SR^2: SUPER-RESOLUTION WITH STRUCTURE-AWARE RECONSTRUCTION

Franziska Schirrmacher[1,*], *Benedikt Lorch*[1,†], *Bernhard Stimpel*[2], *Thomas Köhler*[3], *Christian Riess*[1,†]

[1]IT Security Infrastructures Lab, Computer Science, Univ. of Erlangen-Nürnberg
[2]Pattern Recognition Lab, Computer Science, Univ. of Erlangen-Nürnberg
[3]e.solutions GmbH, Erlangen, Germany

ABSTRACT

Image reconstruction is particularly difficult when the type of image degradations are unknown. This may be the case if the acquisition device is unknown or the images stem from an uncontrolled environment like the internet. Yet, it may be important to reconstruct a specific piece of information from the image, such as digits from signs or vehicle license plates. Existing works incorporate such prior information with a sequential super-resolution and classification pipeline. However, this approach is prone to error propagation.

In this work, we propose a new approach of connecting classification and super-resolution in parallel within a multi-task network. We show that this architecture is able to preserve structures and to remove noisy pixels although the network itself has never been trained on noisy data. We also show that this design allows to transparently trade classification and super-resolution quality. On upsampling by factor 4, we outperform sequential approaches in terms of SSIM by 10% and improve classification by 69%.

***Index Terms*—** Deep learning; Multi-task learning; Super-resolution; Classification

1. INTRODUCTION

Recent advances in deep learning have led to tremendous performance gains of single-image super-resolution. This can partly be attributed to deep convolutional networks [1, 2, 3]. Generative Adversarial Networks further improved the perceptual quality of the reconstructed images [4, 5, 6]. Attention models for super-resolution have shown a great advantage to focus on relevant image information. Here, first-order [7] and second-order [8] attention networks are able to capture details and to bypass irrelevant low-frequency information. Deep neural networks can fully exploit their potential on large and comprehensive training datasets. However, deviations in the characteristics of the training data and test images in the wild can cause considerable performance degradations [9]. More specifically, unseen image degradations like different noise levels or distributions are challenging and impede the performance of deep networks [10].

One way to improve network generalization is multi-task learning [11, 12]. Here, simultaneously learning multiple tasks serves as an inductive bias for the model [13]. In particular, joining super-resolution and classification in one model has shown great benefits. Existing work sequentially connects both tasks. First, a high-resolution image is reconstructed from low-resolution input. Second, an object of interest is classified from the reconstructed image. This approach improved the pixel-wise classification of hyperspectral images on only a limited amount of labeled training data [14] and was also successfully applied to pedestrian detection [15]. A similar design improved scene text recognition, with an end-to-end network of sequential super-resolution and classification [16].

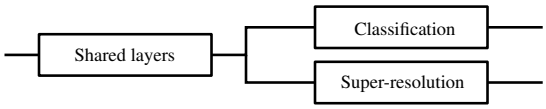


Fig. 1: Schematic illustrations of the proposed super-resolution with a structure-aware reconstruction (SR^2) network. We arrange super-resolution and classification in parallel and connect the tasks with shared layers.

However, there is one notable disadvantage of sequential stacking of super-resolution and classification, namely error propagation. This becomes increasingly severe if additional image degradations negatively affect the super-resolution output, which forces the classifier to operate on a distorted image.

To address this issue, we propose a parallel super-resolution and classification architecture, where both tasks are connected with shared layers, see Fig. 1. As a result, both tasks are performed on the same input. This is a somewhat looser coupling of both tasks. If one task is impaired by distortions, the other task may still succeed. Additionally, both the super-resolved image and the classification result can be useful in several applications. For example, in police investigations it can be important to reconstruct registration numbers or letters from images from uncontrolled, poor-quality sources. Classification can help to reduce the list of suspects, and the super-resolved image is used for validating and defending this assessment. Another application example is face recognition of severely degraded images.

The contributions of this paper are two-fold:

1. We propose a new end-to-end super-resolution with structure-aware reconstruction (SR^2) architecture to connect super-resolution and classification in parallel.

