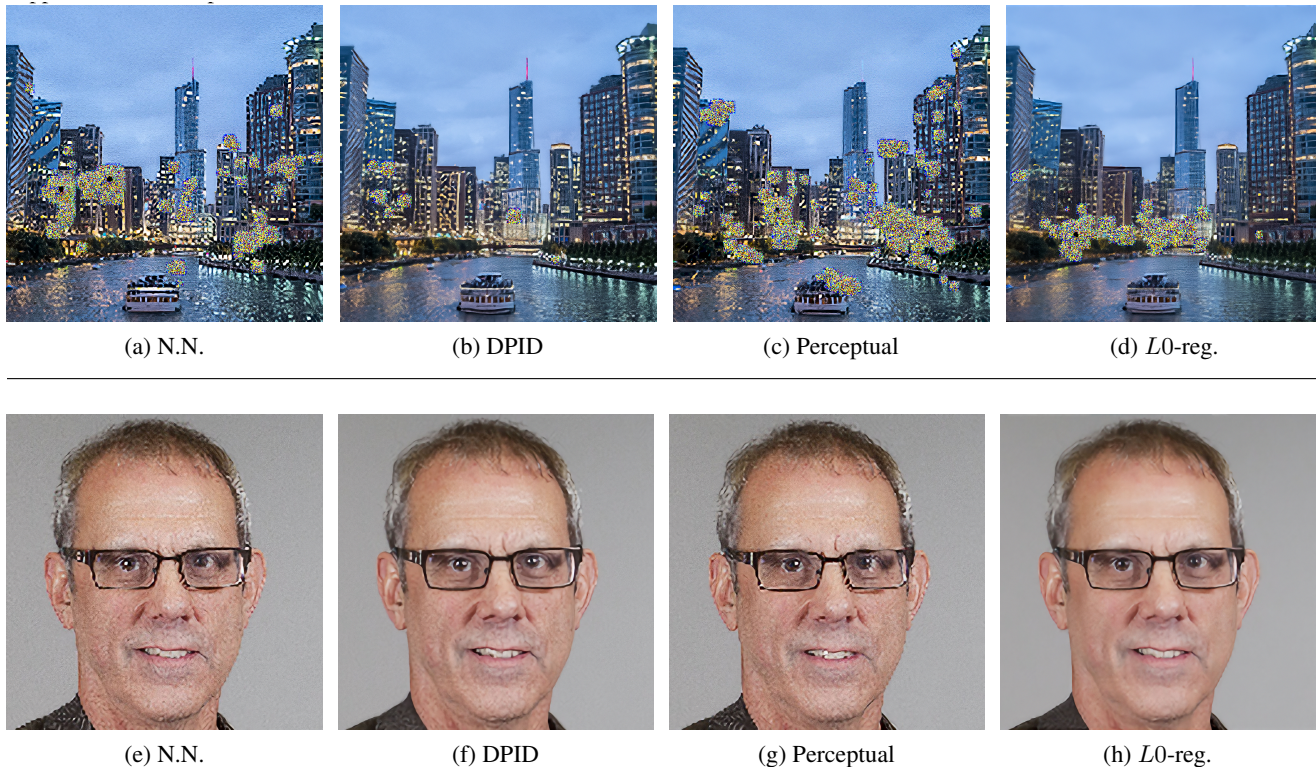

(a) N.N. (b) DPID (c) Perceptual (d) $L0$-reg.

(e) N.N. (f) DPID (g) Perceptual (h) $L0$-reg.

Table 4. Comparison of image downscaling algorithms on the RealSR dataset using its "camera" images as the "ground truth".

