Normalizing Flow as a Flexible Fidelity Objective for Photo-Realistic Super-resolution

Supplementary Material

Andreas Lugmayr Martin Danelljan Fisher Yu Luc Van Gool Radu Timofte
CVL, ETH Zürich, Switzerland

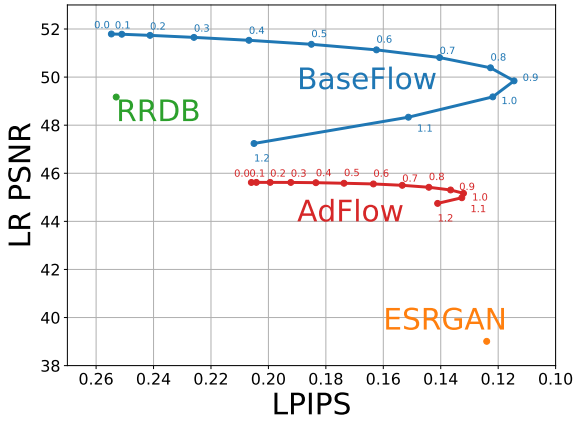

Figure 1. Influence of temperature parameter τ during inference on perception and fidelity. As opposed to AdFlow, ESRGAN and RRDB have only a single operating point. ($4\times$ super-resolution)

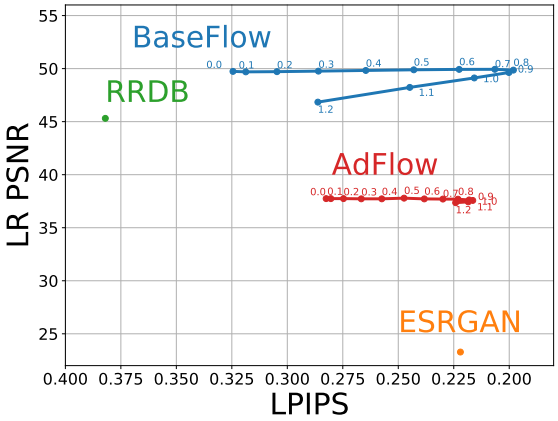

Figure 2. Influence of temperature parameter τ during inference on perception and fidelity. As opposed to AdFlow, ESRGAN and RRDB have only a single operating point. ($6\times$ super-resolution)

In this supplementary material, we first provide further details on the user study in Sec. 1. Secondly, we provide an analysis of the sampling temperature in Sec 2. Third, we present additional details about the minimally generalized L_1 loss in Sec. 3. Finally, we provide a further qualitative and quantitative comparison of AdFlow with other state-of-the-art methods in Sec. 4. Additional visual results, used in our study, will be available on the project page git.io/AdFlow.

1. User Study

As described in Sec. 4.1 in the main paper, we conduct the user study. The GUI interface is shown in Figure 4. We ask the user to evaluate which image of the two looks more realistic. To select the chosen image, the user presses the `1` key for the left and `2` key for the right images. Once a selection was made, the user can see the next image using the arrow right key until they have completed all tasks. Finally, the form is submitted using the button on the top right.

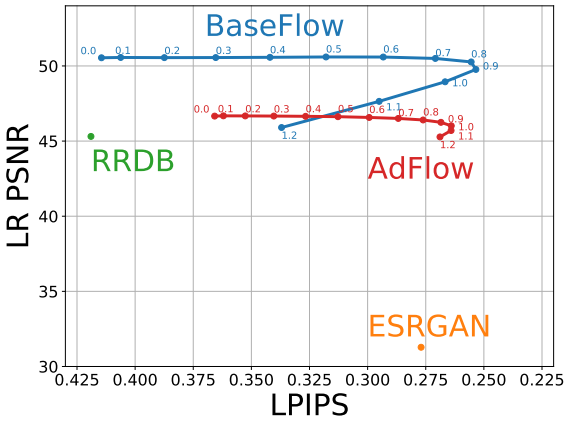

Figure 3. Influence of temperature parameter τ during inference on perception and fidelity. As opposed to AdFlow, ESRGAN and RRDB have only a single operating point. ($8\times$ super-resolution)

