Fast Spatio-Temporal Residual Network for Video Super-Resolution Supplementary Material

Sheng Li[1], Fengxiang He[2], Bo Du*[1], Lefei Zhang*[1], Yonghao Xu[3], and Dacheng Tao[2]

[1]School of Computer Science, Wuhan University, China
[2]UBTECH Sydney AI Centre, SCS, FEIT, the University of Sydney, Australia
[3]The State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China

{shli, remoteking, zhanglefei}@whu.edu.cn {fengxiang.he, dacheng.tao}@sydney.edu.au
yonghaoxu@ieee.org

This supplementary material contains two appendixes. Appendix A collects all the proofs omitted from the main text and Appendix B provides extra empirical results.

A. Proof

This appendix collects all the proofs omitted from the main text.

A.1. Preliminary

This subsection gives the background knowledges necessary to the development of the theoretical analysis.

A tuned FSTRN induces a hypothesis function that maps from low-resolution videos to high-resolution videos. For the brevity, we denote the hypothesis function as

$$F_\theta : \mathbb{R}^{n_{LR}} \to \mathbb{R}^{n_{HR}}, \tag{A.1}$$

$$I_{LR} \mapsto I_{HR}, \tag{A.2}$$

where θ is the tuned parameter, and n_{LR} and n_{LR} are respectively the dimensions of the low-resolution space and the high-resolution space. Suppose all the hypothesis functions F_θ computed by FSTRN constitute a hypothesis space $\mathcal{H}$. To measure the performance of the hypothesis function, we define an object function in the main text as eq. 3.9. The corresponding loss function is defined as follows:

$$\begin{aligned} l\left(I_{SR}, I_{HR}; \theta\right) =& \rho\left(I_{HR} - I_{SR}\right) \\ =& \sqrt{\left(I_{HR} - I_{SR}\right)^2 + \varepsilon^2}, \end{aligned} \tag{A.3}$$

where I_{HR} and I_{LR} are respectively the output (high-resolution image/video) and input (low-resolution image/video), and $\rho\left(x\right) = \sqrt{x^2 + \varepsilon^2}$ is Charbonnier penalty function. Based on the loss function $l(I_{SR}, I_{HR}, F_\theta)$, the expected risk, in term of the hypothesis function F_θ, is defined as follows:

$$\mathcal{R}(F_\theta) = \mathbb{E}_{I_{SR}, I_{HR}} l(I_{SR}, I_{HR}, F_\theta). \tag{A.4}$$

Similarly, the empirical risk is defined as

$$\hat{\mathcal{R}}(F_\theta) = \mathcal{L}(F_\theta) = \frac{1}{N} \sum_{n=1}^{N} l(I_{SR}^n, I_{HR}^n, F_\theta), \tag{A.5}$$

where I_{SR}^n and I_{HR}^n denote the n-th instance in the training set, and N is the sample size, and we redefine the empirical risk as $\hat{\mathcal{R}}$ in accordance with the convention. Finally, the generalization error of hypothesis function $F(\theta)$ is defined as the difference between the expected risk $\mathcal{R}(F_\theta)$ and the corresponding empirical risk $\hat{\mathcal{R}}(F_\theta)$.

As the principle of *Occam's razor* says, the generalization capability of an algorithm is dependent with the complexity of its corresponding hypothesis space (hypothesis complexity): a complex algorithms tend to have a poor generalization ability. In learning theory, three classic measurements of hypothesis complexity are respectively VC-dimension, Rademacher complexity, and covering number (see, respectively, [2], [11], and [5]). An classic result in learning theory expresses the negative correlation between the generalization error of an algorithm and the corresponding Rademacher complexity $\hat{\mathfrak{R}}(\mathcal{H})$ as the following lemma.

