Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 259–264
DOI: 10.15439/2024F7855
ISSN 2300-5963 ACSIS, Vol. 39

Task-driven single-image super-resolution reconstruction of document scans

Maciej Zyrek, Michal Kawulok
0009-0009-4709-2743, 0000-0002-3669-5110
Department of Algorithmics and Software, Silesian University of Technology
Akademicka 16, 44-100 Gliwice, Poland
Email: prozyr@gmail.com; michal.kawulok@polsl.pl

***Abstract*—Super-resolution reconstruction is aimed at generating images of high spatial resolution from low-resolution observations. State-of-the-art super-resolution techniques underpinned with deep learning allow for obtaining results of outstanding visual quality, but it is seldom verified whether they constitute a valuable source for specific computer vision applications. In this paper, we investigate the possibility of employing super-resolution as a preprocessing step to improve optical character recognition from document scans. To achieve that, we propose to train deep networks for single-image super-resolution in a task-driven way to make them better adapted for the purpose of text detection. As problems limited to a specific task are heavily ill-posed, we introduce a multi-task loss function that embraces components related with text detection coupled with those guided by image similarity. The obtained results reported in this paper are encouraging and they constitute an important step towards real-world super-resolution of document images.**

I. Introduction

INSUFFICIENT image spatial resolution is often a bottleneck for computer vision systems that limits the capabilities of image analysis algorithms. In order to address that obstacle, considerable research attention has been paid to developing techniques for image enhancement [1] and super-resolution (SR) [2] aimed at reconstructing high-resolution (HR) images from low-resolution (LR) observations, being either a single image [3] or multiple images presenting the same scene [4].

Potentially, SR algorithms can be extremely valuable in the cases when acquiring an HR image is subject to a trade-off with the acquisition cost (e.g., in remote sensing [5]), speed (e.g., in document scanning [6]), or other factors [7]. However, the attempts to apply SR algorithms as a preprocessing step prior to fulfilling a proper image analysis task are still rather scarce—commonly, the techniques are trained and validated relying on HR reference images, which are downsampled and degraded to simulate the input LR images. As noted in an excellent review by Chen et al. [3], deep networks trained from the simulated data render overoptimistic results and their performance in real-world conditions is much worse, when they are applied to enhancing original, rather than downsampled images. There have been some attempts reported to address this problem relying on the use of real-world data for training [8], [9], but acquiring such data is challenging and costly, and it is not straightforward to exploit the HR references when HR and LR images are captured using different sensors [10]. Another possibility to regularize the training performed from the simulated data is to combine the low-level computer vision task of SR reconstruction with high-level ones like semantic segmentation [11], object detection [12], [13] and recognition [14], [15]. However, this research direction has not been extensively explored so far.

This work was supported by the National Science Centre, Poland, under Research Grant 2022/47/B/ST6/03009.

A. Related Work

Existing SR techniques can be roughly categorized into single-image (SISR) [3] and multi-image (MISR) [4] ones. The latter also embrace methods specialized for video [16] and burst-image SR [17]. While MISR techniques underpinned with information fusion are more successful in recovering the actual HR information, they are also definitely more challenging to apply, as multiple images of the same scene must be acquired and co-registered at subpixel precision. As these restrictions turn out to be impractical in many real-life cases, SISR techniques can be straightforwardly applied and they received much larger research attention. With the advent of deep learning, the field of SISR experienced unprecedented advancements [18] which nowadays allow for generating realistic images even at large magnification ratios of $8\times$ and more [19]. The first convolutional neural network (CNN) for SR (SRCNN) [20] already outperformed the techniques based on sparse coding, despite a relatively simple architecture, which was extended and accelerated to create a faster FSRCNN [21]. The subsequent advancements adopted the achievements in feature representation and nonlinear mapping to modeling the relation between LR and HR images [22]. The larger models included a very deep SR (VDSR) network [23], deep Laplacian pyramid network (LapSRN) with progressive upsampling [24], enhanced deep SR network (EDSR) [25], and SRResNet with residual connections [26] which was used as a generator in a generative adversarial network (GAN) setting. The latest trends in SISR are more focused on reducing the size of the deep models, while preserving the reconstruction quality [27]. Recently, it was demonstrated that SISR can benefit from vision transformers [28] which dynamically adjust the size of the feature maps, thus reducing the model complexity.

