Supplementary Material of SinSR: Diffusion-Based Image Super-Resolution in a Single Step

Yufei Wang[1,2†], Wenhan Yang[3], Xinyuan Chen[2*], Yaohui Wang[2], Lanqing Guo[1], Lap-Pui Chau[4], Ziwei Liu[1], Yu Qiao[2], Alex C. Kot[1], Bihan Wen[1*]
[1]Nanyang Technological University [2]Shanghai Artificial Intelligence Laboratory
[3]PengCheng Laboratory [4]The Hong Kong Polytechnic University

In this supplementary material, we include more mathematical details of the derived deterministic sampling strategy (Section A), more ablation studies on the design of consistency preserving loss (Section B), more experimental results in Section C (comparison with the consistency distillation [11] and more visual comparisons on real-world and synthetic datasets), and the discussion of the limitation in Section D. The code will be released.

A. Mathematical details

In this section, we elaborate on the derivation of the deterministic sampling strategy proposed in Sec. 4.1 in the main paper. In ResShift [15], the target reverse process is defined as follows

$$q(x_{t-1}|x_t, x_0,) = \mathcal{N}(x_{t-1}|\frac{\eta_{t-1}}{\eta_t}x_t + \frac{\alpha_t}{\eta_t}x_0, \kappa^2\frac{\eta_{t-1}}{\eta_t}\alpha_t\mathbf{I}) \tag{1}$$

which is approximated by a deep network as follows

$$p_\theta(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t, y, t), \kappa^2\frac{\eta_{t-1}}{\eta_t}\alpha_t\mathbf{I}). \tag{2}$$

To obtain a deterministic sampling, inspired by DDIM [10], we propose to utilize a new deterministic inference process $q'(x_{t-1}|x_t, x_0, y)$ to reformulate the inference process. Specifically, the deterministic reverse step is defined as

$$q'(x_{t-1}|x_t, x_0, y) := k_t x_0 + m_t x_t + j_t y, \tag{3}$$

where k, m, and j are all unknown variables to be determined. To utilize the pre-trained diffusion model with the parameter θ, we need to make sure that the marginal distribution of the proposed deterministic process is the same as the training one, *i.e.*, $q'(x_{t-1}|x_t, x_0, y) = q(x_{t-1}|x_0, y)$ since the teacher model is trained on the noise images that obey the distribution of $q(x_t|x_0, y)$. Specifically, the $q(x_t|x_0, y)$ that the teacher model trained on is defined as follows,

$$\begin{aligned} q(x_t|x_0, y) &= x_0 + \eta_t(y - x_0) + \sqrt{\kappa^2 + \eta_t}\epsilon \\ &= (1 - \eta_t)x_0 + \eta_t y + \sqrt{\kappa^2 + \eta_t}\epsilon, \end{aligned} \tag{4}$$

where the randomness comes from $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

To keep the marginal distribution unchanged, we solve the variables k, m and j using the method of undetermined coefficients. Specifically, the $q'(x_{t-1}|x_t, x_0, y)$ can be further derived as follows

$$\begin{aligned} q'(x_{t-1}|x_t, x_0, y) &= k_t x_0 + m_t x_t + j_t y \\ &= k_t x_0 + m_t(x_0 + \eta_t(y - x_0) + \sqrt{\kappa^2\eta_t}\epsilon) + j_t y \\ &= (k_t + m_t - m_t\eta_t)x_0 + (m_t\eta_t + j)y + m_t\sqrt{\kappa^2\eta_t}\epsilon, \end{aligned} \tag{5}$$

where we utilize the formulation of $q(x_t|x_0, y) = x_0 + \eta_t(y - x_0) + \sqrt{\kappa^2\eta_t}\epsilon$ defined in [15] to convert x_t to x_0, y and ϵ. To match the marginal distribution of $q'(x_t - 1|x_t, x_0, y)$ and $q(x_t|x_0, y)$, the coefficients of corresponding terms in Eq. 1 and Eq. 5 need to be the same by solving the following equations

$$\begin{cases} k_t + m_t - m_t \cdot \eta_t = 1 - \eta_{t-1} \\ m_t \cdot \eta_t + j_t = \eta_{t-1} \\ m_t^2\kappa^2\eta_t = \kappa^2\eta_{t-1}, \end{cases} \tag{6}$$

where η, κ are hyper-parameters in [15]. We find that the system of equations in Eq. 6 exists a solution, which demonstrates the existence of a deterministic mapping between x_T and x_0. The solution is as follows

$$\begin{cases} m_t = \sqrt{\frac{\eta_{t-1}}{\eta_t}} \\ j_t = \eta_{t-1} - \sqrt{\eta_{t-1}\eta_t} \\ k_t = 1 - \eta_{t-1} + \sqrt{\eta_{t-1}\eta_t} - \sqrt{\frac{\eta_{t-1}}{\eta_t}}. \end{cases} \tag{7}$$

B. More ablation studies

B.1. Detach of the estimated x_T

As shown in Eq. 8 in the paper, the predicted initial state $\hat{x}_T$ of the ground-truth image x_{gt} is detached before being used for generating $\hat{x}_{gt}$. The reason is that if we do not detach the