Lemma 1 (cf. [10], Theorem 3.1)**.** *For any $\delta > 0$, with probability at least $1 - \delta$, the following inequality hold for all $F_\theta \in \mathcal{H}$:*

$$\mathcal{R}(F_\theta) \leq \hat{\mathcal{R}}(F_\theta) + 2\hat{\mathfrak{R}}(l \circ \mathcal{H}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \tag{A.6}$$

*Corresponding author.

where $l \circ \mathcal{H}$ is defined as

$$l \circ \mathcal{H} \triangleq \{l \circ F : \ F \in \mathcal{H}\}. \tag{A.7}$$

Computing the empirical Rademacher complexity of neural network could be extremely difficult and thus still remains an open problem. Fortunately, the empirical Rademacher complexity can be upper bounded by the corresponding ε-covering number $N(\mathcal{H}, \varepsilon, \|\cdot\|_2)$ as the following lemma states.

Lemma 2 (cf. [1], Lemma A.5)**.** *Suppose* $\mathbf{0} \in \mathcal{H}$ *and all conditions in Lemma 1 hold. Then*

$$\hat{\mathfrak{R}}(\mathcal{H}) \leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(l \circ \mathcal{H}, \varepsilon, \| \cdot \|_2) d\varepsilon} \right). \tag{A.8}$$

Some recent works study the hypothesis complexity of deep neural networks and provide upper bounds of the corresponding hypothesis spaces. [1] gives a spectrally-normalised covering bound and a generalization bound for all chain-like neural networks. [6] focuses on the deep neural networks with shortcut connections and gives a covering bound and a corresponding generalization bound. Specifically, for a deep neural network with residual connections, suppose the "stem" is obtained by discarding all residual connections. Apparently, it is a chain-like neural network and can be expressed by the following formula:

$$S = (A_1, \sigma_1, A_2, \sigma_2, \ldots, A_L, \sigma_L), \tag{A.9}$$

where $A_i, i = 1, \ldots, L$ are weight matrices and σ_i are nonlinearities. Meanwhile, we denote all residual connections as V_j, $j \in J$, where J is the index set. Suppose the output of the i-th layer (constituted by the weight matrix A_i and the nonlinearity σ_i) is F_i, and all possible outputs F_i constitute a hypothesis space $\mathcal{H}_i^S$. Similarly, all outputs of the residual connection V_j constitute a hypothesis space $\mathcal{H}_j^V$. In this paper, our theoretical analysis is developed based on the two works stated above. Specifically, the covering bounds given by [6, 1] are respectively as follows.

Lemma 3 (see [6], Theorem 1)**.** *Suppose the ε_i^S-covering number of $\mathcal{H}_i^S$ is $\mathcal{N}_i^S$ and the ε_j^V-covering number of $\mathcal{H}_j^V$ is $\mathcal{N}_i^V$. Then there exists an ε in terms of all ε_i^S and ε_j^V, such that the following inequality holds:*

$$\mathcal{N}(\mathcal{H}) \leq \prod_{i=1}^{L} \mathcal{N}_i^S \prod_{j \in J} \mathcal{N}_V^j. \tag{A.10}$$

Lemma 4 (cf. [1], Lemma A.7)**.** *Suppose there are L weight matrices in a chain-like neural network. Let $(\varepsilon_1, \ldots, \varepsilon_L)$ be given. Suppose the L weight matrices $(A_1, \ldots, A_L)$ lies in $\mathcal{B}_1 \times \ldots \times \mathcal{B}_L$, where $\mathcal{B}_i$ is a ball centered at 0 with the radius of s_i, i.e., $\mathcal{B}_i = \{A_i : \|A_i\| \leq s_i\}$. Furthermore, suppose the input data matrix X is restricted in a ball centred at 0 with the radius of B, i.e., $\|X\| \leq B$. Suppose F is a hypothesis function computed by the neural network. If we define:*

$$\mathcal{H} = \{F(X) : A_i \in \mathcal{B}_i, A_t^{u,v,s} \in \mathcal{B}_t^{u,v,s}\}, \tag{A.11}$$

where $i = 1, \ldots, L$, $(u, v, s) \in I_V$, and $t \in \{1, \ldots, L^{u,v,s}\}$. Let $\varepsilon = \sum_{j=1}^{L} \varepsilon_j \rho_j \prod_{l=j+1}^{L} \rho_l s_l$. Then we have the following inequality:

