Supplementary Materials for "Closed-loop Matters: Dual Regression Networks for Single Image Super-Resolution"

Yong Guo*, Jian Chen*, Jingdong Wang*, Qi Chen, Jiezhang Cao, Zeshuai Deng,
Yanwu Xu†, Mingkui Tan†
South China University of Technology, Guangzhou Laboratory, Microsoft Research Asia, Baidu Inc.
{guo.yong, sechenqi, secaojiezhang, sedengzeshuai}@mail.scut.edu.cn,
{mingkuitan, ellachen}@scut.edu.cn, jingdw@microsoft.com, ywxu@ieee.org

We organize our supplementary materials as follows. First, we provide the derivation of generalization error bound for the dual regression scheme in Section A. Second, we provide more details on the architecture of the proposed DRN model in Section B. Third, we provide more implementation details on the training method for the SR tasks with paired data and unpaired data in Section C. Fourth, we conduct more ablation studies on the proposed dual regression scheme in Section D. Last, we report more visual comparison results in Section E.

A. Theoretical Analysis

In this section, we will analyze the generalization bound for the proposed method. The generalization error of the dual learning scheme is to measure how accurately the algorithm predicts for the unseen test data in the primal and dual tasks. Firstly, we will introduce the definition of the generalization error as follows:

Definition 1 *Given an underlying distribution $\mathcal{S}$ and hypotheses $P \in \mathcal{P}$ and $D \in \mathcal{D}$ for the primal and dual tasks, where $\mathcal{P} = \{P_{\theta_{\mathbf{xy}}}(\mathbf{x}); \theta_{\mathbf{xy}} \in \Theta_{\mathbf{xy}}\}$ and $\mathcal{D} = \{D_{\theta_{\mathbf{yx}}}(\mathbf{y}); \theta_{\mathbf{yx}} \in \Theta_{\mathbf{yx}}\}$, and $\Theta_{\mathbf{xy}}$ and $\Theta_{\mathbf{yx}}$ are parameter spaces, respectively, the generalization error (expected loss) is defined by:*

$$E(P, D) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathcal{L}_P(P(\mathbf{x}), \mathbf{y}) + \lambda\mathcal{L}_D(D(P(\mathbf{x})), \mathbf{x})\right], \ \forall P \in \mathcal{P}, D \in \mathcal{D}.$$

In practice, the goal of the dual learning is to optimize the bi-directional tasks. For any $P \in \mathcal{P}$ and $D \in \mathcal{D}$, we define the empirical loss on the N samples as follows:

$$\hat{E}(P, D) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_P(P(\mathbf{x}_i), \mathbf{y}_i) + \lambda\mathcal{L}_D(D(P(\mathbf{x}_i)), \mathbf{x}_i) \tag{1}$$

Following [14], we define Rademacher complexity for dual learning in this paper. We define the function space as $\mathcal{H}_{dual} \in \mathcal{P} \times \mathcal{D}$, this Rademacher complexity can measure the complexity of the function space, that is it can capture the richness of a family of the primal and the dual models. For our application, we mildly rewrite the definition of Rademacher complexity in [14] as follows:

Definition 2 (Rademacher complexity of dual learning) *Given an underlying distribution $\mathcal{S}$, and its empirical distribution $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\}$, where $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$, then the Rademacher complexity of dual learning is defined as:*

$$R_N^{DL}(\mathcal{H}_{dual}) = \mathbb{E}_{\mathcal{Z}}\left[\hat{R}_{\mathcal{Z}}(P, D)\right], \ \forall P \in \mathcal{P}, D \in \mathcal{D},$$

where $\hat{R}_{\mathcal{Z}}(P, D)$ is its empirical Rademacher complexity defined as:

$$\hat{R}_{\mathcal{Z}}(P, D) = \mathbb{E}_{\sigma}\left[\sup_{(P,D)\in\mathcal{H}_{dual}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i(\mathcal{L}_P(P(\mathbf{x}_i), \mathbf{y}_i) + \lambda\mathcal{L}_D(D(P(\mathbf{x}_i)), \mathbf{x}_i))\right].$$

where $\sigma = \{\sigma_1, \sigma_2, \cdots, \sigma_N\}$ are independent uniform $\{\pm 1\}$-valued random variables with $p(\sigma_i = 1) = p(\sigma_i = -1) = \frac{1}{2}$.