	SSIM↑	PSNR↑	LPIPS↓
Bicubic	0.900 ± 0.046	29.870 ± 2.857	0.167 ± 0.070
Bilinear	0.922 ± 0.036	30.163 ± 2.907	0.132 ± 0.059
Lanczos	0.886 ± 0.053	28.072 ± 2.837	0.191 ± 0.079
N.N.	0.827 ± 0.078	25.713 ± 2.881	0.247 ± 0.105
$L0$-reg.	0.858 ± 0.071	26.278 ± 2.901	0.228 ± 0.099
DPID	0.869 ± 0.065	26.964 ± 2.838	0.225 ± 0.098
Perceptual	0.840 ± 0.085	25.842 ± 2.795	0.239 ± 0.102

Table 5. Additional experiments of the Lanczos algorithm. (a)(b): extension to Table 7 of the main paper; (c) extension to Table 3(a) of the main paper.
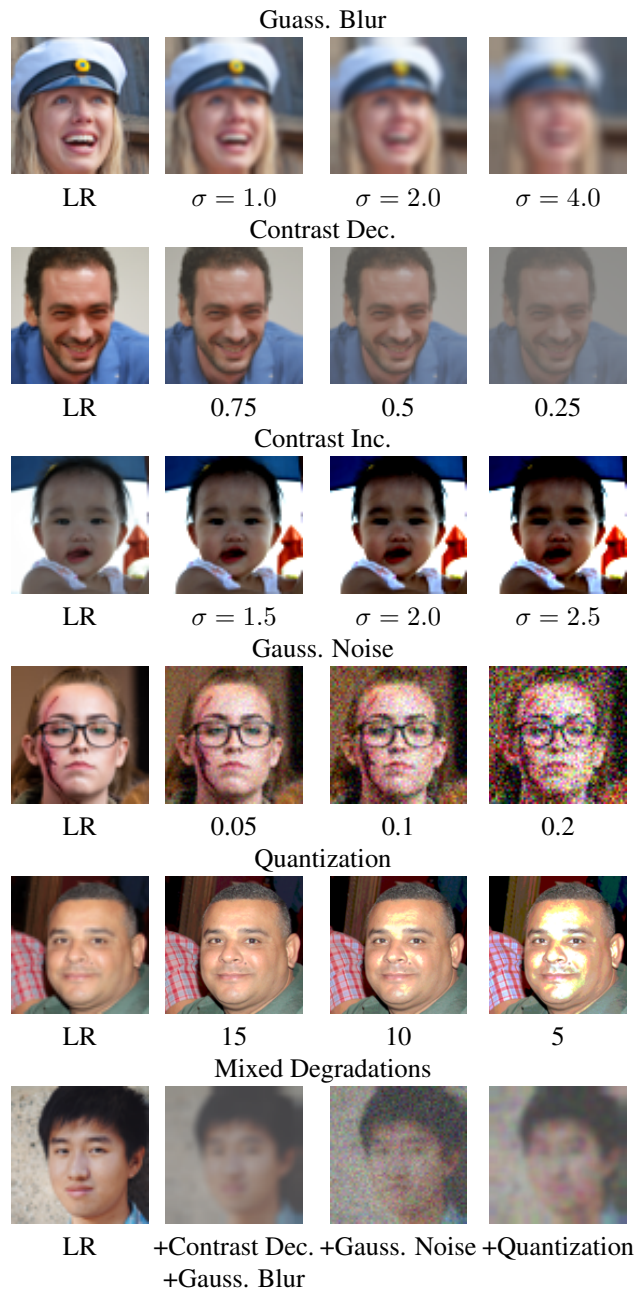
	(a) FFHQ	(b) AFHQ-Cat	(c) RealSR
Lanczos	0.121±0.287	0.142±0.045	0.120±0.133

degradation after downscaling. We therefore conclude that either approach yields valid synthetic downscaling methods.

10. Minimum Degradation that Causes Differences in IDA-RD Values

As Table 10 shows, the minimum degradations that cause differences in IDA-RD values (*e.g.*, for Gauss. Blur, when the degradation parameter changes from 0.0001 to 0.0005, the IDA-RD slightly increases from 0.111±0.034 to

9. Test with Synthetic Downscaling Methods - Degradation Applied Before Downscaling

As Table 9 shows, it can be observed that applying degradation before downscaling yields similar results to applying

Table 6. Examples of images downscaled by synthetic image downscaling methods, *i.e.*, those adds controllable degradations to bicubic-downscaled images (Sec. 4.2 in the main paper). The numbers below images are the degradation parameters. LR: bicubic-downscaled images, Dec.: decrease, Inc.: increase, Gauss.: Gaussian.