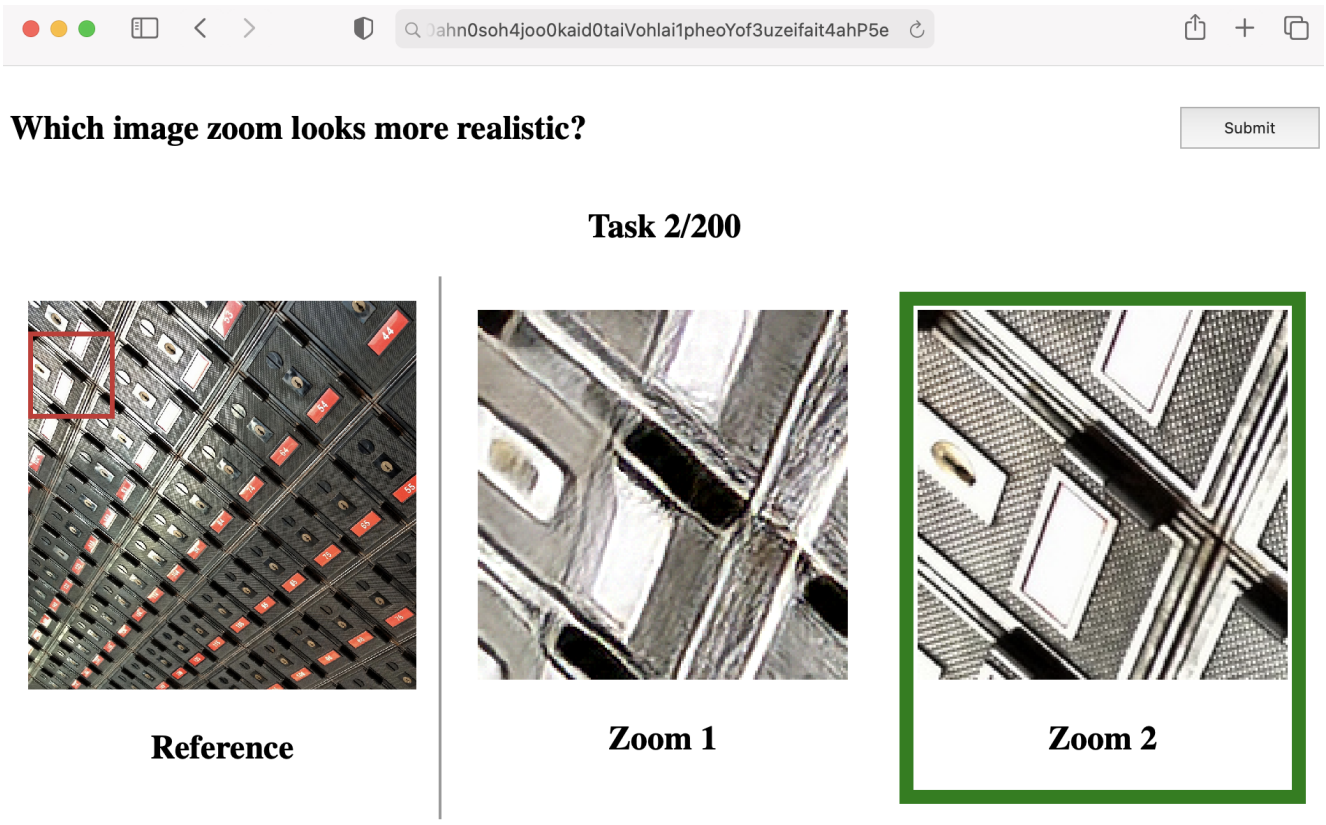

Figure 4. Screenshot of the web GUI of our user study. The 'Reference' provides an overview and indicates which part of an image should be considered. The images 'Zoom 1' and 'Zoom 2' are the two candidates, where the latter was selected to look more realistic.

To increase the data quality, we use a filtering mechanism. For that, we add redundant questions and reject submissions that have a low self-consistency. A visualization of results of the study results is shown in Fig. 5. The green bars display the percentage of votes favoring the photo-realism of AdFlow, while the red bars show the percentage favoring the other method. We display the statistical significance by showing the 95% confidence interval in black. Examples for images used in our study are shown in `visuals.html`.

2. Analysis of Sampling Temperature

Here, we analyze the trade-off between the image quality, in terms of LPIPS [11], and the consistency to the low-resolution input in terms of LR PSNR when varying the sampling temperature. We sample from the latent space with a Gaussian prior distribution $z \sim \mathcal{N}(0, \tau I)$ with variance τ. The latter is usually termed the sampling temperature [4]. Similar to [8] we can set the operation point by adjusting the temperature τ. Figures 1, 2 and 3 show that ESRGAN [6] trades off much more low-resolution consistency to improve the perceptual quality than AdFlow. The best trade-off is achieved at $\tau = 0.9$ for BaseFlow and $\tau = 1$ for AdFlow, as used in the main paper.