*Authors contributed equally.
†Corresponding author.

Generalization bound. Here, we analyze the generalization bound for the proposed dual regression scheme. We first start with a simple case of finite function space. Then, we generalize it to a more general case with infinite function space.

Theorem 1 *Let $\mathcal{L}_P(P(\mathbf{x}),\mathbf{y})+\lambda\mathcal{L}_D(D(P(\mathbf{x})),\mathbf{x})$ be a mapping from $\mathcal{X}\times\mathcal{Y}$ to $[0,C]$, and suppose the function space $\mathcal{H}_{dual}$ is finite, then for any $\delta>0$, with probability at least $1-\delta$, the following inequality holds for all $(P,D)\in\mathcal{H}_{dual}$:*

$$E(P,D)\leq\hat{E}(P,D)+C\sqrt{\frac{\log|\mathcal{H}_{dual}|+\log\frac{1}{\delta}}{2N}}.$$

Proof 1 *Based on Hoeffding's inequality, since $\mathcal{L}_P(P(\mathbf{x}),\mathbf{y})+\lambda\mathcal{L}_D(D(P(\mathbf{x})),\mathbf{x})$ is bounded in $[0,C]$, for any $(P,D)\in\mathcal{H}_{dual}$, then*

$$P\left[E(P,D)-\hat{E}(P,D)>\epsilon\right]\leq e^{-\frac{2N\epsilon^2}{C^2}}$$

Based on the union bound, we have

$$\begin{aligned}&P\left[\exists(P,D)\in\mathcal{H}_{dual}:E(P,D)-\hat{E}(P,D)>\epsilon\right]\\\leq&\sum_{(P,D)\in\mathcal{H}_{dual}}P\left[E(P,D)-\hat{E}(P,D)>\epsilon\right]\\\leq&|\mathcal{H}_{dual}|e^{-\frac{2N\epsilon^2}{C^2}}.\end{aligned}$$

Let $|\mathcal{H}_{dual}|e^{-\frac{2N\epsilon^2}{C^2}}=\delta$, we have $\epsilon=C\sqrt{\frac{\log|\mathcal{H}_{dual}|+\log\frac{1}{\delta}}{2N}}$ and conclude the theorem.

This theorem shows that a larger sample size N and smaller function space can guarantee the generalization. Next, we will give a generalization bound of a general case of an infinite function space using Rademacher complexity.

Theorem 2 *Let $\mathcal{L}_P(P(\mathbf{x}),\mathbf{y})+\lambda\mathcal{L}_D(D(P(\mathbf{x})),\mathbf{x})$ be a mapping from $\mathcal{X}\times\mathcal{Y}$ to $[0,C]$ with the upper bound C, and the function space $\mathcal{H}_{dual}$ be infinite. Then, for any $\delta>0$, with probability at least $1-\delta$, the generalization error $E(P,D)$ (*i.e.*, expected loss) satisfies for all $(P,D)\in\mathcal{H}_{dual}$:*

$$E(P,D)\leq\hat{E}(P,D)+2\hat{R}_{\mathcal{Z}}^{DL}(\mathcal{H}_{dual})+3C\sqrt{\frac{1}{2N}\log\left(\frac{1}{\delta}\right)},\tag{2}$$

