DCDM: Diffusion-Conditioned-Diffusion Model for STISR

Shrey Singh[1] , Prateek Keserwani [1] , Masakazu Iwamura [2] and Partha Pratim Roy[1]

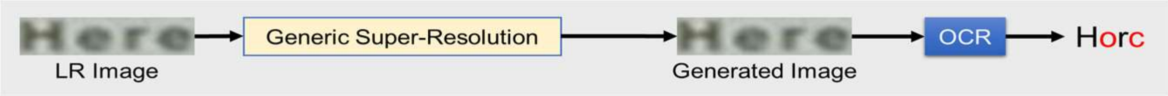
[1] Indian Institute of Technology Roorkee, India
[2] Osaka Metropolitan University, Japan

Motivation

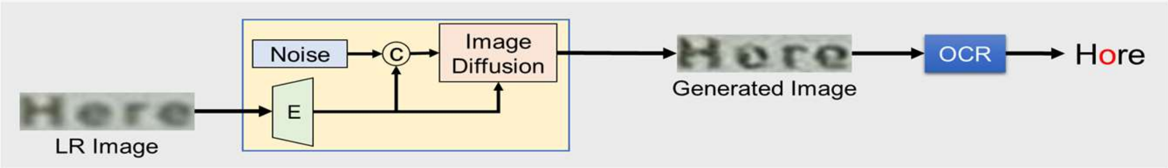
• **Challenges in Scene Text Super-Resolution (STISR):** Severe blurring in scene text images causes loss of essential strokes and textual information, significantly hindering readability and recognizability.

• **Existing Methods Limitations:** Traditional methods rely on deterministic CNNs or integrate recognizers into the SR process, leading to suboptimal performance in handling severe blur and text clarity issues.

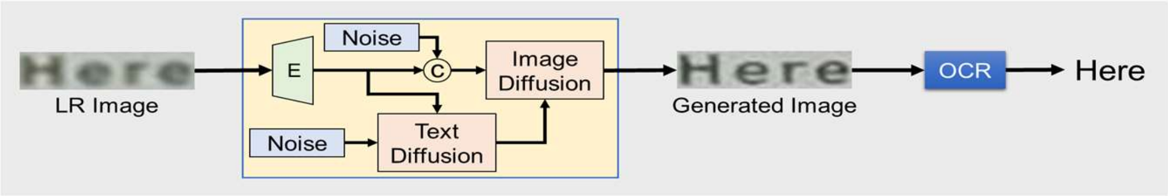
• **Goal:** To enhance the resolution and legibility of low-resolution scene text images, ensuring the output aligns with the distribution of pre-trained text recognizers without needing retraining on a new distribution.



(a) Generic super resolution method



(b) Latent image diffusion model



(c) Diffusion-conditioned-diffusion (Proposed method)

Contributions

• **Introduction of Diffusion-Conditioned-Diffusion Model (DCDM):**

– A novel generative model for STISR that combines two diffusion models:

* Latent text diffusion module for generating character-level text embeddings.
* Image diffusion module for super-resolution.

• **Character-Level CLIP (CL-CLIP) Integration:**

– Aligns high-resolution character-level text embeddings with low-resolution embeddings, ensuring visual coherence and improved text fidelity.

• **Two-Stage Conditioning Approach:**

– Conditioning the model on both the low-resolution image and character-level text embeddings from the latent diffusion model, enabling accurate text recovery.

• **Extensive Evaluation:**

– Achieved superior performance over state-of-the-art methods on TextZoom and Real-CE datasets, demonstrating the effectiveness of the proposed method in text recognition and image quality enhancement.

Proposed Method

- **Character-Level CLIP (CL-CLIP):**

  – Aligns the visual features of high-resolution text with their corresponding character-level semantics, ensuring that the super-resolved images are both visually accurate and textually meaningful.

- **Latent Text Diffusion Model (LTD):**

  – Captures text-specific embeddings from low-resolution images and refines them through a denoising process.

  – Uses cross-attention to effectively align and integrate text and image data for higher-quality text representations.

- **Image Diffusion Model (IDM):**

  – Enhances low-resolution images to produce high-resolution outputs, conditioned on both image and text priors.

  – Utilizes a stepwise denoising process through the Image Denoising UNet (IDUnet) to iteratively refine images.