3. Minimal L_1 generalization

Here we provide further theoretical and empirical analysis when generalizing the L_1 loss with normalizing flows.

3.1. Relation of Flow loss to L_1

We here derive the generalized L_1 objective (Eq. (4) in the main paper) from the Normalizing Flow formulation as a 1-layer special case. Let $z \sim p_z(z) = \frac{1}{2^D} e^{-\|z\|_1}$ be standard Laplace. We use the function f defined as (Eq. (3) of the main paper),

$$z = f(y; x) = \frac{y - g(x)}{b(x)} \,. \tag{1}$$

we obtain the inverse as,

$$y = f^{-1}(z; x) = b(x) \cdot z + g(x) \tag{2}$$

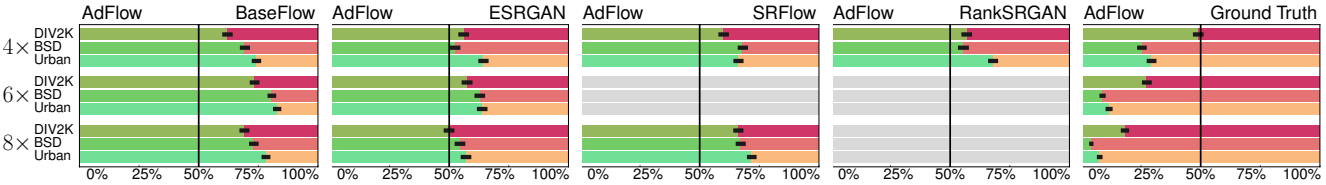

Figure 5. User study results, as the percentage of votes favoring the photo-realism of AdFlow (green) versus each other method (red). A bar represents 1500 user votes. The 95% confidence interval is in black. We compare on DIV2K, BSD100, Urban100 for $4\times$, $6\times$, and $8\times$.

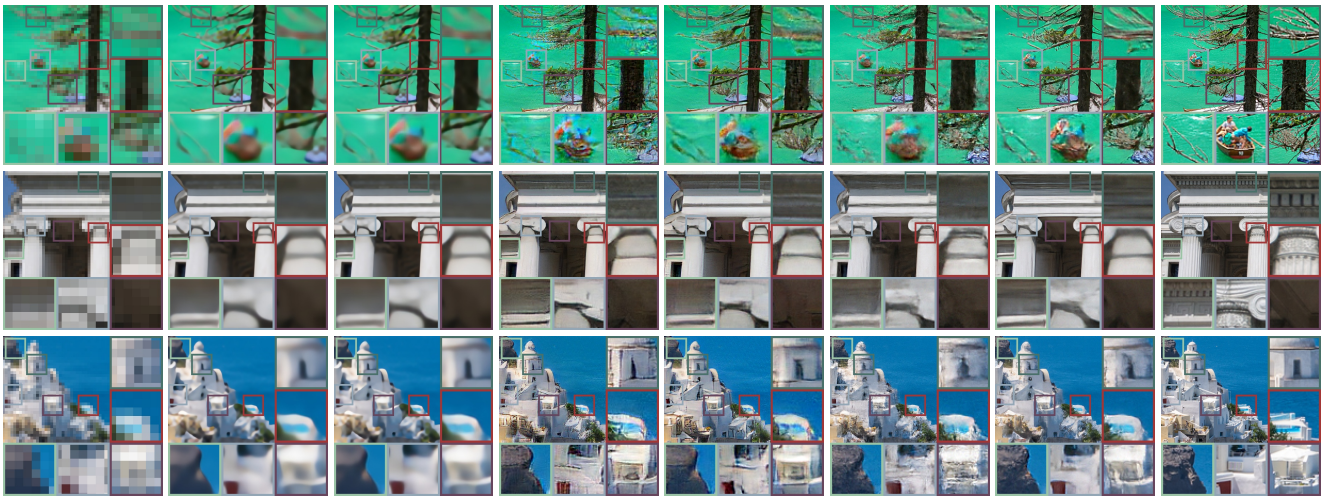

Figure 6. Qualitative comparison of standard approach and L_1 with learned variance on the DIV2K [1] validation set. ($8\times$)