where N is the number of samples and $\hat{R}_{\mathcal{Z}}^{DL}$ is the empirical Rademacher complexity of dual learning. Let $\mathcal{B}(P,D)$ be the generalization bound of the dual regression SR, i.e. *$\mathcal{B}(P,D)=2\hat{R}_{\mathcal{Z}}^{DL}(\mathcal{H}_{dual})+3C\sqrt{\frac{1}{2N}\log\left(\frac{1}{\delta}\right)}$, we have*

$$\mathcal{B}(P,D)\leq\mathcal{B}(P),\tag{3}$$

where $\mathcal{B}(P),P\in\mathcal{H}$ is the generalization bound of standard supervised learning w.r.t. the Rademacher complexity $\hat{R}_{\mathcal{Z}}^{SL}(\mathcal{H})$.

Proof 2 *Based on Theorem 3.1 in [14], we extend a case for $\mathcal{L}_P(P(\mathbf{x}),\mathbf{y})+\lambda\mathcal{L}_D(D(P(\mathbf{x})),\mathbf{x})$ bounded in $[0,C]$, and we have the generalization bound in (2). According to the definition of Rademacher complexity, we have $\hat{R}_{\mathcal{Z}}^{DL}(\mathcal{H}_{dual})\leq\hat{R}_{\mathcal{Z}}^{SL}(\mathcal{H})$ because the capacity of the function space $\mathcal{H}_{dual}\in\mathcal{P}\times\mathcal{D}$ is smaller than the capacity of the function space $\mathcal{H}\in\mathcal{P}$. With the same number of samples, we have $\mathcal{B}(P,D)\leq\mathcal{B}(P)$.*

Theorem 2 shows that with probability at least $1-\delta$, the generalization error is smaller than $2R_N^{DL}+C\sqrt{\frac{1}{2N}\log(\frac{1}{\delta})}$ or $2\hat{R}_{\mathcal{Z}}^{DL}+3C\sqrt{\frac{1}{2N}\log(\frac{1}{\delta})}$. It suggests that using the function space with larger capacity and more samples can guarantee better generalization. Moreover, the generalization bound of dual learning is more general for the case that the loss function $\mathcal{L}_P(P(\mathbf{x}),\mathbf{y})+\lambda\mathcal{L}_D(D(P(\mathbf{x})),\mathbf{x})$ is bounded by $[0,C]$, which is different from [19].

Remark 1 *Based on the definition of Rademacher complexity, the capacity of the function space $\mathcal{H}_{dual}\in\mathcal{P}\times\mathcal{D}$ is smaller than the capacity of the function space $\mathcal{H}\in\mathcal{P}$ or $\mathcal{H}\in\mathcal{D}$ in traditional supervised learning,* i.e., *$\hat{R}_{\mathcal{Z}}^{DL}\leq\hat{R}_{\mathcal{Z}}^{SL}$, where $\hat{R}_{\mathcal{Z}}^{SL}$ is Rademacher complexity defined in supervised learning. In other words, dual learning has a smaller generalization bound than supervised learning and the proposed dual regression model helps the primal model to achieve more accurate SR predictions.*

B. Model Details of Dual Regression Network

Deep neural networks (DNNs) have achieved great success in image classification [9, 4, 8, 10], image generation [6, 3], and image restoration [5, 7]. In this paper, we propose a novel Dual Regression Network (DRN), which contains a primal model and a dual model. Specifically, the primal model contains 2 basic blocks for $4\times$ SR and 3 blocks for $8\times$ SR. To form a closed-loop, according to the architecture design of the primal model, there are 2 dual models for $4\times$ SR and 3 dual models for $8\times$ SR, respectively.

Let B be the number of RCABs [21] and F be the number of base feature channels. For $4\times$ SR, we set $B = 30$ and $F = 16$ for DRN-S and $B = 40$ and $F = 20$ for DRN-L. For $8\times$ SR, we set $B = 30$ and $F = 8$ for DRN-S and $B = 36$ and $F = 10$ for DRN-L. Moreover, we set the reduction ratio $r = 16$ in all RCABs for our DRN model and set the negative slope to 0.2 for all LeakyReLU in DRN. We show the detailed architecture of the $8\times$ DRN model in Table A. To obtain the $4\times$ model, one can simply remove one basic block from the $8\times$ model.