2. We demonstrate the robustness of this approach on two well-known digit datasets. While trained on noise-free data, we test our method on unseen image degradation and report better generalization compared to sequential approaches.

In this work, we focus on digits but there are no inherent restrictions towards the type of object to be super-resolved.

The paper is organized as follows: Section 2 presents the proposed concept and network architectures. Section 3 reports the experiments and discusses the results. Section 4 concludes the paper.

*Supported by the German Research Foundation (146371743/TRR 89)
†Supported by the German Research Foundation, GRK Cybercrime (393541319/GRK2475/1-2019)

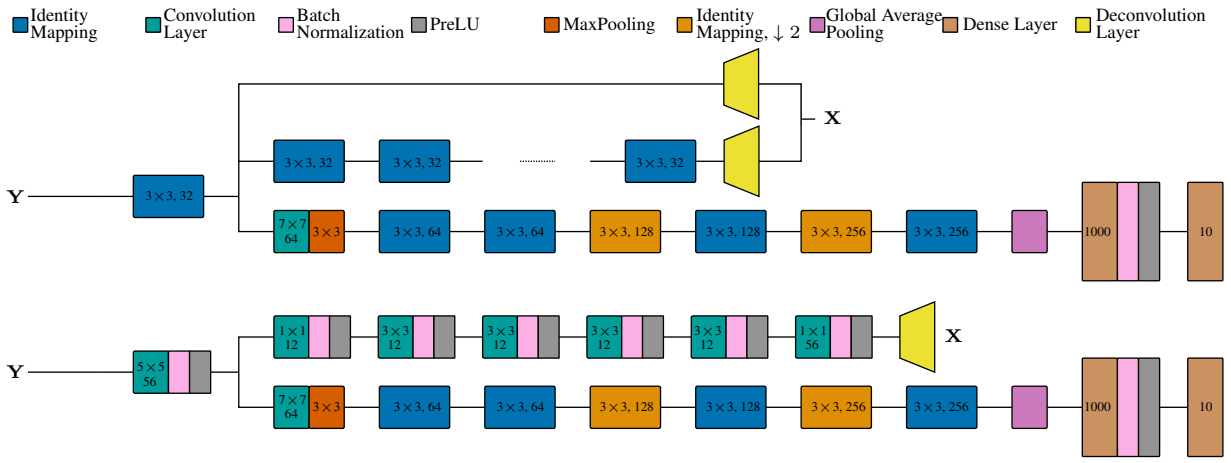


Fig. 2: Example realizations of the proposed architecture. Top: Parallel WDSR and ResNet networks perform super-resolution and classification, respectively. Bottom: Parallel FSRCNN and ResNet networks perform these tasks. In both designs, the first layers are shared to mutually benefit from super-resolution and classification. The annotations in the convolutional layers indicate the size and number of filters. The stride is always 1 except for the identity mapping with downsampling (↓ 2). The black dots in WDSR denote repeated identity mappings.

2. PROPOSED METHOD

The focus of this work is on multi-task learning for image reconstruction, specifically the combination of super-resolution and classification. We aim at recovering a high-resolution (HR) image $\mathbf{X}$ from a low-resolution (LR) image $\mathbf{Y}$. Additionally, the object in $\mathbf{Y}$ is classified. The classification task provides information about the object and improves the performance of the super-resolution. In the proposed approach, both tasks are arranged in parallel. They are connected by shared layers and jointly optimized. The two specific architectures shown in Fig. 2 are example realizations. The architecture on top is denoted as $\text{SR}^2_{\text{WDSR}}$. It connects WDSR [17] for super-resolution and ResNet-18 [18] for classification. The architecture on the bottom is denoted as $\text{SR}^2_{\text{FSRCNN}}$. It connects FSRCNN [19] for super-resolution and ResNet-18 for classification.

