RealViformer: Investigating Attention for Real-World Video Super-Resolution Supplementary Material

Yuehan Zhang and Angela Yao

National University of Singapore
{zyuehan,ayao}@comp.nus.edu.sg

The supplementary features the following sections:

- Section Sec. A: detailed architectures of attention modules used for attention investigation.
- Section Sec. B: more details for the proposed real-world VSR model, Realviformer:
  - Detailed model architecture.
  - Training settings.
  - Reasons for choosing quantitative metrics.
  - Ablations of channel squeeze-excite and rescaling mechanism in ICA module.
  - Visual comparisons.
  - Temporal Consistency.
  - Challenging cases.

A Model Architectures of Attention Modules

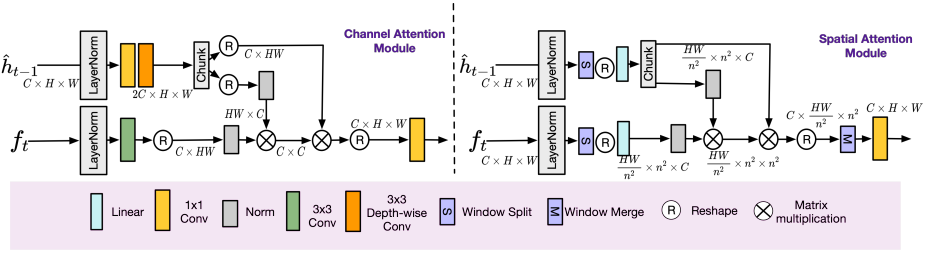


Fig. 12: The architectures of channel (left) and spatial (right) attention modules in Fig.3 (b) of the main paper. The spatial attention is window-based with window-split module S and window-merge module M, which split and merge overlapped windows of size $\mathbb{R}^{\omega^2 \times \omega^2}$.

Fig. 12 shows detailing constructions of channel and spatial attention modules in Fig.4(b) of the main paper. The channel attention module maps layer-normalized f_t to query and $\hat{h}_{t-1}$ to key and value. The attention map is of size

$\mathbb{R}^{C\times C}$, where C is the number of feature channels. The spatial counterpart splits layer-normalized features to windows of size $\mathbb{R}^{\omega^2\times\omega^2}$ spatially, and we use $\omega = 8$ in practice. The correlations are calculated within windows.

B RealViformer

B.1 Model Architecture

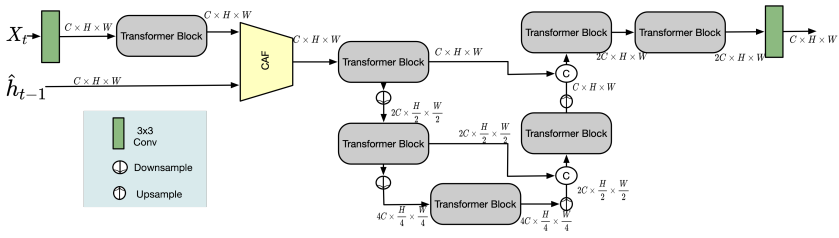


Fig. 13: U-shape architecture for the reconstruction module in RealViformer. The CAF module refers to the Channel Attention Fusion module explained in Sec. 4.2 of the main paper.

Fig. 13 shows the details of the U-shape architecture of the reconstruction module in RealViformer (Fig.9 in the main paper). The shallow feature and aligned hidden state are first merged with the CAF module introduced in the main paper; then, the merged feature is put in a three-level U-shape architecture. The Downsample and Upsample are implemented by PixelUnshuffle and PixelShuffle [13], followed by convolutions.

B.2 Training Settings

We give more details of the two-stage training pipeline. Following RealBasicVSR [2], we perform two-stage training. The first stage trains the model with a Charbonnier loss [7] and SSIM [16] loss for 300K iterations. The weights of charbonnier loss and SSIM loss are 1.0 and 0.01. We use AdamW optimizer [9] with $\beta_1 = 0.9, \beta_2 = 0.999$ and weight decay $1e^{-4}$. The learning rate for the first 80K iterations is $3e^{-4}$, which is then gradually decreased to $1e^{-6}$ by Cosine Annealing scheme [8].