As shown in Table A, we use Conv(1,1) and Conv(3,3) to represent the convolution layer with the kernel size of 1×1 and 3×3, respectively. We use Conv_{s2} to represent the convolution layer with the stride of 2. Following the settings of EDSR [12], we build the Upsampler with one convolution layer and one pixel-shuffle [16] layer to upscale the feature maps. Moreover, we use h and w to represent the height and width of the input LR images. Thus, the shape of output images should be 8h $\times$ 8w for the $8\times$ model.

Table A. Detailed model design of the proposed $8\times$ DRN.

Module	Module details	Input shape	Output shape
Head	Conv(3,3)	(3, 8h, 8w)	($1F$, 8h, 8w)
Down 1	Conv_{s2}-LeakyReLU-Conv	($1F$, 8h, 8w)	($2F$, 4h, 4w)
Down 2	Conv_{s2}-LeakyReLU-Conv	($2F$, 4h, 4w)	($4F$, 2h, 2w)
Down 3	Conv_{s2}-LeakyReLU-Conv	($4F$, 2h, 2w)	($8F$, 1h, 1w)
Up 1	B RCABs	($8F$, 1h, 1w)	($8F$, 1h, 1w)
	$2\times$ Upsampler	($8F$, 1h, 1w)	($8F$, 2h, 2w)
	Conv(1,1)	($8F$, 2h, 2w)	($4F$, 2h, 2w)
Concatenation 1	Concatenation of the output of Up 1 and Down 2	($4F$, 2h, 2w) $\oplus$ ($4F$, 2h, 2w)	($8F$, 2h, 2w)
Up 2	B RCABs	($8F$, 2h, 2w)	($8F$, 2h, 2w)
	$2\times$ Upsampler	($8F$, 2h, 2w)	($8F$, 4h, 4w)
	Conv(1,1)	($8F$, 4h, 4w)	($2F$, 4h, 4w)
Concatenation 2	Concatenation of the output of Up 2 and Down 1	($2F$, 4h, 4w) $\oplus$ ($2F$, 4h, 4w)	($4F$, 4h, 4w)
Up 3	B RCABs	($4F$, 4h, 4w)	($4F$, 4h, 4w)
	$2\times$ Upsampler	($4F$, 4h, 4w)	($4F$, 8h, 8w)
	Conv(1,1)	($4F$, 8h, 8w)	($1F$, 8h, 8w)
Concatenation 3	Concatenation of the output of Up3 and Head	($1F$, 8h, 8w) $\oplus$ ($1F$, 8h, 8w)	($2F$, 8h, 8w)
Tail 0	Conv(3,3)	($8F$, 1h, 1w)	(3, 1h, 1w)
Tail 1	Conv(3,3)	($8F$, 2h, 2w)	(3, 2h, 2w)
Tail 2	Conv(3,3)	($4F$, 4h, 4w)	(3, 4h, 4w)
Tail 3	Conv(3,3)	($2F$, 8h, 8w)	(3, 8h, 8w)
Dual 1	Conv_{s2}-LeakyReLU-Conv	(3, 8h, 8w)	(3, 4h, 4w)
Dual 2	Conv_{s2}-LeakyReLU-Conv	(3, 4h, 4w)	(3, 2h, 2w)
Dual 3	Conv_{s2}-LeakyReLU-Conv	(3, 2h, 2w)	(3, 1h, 1w)

C. More Implementation Details

C.1. Supervised Image Super-Resolution

Training data. Following [18], we train our model on DIV2K [17] and Flickr2K [12] datasets, which contain 800 and 2650 training images separately. We use the RGB input patches of size 48×48 from LR images and the corresponding HR patches as the paired training data, and augment the training data following the method in [12, 21].