Methods	Datasets			
	RealSR		*RealSet65*	
	CLIPIQA↑	MUSIQ↑	CLIPIQA↑	MUSIQ↑
ESRGAN [13]	0.2362	29.048	0.3739	42.369
RealSR-JPEG [1]	0.3615	36.076	0.5282	50.539
BSRGAN [16]	0.5439	<u>63.586</u>	0.6163	**65.582**
SwinIR [3]	0.4654	59.636	0.5782	63.822
RealESRGAN [14]	0.4898	59.678	0.5995	63.220
DASR [4]	0.3629	45.825	0.4965	55.708
LDM-15 [7]	0.3836	49.317	0.4274	47.488
ResShift-15 [15]	0.5958	59.873	0.6537	61.330
SinSR-1 ($\lambda_0 = 1, \lambda_1 = \lambda_2 = 0$)	0.6119	57.118	0.6822	61.267
SinSR-1 ($\lambda_0 = 0, \lambda_1 = \lambda_2 = 1$)	**0.7542**	**63.990**	**0.7637**	<u>64.730</u>
SinSR-1	<u>0.6887</u>	61.582	<u>0.7150</u>	62.169

Table A. Quantitative results of models on two real-world datasets. The best and second best results are highlighted in **bold** and <u>underline</u>.

Methods	Metrics				
	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
ESRGAN [13]	20.67	0.448	0.485	0.451	43.615
RealSR-JPEG [1]	23.11	0.591	0.326	0.537	46.981
BSRGAN [16]	24.42	0.659	0.259	0.581	<u>54.697</u>
SwinIR [3]	23.99	0.667	0.238	0.564	53.790
RealESRGAN [14]	24.04	0.665	0.254	0.523	52.538
DASR [4]	24.75	**0.675**	0.250	0.536	48.337
LDM-30 [7]	24.49	0.651	0.248	0.572	50.895
LDM-15 [7]	<u>24.89</u>	0.670	0.269	0.512	46.419
ResShift-15 [15]	**24.90**	<u>0.673</u>	0.228	0.603	53.897
SinSR-1 ($\lambda_0 = 1, \lambda_1 = \lambda_2 = 0$)	24.69	0.664	<u>0.222</u>	0.607	53.316
SinSR-1 ($\lambda_0 = 0, \lambda_1 = \lambda_2 = 1$)	23.49	0.616	0.226	**0.671**	**55.017**
SinSR-1	24.56	0.657	**0.221**	0.611	53.357

Table B. Quantitative results of models on *ImageNet-Test*. The best and second best results are highlighted in **bold** and <u>underline</u>.

gradient of $\hat{x}_T$, the information of the ground-truth image x_{gt} will leak into the latent code $\hat{x}_T$. The information leak will cause a domain gap between the distribution of $\hat{x}_T$ and the real distribution of $x_T = y + \epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \kappa^2\eta_T\mathbf{I})$. A comparison of $\hat{x}_T$ from models trained w/ and w/o the detach operation can be found in Fig. A and Fig. B. As shown in the figure, the noise $\hat{\epsilon} = \hat{x}_T - y$ predicted from the model trained with the detach operation better alleviates the domain gap with the real noise distribution, therefore better mitigating the gap between training and testing. The quantitative results are in Table C. As shown in the table, using the proposed consistency preserving loss with the detach operation achieves the best performance, demonstrating that a smaller gap between the predicted noise and the real one contributes to the performance improvement. Besides, all the model trained with the proposed consistency preserving distillation achieves stable improvement compared with the baseline, further demonstrating its effectiveness.

B.2. The weights of loss terms

In the main paper, we assign equal weights to each term since we find that it can already achieve satisfactory results. To further evaluate the effect of different weighting strategies, we do ablation studies by assigning different weights for each loss as follows

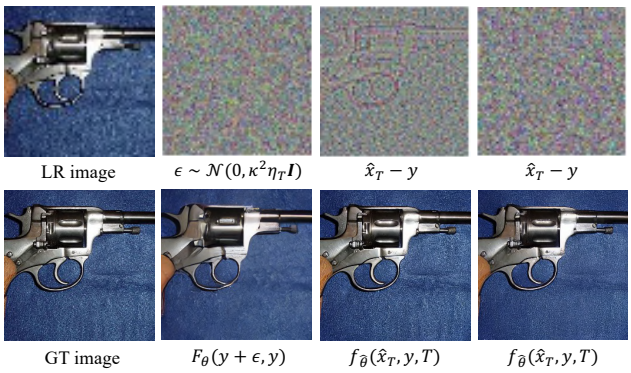
	CLIPIQA ↑	MUSIQ ↑
SinSR (distill only)	0.6536	61.330
SinSR (w/o detach)	<u>0.6994</u>	<u>61.342</u>
SinSR	**0.7150**	**62.169**

Table C. The ablation study of the detach operation on RealSet65.

