PEAN: A Diffusion-Based Prior-Enhanced Attention Network for Scene Text Image Super-Resolution

Zuoyan Zhao

Southeast University

ACM Multimedia 2024
Melbourne, Australia

Outline

ACM Multimedia 2024
Melbourne, Australia

- Introduction
- Prior-Enhanced Attention Network (PEAN)
  - The framework of PEAN
  - Text Prior Enhancement Module
  - Attention-Based Modulation Module
  - Multi-Task Learning
- Experiments
  - Comparing with State-of-the-Art Methods
  - Ablation Study
- Conclusion

Outline

- Introduction
- Prior-Enhanced Attention Network (PEAN)
  - The framework of PEAN
  - Text Prior Enhancement Module
  - Attention-Based Modulation Module
  - Multi-Task Learning
- Experiments
  - Comparing with State-of-the-Art Methods
  - Ablation Study
- Conclusion

Introduction

- To better read text from LR images, researchers formulate the STISR task to reconstruct missing text details in LR images, as a pre-processing step for the scene text recognition task.
- For scene text images, two crucial factors determine whether they could be correctly recognized.
  - **Visual structure:** the restoration of images containing long or deformed text string
  - **Semantic information:** primary text prior prevents the SR network from generating images that contain correct semantic information
- We propose a Prior-Enhanced Attention Network (PEAN) to tackle issues caused by the two factors.