$$\mathcal{N}(\mathcal{H}) \leq \prod_{i=1}^{L} \sup_{\mathbf{A}_{i-1} \in \boldsymbol{\mathcal{B}}_{i-1}} \mathcal{N}_i, \tag{A.12}$$

where $\mathbf{A}_{i-1} = (A_1, \ldots, A_{i-1})$, $\boldsymbol{\mathcal{B}}_{i-1} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_{i-1}$, and

$$\mathcal{N}_i = \mathcal{N}\left(\left\{A_i F_{\mathbf{A}_{i-1}}(X) : A_i \in \mathcal{B}_i\right\} \varepsilon_i, \| \cdot \|\right). \tag{A.13}$$

A.2. Covering bound of FSTRN

This subsection gives a detailed proof for the covering bound of FSTRN. We first recall a result by Bartlett et al. [1].

Lemma 5 (cf. [1], Lemma 3.2)**.** *Let conjugate exponents (p, q) and (r, s) be given with $p \leq 2$, as well as positive reals (a, b, ε) and positive integer m. Let matrix $X \in \mathbb{R}^{n \times d}$ be given with $\|X\|_p \leq b$. Let $\mathcal{H}_A$ denote the family of matrices obtained by evaluating X with all choices of matrix A:*

$$\mathcal{H}_A \triangleq \left\{XA | A \in \mathbb{R}^{d \times m}, \|A\|_{q,s} \leq a\right\}. \tag{A.14}$$

Then

$$\log \mathcal{N}\left(\mathcal{H}_A, \varepsilon, \| \cdot \|_2\right) \leq \left\lceil \frac{a^2 b^2 m^{2/r}}{\varepsilon^2} \right\rceil \log(2dm). \tag{A.15}$$

This covering bound constrains the hypothesis complexity contributed by a single weight matrix.

As Figure 3 shows, suppose all hypothesis functions $F_0^L, F_1^L, \ldots, F_D^L, F_{Up}^L, F_{SR}^L$ respectively constitute a series of hypothesis spaces $\mathcal{H}_0^L, \mathcal{H}_1^L, \ldots, \mathcal{H}_D^L, \mathcal{H}_{Up}^L, \mathcal{H}_{SR}^L$. For the brevity, we rewrite those notations respectively as $F_0^L, F_1^L, \ldots, F_D^L, F_{D+1}^L, F_{D+2}^L$, and $\mathcal{H}_0^L, \mathcal{H}_1^L, \ldots, \mathcal{H}_D^L, \mathcal{H}_{D+1}^L, \mathcal{H}_{D+2}^L$. Also, suppose the covering number respectively with the radiuses $\varepsilon_0^L, \varepsilon_1^L, \ldots, \varepsilon_D^L, \varepsilon_{D+1}^L, \varepsilon_{D+2}^L$ are $\mathcal{N}(\mathcal{H}_0^L), \mathcal{N}(\mathcal{H}_1^L), \ldots, \mathcal{N}(\mathcal{H}_D^L), \mathcal{N}(\mathcal{H}_{D+1}^L), \mathcal{N}(\mathcal{H}_{D+2}^L)$.

Proof of Theorem 1. Employing Lemma 3, we can straight obtain the following inequality.

$$\log \mathcal{N}(\mathcal{H}) \leq \sum_{d=0}^{D} \log \mathcal{N}(\mathcal{H}_d^L). \tag{A.16}$$

Applying eq. (A.15) of Lemma 5, we can obtain the following result. We first calculate the covering bound of FRBs. Denote the PReLU in the d-th FRB as σ^d and denote the weight matrices corresponding to the 2 convolutional layers respectively as A_1^d and A_2^d. Then, for $d = 1, \dots, D$,

$$\begin{aligned}
&\log \mathcal{N}(\mathcal{H}_{d+1}) \\
\leq &\frac{(b_1^{d+1})^2 \|\sigma^d(F_d(X^T)^T)\|_2^2}{(\varepsilon_1^{d+1})^2} \log(2W^2) \\
&+ \frac{(b_2^{d+1})^2 \|A_1^{d+1}\sigma^{d+1}(F_d(X^T)^T)\|_2^2}{(\varepsilon_2^{d+1})^2} \log(2W^2).
\end{aligned} \tag{A.17}$$