Test data. For quantitative comparison on paired data, we evaluate different SR models using five benchmark datasets, including SET5 [2], SET14 [20], BSDS100 [1], URBAN100 [11] and MANGA109 [13].

Implementation details. For training, we apply Adam with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and set minibatch size as 32. The learning rate is initialized to 10^{-4} and decreased to 10^{-7} with a cosine annealing out of 10^6 iterations in total.

C.2. Adaptation to Real-world Scenarios with Unpaired Data

Training data. To obtain the unpaired synthetic data, we randomly choose 3k images from ImageNet [15] (called ImageNet3k) and obtain the LR images using different degradation methods, including Nearest and BD. More specifically, we use Matlab to obtain the Nearest data. The BD data is obtained using the Gaussian kernel with size 7×7 and a standard deviation of 1.6. Note that ImageNet3K HR images are not used in our experiments. Moreover, we collect 3k LR raw video frames from YouTube as the unpaired real-world data to evaluate the proposed DRN in a more general and challenging case. More critically, we use both paired data (DIV2K [17]) and unpaired data to train the proposed models.
Test data. For quantitative comparison on unpaired synthetic data, we obtain the LR images of five benchmark datasets using Nearest and BD degradation methods separately.
Implementation details. We train a DRN-Adapt model for each kind of unpaired data, *i.e.*, Nearest data, BD data, and video frames collected from YouTube. Thus, there are 3 DRN-adapt models in total. And We also train a CinCGAN [22] model for each kind of unpaired data for comparison. Based on pretrained DRN-S, We train our DRN-Adapt models with a learning rate of 10^{-4} and the data ratio of unpaired data $\rho = 30\%$ for a total of 10^5 iterations. Moreover, we apply Adam with $\beta_1 = 0.9$, $\beta_2 = 0.99$ to optimize the models, and set minibatch size as 16.

D. More Ablation Studies on Dual Regression

In this section, we first provide an additional ablation study of the dual regression scheme on other architectures. Then, we investigate the effect of the dual regression scheme on LR images and the effect on both LR and HR images. Last, we compare the performance of the DRN models trained with and without trainable dual models.

D.1. Effect of Dual Regression Scheme on Other Architectures

To verify the impact of the dual regression scheme, we also conduct an ablation study of the dual network for SRResNet (see architecture in Figure A). "SRResNet + Dual" denotes the baseline SRResNet equipped with the dual regression scheme. From Table B, the model with the dual regression scheme consistently outperforms the baseline counterpart, which further demonstrates the effectiveness of our method.

Table B. The impact of the proposed dual regression scheme on the SRResNet model in terms of PSNR score on the five benchmark datasets for $4\times$ SR.

Method	Set5	Set14	BSDS100	Urban100	Manga109
SRResNet	32.26	28.53	27.61	26.24	31.03
SRResNet + Dual	**32.47**	**28.77**	**27.70**	**26.58**	**31.24**

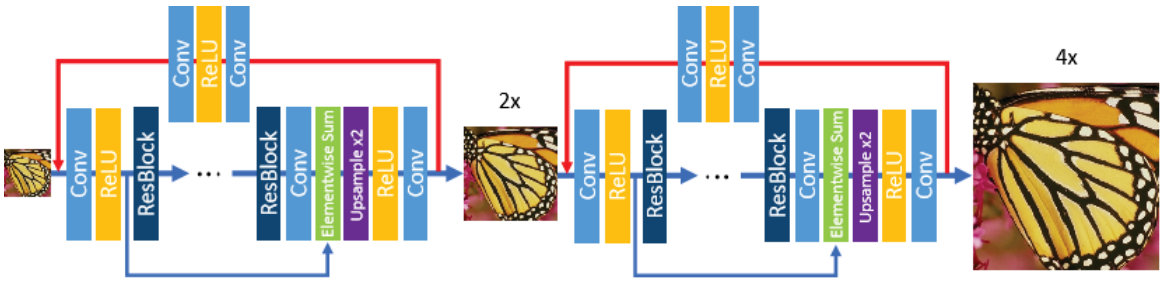

Figure A. The SRResNet architecture equipped with the proposed dual regression scheme for $4\times$ SR.