Super-Resolution. WDSR consists of two branches. One is a concatenation of residual building blocks followed by a deconvolution operation. The other branch only performs deconvolution. We implement a slightly modified version of the architecture compared to the original paper. We use the identity mappings, residual building blocks with changed order of operations, proposed by He *et al.* [20]. Additionally, we replace the weight normalization by batch normalization. We use 16 identity mappings for WDSR in total, including the shared layers. FSRCNN is a concatenation of convolution layers, comprising feature extraction, shrinking, mapping, expanding, and deconvolution.

Classification. ResNet is a common classification architecture. There are several ResNet variants, from which we use ResNet-18. However, any other classification network could in principle also be used. We remove the last two identity mappings from ResNet-18 and perform less downsampling of the feature maps due to the small input size of only $\leq 16 \times 16$ pixels in our datasets. The classification output is an n-element vector, where n denotes the number of classes.

Joint Super-Resolution and Classification. As shown in previous works, the joint optimization of super-resolution and classification is of advantage [15]. In contrast to previous works, we propose to use shared layers to which both tasks contribute, followed by a split into two branches. The shared layer is taken from the super-resolution network. For WDSR, it is the first identity mapping. For FSRCNN, it is the first convolution layer. Experiments regarding more shared layers are reported in Sec. 3.3.

Fig. 3: Example images from MNIST (left) and SVHN (right).

After the split, one branch performs super-resolution, and the second branch performs classification. Thus, the gradient of the classification network contributes to the weight update of the first layer(s) of the super-resolution network. In the end, both loss functions are combined into a total loss $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SR}} + \lambda \mathcal{L}_{\text{Cl}} \ , \tag{1}$$

where λ denotes the weight of the classification loss $\mathcal{L}_{\text{Cl}}$ and $\mathcal{L}_{\text{SR}}$ denotes the super-resolution loss. We use the cross-entropy for classification and the mean-squared error for super-resolution.

3. EXPERIMENTAL PROTOCOL AND RESULTS

Datasets. The experiments are performed on the MNIST [21] and SVHN [22] datasets. Both datasets contain digits in various representations and of varying difficulty. Figure 3 shows example data.

MNIST contains white handwritten digits on a black background. It consists of 60 000 training images and 10 000 test images with a size of 28×28 pixels.

SVHN contains street view house numbers. We use the MNIST-like color images, consisting of 73 257 training images and 26 032 test images with a size of 32×32 pixels. We also use 126 743 images of the additional training data in the dataset. SVHN is more challenging than MNIST. MNIST shows only a single digit per image. SVHN images may exhibit additional rotation and distractors next to the digit. Backgrounds in MNIST are black. SVHN backgrounds have different colors, may vary within the same image, and the overall contrast between digit and background may be considerably lower than in MNIST.

	2x	4x
SRMD [23]	0.917	0.803
FSRCNN [19]	0.944	0.796
WDSR [17]	0.950	0.768
SR^2_{FSRCNN}	0.942	**0.828**
SR^2_{WDSR}	**0.959**	0.801

Table 1: The average SSIM on the MNIST test data. Gaussian noise with $\sigma = 10^{-4}$ is added to the LR images.

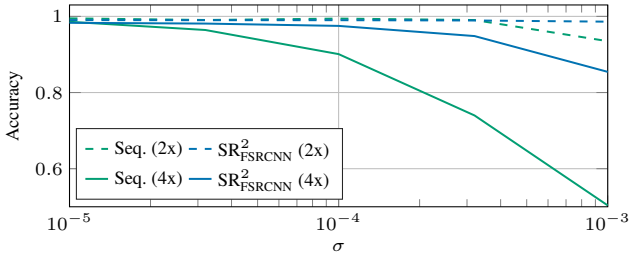


Fig. 4: Comparison of sequential (seq.) connection of FSRCNN and ResNet-18 and SR^2_{FSRCNN} on the MNIST test data. We report the average accuracy of the predicted digits displayed in the LR images. Gaussian noise with standard deviation σ is added to the LR images.