Apparently,

$$\|\sigma^{d+1}(F_d(X^T)^T)\|_2^2 \leq (\rho^{d+1})^2 \|F_d(X^T)^T\|_2^2, \tag{A.18}$$

and

$$\begin{aligned}
&\|A_1^{d+1}\sigma^{d+1}(F_d(X^T)^T)\|_2^2 \\
\leq &(s_1^{d+1})^2 \|\sigma^{d+1}(F_d(X^T)^T)\|_2^2 \\
\leq &(s_1^{d+1}\rho^{d+1})^2 \|F_d(X^T)^T\|_2^2.
\end{aligned} \tag{A.19}$$

Also, motivated by the proof of Lemma 4.3 of [6], we can obtain the following equations.

$$\varepsilon_1^{d+1} = \varepsilon_d^L \rho^{d+1}, \tag{A.20}$$

$$\varepsilon_2^{d+1} = \varepsilon_1^{d+1}(1 + s_1^{d+1}) = \varepsilon_d^L \rho^{d+1}(1 + s_1^{d+1}), \tag{A.21}$$

and

$$\begin{aligned}
\varepsilon_{d+1}^L &= \varepsilon_2^{d+1}(1 + s_2^{d+1}) \\
&= \varepsilon_d^L \rho^{d+1}(1 + s_1^{d+1})(1 + s_2^{d+1}).
\end{aligned} \tag{A.22}$$

Applying eqs. (A.18), (A.19), (A.20), (A.21), (A.22) to eq. (A.17), we can obtain a covering bound for FRBs as follows.

$$\begin{aligned}
&\log \mathcal{N}(\mathcal{H}_{d+1}) \\
\leq &\frac{\|F_d(X)\|_2^2}{(\varepsilon_{d+1}^L)^2} \log\left(2W^2\right) (\rho^{d+1})^2 \\
&\left[(b_1^{d+1})^2(1 + s_1^{d+1})^2 + (b_2^{d+1})^2 (s_1^{d+1})^2\right].
\end{aligned} \tag{A.23}$$

By applying eq. (A.19) and the induction method, we can straight get the following inequality:

$$\begin{aligned}
&\log \mathcal{N}(\mathcal{H}_{d+1}) \\
\leq &\prod_{i=1}^{d} \left[\left(\rho^i s_1^i s_2^i\right)^2 + 1\right] \left[\left(b_1^d\right)^2 \left(1 + s_2^d\right)^2 + \left(b_2^d s_1^d\right)^2\right] \\
&\left(\frac{\|X\|_2 s_1 \rho^d}{\varepsilon^d}\right)^2.
\end{aligned} \tag{A.24}$$

Similarly, we can also get the following inequalities.

$$\log \mathcal{N}(\mathcal{H}_1) \leq \frac{b_1^2 \|X\|_2^2 \bar{\alpha}}{\varepsilon^2} \log\left(2W^2\right), \tag{A.25}$$

$$\begin{aligned}
\log \mathcal{N}(\mathcal{H}_{D+1}) \leq &\left\{1 + \sum_{i=1}^{D} \prod_{j=1}^{i} \left[(\rho^j)^2 (s_1^j s_2^j)^2 + 1\right]\right\} \\
&\|X\|_2^2 \rho_1^2 \frac{b_2^2}{\varepsilon_2^2} \log\left(2W^2\right) \frac{b_2^2}{\varepsilon_2^2},
\end{aligned} \tag{A.26}$$

$$\begin{aligned}
\log \mathcal{N}(\mathcal{H}_{D+2}) \leq &\left\{1 + \sum_{i=1}^{D} \prod_{j=1}^{i} \left[(\rho^j)^2 (s_1^j s_2^j)^2 + 1\right]\right\} \\
&\|X\|_2^2 \rho_1^2 \frac{b_2^2}{\varepsilon_2^2} \log\left(2W^2\right) s_2^2 \frac{b_3^2}{\varepsilon_3^2} \\
&+ \frac{b_1^2 \|X\|_2^2}{\varepsilon^2} \log\left(2W^2\right).
\end{aligned} \tag{A.27}$$