Since $z \sim p_z(z)$ is a standard Laplace distribution, it is easy to see that $y \sim p(y|x;\theta) = \mathcal{L}(y; g(x), b(x))$ as given by Eq. (4) in the main paper, that is

$$-\log p(y|x;\theta) \propto \left\| \frac{y - g(x)}{b(x)} \right\|_1 + \sum_{ijc} \log b(x)_{ijc} \,. \quad (3)$$

Hence, (1) is the flow f of (3). Inserting (1) into the NLL formula for flows (Eq. (6a) in the main paper) gives,

$$\begin{aligned} -\log p(y|x,\theta) &= -\log p(z) - \log \left| \det \frac{\partial f_\theta}{\partial y}(y;x) \right| \\ &= -\log \frac{1}{2^D} e^{-\| \frac{y-g(x)}{b(x)} \|_1} - \log |\det \operatorname{diag}(\frac{1}{b(x)})| \\ &= D \log 2 + \left\| \frac{y - g(x)}{b(x)} \right\|_1 - \log |\prod_{ijc} \frac{1}{b(x)_{ijc}}| \\ &\propto \left\| \frac{y - g(x)}{b(x)} \right\|_1 + \sum_{ijc} \log b(x)_{ijc} \,. \end{aligned} \quad (4)$$

Here, the Jacobian $\frac{\partial f_\theta}{\partial y} = \operatorname{diag}(\frac{1}{b(x)})$ is a diagonal matrix with elements $\frac{1}{b(x)_{ijk}}$. The final result thus corresponds to the NLL derived directly in (3). We therefore conclude that the generalized L_1 objective is a special case given by the 1-layer normalizing flow defined in (1).

3.2. Empirical Analysis

We report results for the intermediate step of predicting an adaptive variance according to the Laplacian model described in Section 3.1, Equations (3)-(4) of the main paper. Those three channels predict the log-scale $a(x) = \log(b(x))$ of the Laplace distribution. The loss in Eq. (4) of the main paper can thus be written as,

$$-\log p(y|x;\theta) \propto \left\| \frac{y - g(x)}{\exp(a(x))} \right\|_1 + \sum_{ijc} a(x)_{ijc} \,. \quad (5)$$

We notice that even this extension of the L_1 objective reduces the conflict with the adversarial loss to some extent. The effective removal of artifacts for $8\times$ super-resolution is especially apparent in the first row of Figure 6 between ESRGAN [10] and ESRGAN + Adaptive Variance. Our further generalization of L_1 loss continues to improve the quality of the super-resolutions.

As the increase in visual quality alone would not be a good indicator for a reduced conflict of objectives, we also report the low-resolution consistency in Table 1 which improves by 2.91dB from ESRGAN [10] to ESRGAN + Adaptive Variance. An additional generalization to BaseFlow and AdFlow leads to a further improved low-resolution consistency. Based on observing an improved

	PSNR ↑	SSIM ↑	LPIPS ↓	LR-PSNR ↑
RRDB [10]	25.52	0.697	0.419	45.31
RRDB [10] + Adaptive Variance	25.47	0.696	0.418	44.51
ESRGAN [10]	22.14	0.578	0.277	31.28
ESRGAN [10] + Adaptive Variance	22.94	0.593	0.280	34.19
$\text{BaseFlow}_{\text{RRDB}}$	23.58	0.595	0.253	49.78
$\text{AdFlow}_{\text{RRDB}}$	23.45	0.602	0.253	47.54

Table 1. Quantitative results of standard approach and L_1 with learned variance on the DIV2K [1] validation set. "Adaptive variance" indicates the generalized L_1 loss with predicted variance, as described in Sec. 3. ($8\times$)

		PSNR ↑	SSIM ↑	LPIPS ↓	LR-PSNR ↑
DIV2K	Bicubic	26.69	0.766	0.409	38.69
	RRDB [10]	29.44	0.844	0.253	49.17
	ESRGAN [10]	26.20	0.747	0.124	39.01
	RankSRGAN [12]	26.55	0.750	0.128	42.33
	SRFlow [8]	27.08	0.756	0.120	49.97
	BaseFlow	27.21	0.760	0.118	49.88
	AdFlow	27.02	0.768	0.132	45.17
BSD100	Bicubic	22.40	0.508	0.713	37.13
	RRDB [10]	23.58	0.572	0.554	45.26
	ESRGAN [10]	20.99	0.462	0.332	31.68
	SRFlow [8]	21.76	0.467	0.335	51.01
	BaseFlow	22.03	0.478	0.325	50.17
	AdFlow	22.01	0.486	0.327	48.78
Urban100	Bicubic	19.31	0.477	0.686	33.93
	RRDB [10]	21.15	0.603	0.401	43.33
	ESRGAN [10]	18.43	0.475	0.306	28.88
	SRFlow [8]	19.29	0.501	0.309	48.11
	BaseFlow	19.72	0.513	0.304	48.71
	AdFlow	19.04	0.506	0.278	44.67