There have been also some reported attempts to employ SISR to improve text detection and optical character recog-

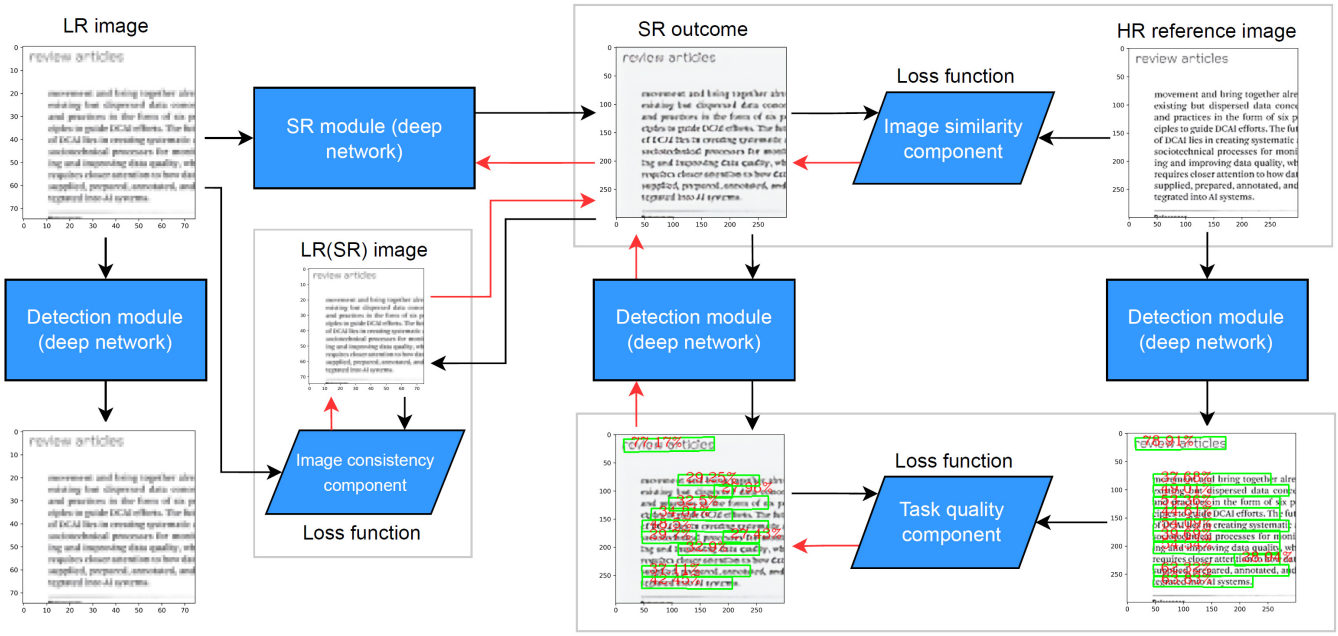

Fig. 1. Outline of the proposed self-supervised task-driven training underpinned with text detection. Red arrows indicate the propagation of the loss functions, and the black arrows show the data flow.

nition (OCR). Dong et al. adapted their SRCNN for that purpose [29] and in [30] a network with a fairly simple architecture with three convolutional layers was employed for super-resolving document scans. Wang et al. proposed to enrich a GAN-based approach with text perceptual loss to help the generator produce recognition-friendly information [31] and later they introduced a TextZoom dataset [6] composed of cropped texts from the RealSR dataset [32] with natural images captured in uncontrolled environment. In [33], a text-focused SR method was introduced which employs a vision transformer to extract sequential information. Inspired by the Gestalt psychology, stroke-based text priors were proposed in [34] and text priors were exploited for training an SR network in [35].