Applying eqs. (A.24, A.25, A.26, and A.27) to eq. (A.16), we eventually prove the theorem. □

A.3. Generalization Bound for FSTRN

The Theorem 2 is the same as Theorem 4.4 of [6]. For the completeness of this paper, we restate the proof here.

Proof of Theorem 2. We prove this theorem in 2 steps: (1) First apply Lemma 2 to get an upper bound on the Rademacher complexity; and then (2) Apply the result of (1) to Lemma 1 in order to get a generalization bound.

(1) *Upper bound on the Rademacher complexity.*

Applying eq. (A.8) of Lemma 2, we can get the following inequality:

$$\begin{aligned}
\mathfrak{R}(\mathcal{H}_\lambda|_D) &\leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{H})} \mathrm{d}\varepsilon\right) \\
&\leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \frac{\sqrt{R}}{\varepsilon} \mathrm{d}\varepsilon\right) \\
&\leq \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \sqrt{R} \log \frac{\sqrt{n}}{\alpha}\right).
\end{aligned} \tag{A.28}$$

Apparently, the infinimum is reached uniquely at $\alpha = 3\sqrt{\frac{R}{n}}$. Here, we use a choice $\alpha = \frac{1}{n}$, and get the following inequality:

$$\mathfrak{R}(\mathcal{H}_\lambda|_D) \leq \frac{4}{n^{\frac{3}{2}}} + \frac{18}{n} \sqrt{R} \log n. \tag{A.29}$$

(2) *Upper bound on the generalization error.*

Combining with Lemma 1, we get the following inequality:

$$\Pr\{\arg\max_i F(x)_i \neq y\}$$

$$\leq \hat{\mathcal{R}}_\lambda(F) + \frac{8}{n^{\frac{3}{2}}} + \frac{36}{n}\sqrt{R}\log n + 3\sqrt{\frac{\log(1/\delta)}{2n}}. \quad \text{(A.30)}$$

The proof is completed. □

B. Empirical Results

This appendix collects all empirical results omitted from the main text. Our algorithm outperforms the state-of-the-art methods in both qualitative and quantitative aspects.

B.1. Quantitatively Results

The quantitative results of all the methods on Vid4 [9] are summarized in Table 1, where the evaluation measures are the PSNR and SSIM indices. As demonstrated in Table 1, our algorithm has excellent robustness in different scenarios and outperforms all other methods.

B.2. Qualitatitve Results

We also qualitatively compare our algorithm with several existing algorithms, Bicubic, SRCNN[4], SRGAN[8], RDN[12], BRCN[7], VESPCN[3], and our FSTRN. The comparison experiments are all with scale factor 4. The qualitative results also illustrate the excellent performance of our algorithm.

References

[1] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NIPS*, pages 6240–6249, 2017.

[2] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(Nov):463–482, 2002.

[3] Jose Caballero, Christian Ledig, Andrew P. Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, pages 2848–2857, 2017.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, volume 8692 of *Lecture Notes in Computer Science*, pages 184–199. Springer, 2014.

[5] Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. In *Selected Works of RM Dudley*, pages 125–165. Springer, 2010.

[6] Fengxiang He, Tongliang Liu, and Dacheng Tao. Why resnet works? residuals generalize. *CoRR*, abs/1904.01367, 2019.

[7] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *NIPS*, pages 235–243, 2015.

[8] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 105–114, 2017.

[9] Ce Liu and Deqing Sun. A bayesian approach to adaptive video super resolution. In *CVPR*, pages 209–216. IEEE Computer Society, 2011.

[10] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[11] Vladimir N Vapnik and Alexey J Chervonenkis. *Theory of pattern recognition*. Nauka, 1974.