Guass. Blur

| LR | $\sigma = 1.0$ | $\sigma = 2.0$ | $\sigma = 4.0$ |

Contrast Dec.

| LR | 0.75 | 0.5 | 0.25 |

Contrast Inc.

| LR | $\sigma = 1.5$ | $\sigma = 2.0$ | $\sigma = 2.5$ |

Gauss. Noise

| LR | 0.05 | 0.1 | 0.2 |

Quantization

| LR | 15 | 10 | 5 |

Mixed Degradations

| LR | +Contrast Dec. +Gauss. Blur | +Gauss. Noise | +Quantization |

0.112±0.034), indicating that our IDA-RD is stable against small degradations. Note that the baseline IDA-RD, *i.e.*, no degradation, is 0.110.

11. Motivation Justification

As Table 11 shows, non-blind or non-stochastic SR methods are slightly better but still not comparable to SRFlow.

As Table 12 shows, existing NR-IQA metrics are not suitable for the image downscaling problem, especially extreme

Table 7. Examples of images downscaled by real-world image downscaling methods. N.N.: Nearest Neighbour; $L0$-reg.: $L0$-regularized.

| Bicubic | Bilinear | N.N. | DPID | Perceptual | $L0$-reg. |

Table 8. Statistics of our balanced FFHQ subsets. MI: Minors, Y: Youth, MA: Middle Aged, S: Senior; A: Asian, W: White, B: Black; M: Male, F: Female. J.E.: Joint Entropy, which measures the extent to which a subset is balanced. As a reference, a fully-balanced subset has a joint entropy of $-24 * (1/24) * \log_2(1/24) \approx 4.5850$.

Size	Age				Ethnicity			Gender		J.E.
	MI	Y	MA	S	A	W	B	M	F	
30	6	9	7	8	10	10	10	15	15	4.2817
300	76	75	70	79	102	100	98	150	150	4.4998
600	168	142	141	149	200	194	206	329	271	4.5245
900	222	227	215	236	304	295	301	452	448	4.5343
1200	445	442	453	460	608	591	601	902	898	4.5375
1500	684	664	673	679	909	887	904	1352	1348	4.5386

downscaling.

12. Visualization of Existing Downscaling Methods

As Fig. 2 shows, state-of-the-art (SOTA) image downscaling methods improve the perceptual quality by selectively "enhancing" image features (DPID explicitly mentioned that it "assigns larger weights to pixels that deviate more from their local image neighborhood" [44]), *e.g.*, the glasses frames and clothes patterns in Fig. 2 (i-c,d,e,f); the tessellation gaps in Fig. 2 (ii-c,d,e,f); the hair and watermelon seeds (clothes pattern) in Fig. 2 (iii-c,d,e,f). Nevertheless, selectively "enhancing" perceptually-important features means downweighting all other features, resulting in higher uncertainty (*i.e.*, information loss) when reconstructing other features during SR. Since the number of perceptually-important features is typically less than the number of other features, SOTA image downscaling methods lose more information, resulting in higher IDA-RD scores. Please note that N. N. shares a similar idea but uses a very simple "selection" method, thus losing a large amount of information as well.

13. Qualitative Evaluation of Existing Downscaling Methods

As Fig. 3 shows, state-of-the-art image downscaling methods achieve better perceptual quality by "exaggerating" perceptually important features in the original image (*e.g.*, building lights, water reflections), thus leading to over-exaggeration

in the upscaled images. As a result, they have lower IDA-RD scores than bicubic and bilinear downscaling.

14. Limitation and Future Work

Limitations. Since our measure makes use of GAN- and Flow-based super-resolution (SR) models, the limitations of these models are carried over as well. First of all, we cannot use test data beyond the learnt distribution of the SR model. For example, unlike the SRFlow [24] model trained on general images that are used in the main paper, our GAN-based implementation uses a StyleGAN generator pre-trained on portrait images, which only allows for the use of portrait face images to evaluate downscaling algorithms. Also, although highly unlikely to occur, we cannot evaluate downscaling algorithms whose output images are of higher quality than those generated by the SR model (*i.e.*, no distortion).