$$\mathcal{L} = \lambda_0\mathcal{L}_{distill} + \lambda_1\mathcal{L}_{inverse} + \lambda_2\mathcal{L}_{gt}, \tag{8}$$

i.e., $\lambda_0 = \lambda_1 = \lambda_2 = 1$ by default. As for the $\mathcal{L}_{inverse}$, it is only responsible for providing a detached prediction $\hat{x}_T$, and its weight is empirically found to have less impact on the final results. Therefore, we keep $\lambda_1 = 1$ for all experiments. As for λ_2, we discuss two extreme cases: when $\lambda_2 = 0$, the model can be regarded as degrading to our baseline model *SinSR* (distill only) since the information from GT images is not utilized; when $\lambda_2 = 1, \lambda_0 = 0$, *i.e.*, by removing the $\mathcal{L}_{distill}$ in Eq. 8, we find that the model can achieve better results measured by metrics of perceptual quality, while suffers from degradation in fidelity performance. The results are shown in Table A and B.

As shown in the tables, setting $\lambda_2 = \lambda_1 = 1, \lambda_0 = 0$ can



(a) x and y (b) random ϵ (c) w/o detach (d) w/ detach

Figure A. A comparison between the models trained w/ and w/o the detach operation. (b) The HR image generated from a randomly generated noise fails to recover all the details. (c) The model trained w/o the detach operation will encode the residual information into the random noise, causing a serious domain gap between $\hat{x}_T$ and the real one. (d) By utilizing the detach operation, $\hat{\epsilon} = \hat{x}_T - y$ well obey a similar distribution of ϵ. Since the reconstruction of x is still not perfect even using a predicted initial state $\hat{x}_T$, the reconstruction loss $\mathcal{L}_{gt}$ in Eq. 8 of the main paper can be used as a good regularization term to further improve the performance.

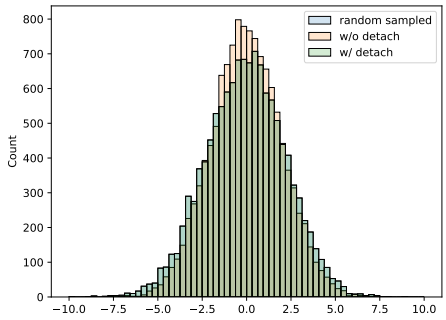
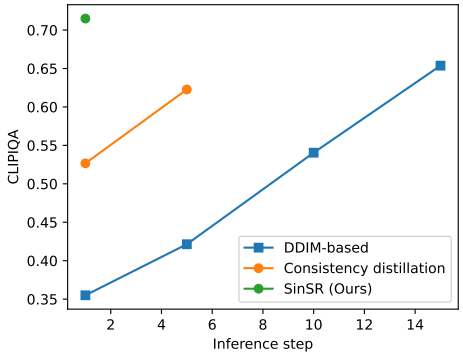
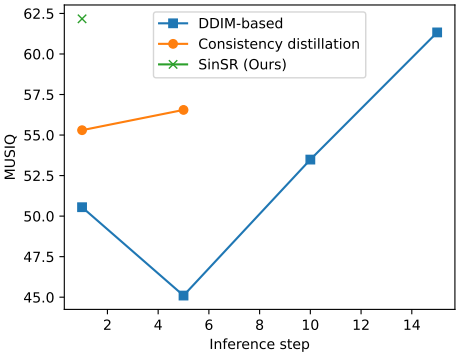


Figure B. The histograms of the random sampled ϵ, the predicted $\hat{\epsilon}$ from the model trained w/ and w/o the detach operation. An obvious domain gap can be observed in the predicted $\hat{\epsilon}$ using the model trained w/o the detach operation.

lead to significant improvements in terms of CLIPIQA [12] and MUSIQ [2], *e.g.*, it achieves the best performance in RealSR set among all the competitors, in only one inference step. However, using $\mathcal{L}_{inverse}, \mathcal{L}_{gt}$ only leads to the degradation of the fidelity performance as shown in Table B. This degradation, *i.e.*, the contradiction between perceptual quality and fidelity performance, can also be observed in LDM [7] and ResShift [15]. For example, in Table B, by increasing the inference steps, LDM-30 has better perceptual quality than LDM-15 while having worse PSNR and SSIM. This ablation study further demonstrates the effectiveness of the proposed consistency preserving loss.



(a) The performance measured by CLIPIQA↑



(b) The performance measured by MUSIQ↑

Figure C. A comparison of the proposed method *SinSR* with ResShift accelerated by DDIM-based sampling and consistency distillation (CD) [11] on RealSet65.

C. More experimental results

C.1. Comparison with consistency distillation

Similar to the progressive distillation [9], to avoid the generation of samples from solving the ODE of the diffusion models [6, 17], a new family of models named consistency models [11] are recently proposed which are proved to have good performance in generation tasks. To further demonstrate the superiority of the proposed method in the SR task, we further compare the proposed method *SinSR* with the ResShift model accelerated by [11]. The results can be found in Fig. C. As shown in the figure, the consistency distillation achieves significant improvement in terms of CLIPIQA and MUSIQ under the same number of inference steps compared with DDIM-based acceleration. However, the proposed method still achieves the best result among all the competitors. A visual comparison between SOTA models for acceleration in one step can be found in Fig. D, where the proposed method achieves the best perceptual qualities and rich details.

