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 DeepFace[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 (4×3×2) 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 SRFlow, 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×) on Real-world Datasets (Unstable)

As Fig. 1 shows, SRFlow becomes unstable for a scaling factor of 8×. 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($\times$100)	0.349±0.081	0.343±0.097	0.553±0.329	0.356±0.086	0.537±0.201	0.483±0.129
Diffbir($\times$100)	0.340±0.167	0.333±0.163	0.703±0.353	0.313±0.136	0.681±0.217	0.437±0.192
StableSR($\times$100)	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

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

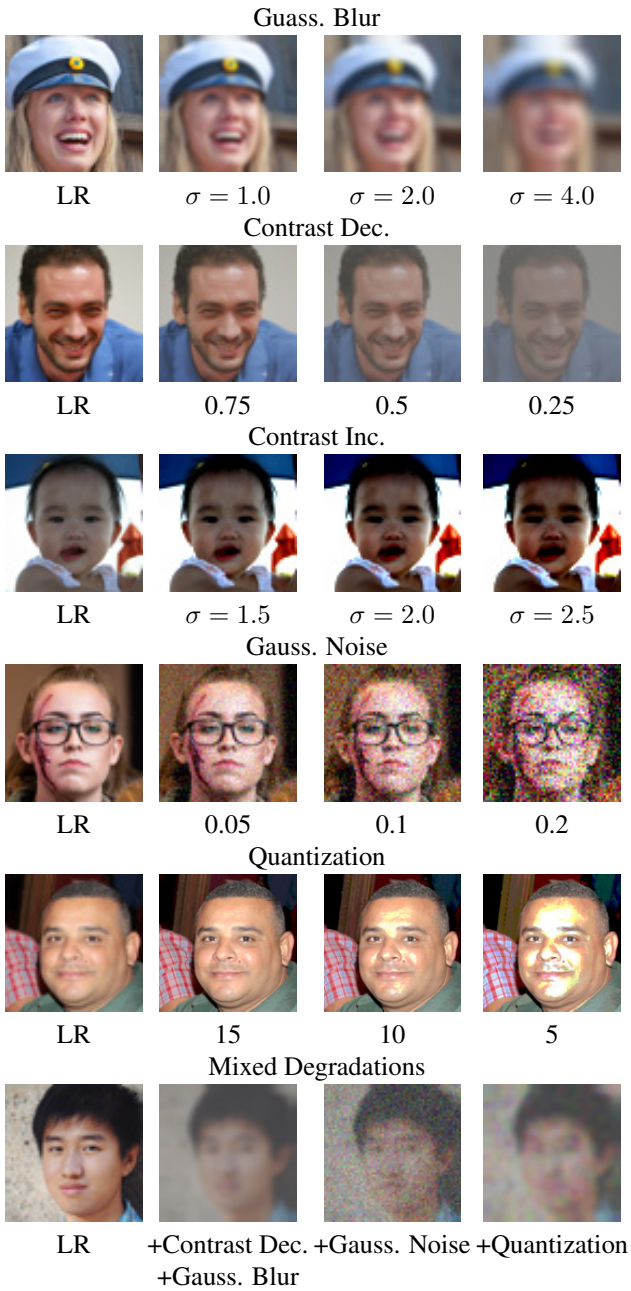
Table 5. Additional experiments of the Lanczos algorithm. (a)(b): extension to Table 7 of the main paper; (c) extension to Table 3(a) of the main paper.

	(a) FFHQ	(b) AFHQ-Cat	(c) RealSR
Lanczos	0.121±0.287	0.142±0.045	0.120±0.133

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

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.