[12] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018.

Methods	City PSNR / SSIM	Calendar PSNR / SSIM	Walk PSNR / SSIM	Foliage PSNR / SSIM	Average PSNR / SSIM
Bicubic	24.82 / 0.58	19.98 / 0.55	25.33 / 0.78	22.91 / 0.54	23.25 / 0.62
SRCNN[4]	25.46 / 0.65	21.08 / 0.65	27.16 / 0.84	24.05 / 0.66	24.47 / 0.71
SRGAN[8]	25.30 / 0.64	21.04 / 0.64	26.55 / 0.81	23.69 / 0.62	24.16 / 0.68
RDN[12]	25.59 / 0.66	20.99 / 0.63	27.19 / 0.83	24.05 / 0.66	24.49 / 0.70
BRCN[7]	25.46 / 0.64	21.10 / 0.64	27.06 / 0.84	24.03 / 0.65	24.44 / 0.70
VESPCN[3]	25.55 / 0.66	21.07 / 0.65	27.17 / 0.84	24.08 / 0.67	24.50 / 0.71
FSTRN(ours)	**25.76 / 0.68**	**21.36 / 0.68**	**27.57 / 0.85**	**24.21 / 0.67**	**24.76 / 0.72**

Table 1: Comparison of the PSNR and SSIM results on vid4 [9] sequences by Bicubic, SRCNN[4], SRGAN[8], RDN[12], BRCN[7], VESPCN[3], and our FSTRN with scale factor 4.

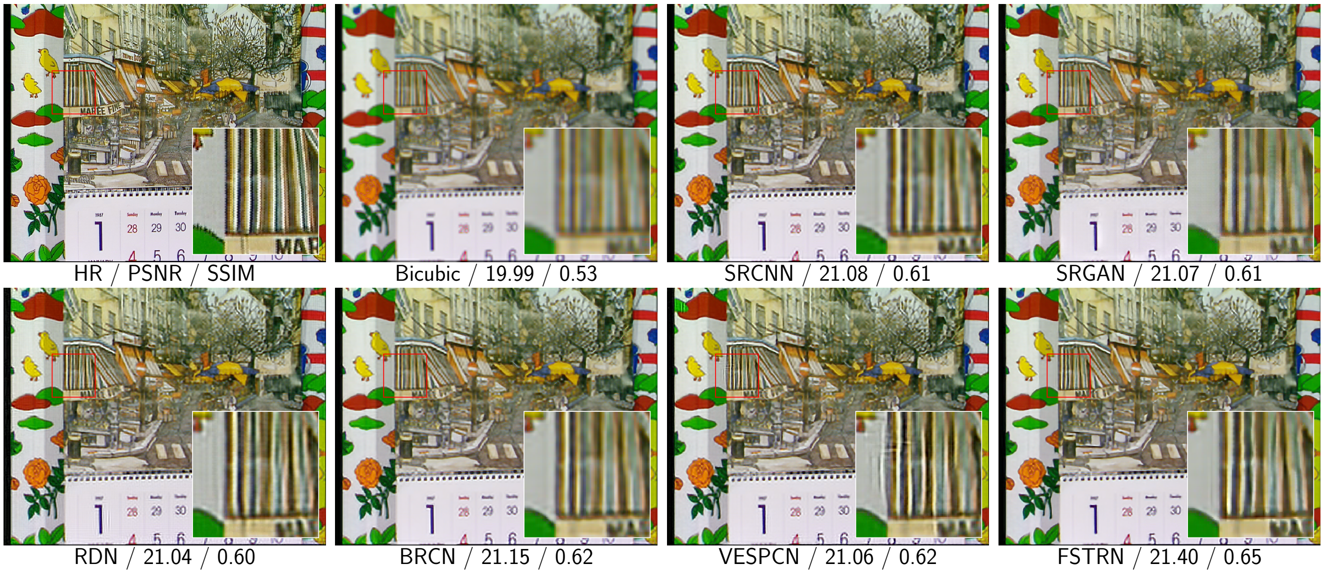


Figure 1: Visual comparisons of the super-resolution results for video **Calendar** on $\times 4$ upscaling factor.