C.2. Comparison with SOTA methods

We provide more visual comparison with the SOTA methods on real-world and synthetic datasets. Some examples are shown in Fig. E, Fig. F, Fig. G, Fig. H and Fig. I.

D. Limitations

While the proposed method achieves promising results in only one inference step. Some limitations still exist. The main limitation is the contradiction between the fidelity performance and the perceptual quality, *i.e.*, there exists a slight performance drop in terms of PSNR and SSIM compared with the teacher model. However, this phenomenon can be widely observed in other diffusion-based methods [7, 8, 15]. Besides, while decreasing the number of inference steps can greatly improve PSNR and SSIM, it will cause serious degradation in terms of perceptual quality. How to obtain a better trade-off may still be an open problem and left to be explored in the future.

References

[1] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPR*, pages 466–467, 2020. 2

[2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 3

[3] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 2

[4] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, pages 574–591. Springer, 2022. 2

[5] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022. 5

[6] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021. 3

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4

[8] 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*, 45(4):4713–4726, 2022. 4

[9] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021. 3

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 1

[11] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 1, 3, 5

[12] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 3

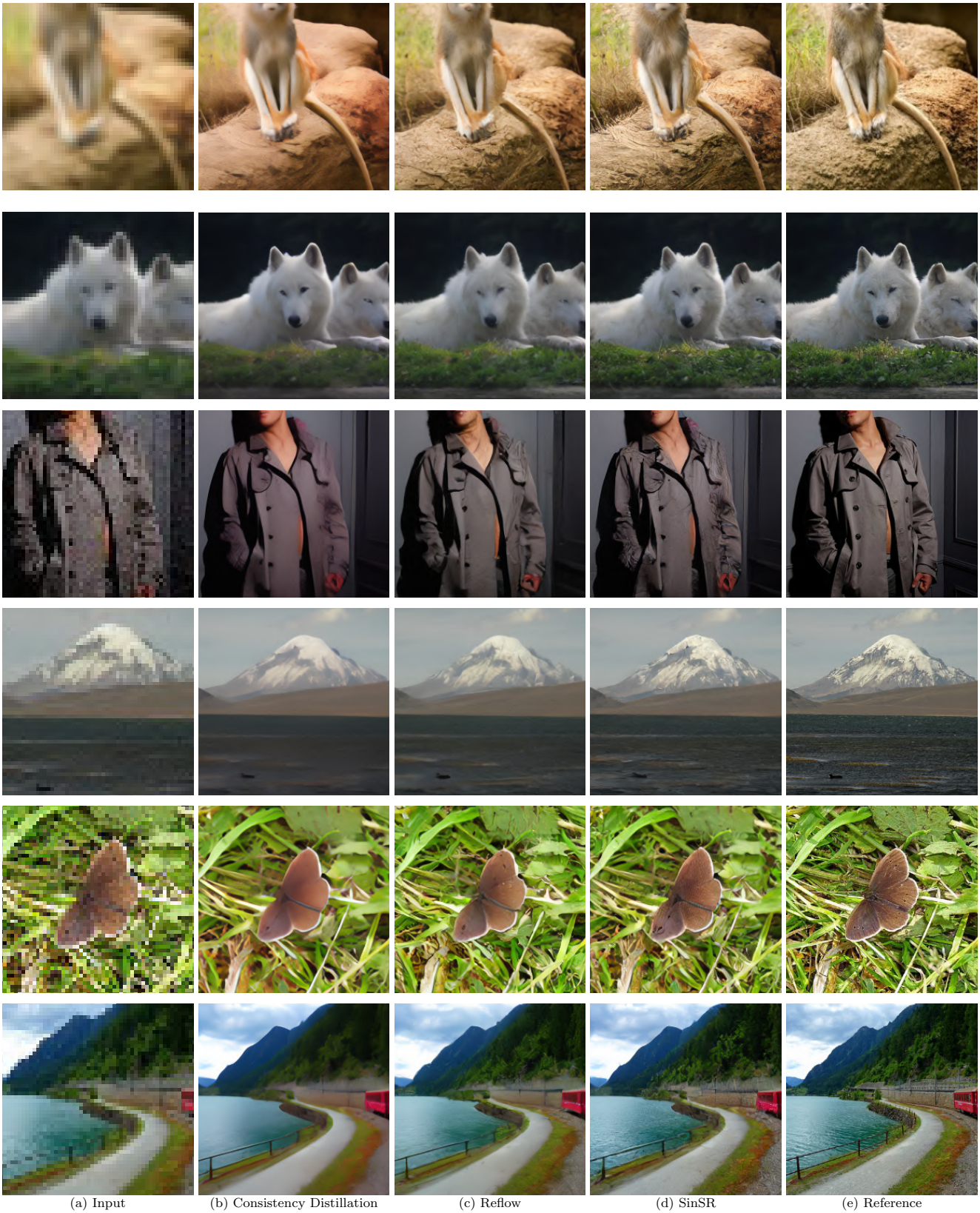
[13] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2

[14] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 2

[15] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 4

[16] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[17] Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *International Conference on Machine Learning*, pages 42390–42402. PMLR, 2023. 3



(a) Input (b) Consistency Distillation (c) Reflow (d) SinSR (e) Reference

Figure D. Comparison with SOTA acceleration methods in a single step, including Reflow [5] and consistency distillation [11].