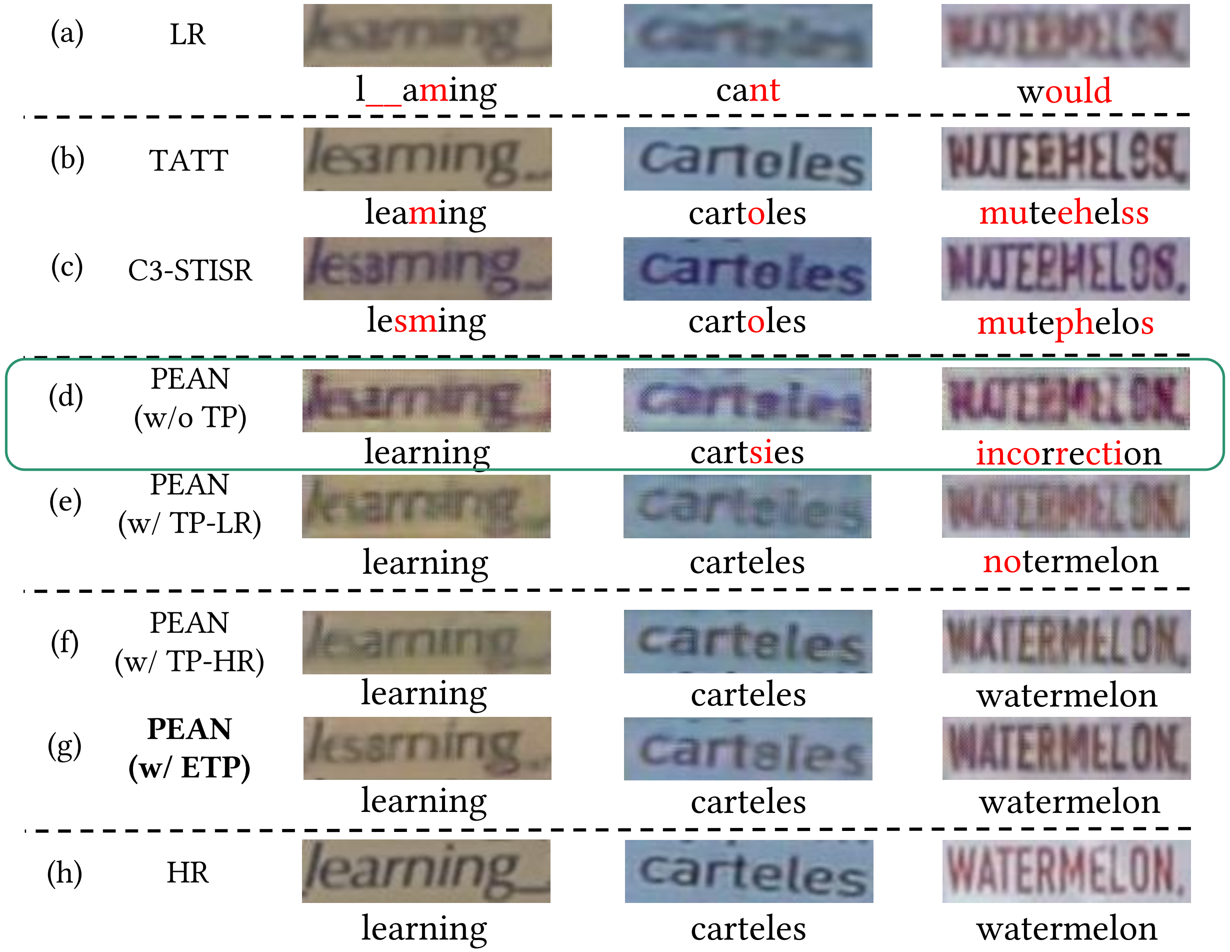
Introduction

ACM Multimedia 2024 Melbourne, Australia

- An Attention-based Modulation Module (AMM) is proposed to substitute the SRB, endowing the network with a larger receptive feld to images, thereby restoring the visual structure of images with text in various shapes and lengths.

- However, the lack of semantic information limits the capability of such model.

- Text prior derived from high-resolution (HR) images is a robust choice for STISR, in view of the high recognition accuracy of HR images.