Future work. Our framework still requires a ground truth HR image. However, we believe the distortion can be calculated without such a ground truth image. To further validate our IDA-RD measure, in the future we will we use the *meta-measure* methodology [11, 32], in which secondary, easily quantifiable measures are constructed to quantify the performance of a less easily quantifiable measure.

Table 9. IDA-RD scores for synthetic image downscaling with different types and levels of degradations (degradation applied before downscaling). The numbers in parentheses denote degradation parameters.

	Gauss. Blur		Gauss. Noise		Contrast Inc.		Contrast dec.		Quantization
(1.0)	0.321±0.048	(0.05)	0.480±0.031	(1.5)	0.234±0.042	(0.75)	0.330±0.047	(15)	0.162±0.015
(2.0)	0.432±0.050	(0.10)	0.64±0.052	(2.0)	0.317±0.043	(0.50)	0.644±0.070	(10)	0.205±0.013
(3.0)	0.579±0.055	(0.20)	0.658±0.052	(2.5)	0.462±0.043	(0.25)	0.669±0.034	(5)	0.464±0.054
Spear.	1.000		1.000		1.000		-1.000		-1.000

Table 10. The minimum degradations that cause differences in IDA-RD values. The numbers in parentheses denote degradation parameters.

	Gauss. Blur		Gauss. Noise		Contrast Inc.		Contrast dec.		Quantization
(0.0001)	0.111±0.034	(0.0001)	0.110±0.029	(1.001)	0.111±0.034	(0.999)	0.111±0.034	(19)	0.111±0.035
(0.0005)	0.112±0.034	(0.0005)	0.110±0.029	(1.005)	0.111±0.034	(0.995)	0.111±0.034	(18)	0.182±0.038
(0.0010)	0.113±0.035	(0.0010)	0.118±0.054	(1.010)	0.115±0.029	(0.990)	0.112±0.031	(17)	0.193±0.041
(0.0050)	0.113±0.035	(0.0030)	0.118±0.062	(1.050)	0.120±0.032	(0.950)	0.113±0.032		—-
(0.0100)	0.113±0.036	(0.0040)	0.203±0.062	(1.100)	0.126±0.029	(0.900)	0.119±0.032		—-
(0.0500)	0.114±0.034	(0.0050)	0.291±0.062	(1.150)	0.126±0.029	(0.850)	0.123±0.031		—-
(0.1000)	0.118±0.042	(0.0100)	0.318±0.061	(1.200)	0.130±0.029	(0.800)	0.131±0.032		—-
(0.2500)	0.202±0.043		—-		—-		—-		—-
(0.3000)	0.214±0.044		—-		—-		—-		—-
Spear.	0.983		0.982		0.982		-0.991		-1.000

Table 11. Invalidity of using ESRGAN, SR3, BSRGAN, RSR and Real-ESRGAN in our IDA-RD measure.

	Bicubic	Bilinear	N.N.	DPID	Perceptual	$L0$-reg.
ESRGAN	0.022±0.012	0.017±0.006	0.058±0.016	0.025±0.009	0.024±0.004	0.024±0.007
BSRGAN	0.010±0.008	0.011±0.008	0.024±0.022	0.013±0.011	0.025±0.018	0.011±0.008
Real-ESRGAN	0.014±0.010	0.015±0.011	0.026±0.022	0.016±0.012	0.026±0.017	0.017±0.013
SR3	0.169±0.048	0.164±0.047	0.179±0.040	0.171±0.044	0.172±0.043	0.171±0.049
RSR	0.231±0.071	0.208±0.095	0.423±0.132	0.288±0.099	0.379±0.123	0.231±0.071

Table 12. Results of NIQE, BRISQUE, MANIQA and CONTRIQUE at higher resolutions.