The aforementioned SR techniques were trained to enhance images for OCR, relying on loss functions that are correlated with that specific computer vision task. In addition to that, it is also possible to train an SR network in a task-driven manner, in which the task itself is exploited as a loss function to optimize the network's parameters. Haris et al. applied an object detection loss [12] which although leads to worse peak signal-to-noise ratio (PSNR) scores than relying on the image-similarity L1 loss, but object detection from the super-resolved images is much more effective. Similar task-driven loss functions were also defined for semantic image segmentation [36], [37]. However, in all these cases the ground-truth references related with the specific task are required for the training data.

B. Contribution

In this paper, we report our work on task-driven SISR aimed at improving text detection for an OCR system. Our contribution can be summarized as follows:

1) We propose a multi-task training underpinned with a loss function composed of image-similarity and text detection-based components.
2) The individual components of the loss function are dynamically balanced during the network training to ensure that all components are optimized at a similar pace, even though they have different magnitudes and learning speeds.
3) We propose a self-supervised approach to task-driven training, with the reference labels automatically extracted from the HR reference images.
4) We report the results of an extensive experimental study which demonstrates that the proposed technique enhances text detection accuracy from document scans super-resolved using three different SR methods.

II. Proposed Approach

The proposed task-driven training scheme is outlined in Fig. 1. An SR network is trained using three types of loss functions: (*i*) similarity with the HR reference image, (*ii*) consistency component—similarity between the downsampled SR outcome and the input LR image, and (*iii*) task quality components—the similarity in the space of deep features extracted using a network that performs text detection and recognition. For image-based loss components, we employ the L2 metric (they are termed L2-HR and L2-LR for reference-based and consistency components, respectively), while for computing the task-based loss, we rely on the L1 distance.

In our study, we employed[1] the connectionist text proposal network (CTPN) for text detection [38]. For CTPN, we exploit 512 features from the final fully-connected layer (we term them as CTPN-deep), as well as the final outputs that encode the coordinates and confidence scores (20 features each, termed CTPN-out). During training, we compute the distances for these three feature spaces and we treat them as different tasks in our setup. For training SR networks, we used a CTPN model that has been already trained—its parameters are frozen during task-driven training and the gradient is propagated to optimize the SR network. Importantly, we establish the target text positions based on the outcome of text detection in the HR reference images. In this way, we do not need the text positions to be annotated, making the training self-supervised.

Our initial attempts to exploit a loss function constructed from multiple components revealed that it is quite challenging to ensure the stability between them during training. Even if we weigh these components to provide a proper initial balance, the training is becoming focused on those that are easier to be optimized and the problem turns into an imbalanced one over time. In order to address that issue, we employed a dynamic weight averaging (DWA) algorithm that adjusts the weights assigned to the particular tasks based on their individual improvements observed in subsequent training steps [39]. In DWA, for N tasks, the weight assigned to an x-th task at t-th step is determined as:

$$w_x(t) = N \exp \frac{r_x(t-1)}{T} \Bigg/ \sum_{i=1}^{N} \exp \frac{r_i(t-1)}{T} , \quad (1)$$

where

$$r_i(t) = L_i(t)/L_i(t-1). \quad (2)$$

L_i is the value of the i-th loss component and T is the temperature controlling the softness of the task weighting (here, $T = 1$). In this way, the larger weights are assigned to these tasks in the t-th step whose losses decreases less in the preceding $(t-1)$-th step. This makes the training more focused on these tasks that are more difficult to optimize and it prevents a single component from dominating the training process or being neglected.

III. Experiments

In our experiments, we exploited three types of datasets: (*i*) natural MS COCO images [40] for training baseline SR models, (*ii*) scans from the benchmark datasets: Old Books[2] and LRDE Document Binarization Dataset (LRDE-DBD)[3] [41], and (*iii*) our *scanned documents* dataset with real-world scans performed using a Canon LiDE 400 scanner. In our study, we investigated the SRCNN [20], FSRCNN [21] and SRResNet [26] techniques for SR at $4\times$ magnification factor. We selected these networks, as they are easy to train, while having a different level of architecture complexity. For training these methods using the regular image-based loss function (L2-HR), we exploited the MS COCO images (LR images were obtained by downsampling the HR images) and for task-based training, we exploited a training set extracted from the Old Books and LRDE-DBD datasets (70% images). The test sets were formed from the remaining 30% of Old Books and LRDE-DBD datasets, as well as from all the scanned documents (we used five different scans split into 864 patches with 512×512 pixels). The CTPN model was trained beforehand from the ICDAR2017 dataset [42] and its parameters were frozen during the task-driven trainings.