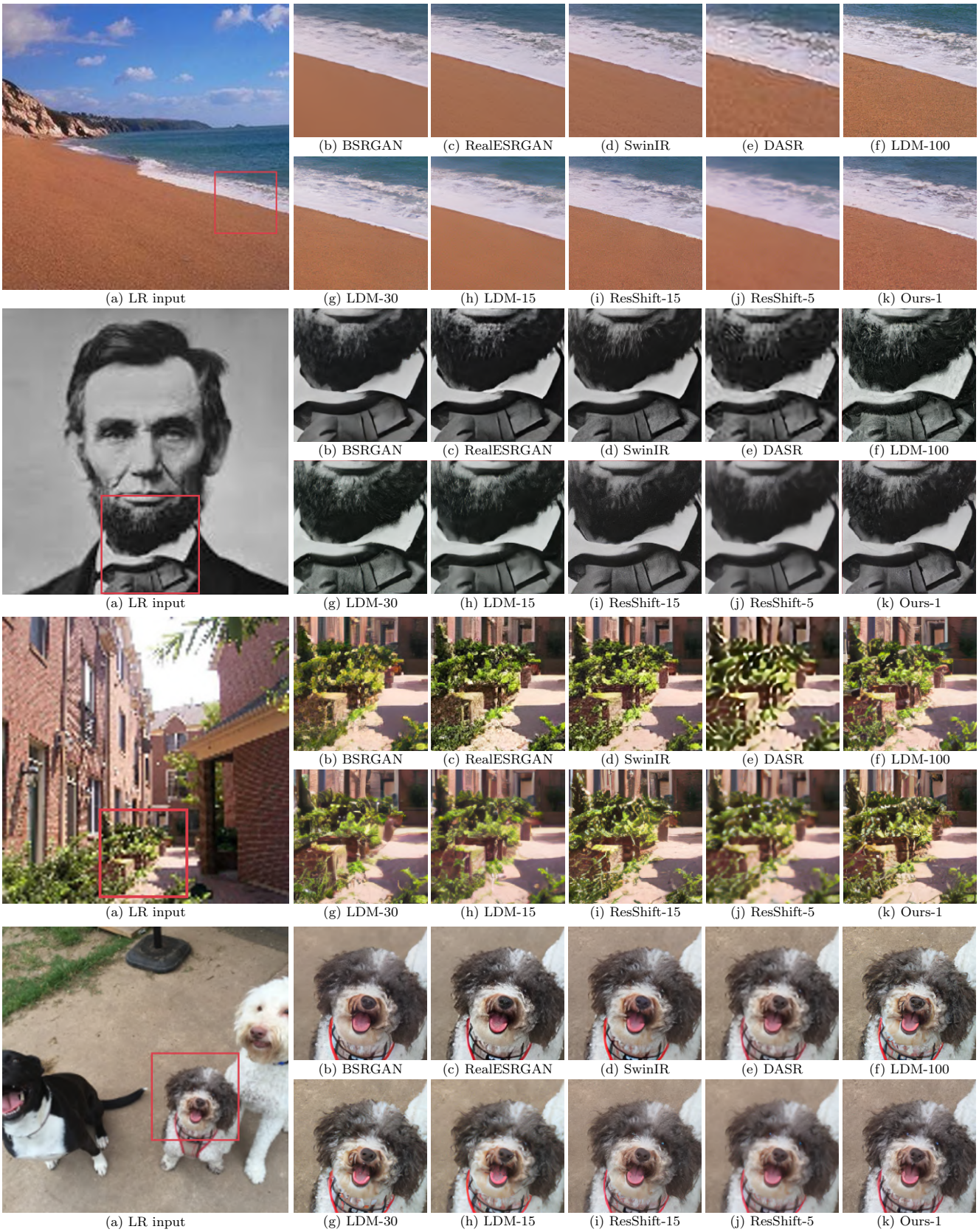


Figure E. Visual comparison on real-world samples. Please zoom in for more details.