Table 2. Approximations for perceptual quality on the sets DIV2K (val.), BSD100, and Urban100 ($4\times$). Since [2, 3, 5, 7, 8, 9, 10] showed the limitations of calculated metrics for SR our main metric is the human study.

visual quality and low-resolution consistency, we conclude that the minimally generalized L_1 loss reduces the conflict in objectives, which further validates our strategy of replacing the L_1 with a more flexible generalization.

4. Detailed Results

In this section, we provide an extended quantitative and qualitative analysis of the same BaseFlow and AdFlow networks evaluated in the main paper. For completeness, we here provide the PSNR, SSIM and LPIPS on the DIV2K, BSD100, and Urban100 datasets. Results are reported in Tables 2, 3 and 4. However, note that these metrics do not well reflect photo-realism, as discussed in Sec. 4.1 in the main paper.

Further qualitative results for the scale levels $4\times$, $6\times$ and $8\times$ are provided in Figures 7, 8 and 9 respectively.

		PSNR ↑	SSIM ↑	LPIPS ↓	LR-PSNR ↑
DIV2K	Bicubic	24.87	0.680	0.519	37.78
	RRDB	26.51	0.741	0.382	46.86
	ESRGAN [10]	23.16	0.629	0.222	33.21
	BaseFlow	24.04	0.621	0.216	49.12
	AdFlow	23.94	0.6505	0.216	37.57
BSD100	Bicubic	23.26	0.564	0.645	37.51
	RRDB	24.42	0.625	0.507	46.41
	ESRGAN [10]	21.42	0.501	0.288	32.76
	BaseFlow	21.98	0.500	0.274	48.60
	AdFlow	22.17	0.533	0.269	38.52
Urban100	Bicubic	20.20	0.541	0.606	34.24
	RRDB	21.95	0.650	0.371	44.81
	ESRGAN [10]	19.43	0.541	0.251	30.87
	BaseFlow	20.43	0.564	0.255	47.90
	AdFlow	20.26	0.583	0.235	36.01

Table 3. Approximations for perceptual quality on the sets DIV2K (val.), BSD100, and Urban100 ($6\times$). Since [2, 3, 5, 7, 8, 9, 10] showed the limitations of calculated metrics for SR our main metric is the human study.

		PSNR ↑	SSIM ↑	LPIPS ↓	LR-PSNR ↑
DIV2K	Bicubic	23.74	0.627	0.584	37.14
	RRDB [10]	25.52	0.697	0.419	45.31
	ESRGAN [10]	22.14	0.578	0.277	31.28
	SRFlow [8]	23.04	0.578	0.275	49.02
	BaseFlow	23.58	0.595	0.253	49.78
	AdFlow	23.38	0.600	0.264	46.02
BSD100	Bicubic	22.40	0.508	0.713	37.13
	RRDB [10]	23.58	0.572	0.554	45.26
	ESRGAN [10]	20.99	0.462	0.332	31.68
	SRFlow [8]	21.76	0.467	0.335	51.01
	BaseFlow	22.03	0.478	0.325	50.17
	AdFlow	22.01	0.486	0.327	48.78
Urban100	Bicubic	19.31	0.477	0.686	33.93
	RRDB [10]	21.15	0.603	0.401	43.33
	ESRGAN [10]	18.43	0.475	0.306	28.88
	SRFlow [8]	19.29	0.501	0.309	48.11
	BaseFlow	19.72	0.513	0.304	48.71
	AdFlow	19.04	0.506	0.278	44.67

Table 4. Approximations for perceptual quality on the sets DIV2K (val.), BSD100, and Urban100 ($8\times$). Since [2, 3, 5, 7, 8, 9, 10] showed the limitations of calculated metrics for SR our main metric is the human study.

References

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

[2] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 3599–3608. IEEE, 2019.