The reconstruction quality was measured relying on image similarity metrics, namely PSNR, structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) [43], computed between the super-resolved image and the HR reference (thus, reflecting the L2-HR loss function). For assessing the text detection quality, we employed intersection over union (IoU) between the text positions detected in the super-resolved image and in the corresponding HR reference. We also report the distances in the CTPN-deep and CTPN-out feature spaces that are used for computing the task-based components of the loss function.

First, we trained each network from scratch (60 epochs), guiding the training using a standard baseline configuration (with the L2-HR loss) and using all loss components, including L2-HR, the consistency (L2-LR) and task-based CTPN-deep and CTPN-out components. For FSRCNN and SRResNet, we fine-tuned the baseline models (100 epochs) relying on (*i*) L2-HR loss combined with the task-based loss components, (*ii*) the task component coupled with the consistency loss, and (*iii*) using all loss components. In addition to that, we trained SRResNet (as the best performing model) from scratch relying only on the task-based components (hence without using the image similarity at all). In Table I, we report the scores obtained for two test sets (unseen during training): for the test set of the benchmark datasets and for our dataset with the scanned documents. It can be observed that incorporating the task-based components improves the scores in terms of the image-based metrics in most cases and it always improves the quality of the text detection task (the differences are definitely higher for our scanned documents). It is also clear that the models cannot be trained from scratch without using the image-based components—apparently the problem is not convex enough and the training gets stuck in a local minimum.

A sample of the qualitative results is presented in Fig. 2 for a benchmark image and one of our scans (the configurations presented in the figure are referenced from Table I). While for the benchmark image (two upper rows), the text quality is consistently good across all configurations, for our scan, it is definitely better for the model fine-tuned in a task-driven way (d), and it is actually quite close to the result obtained in the HR reference. It can also be seen that the texts are quite clear when SRResNet is trained without using the image-based loss components (which also leads to good detection outcome), but the stability in the color space is not preserved, leading to extremely poor quantitative scores reported earlier in Table I.

[1]For CTPN, we use implementation available at https://github.com/courao/ocr.pytorch

[2]Available at https://github.com/PedroBarcha/old-books-dataset

[3]Available at https://www.lrde.epita.fr/wiki/Olena/DatasetDBD

TABLE I

QUANTITATIVE SCORES OBTAINED FOR THE IMAGES FROM THE OLD BOOKS AND LRDE-DBD BENCHMARKS AND FROM OUR DATASET WITH DOCUMENT SCANS, OBTAINED USING DIFFERENT SR TECHNIQUES TRAINED WITH A VARIETY OF LOSS FUNCTIONS. FOR EACH METRIC AND CATEGORY, THE BEST RESULT IS BOLDFACED.