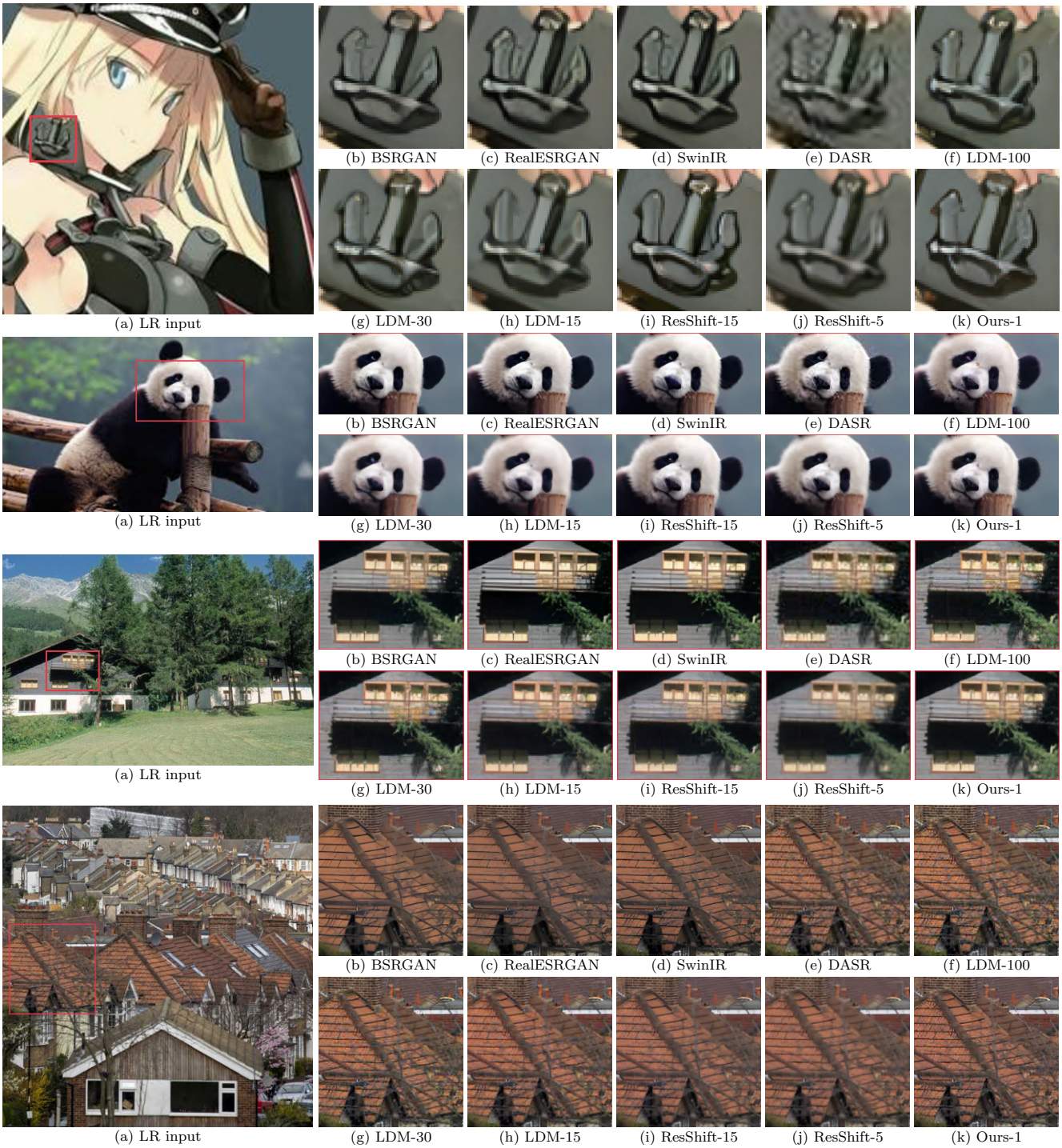


Figure F. Visual comparison on real-world samples. Please zoom in for more details.