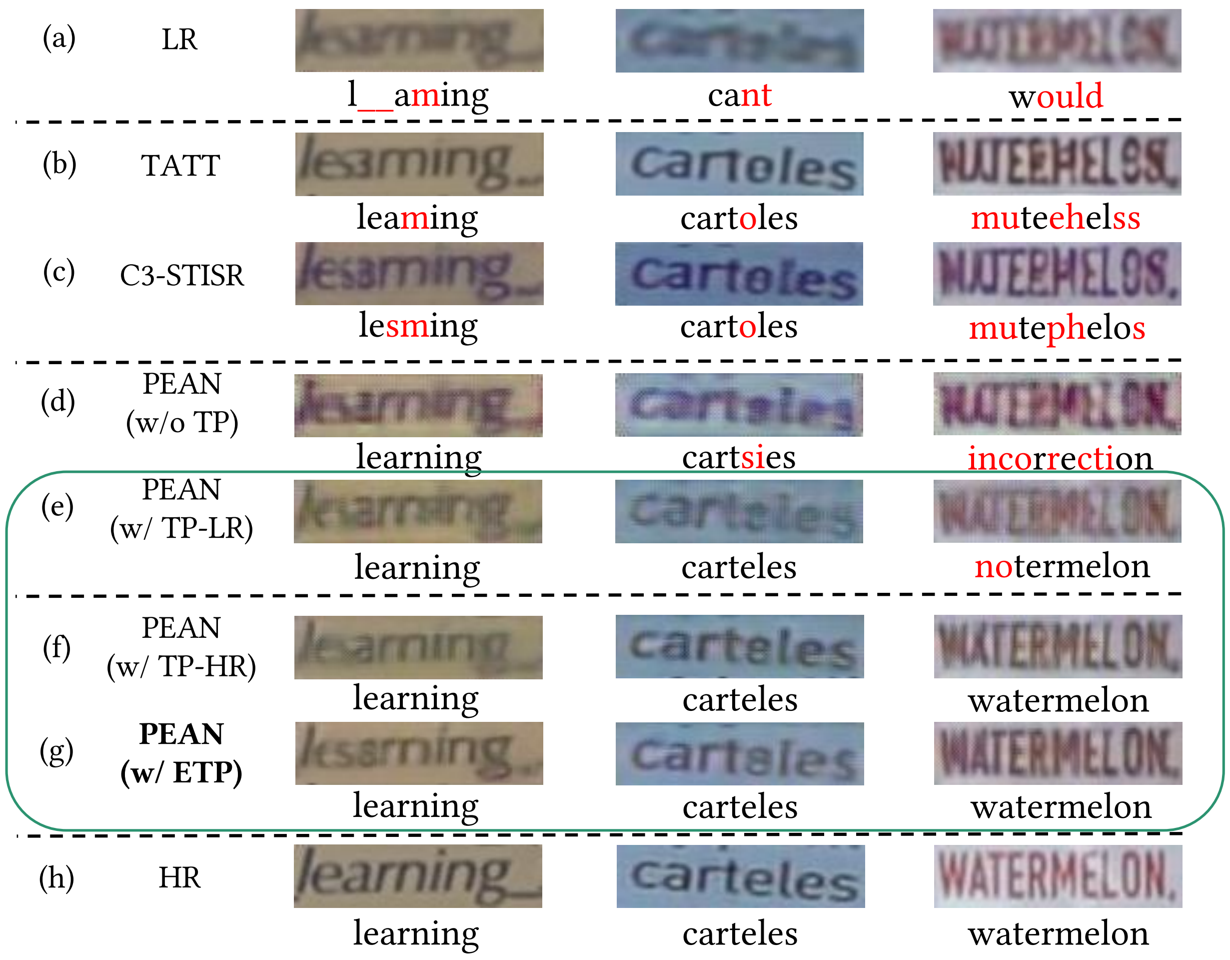


Introduction

- We conduct an exploratory experiment wherein we substitute the text prior from LR images (TP-LR) with the text prior from HR images (TP-HR) within such model, yielding superior outcomes.

TP-LR	TPEM	TP-HR	Easy	Medium	Hard	Average
			75.7	60.2	42.1	60.4
✓			79.7	62.3	46.1	63.8
✓	✓		**84.5**	**71.4**	**52.9**	**70.6**
		✓	88.4	75.5	61.3	75.9

- This inspires the design of a module for enhancing the primary text prior, resulting in the creation of the Enhanced Text Prior (ETP), which is comparable in effectiveness to TP-HR.



Introduction

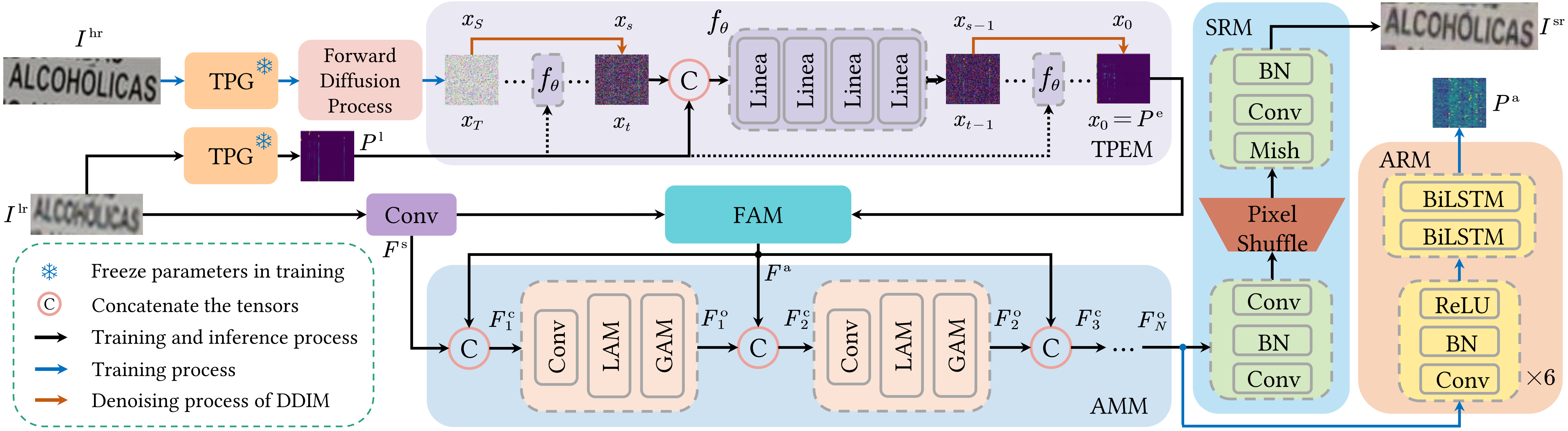
- The ETP provides valuable guidance to the SR network, promoting the generation of SR images with high semantic accuracy.
- Given the remarkable performance of diffusion models, we propose a diffusion-based Text Prior Enhancement Module (TPEM) to obtain the ETP owing to their ability to map complex distributions.
- We adopt the Multi-Task Learning (MTL) paradigm in the training phase.
  - **Image restoration task**: focuses on generating high-quality SR images.
  - **Text recognition task**: stimulates the model to generate more readable SR results.