- **Hybrid Conditioning:**

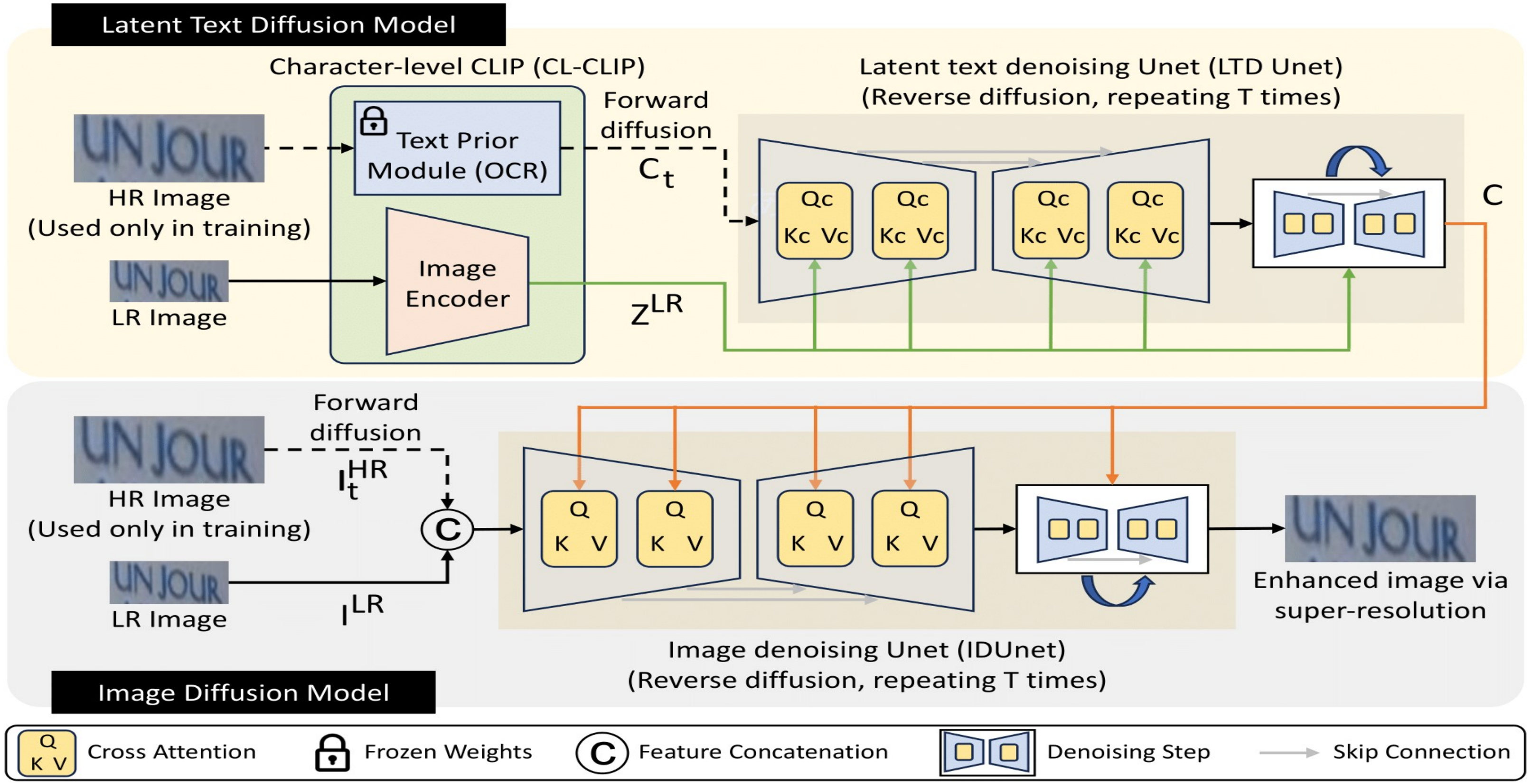
  – Combines low-resolution image data and text embeddings to guide the model, enabling the generation of clearer and more accurate scene text.

Proposed Method

- **Objective Function:** – The model minimizes the error between predicted and true noise during the diffusion process, formulated as:

$$\mathscr{L}_{I^{\mathrm{LR}} \to I^{\mathrm{HR}}} = \mathbb{E}_{I^{\mathrm{HR}}, I^{\mathrm{LR}}, C, \varepsilon \sim \mathcal{N}(0,1), t}\left[\left\| \varepsilon - \varepsilon_{\theta}(I_t^{\mathrm{HR}}, I^{\mathrm{LR}}, C, t) \right\|_2^2 \right].$$

- **Proposed Method Architecture DCDM**



Experimental Results

• Quantitative Results on Textzoom Dataset (Accuracy)