Training. The training procedures are similar for both datasets and performed in two stages, analogously to previous work [15]. First, the super-resolution network is trained using the ADAM optimizer with a learning rate of 10^{-3}. All network weights are initialized using the method proposed by He *et al.* [24]. In a second step, the classification branch is attached to the network. The weights from the previous step are used to initialize the super-resolution branch and the shared layer(s). In the second training step, we use a learning rate of 10^{-5} and jointly optimize both tasks. All weights are subject to L2 regularization with regularization weight 0.01.

The training data is split into training (80%) and validation (20%) to tune the hyperparameter of the networks. Each image is prepared as follows: The low-resolution network input $\mathbf{Y}$ is obtained via Gaussian blur and subsequent downsampling by factors 2 or 4, respectively. To prevent overfitting, we additionally augment the training data before blurring and downsampling using horizontal and vertical image shifts by up to ± 3 pixels and rotation by up to 10°.

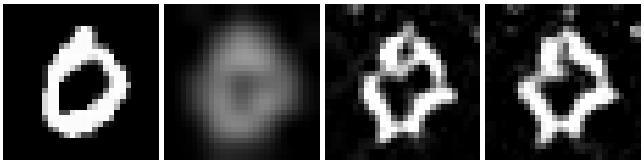
Evaluation Protocol. Tests are is performed on noisy data to investigate the robustness of the proposed method. To this end, Gaussian noise with varying standard deviations σ is added to the low-resolution test images. Since training of deep neural networks is a non-convex optimization problem, different initialization yields a variation in performance. To account for this, we train the network three times and report the mean Structural Similarity Index (SSIM) [25] and the accuracy of the digit classification over those runs.

3.1. Comparison of Architectures

The proposed method is compared to the original WDSR [17] and FSRCNN [19] methods, and to the pretrained super-resolution network for multiple degradations (SRMD) [23]. SRMD was trained on data with varying blur, downsampling, and noise.

We first report results for super-resolution by comparing the average structural similarity measure (SSIM) on MNIST in Tab. 1. For SRMD we report the average SSIM over one test run using the provided pretrained weights. All other results are averaged over three training and test runs. The first and second column report results on magnification factors 2 and 4. The test images for this experiment are distorted with additive Gaussian noise with $\sigma = 10^{-4}$. Only SRMD has seen noisy data during training, which makes this task difficult for the other networks. For a magnification of 2, SR^2_{WDSR} outperforms WDSR, FSRCNN, and SRMD. However, for the more challenging case of magnification of 4, both proposed architectures perform substantially better than FSRCNN and WDSR. SR^2_{FSRCNN} outperforms SRMD, although not trained on any noisy data.

In a second experiment, we show that our parallel architecture SR^2_{FSRCNN} outperforms the sequential architecture. Similar to prior work [15] we construct the sequential architecture by concatenating FSRCNN and ResNet-18. Thus, the digit is classified based on the high-resolution reconstructed image. The weighting of the classification is chosen for both approaches individually and based on the validation data. For sequential connection we chose $\lambda = 0.11$ and for SR^2_{FSRCNN} we chose $\lambda = 9$. To show the disadvantage of sequential connection, we evaluate the accuracy of the predicted digits. The accuracy heavily drops with increasing noise levels and for magnification factor 4, see Fig. 4. For magnification factor 2, the results are comparable for low noise but the gap increases for strong noise. Thus, the proposed method reliably classifies the input image even under strong distortion. This is shown in Fig. 5 for magnification factor 4: (a) shows the ground truth and (b) the input image. While the sequential architecture links the two sides of the 0, the proposed SR^2_{FSRCNN} can suppress the noisy pixel, see (c) and (d).



(a) Ground truth (b) Bicubic (c) Sequential (d) SR^2_{FSRCNN}

Fig. 5: Comparison of sequential connection of FSRCNN and ResNet-18 and SR^2_{FSRCNN}. Gaussian noise with $\sigma = 10^{-3}$ is added to the MNIST test image and perform 4x magnification.