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)

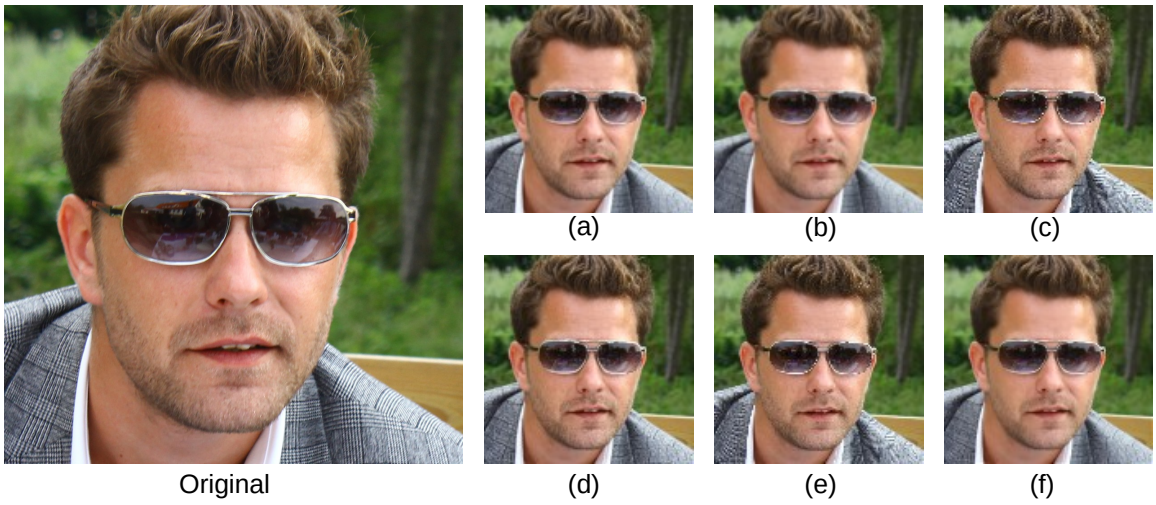
Resolution	LR	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 4.0$
1024×1024	0.513	0.481	0.475	0.475
512×512	0.624	0.614	0.612	0.612
256×256	0.679	0.676	0.6762	0.676

(c) MANIQA scores (higher is better)

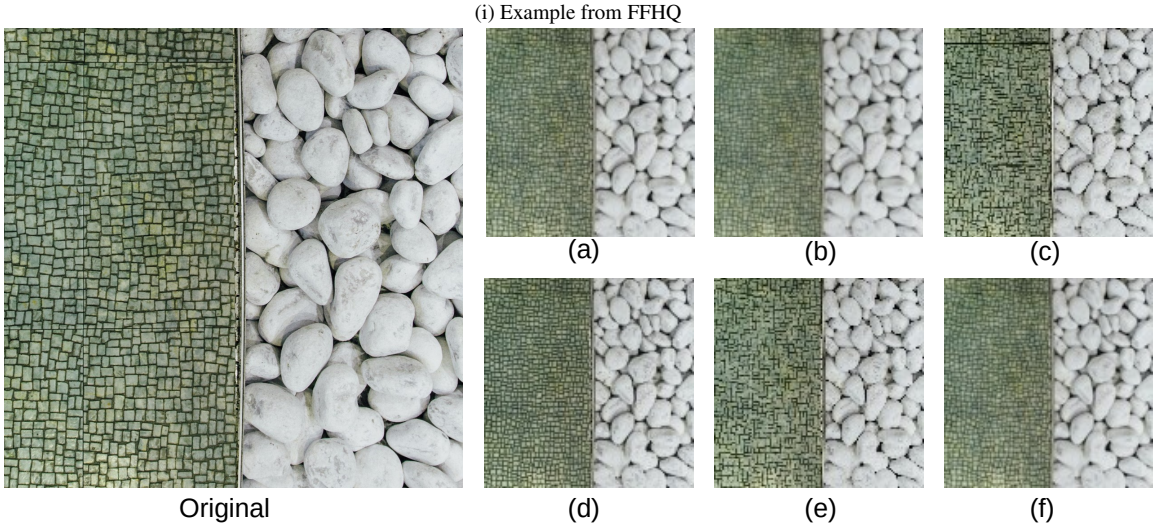
Resolution	LR	$\sigma = 1.0$	$\sigma = 2.0$	$\sigma = 4.0$
1024×1024	54.989	33.965	32.037	32.114
512×512	64.600	52.143	49.588	49.588
256×256	57.145	55.847	55.538	55.538
128×128	50.782	50.595	50.557	50.557
64×64	55.608	55.591	55.577	55.577
32×32	54.569	54.572	54.568	54.568