[3] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[5] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero,

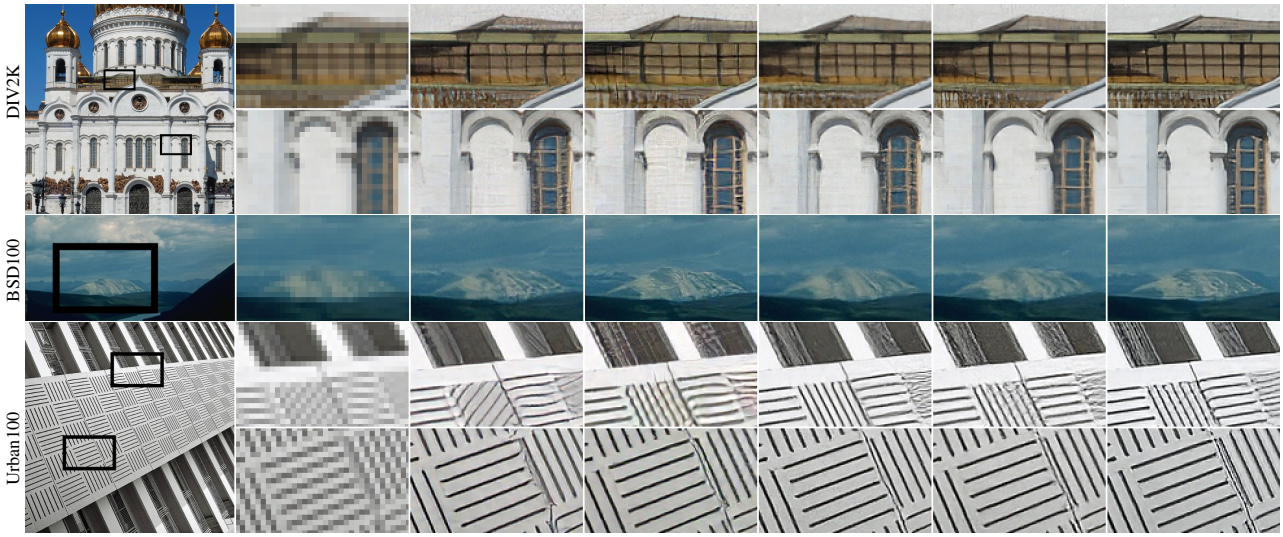

Figure 7. Qualitative comparison with state-of-the-art approaches on the DIV2K (val), BSD100 and Urban100 set for $4\times$ SR.

