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

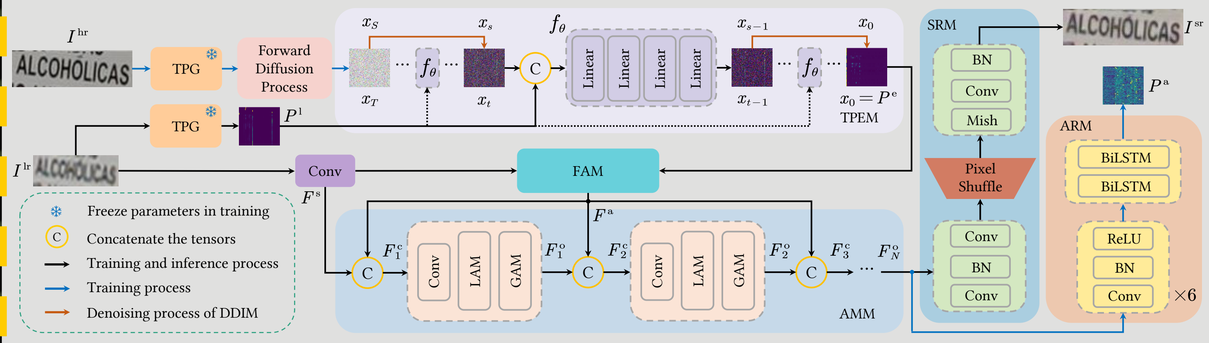
Zuoyan Zhao, Hui Xue, Pengfei Fang, Shipeng Zhu

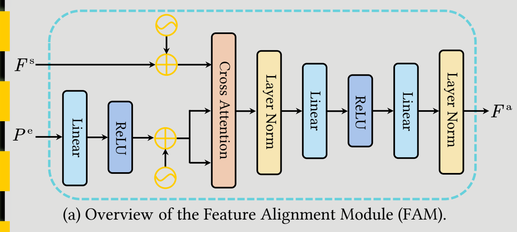
ACM Multimedia 2024
Melbourne, Australia

1 Background

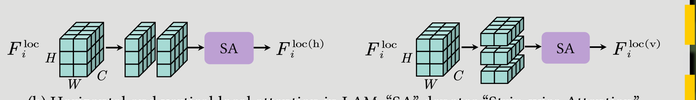
- For scene text images, two crucial factors determine whether they could be correctly recognized.
  - **Visual structure:** the restoration of images containing long or deformed text string
  - **Semantic information:** primary text prior prevents the SR network from generating images that contain correct semantic information
- We propose a Prior-Enhanced Attention Network (PEAN) to tackle issues caused by the two factors.

2 Methodology

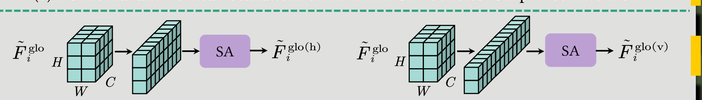




(a) Overview of the Feature Alignment Module (FAM).



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



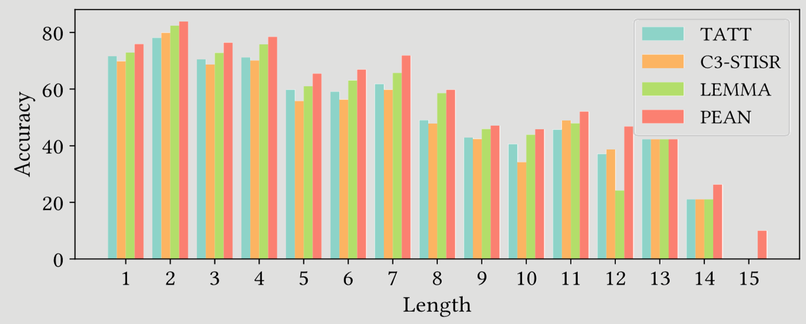
(c) Horizontal and vertical global attention in GAM. "SA" denotes "Strip-wise Attention".

- **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.
- **TPEM:** is proposed to obtain the ETP owing to their ability to map complex distributions considering the remarkable performance of diffusion models.
  - Text prior derived from HR images (TP-HR) is a robust choice for STISR, given the high recognition accuracy of HR images.
  - The ETP provides valuable guidance, similar to TP-HR, to the SR network, promoting the generation of SR images with high semantic accuracy.
- **MTL paradigm:** is adopted 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.

3 Experimental Results

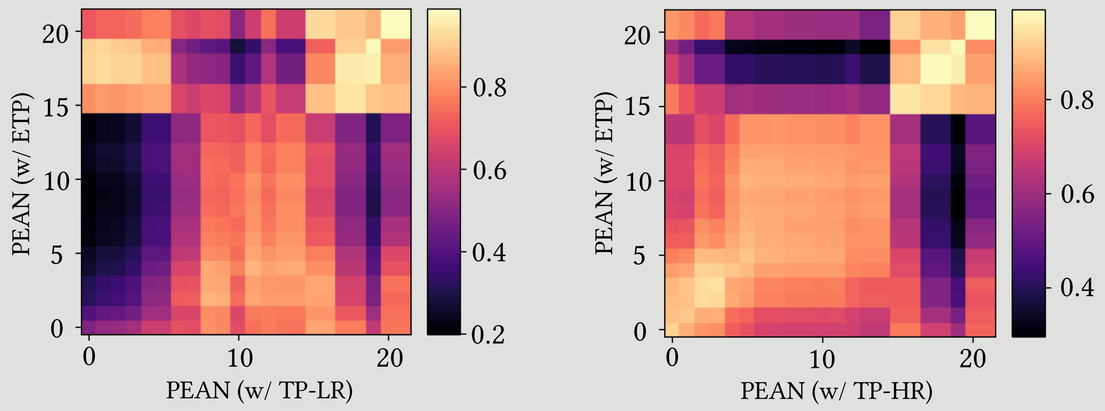
- **Datasets:** TextZoom (main), IIIT5K, SVTP, IC15, *et al.* (auxiliary).
- **Evaluation Metrics:** the recognition accuracy of ASTER, CRNN and MORAN (main), PSNR, SSIM, *et al.* (auxiliary).

Methods	Accuracy of ASTER (%)				Accuracy of MORAN (%)				Accuracy of CRNN (%)			
	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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



4 Visualization Results

5 Representation Analysis



- The power of the ETP lies in its ability to make the representations learned by AMM and SRM more similar to those learned by the corresponding modules in PEAN (w/ TP-HR), which is known for its superior performance.

Full Paper

Code and Model

OpenReview

Poster design: *Zuoyan Zhao*

E-mail: zuoyanzhao@seu.edu.cn