(d) CONTRIQUE scores (higher is better)

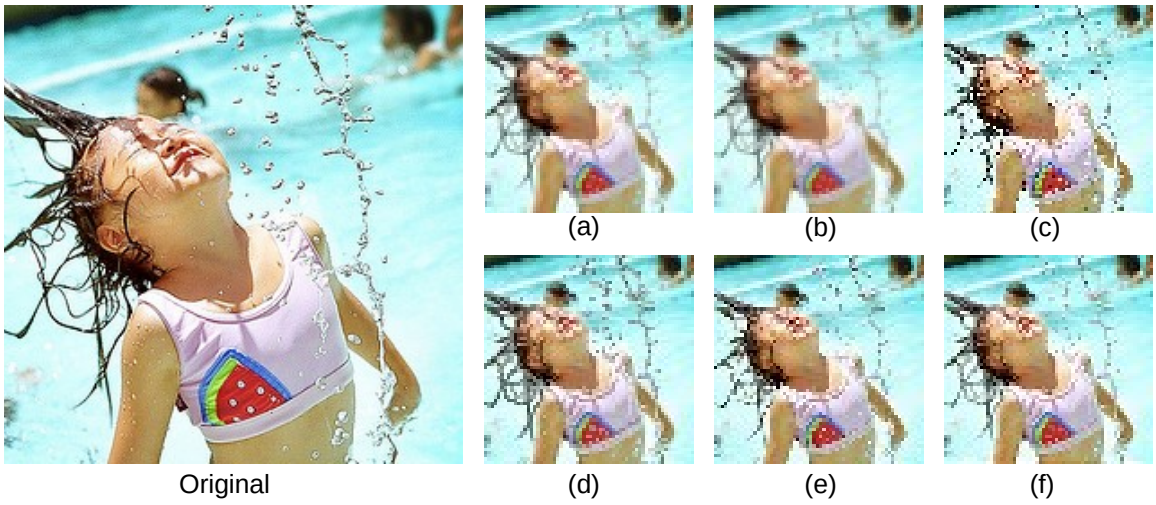
Figure 2. Examples of images ($\times 8$) from FFHQ, DIV2K and Flickr30K datasets downscaled by real-world image downscaling methods. (a) Bicubic (b) Bilinear (c) Nearest Neighbor (N.N.) (d) DPID (e) Perceptual (f) $L0$-regularized