Model and training type (a reference in Fig. 2)		Loss function: L2-HR	L2-LR	CTPN-deep	CTPN-out	Image similarity metrics: PSNR↑ (dB)	SSIM↑	LPIPS↓	Text detection metrics: IoU ↑	CTPN-deep↓ ($\cdot 10^{-2}$)	CTPN-out↓ ($\cdot 10^{-2}$)
Test set from the simulated benchmark images (Old Books and LRDE-DBD):											
SRCNN (from scratch)	(a)	✓				**21.16**	0.8481	**0.1818**	0.8923	—	—
SRCNN (from scratch)		✓	✓	✓	✓	21.08	**0.8489**	0.1897	**0.9290**	**1.831**	**3.366**
FSRCNN (from scratch)	(b)	✓				24.17	**0.9134**	**0.1790**	0.9332	—	—
–fine-tuned		✓		✓	✓	**25.00**	0.9071	0.2982	**0.9604**	1.005	1.750
–fine-tuned			✓	✓	✓	20.06	0.6394	0.4681	0.9559	**0.993**	**1.742**
–fine-tuned		✓	✓	✓	✓	24.59	0.8848	0.3471	0.9560	1.113	1.939
FSRCNN (from scratch)		✓	✓	✓	✓	24.54	0.8880	0.3245	0.9588	1.097	1.919
SRResNet (from scratch)	(c)	✓				24.10	0.9147	0.1553	0.9392	—	—
–fine-tuned		✓		✓	✓	28.16	0.9537	0.1037	0.9614	**0.676**	**1.177**
–fine-tuned			✓	✓	✓	24.67	0.8404	0.3048	0.9676	0.694	1.198
–fine-tuned	(d)	✓	✓	✓	✓	**28.04**	**0.9578**	**0.0993**	**0.9761**	0.714	1.235
SRResNet (from scratch)		✓	✓	✓	✓	25.49	0.9302	0.1614	0.9590	1.036	1.802
SRResNet (from scratch)	(e)			✓	✓	2.97	−0.1832	0.7714	0.9197	2.564	4.737
Scanned documents:											
SRCNN (from scratch)	(a)	✓				16.83	0.5709	**0.4301**	0.7103	—	—
SRCNN (from scratch)		✓	✓	✓	✓	**17.06**	**0.5782**	0.4344	**0.7275**	**2.827**	**5.342**
FSRCNN (from scratch)	(b)	✓				18.68	**0.6542**	0.3681	0.7341	—	—
–fine-tuned		✓		✓	✓	**18.84**	0.6467	**0.3585**	0.7608	2.363	4.290
–fine-tuned			✓	✓	✓	16.39	0.4796	0.4343	**0.7688**	**2.294**	**4.174**
–fine-tuned		✓	✓	✓	✓	18.82	0.6429	0.3588	0.7635	2.328	4.219
FSRCNN (from scratch)		✓	✓	✓	✓	18.82	0.6444	0.3644	0.7641	2.414	4.348
SRResNet (from scratch)	(c)	✓				18.70	0.6634	0.3798	0.7264	—	—
–fine-tuned		✓		✓	✓	19.62	0.7075	0.3189	**0.7910**	**1.985**	3.397
–fine-tuned			✓	✓	✓	19.00	0.6327	0.3261	0.7886	1.994	3.520
–fine-tuned	(d)	✓	✓	✓	✓	**19.81**	**0.7076**	**0.3164**	0.7807	2.023	**3.483**
SRResNet (from scratch)		✓	✓	✓	✓	19.23	0.6731	0.3591	0.7576	2.361	4.280
SRResNet (from scratch)	(e)			✓	✓	2.91	−0.0929	0.9250	0.7186	3.170	5.592