Category	Method	Accuracy of CRNN				Accuracy of MORAN				Accuracy of ASTER			
		Easy	*Medium*	*Hard*	*Avg.*	*Easy*	*Medium*	*Hard*	*Avg.*	*Easy*	*Medium*	*Hard*	*Avg.*
Baseline	BICUBIC	36.4%	21.1%	21.1%	26.8%	60.6%	37.9%	30.8%	44.1%	67.4%	42.4%	31.2%	48.2%
Generic image super-resolution	SRCNN	41.1%	22.3%	22.0%	29.2%	63.9%	40.0%	29.4%	45.6%	70.6%	44.0%	31.5%	50.0%
	SRResNet	45.2%	32.6%	25.5%	35.1%	66.0%	47.1%	33.4%	49.9%	69.4%	50.5%	35.7%	53.0%
	RCAN	46.8%	27.9%	26.5%	34.5%	63.1%	42.9%	33.6%	47.5%	67.3%	46.6%	35.1%	50.7%
	SAN	50.1%	31.2%	28.1%	37.2%	65.6%	44.4%	35.2%	49.4%	68.1%	48.7%	36.2%	52.0%
	HAN	51.6%	35.8%	29.0%	39.6%	67.4%	48.5%	35.4%	51.5%	71.1%	52.8%	39.0%	55.3%
Text-based backbone	TSRN	52.5%	38.2%	31.4%	41.4%	70.1%	55.3%	37.9%	55.4%	75.1%	56.3%	40.1%	58.3%
	TBSRN	59.6%	47.1%	35.3%	48.1%	74.1%	57.0%	40.8%	58.4%	75.7%	59.9%	41.6%	60.0%
	PCAN	59.6%	45.4%	34.8%	47.4%	73.7%	57.6%	41.0%	58.5%	77.5%	60.7%	43.1%	61.5%
Stroke-aware	TG	61.2%	47.6%	35.5%	48.9%	75.8%	57.8%	41.4%	59.4%	77.9%	60.2%	42.4%	61.3%
Text-prior	TPGSR	63.1%	52.0%	38.6%	51.8%	74.9%	60.5%	44.1%	60.5%	78.9%	62.7%	44.5%	62.8%
	TATT	62.6%	53.4%	39.8%	52.6%	72.5%	60.2%	43.1%	59.5%	78.9%	63.4%	45.4%	63.6%
	C3-STISR	65.2%	53.6%	39.8%	53.7%	74.2%	61.0%	43.2%	60.5%	79.1%	63.3%	46.8%	64.1%
Diffusion +	TCDM	**67.3%**	**57.3%**	**42.7%**	**55.7%**	77.6%	62.9%	**45.9%**	62.2%	81.3%	**65.1%**	**50.1%**	65.5%
DCDM	Proposed	65.7%	**57.3%**	41.4%	55.5%	**78.4%**	**63.5%**	45.3%	**63.4%**	**81.8%**	**65.1%**	47.4%	**65.8%**
Ground truth	BICUBIC (HR) ↓	76.4%	75.1%	64.6%	72.4%	91.2%	85.3%	74.2%	84.1%	94.2%	87.7%	76.2%	86.6%

• PSNR/SSMI Results on Textzoom

Category	Method	PSNR				SSMI ($\times 10^{-2}$)			
		Easy	*Medium*	*Hard*	*Avg.*	*Easy*	*Medium*	*Hard*	*Avg.*
Baseline	BICUBIC (LR)↑	22.35	18.98	19.39	20.35	78.84	62.54	65.92	69.61
Generic image	SRCNN	23.48	19.06	19.34	20.78	83.79	63.23	67.91	72.27
	SRResNet	24.36	18.88	19.29	21.03	86.81	64.06	69.11	74.03
	HAN	23.30	19.02	20.16	20.95	86.91	65.37	73.87	75.96
Text-based backbone	TSRN	25.07	18.86	19.71	19.70	88.97	66.76	73.02	71.57
	TBSRN	23.46	19.17	19.68	19.10	87.29	64.55	74.52	70.66
	PCAN	24.57	19.14	20.26	21.49	88.30	67.81	74.75	77.52
Stroke-aware	TG	-	-	-	21.40	-	-	-	74.56
Text-prior	TPGSR	24.35	18.73	19.93	19.79	88.60	67.84	75.07	72.93
	TATT	24.72	19.02	20.31	21.52	90.06	**69.11**	77.03	79.30
	C3-STISR	-	-	-	21.51	-	-	-	77.21
Diffusion + Text-prior + Synthesized	TCDM	-	-	-	22.83	-	-	-	**79.58**
DCDM	Proposed	**26.47**	**20.29**	**21.25**	**22.87**	**90.80**	68.73	**77.34**	79.54