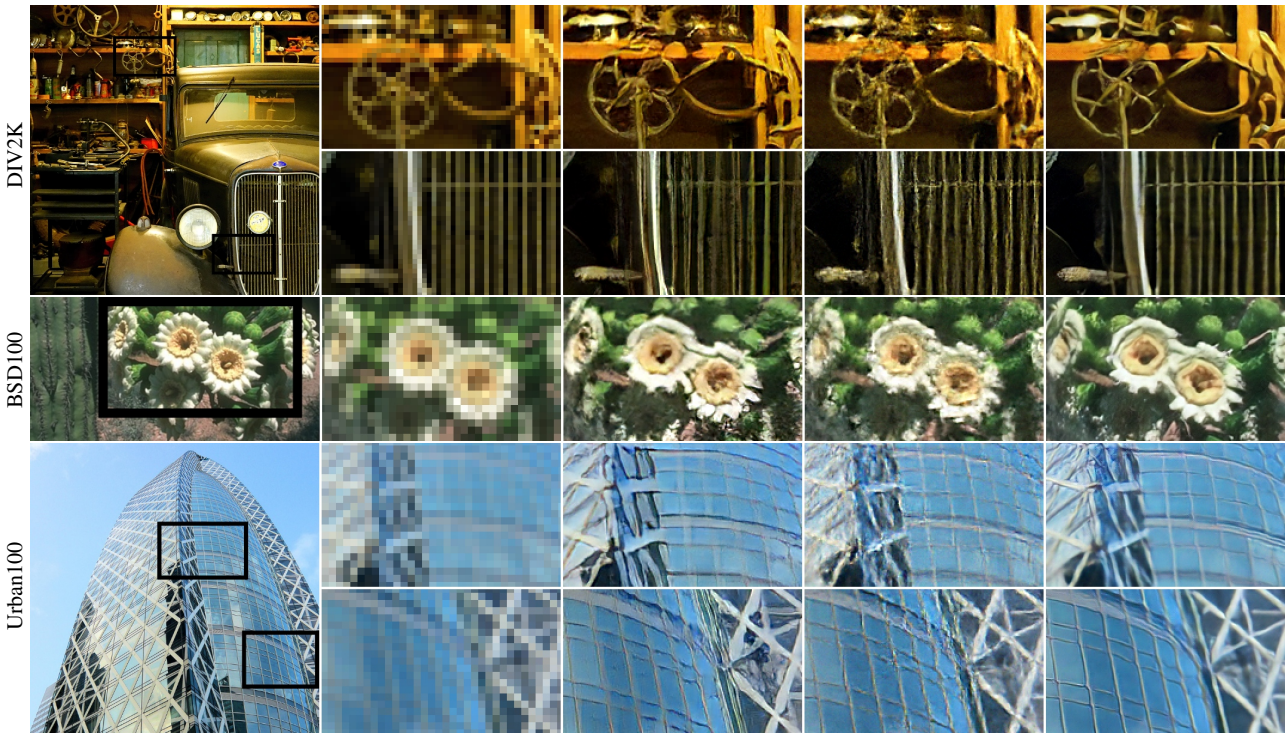

Figure 8. Qualitative comparison with state-of-the-art approaches on the DIV2K (val), BSD100 and Urban100 set for $6\times$ SR.

Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*, 2017.

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

[7] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. In *ICCV Workshops*, 2019.

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

[9] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4501–4510. IEEE Computer Society,

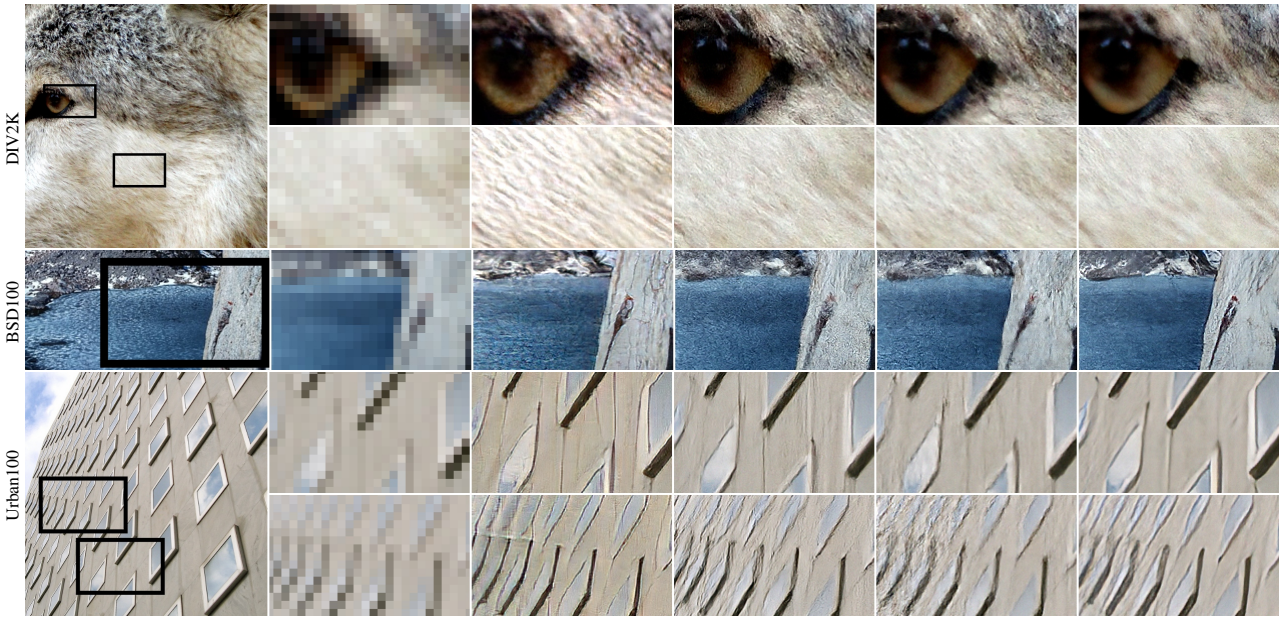

Figure 9. Qualitative comparison with state-of-the-art approaches on the DIV2K (val), BSD100 and Urban100 set for $8\times$ SR.

2017.

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

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

[12] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3096–3105, 2019.