Outline

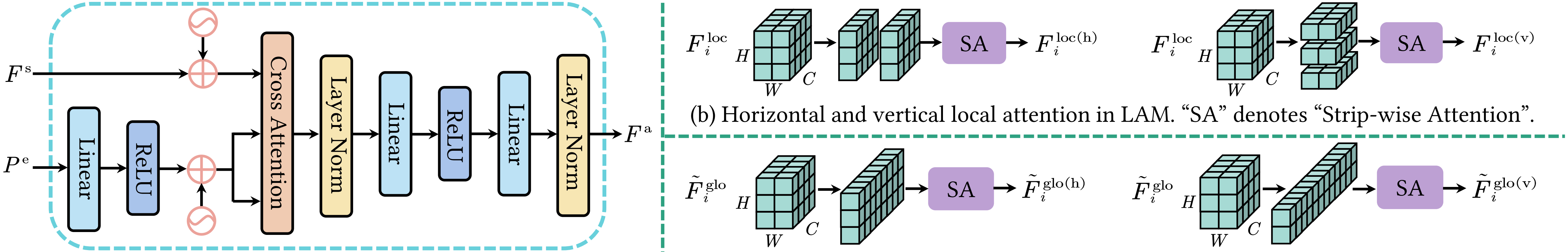
ACM Multimedia 2024
Melbourne, Australia

- Introduction
- **Prior-Enhanced Attention Network (PEAN)**
  - The framework of PEAN
  - Text Prior Enhancement Module
  - Attention-Based Modulation Module
  - Multi-Task Learning
- Experiments
  - Comparing with State-of-the-Art Methods
  - Ablation Study
- Conclusion

Prior-Enhanced Attention Network



Prior-Enhanced Attention Network



(a) Overview of the Feature Alignment Module (FAM).

(b) Horizontal and vertical local attention in LAM. "SA" denotes "Strip-wise Attention".

(c) Horizontal and vertical global attention in GAM. "SA" denotes "Strip-wise Attention".

Outline

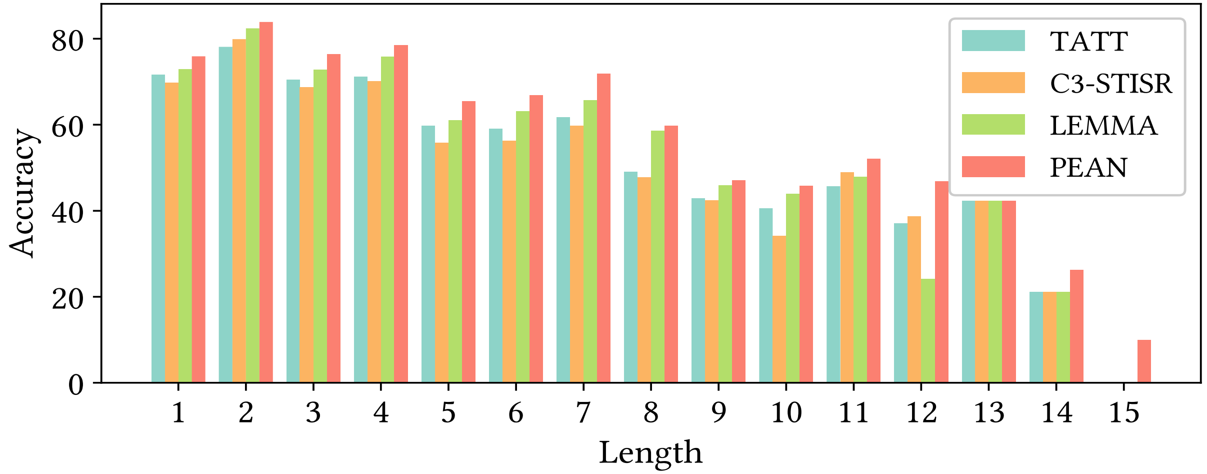
- Introduction
- Prior-Enhanced Attention Network (PEAN)
  - The framework of PEAN
  - Text Prior Enhancement Module
  - Attention-Based Modulation Module
  - Multi-Task Learning
- **Experiments**
  - Comparing with State-of-the-Art Methods
  - Ablation Study
- Conclusion