In the second stage, the model is trained for another 130K iterations with the Charbonnier loss, SSIM loss, perceptual loss [5] and GAN loss [3] together, weighted by 1, 0.001, 1, and 0.005, respectively. The implementations of perceptual loss, GAN loss, and discriminator follow RealBasicVSR [2]. We renewed the batch size to 8 and the optimizer to Adam [6]. The learning rates of the generator and discriminator are $5e^{-5}$ and $1e^{-4}$, which remain unchanged.

B.3 Choice of Quantitative Metrics

No-reference evaluation is an open research question, and we are still on the way to finding good no-reference metrics. Although commonly used in previous papers, NIQE [12] and PI [14] have a severe bias to over-sharpened images, as shown in Tab. 4. Thus, we choose ILNIQE [20], an improved version of NIQE, and NRQM [10] for evaluation. The other metric, BRISQUE [11], used in previous SOTAs, was proposed in 2011 without considering any artifacts from deep models. It has been shown to correlate poorly with human opinions when evaluated images are not of very low quality [1]. Thus, we substitute this unreliable quantitative metric with a user study in the main paper.

Table 4: Scores for ground-truth and over-sharpened ground-truth REDS4. The better score is colored in Red.

	NIQE ↓	PI↓	ILNIQE↓	NRQM↑
Ground-truth REDS4	2.22	2.62	17.57	6.95
Sharpened ground-truth REDS4	1.89	2.51	27.45	6.90

B.4 Ablations

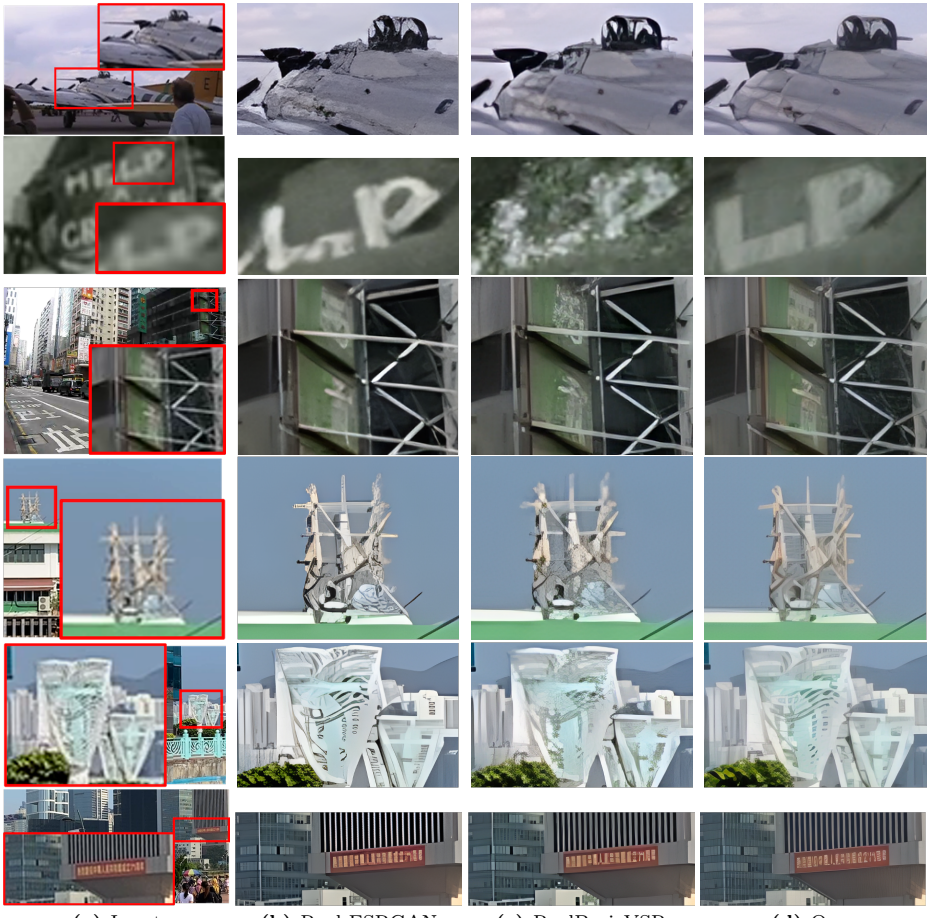
This section separately checks the impact of channel squeeze-excite and rescaling mechanisms in ICA. As shown in Tab. 5, the achievement of ICA is a compound effect of channel squeeze-excite and rescaling mechanisms. Using them separately decreases the performance.