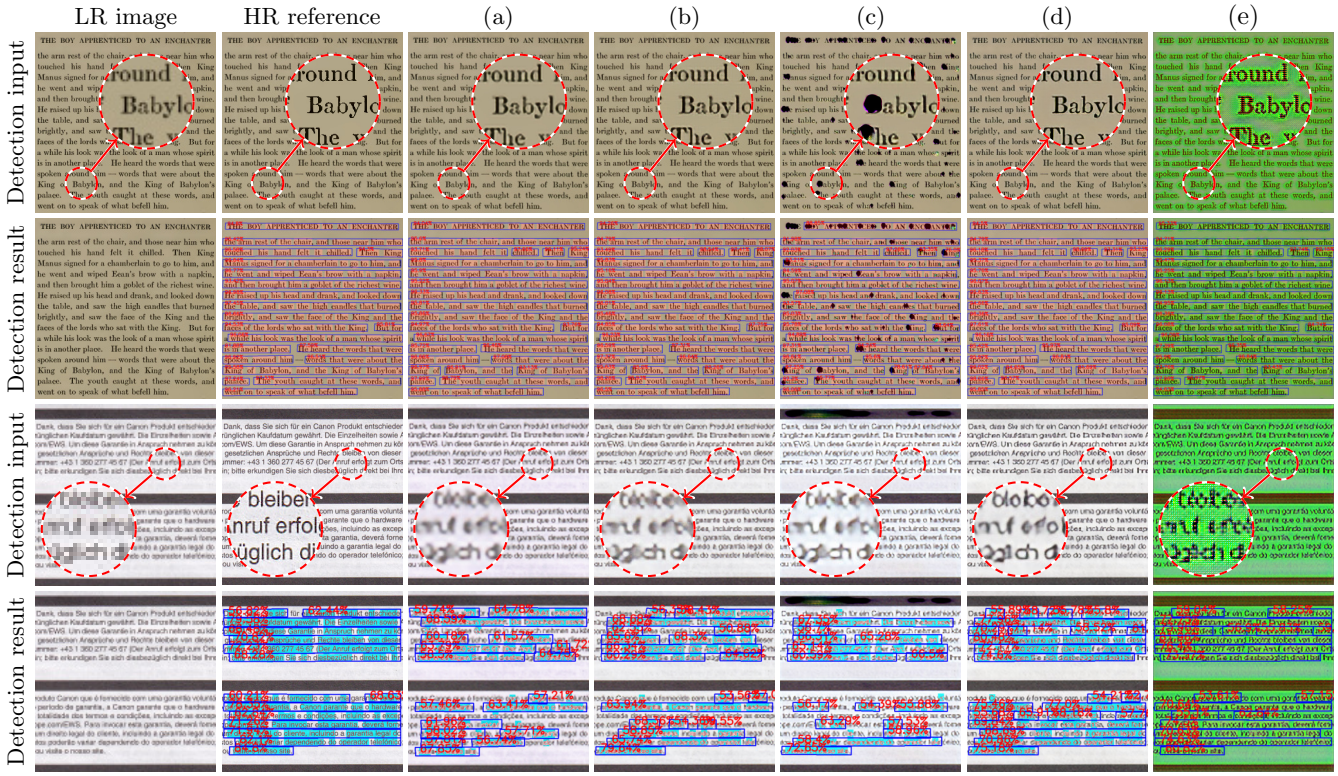

Fig. 2. Examples of SR reconstruction with: (a) SRCNN, (b) FSRCNN and (c) SRResNet (all with L2-HR loss), (d) fine-tuned SRResNet (L2-HR, L2-LR, CTPN-deep and CTPN-out loss functions), and (e) SRResNet trained from scratch (CTPN-deep and CTPN-out loss functions). These settings are also referred to in Table I. For each example (top: Old Books; bottom: our dataset), we include the detection input (i.e., the SR outcome) and the detected text.

IV. Conclusions and outlook

In this paper, we reported our initial attempts to apply task-driven training for SISR, guided by text detection. The results are highly encouraging, revealing high potential of task-based loss functions. Importantly, in contrast to the earlier works concerned with task-driven SR, we train the models in a self-supervised way, as we retrieve the annotations by processing the HR reference images.

Our ongoing research is focused on including the text recognition components that may improve the guidance during training. Also, we plan to adapt our approach to MISR problems and to create a dataset embracing samples composed of multiple scans of the same document.

References

[1] F. Jelowicki, "Enhancing image quality through automated projector stacking," in *Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, FedCSIS 2023, Warsaw, Poland, September 17-20, 2023*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, M. Paprzycki, and D. Slezak, Eds., vol. 37, 2023, pp. 153–156. [Online]. Available: https://doi.org/10.15439/2023F9900

[2] W. Yang, X. Zhang, Y. Tian, W. Wang, J. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transaction on Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec 2019. [Online]. Available: https://doi.org/10.1109/TMM.2019.2919431

[3] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, "Real-world single image super-resolution: A brief review," *Information Fusion*, vol. 79, pp. 124–145, 2022. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.09.005

[4] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Processing*, vol. 128, pp. 389–408, 2016. [Online]. Available: https://doi.org/10.1016/j.sigpro.2016.05.002

[5] T. Tarasiewicz, J. Nalepa, R. A. Farrugia, G. Valentino, M. Chen, J. A. Briffa, and M. Kawulok, "Multitemporal and multispectral data fusion for super-resolution of Sentinel-2 images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–19, 2023. [Online]. Available: https://doi.org/10.1109/TGRS.2023.3311622

[6] W. Wang, E. Xie, X. Liu, W. Wang, D. Liang, C. Shen, and X. Bai, "Scene text image super-resolution in the wild," in *Proc. IEEE/CVF ECCV*. Springer, 2020, pp. 650–666. [Online]. Available: https://doi.org/10.1007/978-3-030-58607-2_38