Fig. 6 illustrates the SSIM values of FSRCNN, sequential (seq.) connection of FSRCNN and ResNet-18 as well as the proposed SR^2_{FSRCNN}. The gap between the proposed approach and the competing methods increases with increasing noise level. Also, for magnification factor 4, the performance of the proposed approach is substantially better than the sequential connection.

3.2. Influence of the Classification-driven Regularization

Interestingly, the classification branch actively supports the super-resolution branch. This can be observed in the classification-driven suppression of noisy pixels by constraining the shape of the digits. Thus, the super-resolution branch benefits from slight denoising.

This influence of the classification on the super-resolution is controlled by λ. Increasing λ leads to better structure preservation in the super-resolution.

This is shown in Fig. 7 for magnification factor 4: (a) shows the ground truth, and (b)-(c) show baselines for bicubic upsampling and

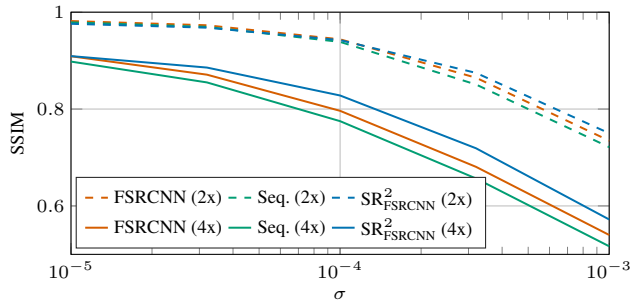


Fig. 6: The average SSIM value of FSRCNN compared to the sequential (seq.) connection of FSRCNN and ResNet-18 and SR^2_{FSRCNN} on the MNIST test data. Gaussian noise with standard deviation σ is added to the LR images.

(a) Ground truth (b) Bicubic (c) FSRCNN (d) SR^2_{FSRCNN} ($\lambda = 0.1$) (e) SR^2_{FSRCNN} ($\lambda = 9$)

Fig. 7: Qualitative results of a MNIST test image for 4x magnification and Gaussian noise with $\sigma = 0.00032$. We compare SR^2_{FSRCNN} with increasing weight λ and FSRCNN.

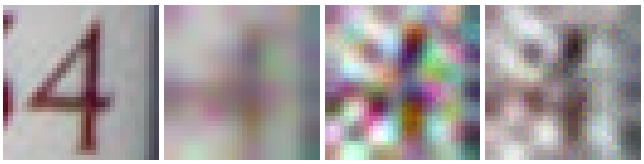
(a) Ground truth (b) Bicubic (c) WDSR (d) SR^2_{WDSR} (1 sl.) (e) SR^2_{WDSR} (4 sl.)

Fig. 8: Comparison of SR^2_{WDSR} with varying number of shared layers (sl.). Gaussian noise with $\sigma = 0.0001$ is added to the input image and super-resolution with 2x magnification is performed.

FSRCNN. Fig. 7 (d)-(e) show the results of SR^2_{FSRCNN} for increasing values of λ. It can be observed that the noise pixel in the center of the image is suppressed with increasing λ. At the same time, the noise pixel at the top arc of the digit is not removed. The network treats these two pixels differently since the first pixel might potentially lead to confusion with another digit in the classification, while the second pixel does not.

3.3. Influence of the Number of Shared Layers

To adjust the denoising performance of SR^2, the user can set the number of shared layers. With an increasing number of shared layers, the network provides higher-fidelity reconstructions for low-quality input images, but slightly blurrier reconstructions for higher-quality input images. To show this property, we use the repetitive residual building blocks in the WDSR network. In this experiment, the overall number of residual units for super-resolution is set to 16, including the shared layers. The experiments are performed on the SVHN dataset to show that the proposed approach is able to reconstruct more challenging images. In Fig. 8, qualitative results for magnification factor 2 and noise level $\sigma = 0.0001$ are shown.



(a) Ground truth (b) Bicubic (c) WDSR (d) SR^2_{WDSR}

Fig. 9: Comparison of SR^2_{WDSR} with 4 shared layers and WDSR. Gaussian noise with $\sigma = 0.001$ is applied to the SVHN test data and super-resolution with 4x magnification is performed.