Resolution	LR	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 4.0$
1024×1024	3.700	4.158	5.173	6.471
512×512	2.406	3.959	5.574	6.299
256×256	3.047	4.611	7.133	6.792
128×128	18.873	18.872	18.870	18.869
64×64	18.872	18.872	18.870	18.869
32×32	18.873	18.869	18.870	18.867

(a) NIQE scores (lower is better)

Resolution	LR	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 4.0$
1024×1024	26.792	32.827	48.971	59.043
512×512	19.536	33.391	57.447	63.144
256×256	28.582	39.282	55.747	65.990
128×128	16.045	34.423	47.017	55.166
64×64	41.360	42.417	43.346	54.344
32×32	43.458	43.458	44.015	43.668

(b) BRISQUE scores (lower is better)

Resolution	LR	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 4.0$
1024×1024	0.513	0.481	0.475	0.475
512×512	0.624	0.614	0.612	0.612
256×256	0.679	0.676	0.6762	0.676

(c) MANIQA scores (higher is better)

| Resolution	LR	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 4.0$
1024×1024	54.989	33.965	32.037	32.114
512×512	64.600	52.143	49.588	49.588
256×256	57.145	55.847	55.538	55.538
128×128	50.782	50.595	50.557	50.557
64×64	55.608	55.591	55.577	55.577
32×32	54.569	54.572	54.568	54.568

(d) CONTRIQUE scores (higher is better)

Figure 2. Examples of images ($\times 8$) from FFHQ, DIV2K and Flickr30K datasets downscaled by real-world image downscaling methods. (a) Bicubic (b) Bilinear (c) Nearest Neighbor (N.N.) (d) DPID (e) Perceptual (f) $L0$-regularized
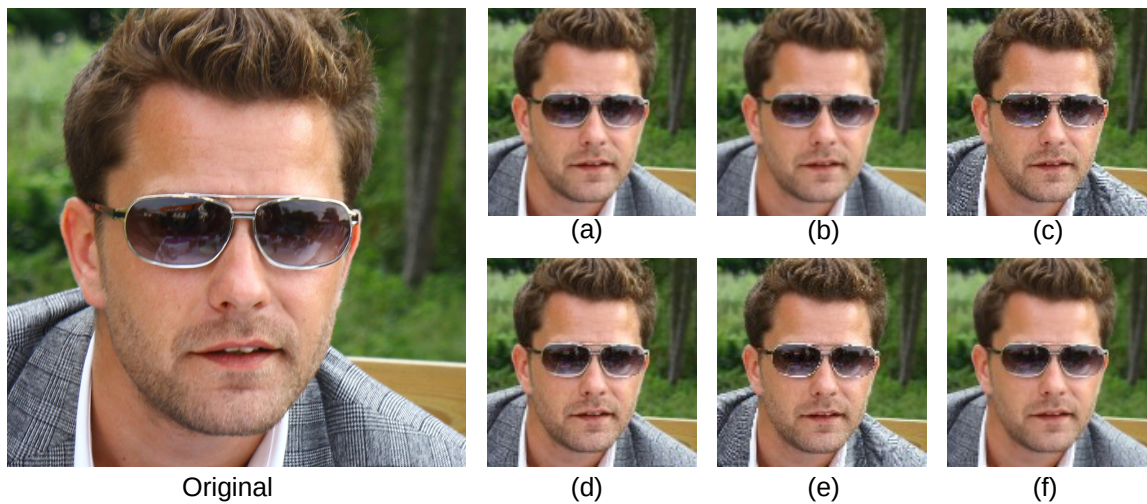