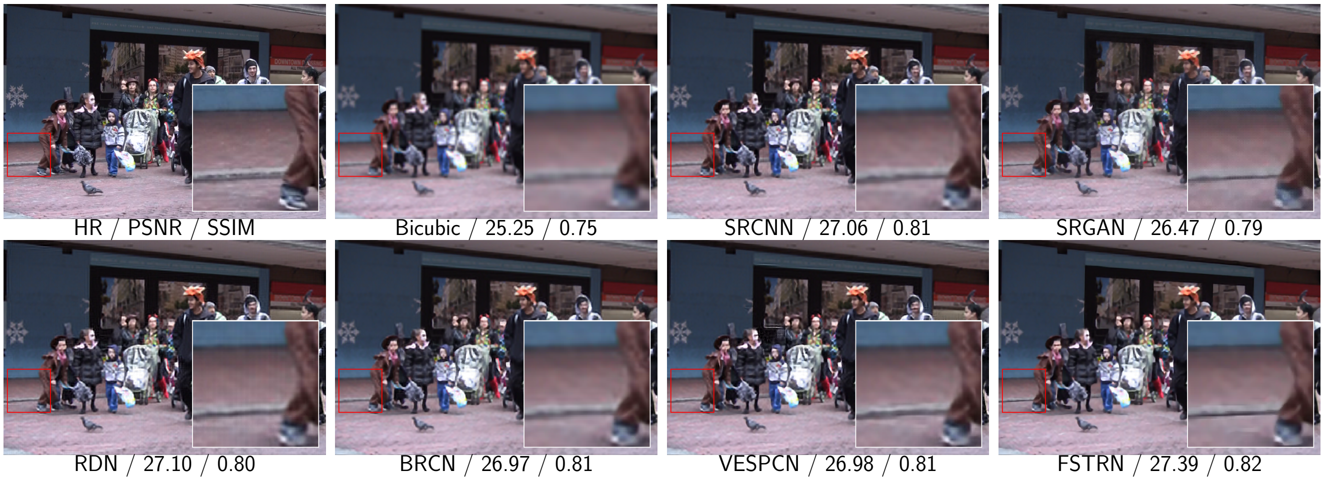
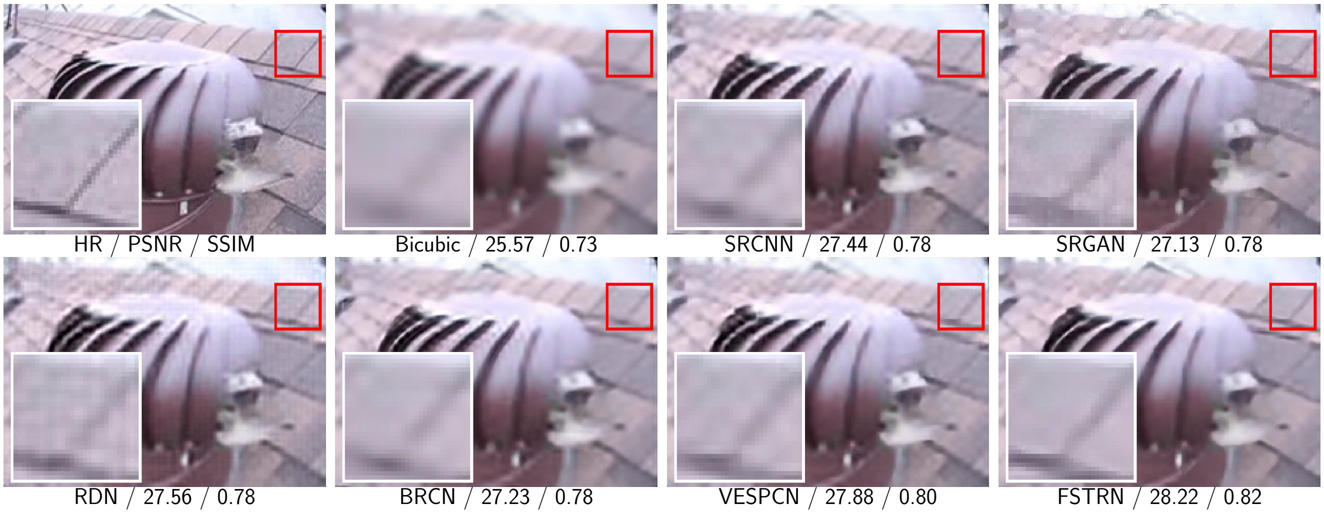