[7] T. Balon, M. Knapik, and B. Cyganek, "Real-time detection of small objects in automotive thermal images with modern deep neural architectures," in *Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, FedCSIS 2023, Warsaw, Poland, September 17-20, 2023*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, M. Paprzycki, and D. Slezak, Eds., vol. 37, 2023, pp. 29–35. [Online]. Available: https://doi.org/10.15439/2023F8409

[8] J. Cai, S. Gu, R. Timofte, and L. Zhang, "NTIRE 2019 Challenge on real image super-resolution: Methods and results," in *Proc. IEEE/CVF CVPR*, 2019, pp. 1–13. [Online]. Available: https://doi.org/10.1109/CVPRW.2019.00274

[9] M. Märtens, D. Izzo, A. Krzic, and D. Cox, "Super-resolution of PROBA-V images using convolutional neural networks," *Astrodynamics*, vol. 3, no. 4, pp. 387–402, 2019. [Online]. Available: https://doi.org/10.1007/s42064-019-0059-8

[10] P. Kowaleczko, T. Tarasiewicz, M. Ziaja, D. Kostrzewa, J. Nalepa, P. Rokita, and M. Kawulok, "A real-world benchmark for Sentinel-2 multi-image super-resolution," *Scientific Data*, vol. 10, no. 1, p. 644, 2023. [Online]. Available: https://doi.org/10.1038/s41597-023-02538-9

[11] Z. Guo, G. Wu, X. Song, W. Yuan, Q. Chen, H. Zhang, X. Shi, M. Xu, Y. Xu, R. Shibasaki *et al.*, "Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery," *IEEE Access*, vol. 7, pp. 99 381–99 397, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2928646

[12] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," in *Proc. ICONIP*. Springer, 2021, pp. 387–395. [Online]. Available: https://doi.org/10.1007/978-3-030-92307-5_45

[13] T. Balon, M. Knapik, and B. Cyganek, "New thermal automotive dataset for object detection," in *Position Papers of the 17th Conference on Computer Science and Intelligence Systems, FedCSIS 2022, Sofia, Bulgaria, September 4-7, 2022*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, M. Paprzycki, and D. Slezak, Eds., vol. 31, 2022, pp. 43–48. [Online]. Available: https://doi.org/10.15439/2022F283

[14] X. Yang, W. Wu, K. Liu, P. W. Kim, A. K. Sangaiah, and G. Jeon, "Long-distance object recognition with image super resolution: A comparative study," *IEEE Access*, vol. 6, pp. 13 429–13 438, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2799861

[15] M. Wlodarczyk-Sielicka and D. Polap, "Interpolation merge as augmentation technique in the problem of ship classification," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020, Sofia, Bulgaria, September 6-9, 2020*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 21, 2020, pp. 443–446. [Online]. Available: https://doi.org/10.15439/2020F11

[16] H. Liu, Z. Ruan, P. Zhao, C. Dong, F. Shang, Y. Liu, L. Yang, and R. Timofte, "Video super-resolution based on deep learning: a comprehensive survey," *Artificial Intelligence Review*, pp. 1–55, 2022. [Online]. Available: https://doi.org/10.1007/s10462-022-10147-y

[17] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Deep burst super-resolution," in *Proc. IEEE/CVF CVPR*, 2021, pp. 9209–9218. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.00909

[18] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2021. [Online]. Available: https://doi.org/10.1109/TPAMI.2020.2982166

[19] R. Abiantun, F. Juefei-Xu, U. Prabhu, and M. Savvides, "SSR2: Sparse signal recovery for single-image super-resolution on faces with extreme low resolutions," *Pattern Recognition*, vol. 90, pp. 308–324, 2019. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.01.032

[20] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. IEEE/CVF ECCV*. Springer, 2014, pp. 184–199. [Online]. Available: https://doi.org/10.1007/978-3-319-10593-2_13

[21] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. IEEE/CVF ECCV*. Springer, 2016, pp. 391–407. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_25