(i) Example from FFHQ
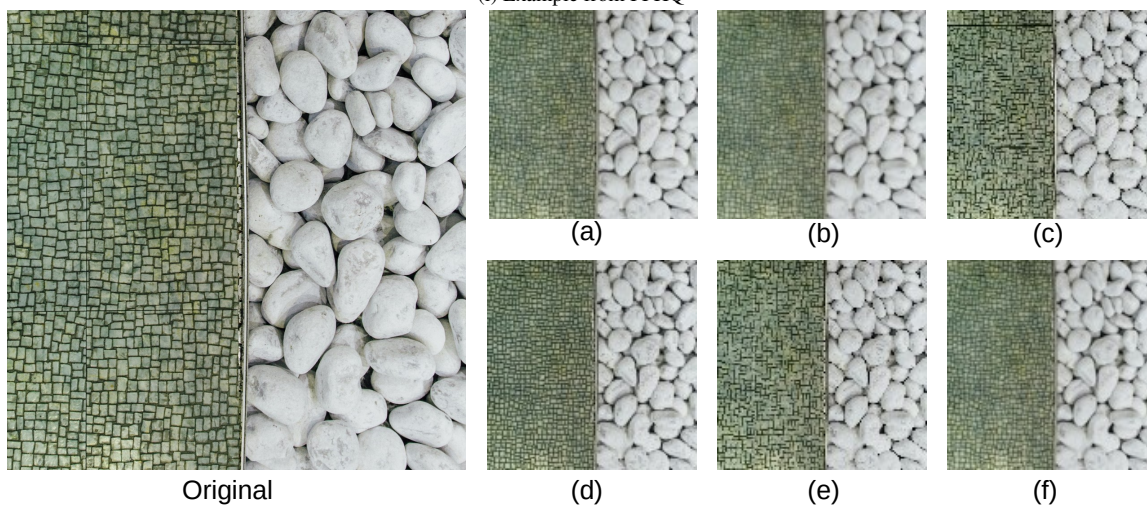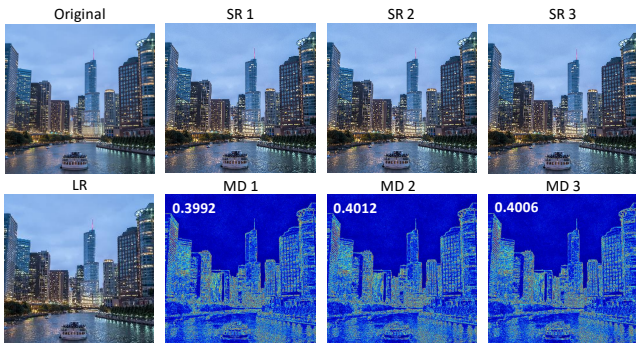
(ii) Example from DIV2K
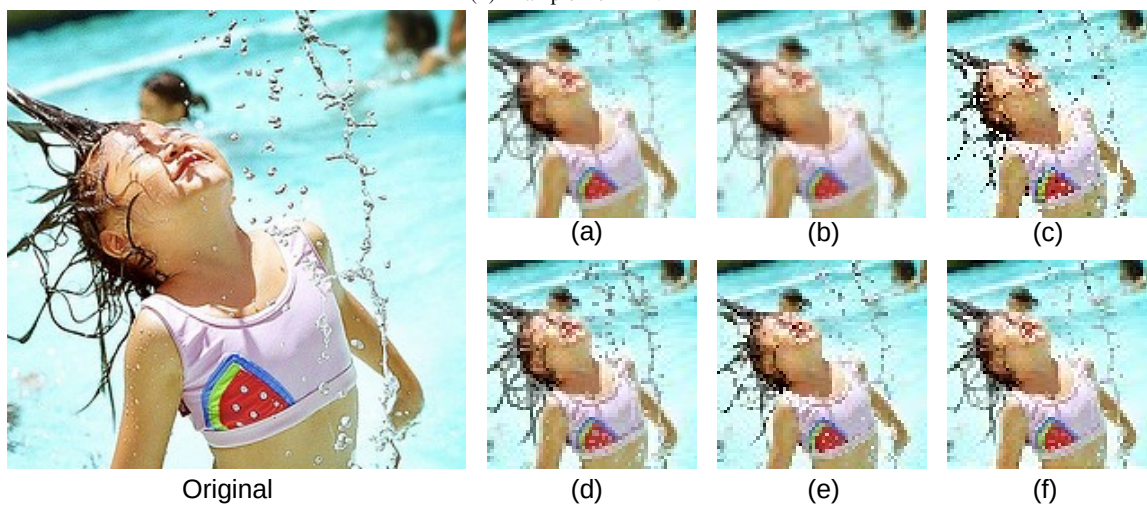
(iii) Example from Flickr30K