Table 5: Ablation study of channel squeeze-excite and rescaling mechanisms. The channel squeeze-excite mechanism is denoted as SE in the table. Adding channel squeeze-excite or rescaling mechanism separately to the original channel attention module [19] will decrease the NRQM score. Their compound effect helps achieve state-of-the-art performance.

	SE	Rescaling	NRQM
Realviformer$^-$			6.196
+ squeeze-excite	✓		6.099
+ rescaling		✓	6.05
Realviformer	✓	✓	6.338

B.5 Visual Comparisons

This section shows more visual comparisons with RealSR [4], Real-ESRGAN [15], and RealBasicVSR [2] in Fig. 14. Our method predicts clear patterns without obvious high-frequency artifacts. We also provide video examples of our method in the supplementary attachments.



(a) Input (b) Real-ESRGAN (c) RealBasicVSR (d) Ours

Fig. 14: Visual comparisons between our method, Real-ESRGAN [15], and RealBasicVSR [2]. The first three rows are test cases from VideoLQ [2], and the last three are from RealVSR [18]. Our method produces clear structural patterns without high-frequency artifacts; others are blurry or sacrifice structural information for high sharpness.

B.6 Temporal Consistency

We compare temporal profiles [17] on synthetic data in Fig. 15, where the horizontal dimension is time. Ours shows smoother profiles that are close to ground truths.

B.7 Challenging Cases

Although RealViformer shows state-of-the-art performance on two real-world video datasets overall, its prediction at the boundary in low-contrast regions has

(a) RealBasicVSR **(b)** ours **(c)** GT

Fig. 15: Temporal profiles of RealBasicVSR, our method, and ground-truth (GT).

lower sharpness, as shown in Fig. 16. Achieving sharpness without introducing high-frequency artifacts is a challenging problem for super-resolution. We aim at future work to improve the sharpness of RealViformer.

(a) Input **(b)** Real-ESRGAN **(c)** RealBasicVSR **(d)** Ours

Fig. 16: Challenging cases of our method. The boundaries in low-contrast regions are not as sharp as RealBasicVSR [2] and Real-ESRGAN [15].

References

1. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 3
2. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5962–5971 (2022) 2, 3, 4, 5
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020) 2
4. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 466–467 (2020) 3
5. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016) 2

6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 2
7. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2599–2613 (2018) 2
8. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 2
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 2
10. Ma, C., Yang, C.Y., Yang, X., Yang, M.H.: Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding **158**, 1–16 (2017) 3
11. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012) 3
12. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012) 3
13. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016) 2
14. Vasu, S., Thekke Madam, N., Rajagopalan, A.: Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018) 3
15. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1905–1914 (2021) 3, 4, 5
16. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 2
17. Xiao, Z., Xiong, Z., Fu, X., Liu, D., Zha, Z.J.: Space-time video super-resolution using temporal profiles. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 664–672 (2020) 4
18. Yang, X., Xiang, W., Zeng, H., Zhang, L.: Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4781–4790 (2021) 4
19. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022) 3
20. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE Transactions on Image Processing **24**(8), 2579–2591 (2015) 3