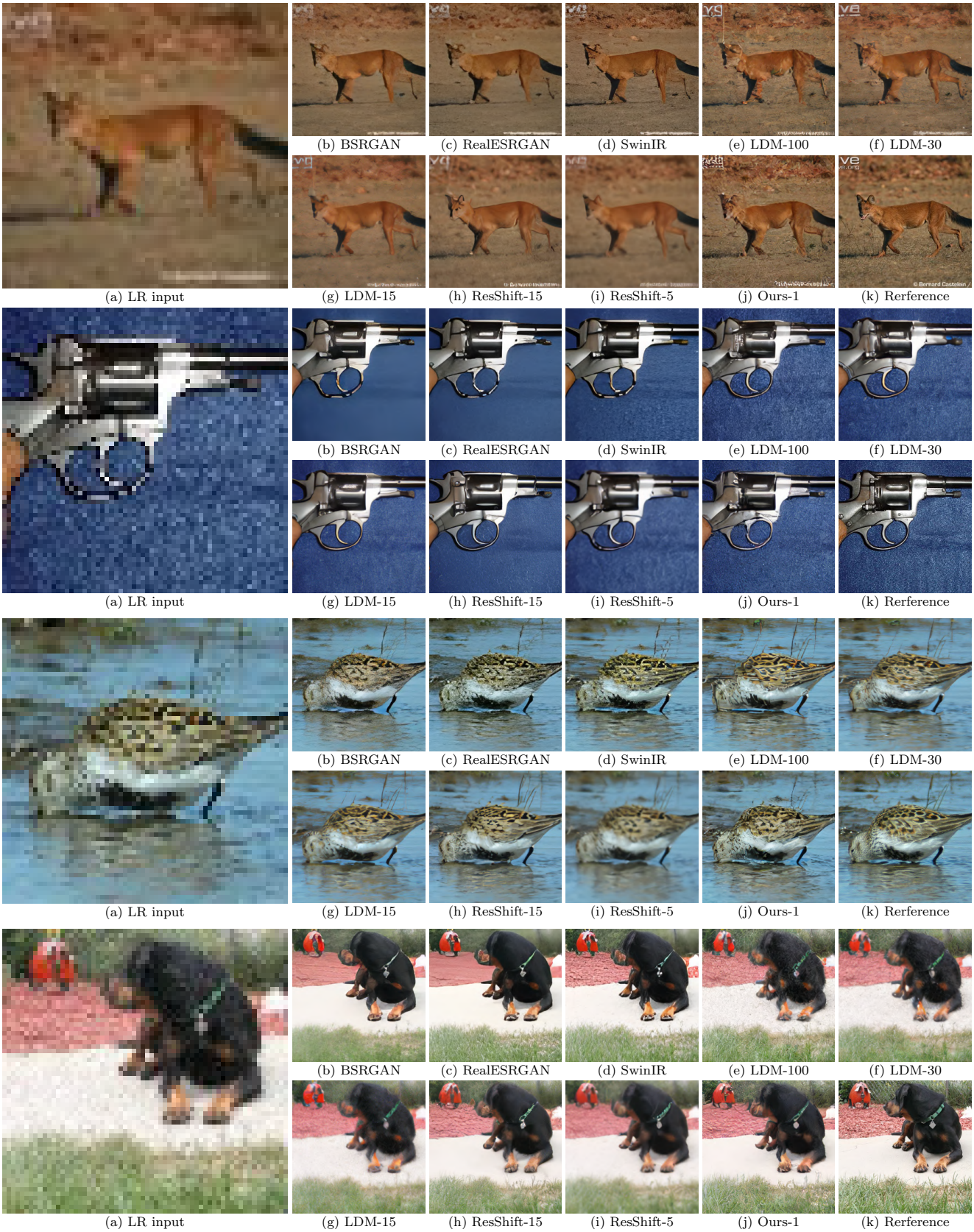


Figure G. Visual comparison on the synthetic test set ImageNet-Test. Please zoom in for more details.