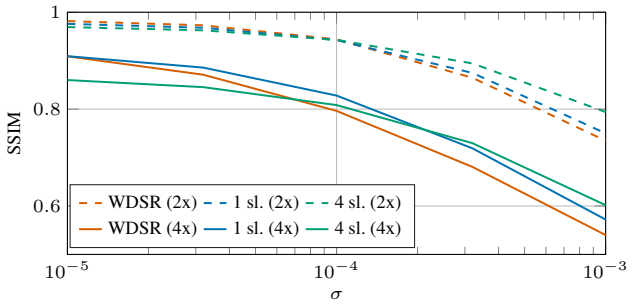


Fig. 10: Performance comparison of SR^2_{WDSR} with 1 and 4 shared layers (sl.) on the SVHN dataset. Gaussian noise with standard deviation σ is added to the LR images.

Fig. 8(a) shows the ground truth and (b)-(c) show baselines for bicubic upsampling and WDSR. The results for SR^2_{WDSR} with 1 and 4 shared layers are shown in (d)-(e). With an increasing number of shared layers, the proposed architecture is able to substantially reduce the background noise. However, a lower number of shared layers make the digit 3 slightly sharper. A more challenging case is shown in Fig. 9. The input image is severely distorted and WDSR is not able to remove the background noise. The proposed approach, however, reconstructs relevant parts of the digit 4 and reduces the background noise. Fig. 10 illustrates the SSIM values of WDSR and SR^2_{WDSR} with 1 and 4 shared layers. Using 4 shared layers, the performance of the network is not affected by lower noise levels and barely decreases. With increasing noise level, SR^2_{WDSR} with 4 shared layers outperforms WDSR and SR^2_{WDSR} with 1 shared layer.

4. CONCLUSION

This paper proposes super-resolution with structure-aware reconstruction (SR^2) via parallel branches of super-resolution and classification in a multi-task network. Both tasks share the first layers of the network and then split into branches. This enables the high-fidelity reconstruction of images distorted by unseen degradations. The object information, given by the classification, constrains the shape of the reconstructed object. Experiments on the reconstruction of digits confirm these properties. The number of shared layers is a design choice and allows to adjust the denoising performance of SR^2. A lower number produces slightly sharper reconstructions, while a higher number improves the noise reduction of background pixels. Future work aims at reconstructing other object types, such as faces or text, to further expand the proposed approach.

5. REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[2] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 1646–1654.

[3] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.

[4] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 105–114.

[5] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4570–4580.

[6] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao, "Ranksrgan: Generative adversarial networks with ranker for image super-resolution," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 286–301.

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, "Second-order attention network for single image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[9] Thomas Köhler, Michel Bätz, Farzad Naderi, André Kaup, Andreas Maier, and Christian Riess, "Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[10] F. Schirrmacher, C. Riess, and T. Köhler, "Adaptive quantile sparse image (aquasi) prior for inverse imaging problems," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 503–517, 2020.

[11] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[12] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7345–7353.

[13] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[14] S. Hao, W. Wang, Y. Ye, E. Li, and L. Bruzzone, "A deep network architecture for super-resolution-aided hyperspectral image classification with classwise loss," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4650–4663, 2018.

[15] Y. Pang, J. Cao, J. Wang, and J. Han, "Jcs-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3322–3331, 2019.

[16] Wenjia Wang, Enze Xie, Peize Sun, Wenhai Wang, Lixun Tian, Chunhua Shen, and Ping Luo, "Textsr: Content-aware text super-resolution guided by recognition," *arXiv preprint arXiv:1909.07113*, 2019.

[17] Yuchen Fan, Jiahui Yu, and Thomas S Huang, "Wide-activated deep residual networks based restoration for bpg-compressed images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2621–2624.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[19] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*, 2016, pp. 391–407.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, 2016, pp. 630–645.

[21] Yann LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[23] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3262–3271.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[25] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.