D.2. Effect of the Dual Regression on HR Data

As mentioned in Section 3.1, one can also add a dual regression constraint on the HR domain, *i.e.*, downscaling and upscaling to reconstruct the original HR images. In this experiment, we investigate the impact of dual regression loss on HR data and show the results in Table C. For convenience, we use "DRN-S with dual HR" to represent the model with the regression on both LR and HR images. From Table C, DRN-S yields comparable performance with "DRN-S with dual HR" while only needs half the computation cost. Thus, it is not necessary to apply the dual regression on HR images in practice.

Table C. The impact of the dual regression loss on HR data for 4× SR. We take DRN-S as the baseline model.

Method	MAdds	Set5	Set14	BSDS100	Urban100	Manga109
DRN-S with dual HR	51.20G	32.69	28.93	27.79	26.85	31.54
DRN-S (Ours)	25.60G	32.68	28.93	27.78	26.84	31.52

D.3. Effect of the Trainable Dual Models

To verify the impact of the trainable dual models, we conduct an ablation study on the dual model whether it is trainable or not (see results in Table D). "DRN-S with fixed dual" denotes the model using a fixed degradation method (*i.e.*, Bicubic) to form the close-loop. From Table D, the model with trainable dual models significantly outperforms the counterpart using the fixed degradation method, which demonstrates the necessity of the trainable dual models.

Table D. Comparison of the DRN-S models equipped with the trainable dual model and the fixed dual degradation method for 4× SR.

Method	Set5	Set14	BSDS100	Urban100	Manga109
DRN-S with fixed dual	32.31	28.51	27.45	26.27	31.04
DRN-S (Ours)	**32.68**	**28.93**	**27.78**	**26.84**	**31.52**

D.4. Impact of Different Degradation Methods to Obtain Paired Synthetic Data

In this experiment, we investigate the impact of different degradation methods to obtain paired synthetic data. We change kernel from Bicubic to Nearest and evaluate the adaptation models on BD data. From Table F, DRN-Adapt obtains similar results when we use different degradation methods to obtain the paired synthetic data.

Table F. The impact of different degradation methods on DRN-Adapt for 8× SR.

Degradation Method	Set5	Set14	BSDS100	Urban100	Manga109
Nearest	24.60	23.03	**23.60**	**20.61**	21.46
Bicubic	**24.62**	**23.07**	23.59	20.57	**21.52**

E. More Comparisons and Results

For supervised super-resolution, we put more visual results in this section shown in Figures C and D, respectively. Considering the scenario with unpaired data, we put more visual results on real-world unpaired data (See Figure E). From these results, our models are able to produce the images with sharper edges and clearer textures than state-of-the-art methods.