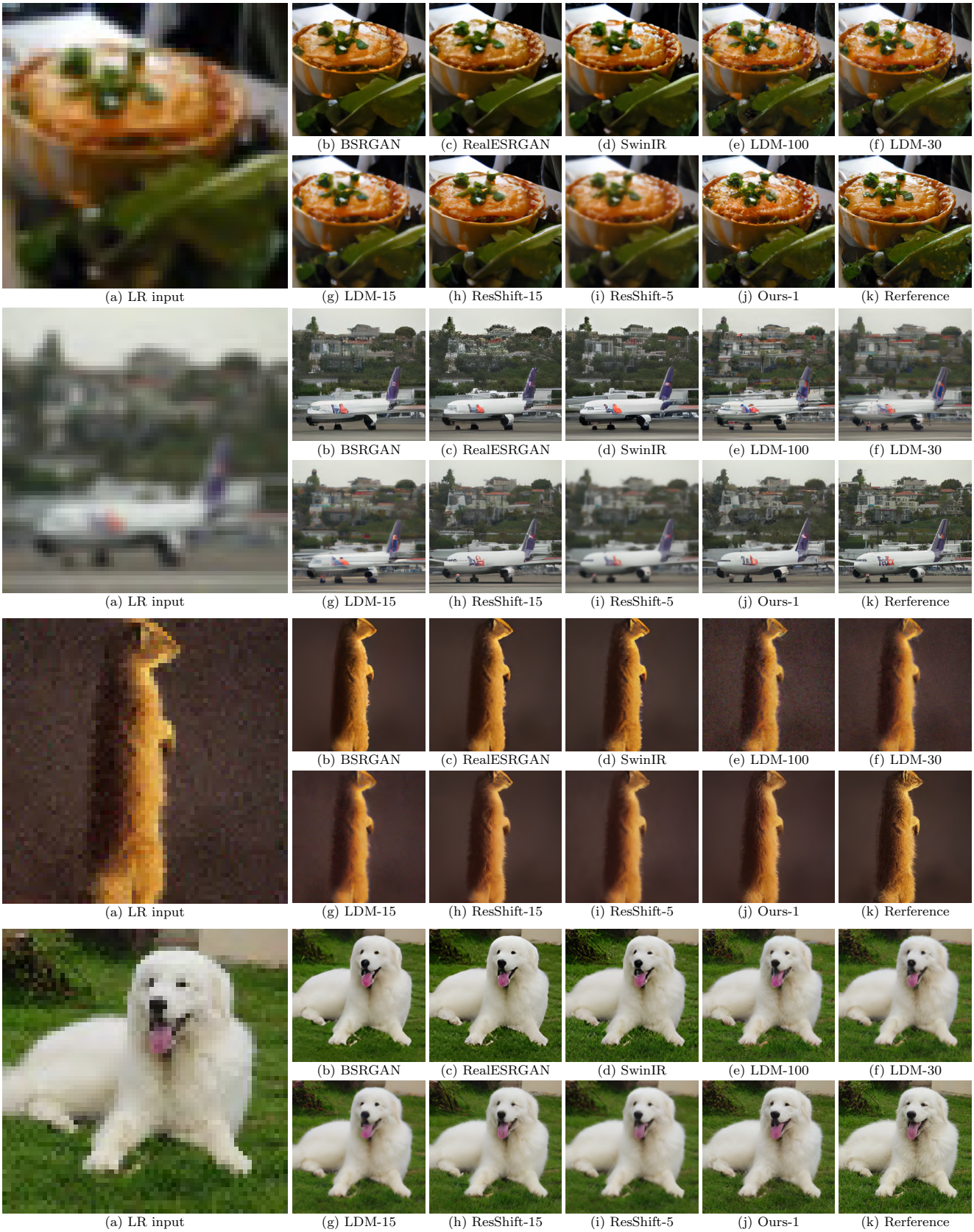


Figure H. Visual comparison on the synthetic test set ImageNet-Test. Please zoom in for more details.

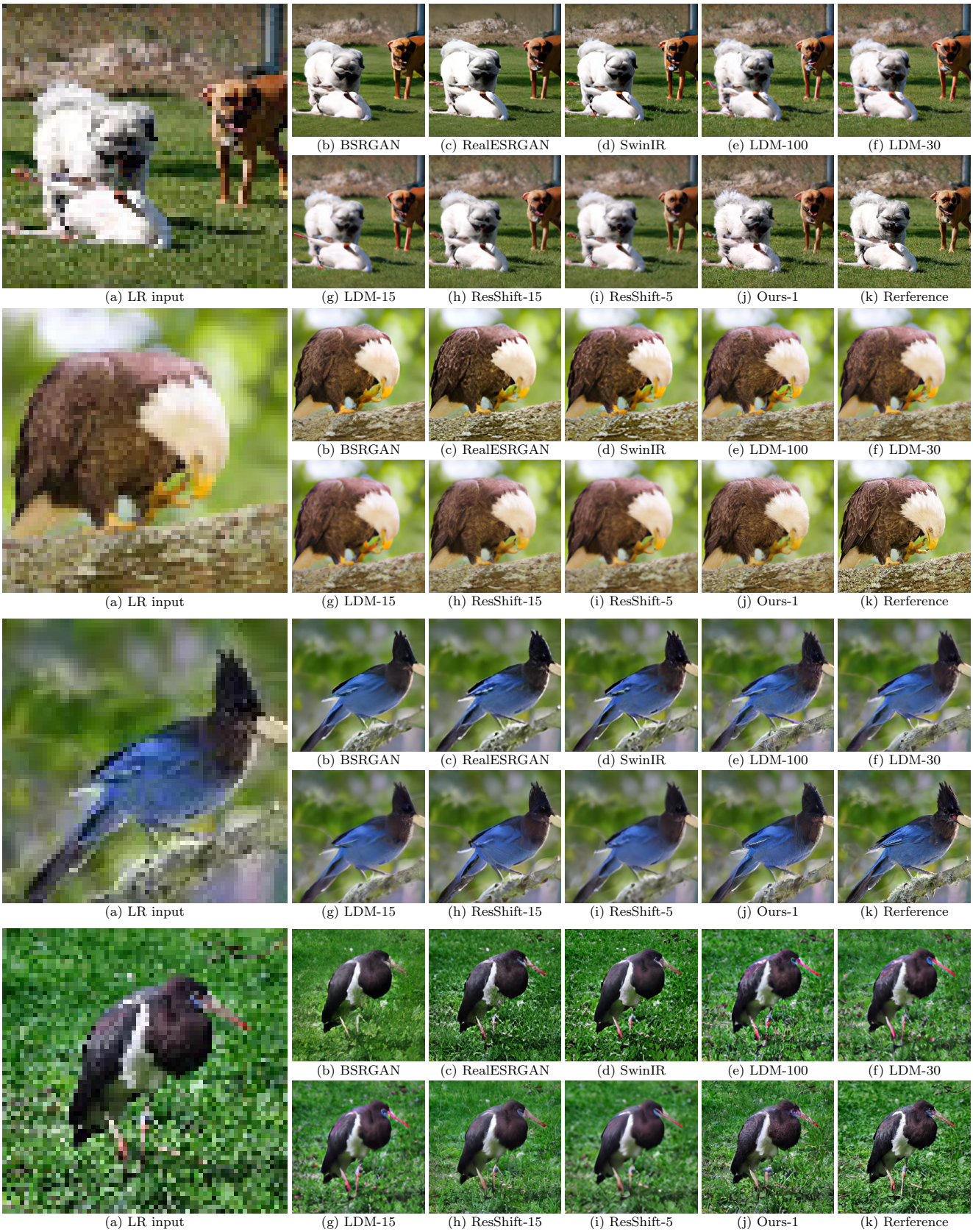


Figure I. Visual comparison on the synthetic test set ImageNet-Test. Please zoom in for more details.