Figure 2: Visual comparisons of the super-resolution results for video **Walk** on $\times 4$ upscaling factor.



HR / PSNR / SSIM

Bicubic / 25.57 / 0.73

SRCNN / 27.44 / 0.78

SRGAN / 27.13 / 0.78

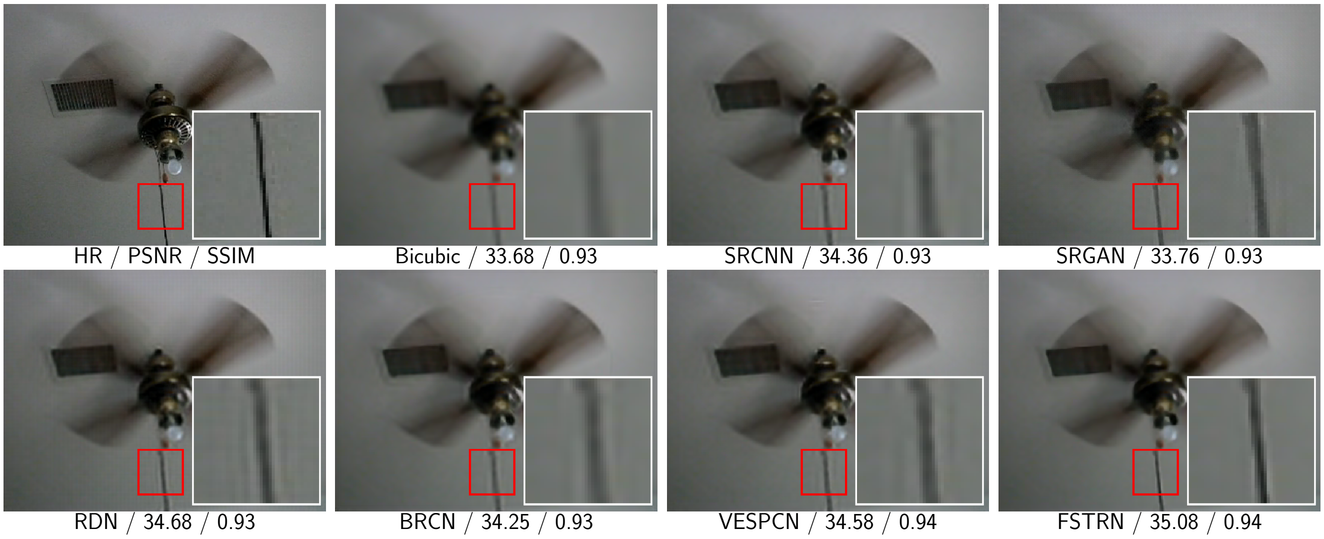
RDN / 27.56 / 0.78

BRCN / 27.23 / 0.78

VESPCN / 27.88 / 0.80

FSTRN / 28.22 / 0.82

Figure 3: Visual comparisons of the super-resolution results for video **Turbine** on ×4 upscaling factor.



HR / PSNR / SSIM

Bicubic / 33.68 / 0.93

SRCNN / 34.36 / 0.93

SRGAN / 33.76 / 0.93

RDN / 34.68 / 0.93

BRCN / 34.25 / 0.93

VESPCN / 34.58 / 0.94

FSTRN / 35.08 / 0.94

Figure 4: Visual comparisons of the super-resolution results for video **Fan** on ×4 upscaling factor.