[22] Y. Huang, J. Li, X. Gao, Y. Hu, and W. Lu, "Interpretable detail-fidelity attention network for single image super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 2325–2339, 2021. [Online]. Available: https://doi.org/10.1109/TIP.2021.3050856

[23] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE/CVF CVPR*, 2016, pp. 1646–1654. [Online]. Available: https://doi.org/10.1109/CVPR.2016.182

[24] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2599–2613, 2019. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2865304

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE/CVF CVPR Workshops*, 2017, pp. 136–144. [Online]. Available: https://doi.org/10.1109/CVPRW.2017.151

[26] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE/CVF CVPR*, 2017, pp. 4681–4690. [Online]. Available: https://doi.org/10.1109/CVPR.2017.19

[27] M. Ayazoglu, "Extremely lightweight quantization robust real-time single-image super resolution for mobile devices," in *Proc. IEEE/CVF CVPR*, 2021, pp. 2472–2479. [Online]. Available: https://doi.org/10.1109/CVPRW53098.2021.00280

[28] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proc. IEEE/CVF CVPR*, 2022, pp. 457–466. [Online]. Available: https://doi.org/10.1109/CVPRW56347.2022.00061

[29] C. Dong, X. Zhu, Y. Deng, C. C. Loy, and Y. Qiao, "Boosting optical character recognition: A super-resolution approach," *arXiv preprint arXiv:1506.02211*, 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1506.02211

[30] R. K. Pandey and A. Ramakrishnan, "Efficient document-image super-resolution using convolutional neural network," *Sādhanā*, vol. 43, pp. 1–6, 2018. [Online]. Available: https://doi.org/10.1007/s12046-018-0794-1

[31] W. Wang, E. Xie, P. Sun, W. Wang, L. Tian, C. Shen, and P. Luo, "TextSR: Content-aware text super-resolution guided by recognition," *arXiv preprint arXiv:1909.07113*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1909.07113

[32] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proc. IEEE ICCV*, 2019. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00318

[33] J. Chen, B. Li, and X. Xue, "Scene text telescope: Text-focused scene image super-resolution," in *Proc. IEEE/CVF CVPR*, 2021, pp. 12 026–12 035. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.01185

[34] J. Chen, H. Yu, J. Ma, B. Li, and X. Xue, "Text Gestalt: Stroke-aware scene text image super-resolution," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 285–293. [Online]. Available: https://doi.org/10.1609/aaai.v36i1.19904

[35] J. Ma, S. Guo, and L. Zhang, "Text prior guided scene text image super-resolution," *IEEE Transactions on Image Processing*, vol. 32, pp. 1341–1353, 2023. [Online]. Available: https://doi.org/10.1109/TIP.2023.3237002

[36] T. Frizza, D. G. Dansereau, N. M. Seresht, and M. Bewley, "Semantically accurate super-resolution generative adversarial networks," *Computer Vision and Image Understanding*, p. 103464, 2022. [Online]. Available: https://doi.org/10.1016/j.cviu.2022.103464

[37] M. S. Rad, B. Bozorgtabar, C. Musat, U.-V. Marti, M. Basler, H. K. Ekenel, and J.-P. Thiran, "Benefiting from multitask learning to improve single image super-resolution," *Neurocomputing*, vol. 398, pp. 304–313, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.07.107

[38] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. IEEE/CVF ECCV*. Springer, 2016, pp. 56–72. [Online]. Available: https://doi.org/10.1007/978-3-319-46484-8_4

[39] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3614–3633, 2021. [Online]. Available: https://doi.org/10.1109/TPAMI.2021.3054719

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. IEEE/CVF ECCV*. Springer, 2014, pp. 740–755. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48

[41] G. Lazzara and T. Géraud, "Efficient multiscale sauvola's binarization," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 2, pp. 105–123, 2014. [Online]. Available: https://doi.org/10.1007/s10032-013-0209-0

[42] R. Gomez, B. Shi, L. Gomez, L. Numann, A. Veit, J. Matas, S. Belongie, and D. Karatzas, "ICDAR2017 robust reading challenge on COCO-text," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 1435–1443. [Online]. Available: https://doi.org/10.1109/ICDAR.2017.234

[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF CVPR*, 2018. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00068