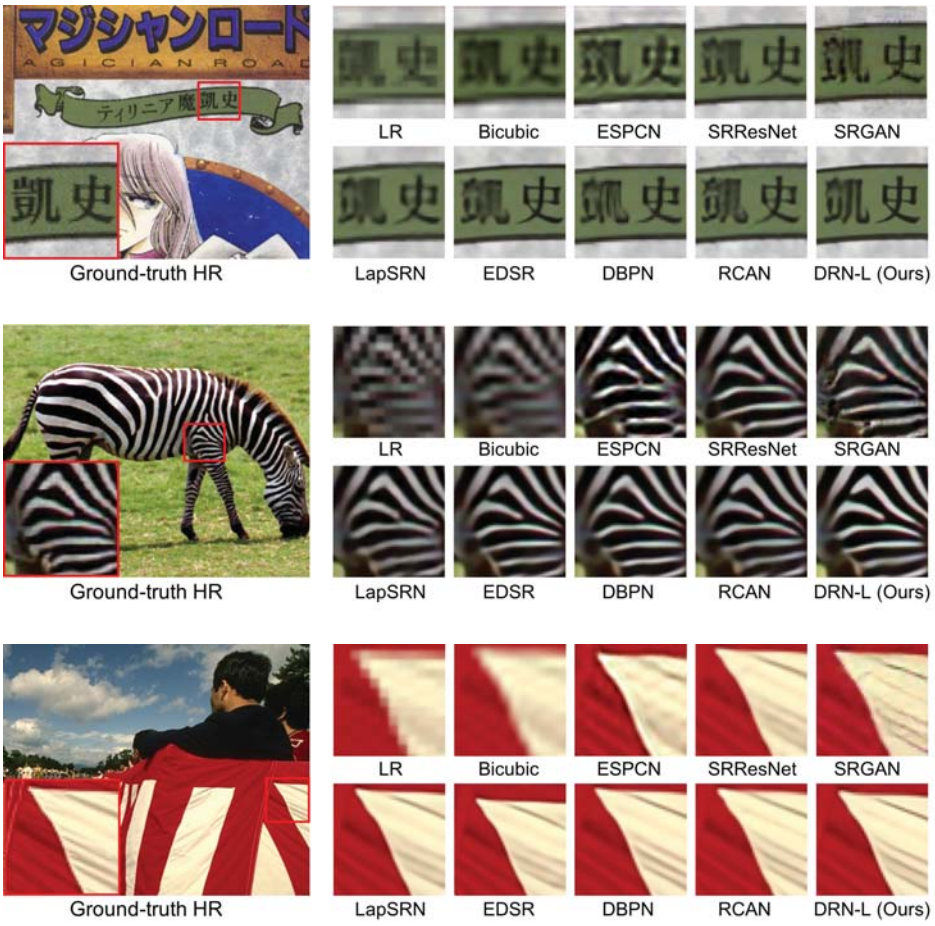

Figure C. Visual comparison for $4\times$ image super-resolution on benchmark datasets.