(i) Example from FFHQ

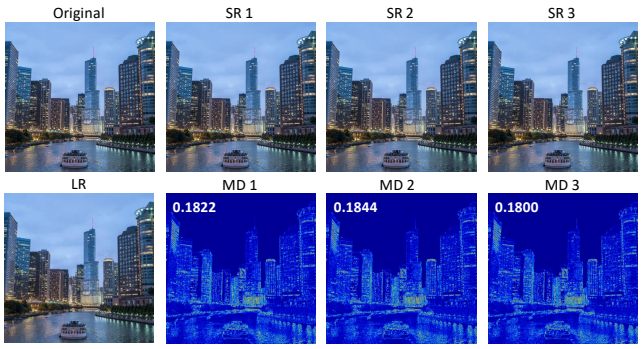


(ii) Example from DIV2K

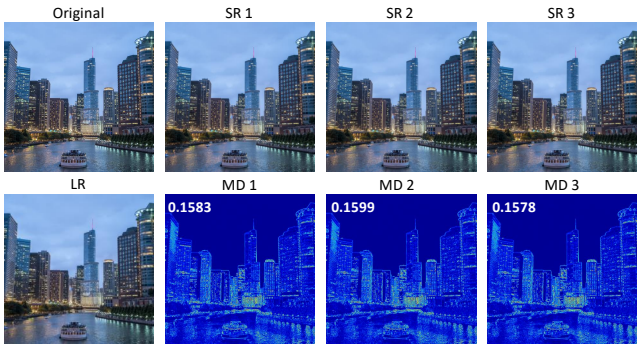


(iii) Example from Flickr30K

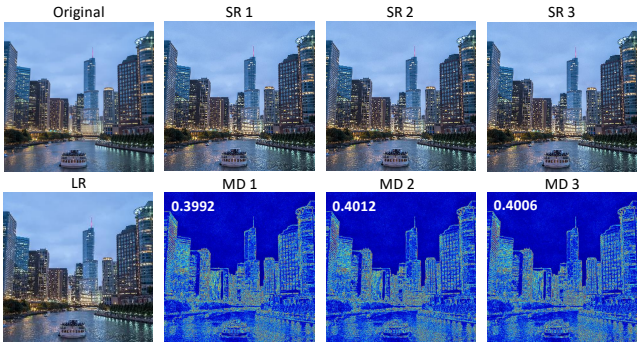
Figure 3. Qualitative evaluation of existing image downscaling methods. Original: the input HR image; LR: the downscaled LR image; SR1, SR2, SR3: three instances of upscaled images; MD1, MD2, MD3: difference map visualizations of (SR1, Original), (SR2, Original), and (SR3, Original), respectively. The 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.



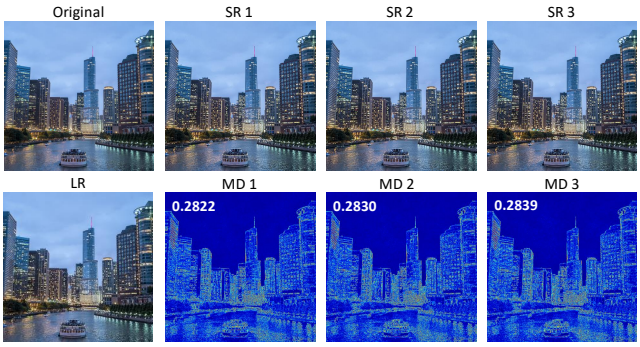
(a) Bicubic



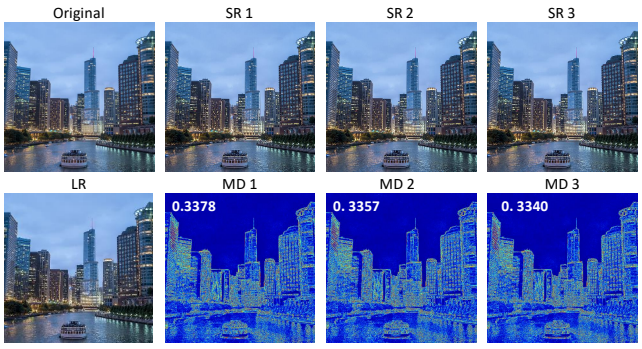
(b) Bilinear



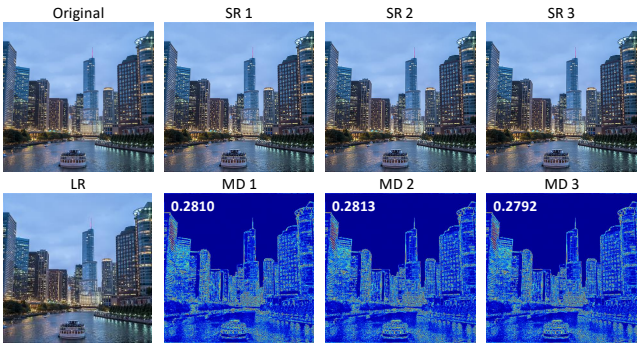
(c) N.N.



(d) DPID



(e) Perceptual



(f) $L0$-reg.