Experiments

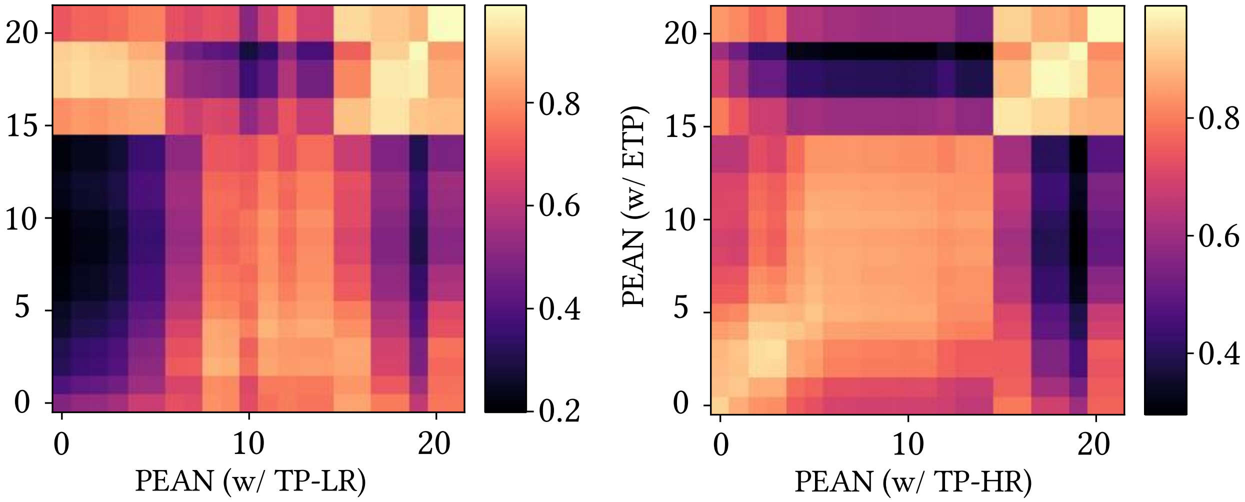
Methods	Accuracy of ASTER [53] (%)				Accuracy of MORAN [33] (%)				Accuracy of CRNN [52] (%)			
	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average
LR	62.4	42.7	31.6	46.6	59.4	36.0	28.2	42.3	37.5	21.2	21.4	27.3
SRCNN [10]	69.4	43.4	32.2	49.5	63.2	39.0	30.2	45.3	38.7	21.6	20.9	27.7
SRResNet [27]	69.6	47.6	34.3	51.3	60.7	42.9	32.6	46.3	39.7	27.6	22.7	30.6
RDN [73]	70.0	47.0	34.0	51.5	61.7	42.0	31.6	46.1	41.6	24.4	23.5	30.5
RRDB [63]	70.9	44.4	32.5	50.6	63.9	41.0	30.8	46.3	40.6	22.1	21.9	28.9
LapSRN [26]	71.5	48.6	35.2	53.0	64.6	44.9	32.2	48.3	46.1	27.9	23.6	33.3
ESRT [10]	69.8	49.1	35.2	52.5	61.9	41.7	32.2	46.3	48.2	27.9	25.8	34.8
Omni-SR [60]	71.2	52.3	38.1	54.9	66.7	47.9	36.5	51.4	54.8	37.4	29.4	41.4
SRFormer [78]	69.0	45.1	32.8	50.2	61.3	39.6	29.9	44.7	41.0	22.8	22.9	29.6
TSRN [62]	75.1	56.3	40.1	58.3	70.1	53.3	37.9	54.8	52.5	38.2	31.4	41.4
TBSRN [5]	75.7	59.9	41.6	60.1	74.1	57.0	40.8	58.4	59.6	47.1	35.3	48.1
PCAN [74]	77.5	60.7	43.1	61.5	73.7	57.6	41.0	58.5	59.6	45.4	34.8	47.4
TG [6]	77.9	60.2	42.4	61.3	75.8	57.8	41.4	59.4	61.2	47.6	35.5	48.9
SGENet [57]	75.8	60.7	45.0	61.4	71.5	56.2	41.4	57.3	59.4	47.9	37.7	49.0
TPGSR [34]	78.9	62.7	44.5	62.8	74.9	60.5	44.1	60.5	63.1	52.0	38.6	51.8
TATT [35]	78.9	63.4	45.4	63.6	72.5	60.2	43.1	59.5	62.6	53.4	39.8	52.6
C3-STISR [75]	79.1	63.3	46.8	64.1	74.2	61.0	43.2	60.5	65.2	53.6	39.8	53.7
TATT + DPMN [81]	79.3	64.1	45.2	63.9	73.3	61.5	43.9	60.4	64.4	54.2	39.2	53.4
TSAN [82]	79.6	64.1	45.3	64.1	78.4	61.3	45.1	62.7	64.6	53.3	38.8	53.0
TEAN [55]	80.4	64.5	45.6	64.6	76.8	60.8	43.4	61.4	63.7	52.5	38.1	52.2
MSPIE [83]	80.4	63.4	46.3	64.4	74.0	61.4	44.4	60.8	64.5	54.2	39.6	53.5
TCDM [39]	81.3	65.1	50.1	65.5	77.6	62.9	45.9	62.2	67.3	57.3	42.7	55.7
LEMMA [19]	81.1	66.3	47.4	66.0	77.7	64.4	44.6	63.2	67.1	58.8	40.6	56.3
RTSRN [70]	80.4	66.1	49.1	66.2	77.1	63.3	46.5	63.2	67.0	59.2	42.6	57.0
RGDiffSR [77]	81.1	65.4	49.1	66.2	78.6	62.1	45.4	63.1	67.6	56.5	42.7	56.4
TextDiff [29]	80.8	66.5	48.7	66.4	77.7	62.5	44.6	62.7	64.8	55.4	39.9	54.2
PEAN	**84.5**	**71.4**	**52.9**	**70.6**	**79.4**	**67.0**	**49.1**	**66.1**	**68.9**	**60.2**	**45.9**	**59.0**
HR	94.2	87.7	76.2	86.6	91.2	85.3	74.2	84.1	76.4	75.1	64.6	72.4