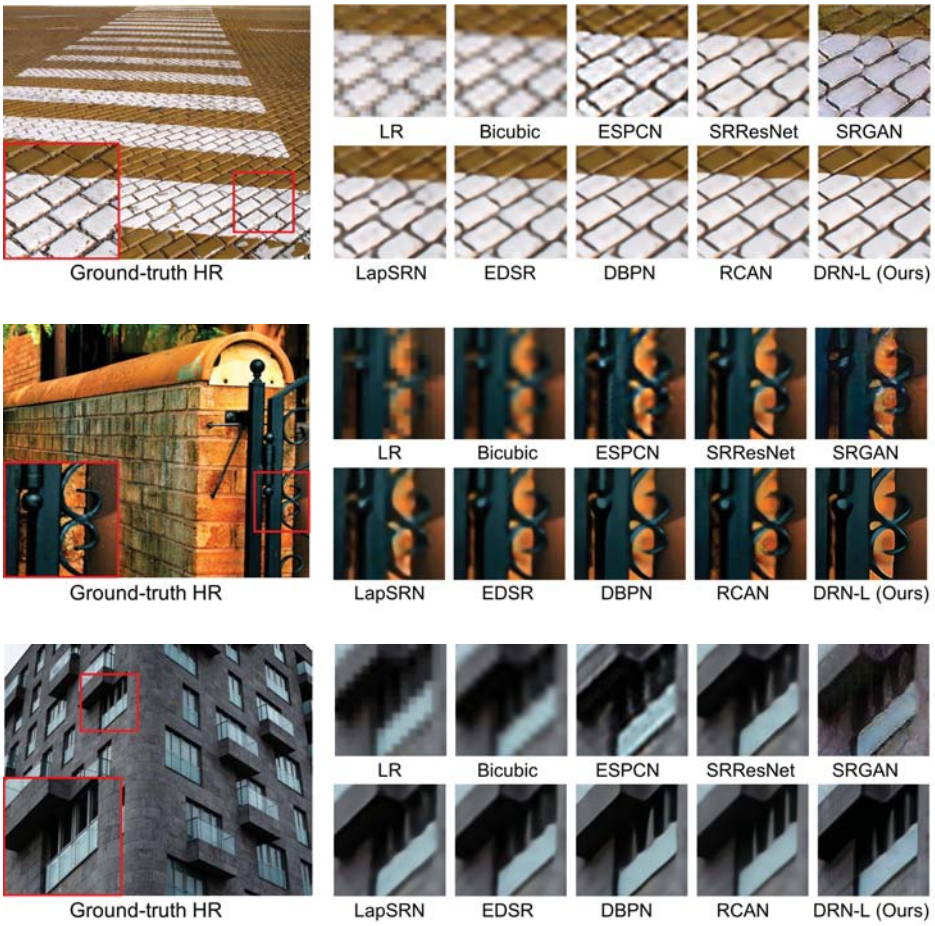

Figure D. Visual comparison for $8\times$ image super-resolution on benchmark datasets.

Figure E. Visual comparison of model adaptation for $8\times$ super-resolution on real-world video frames (from YouTube).

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. 3

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012. 3

[3] Jiezhang Cao, Yong Guo, Qingyao Wu, Chunhua Shen, Junzhou Huang, and Mingkui Tan. Adversarial learning with local coordinate coding. In *International Conference on Machine Learning*, 2018. 3

[4] Yong Guo, Jian Chen, Qing Du, Anton Van Den Hengel, Qinfeng Shi, and Mingkui Tan. Multi-way backpropagation for training compact deep neural networks. *Neural networks*, 2020. 3

[5] Yong Guo, Qi Chen, Jian Chen, Junzhou Huang, Yanwu Xu, Jiezhang Cao, Peilin Zhao, and Mingkui Tan. Dual reconstruction nets for image super-resolution with gradient sensitive loss. *arXiv preprint arXiv:1809.07099*, 2018. 3

[6] Yong Guo, Qi Chen, Jian Chen, Qingyao Wu, Qinfeng Shi, and Mingkui Tan. Auto-embedding generative adversarial networks for high resolution image synthesis. *IEEE Transactions on Multimedia*, 2019. 3

[7] Yong Guo, Yongsheng Luo, Zhenhao He, Jin Huang, and Jian Chen. Hierarchical neural architecture search for single image super-resolution. *arXiv preprint arXiv:2003.04619*, 2020. 3

[8] Yong Guo, Mingkui Tan, Qingyao Wu, Jian Chen, Anton Van Den Hengel, and Qinfeng Shi. The shallow end: Empowering shallower deep-convolutional networks through auxiliary outputs. *arXiv preprint arXiv:1611.01773*, 2016. 3

[9] Yong Guo, Qingyao Wu, Chaorui Deng, Jian Chen, and Mingkui Tan. Double forward propagation for memorized batch normalization. In *AAAI*, 2018. 3

[10] Yong Guo, Yin Zheng, Mingkui Tan, Qi Chen, Jian Chen, Peilin Zhao, and Junzhou Huang. Nat: Neural architecture transformer for accurate and compact architectures. In *Advances in Neural Information Processing Systems*, pages 735–747, 2019. 3

[11] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 3

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 3

[13] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20), 2017. 3

[14] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2012. 1, 2

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 2015. 4

[16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 3

[17] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1110–1121. IEEE, 2017. 3, 4

[18] 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 *European Conference on Computer Vision Workshops*, pages 0–0, 2018. 3

[19] Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. Dual supervised learning. In *International Conference on Machine Learning*, 2017. 2

[20] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pages 711–730. Springer, 2010. 3

[21] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, 2018. 3

[22] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 4