Figure 3. Qualitative evaluation of existing image downscaling methods. Original: the input HR image; LR: the downscaled LR image; SR1, SR2, SR3: three instances of upscaled images; MD1, MD2, MD3: difference map visualizations of (SR1, Original), (SR2, Original), and (SR3, Original), respectively. The white numbers on the left-top corners: the corresponding LPIPS scores of the difference map visualizations. State-of-the-art image downscaling methods (DPID, Perceptual and $L0$-reg.) achieve better perceptual quality by "exaggerating" perceptually important features in the original image (*e.g.*, building lights, water reflections), thus leading to over-exaggeration in the upscaled images and lower IDA-RD scores.
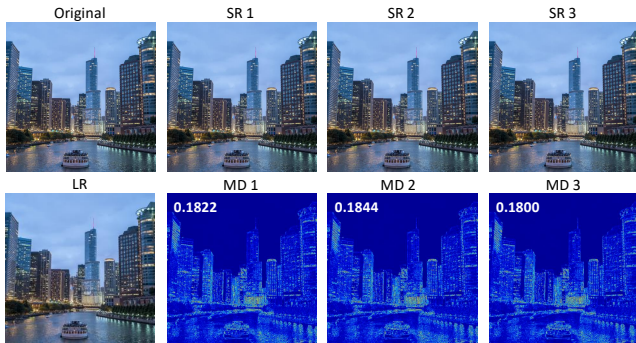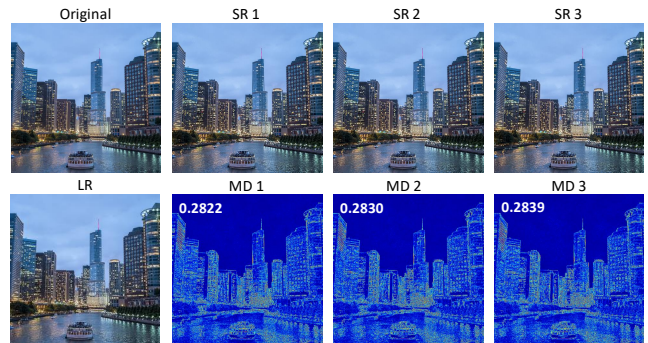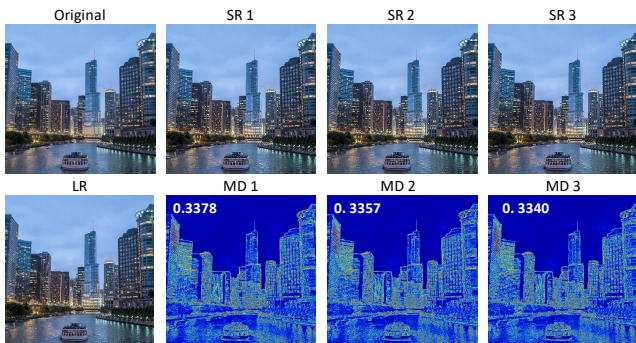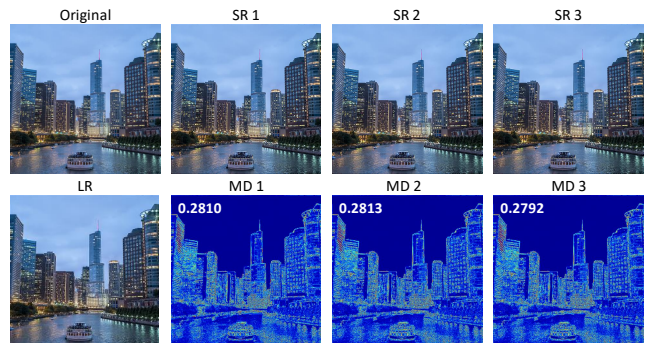

(a) Bicubic
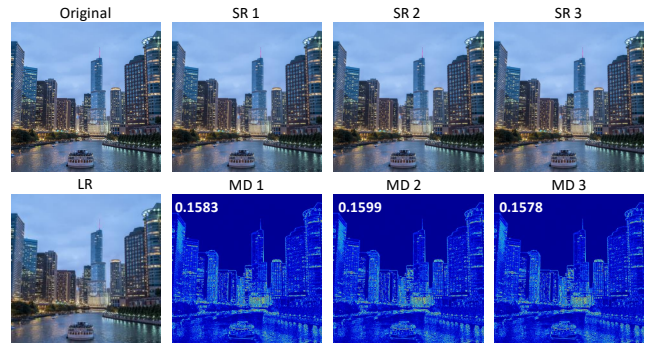
(b) Bilinear

(c) N.N.

(d) DPID

(e) Perceptual

(f) $L0$-reg.