Experiments

Experiments



Methods	Easy	Medium	Hard	Average
SRB [62]	80.1	64.4	46.4	64.7
ViT [12]	81.8	65.7	49.5	66.7
Swin [30]	73.8	55.1	39.0	57.1
CSWin [11]	70.2	52.9	37.2	54.5
Stripformer [58]	72.9	53.6	37.3	55.7
AMM	**84.5**	**71.4**	**52.9**	**70.6**



Loss Functions	Easy	Medium	Hard	Average
$\mathcal{L}_{mse}$	76.2	58.8	41.5	59.9
$+ \mathcal{L}_{sfm}$	79.2	64.3	47.0	64.5
$+ \mathcal{L}_{mae}$	79.6	65.1	47.1	64.9
$+ \mathcal{L}_{ctc}^{t}$	81.4	68.8	50.7	67.9
$+ \mathcal{L}_{ctc}^{a}$	**84.5**	**71.4**	**52.9**	**70.6**

Outline

- Introduction
- Prior-Enhanced Attention Network (PEAN)
  - The framework of PEAN
  - Text Prior Enhancement Module
  - Attention-Based Modulation Module
  - Multi-Task Learning
- Experiments
  - Comparing with State-of-the-Art Methods
  - Ablation Study
- **Conclusion**

Melbourne, Australia

Conclusion

- We propose a Prior-Enhanced Attention Network (PEAN) for scene text image super-resolution (STISR).
- A Text Prior Enhancement Module (TPEM) is designed to provide the ETP for the subsequent SR process, enabling SR images to contain accurate semantic information.
- An Attention-based Modulation Module (AMM) is devised to obtain local and global coherence in scene text images, which can recover the visual structure of images with text in various sizes and deformations.
- We introduce the Multi-Task Learning (MTL) paradigm to improve the legibility of LR images.
- Experiments demonstrate that our proposed PEAN achieves SOTA performance.
- We believe our work will serve as a strong baseline for future works, and will push forward the research of STISR as well as other sub-felds of scene text images.

References

- Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *Proceedings of the European Conference on Computer Vision* (***ECCV***), 2020.
- Jianqi Ma, Zhetong Liang, and Lei Zhang. A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (***CVPR***), 2022.
- Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. C3-STISR: Scene Text Image Super-resolution with Triple Clues. In *Proceedings of the International Joint Conference on Artificial Intelligence* (***IJCAI***), 2022.
- Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. DiffIR: Efficient Diffusion Model for Image Restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (***ICCV***), 2023.
- Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip Transformer for Fast Image Deblurring. In *Proceedings of the European Conference on Computer Vision* (***ECCV***), 2022.
- Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* (***TPAMI***), 2023.

Authors & Contact Information

Zuoyan Zhao

Hui Xue

Pengfei Fang

Shipeng Zhu

- Main Paper: https://doi.org/10.1145/3664647.3680974
- Full Paper (with Supplementary Material): https://arxiv.org/abs/2311.17955
- Code: https://github.com/jdfxzzy/PEAN
- OpenReview: https://openreview.net/forum?id=IxSKhO7ed6
- E-mail: zuoyanzhao@seu.edu.cn

Thank you!

Zuoyan Zhao

Southeast University

ACM Multimedia 2024
Melbourne, Australia