• Quantitative Results on Real-CE Dataset

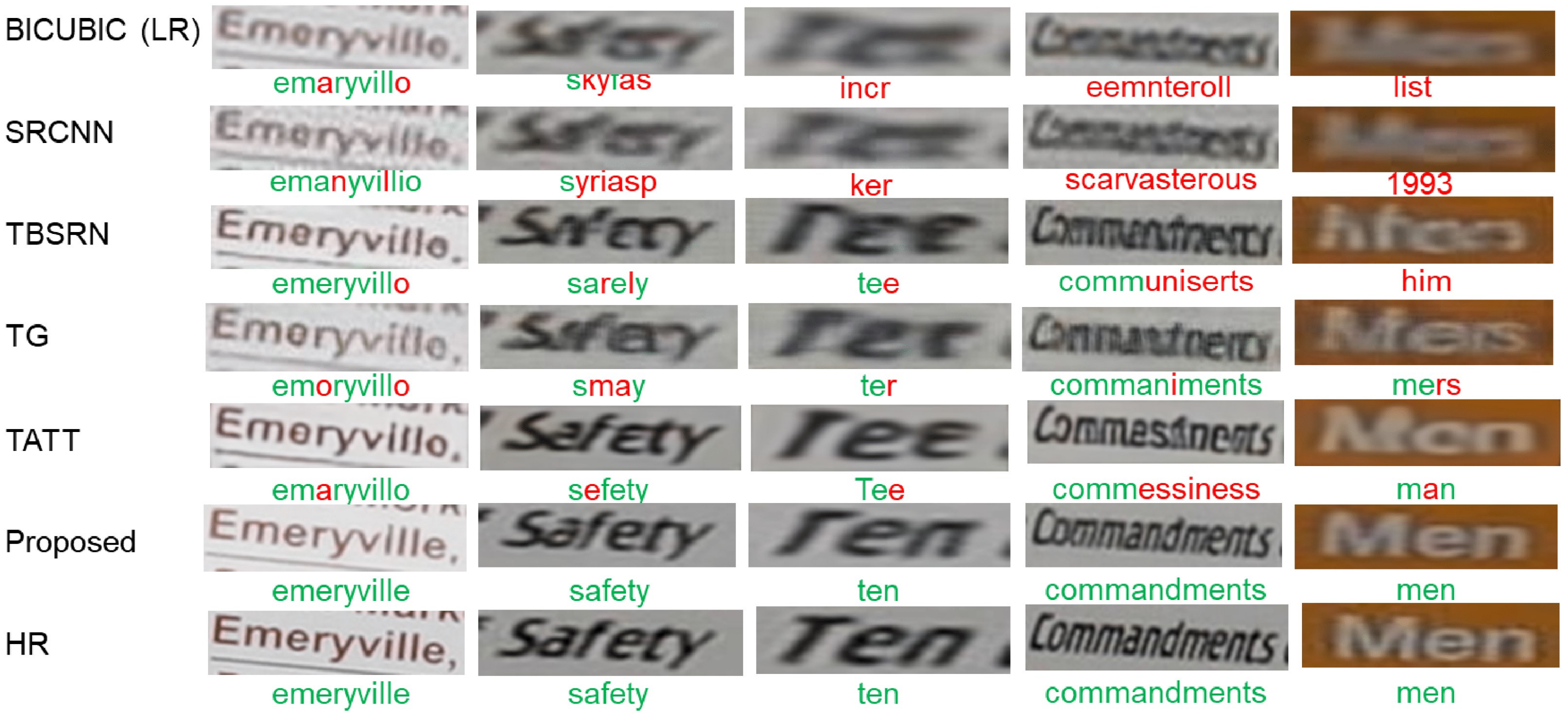
Method	×4						×2					
	Trained on TextZoom			Trained on Real-CE			Trained on TextZoom			Trained on Real-CE		
	PSNR	**SSIM**	**ACC**	**PSNR**	**SSIM**	**ACC**	**PSNR**	**SSIM**	**ACC**	**PSNR**	**SSIM**	**ACC**
TSRN	17.47	48.53	17.96	18.11	48.50	23.16	18.73	56.76	24.71	18.99	52.33	28.54
TPGSR	17.37	49.13	20.76	18.07	47.58	23.26	17.99	53.12	26.55	18.83	55.62	30.07
TBSRN	17.59	49.19	22.46	18.33	48.26	25.27	18.41	54.56	29.05	19.01	53.66	31.81
TATT	17.43	50.10	21.00	17.96	49.04	23.30	18.24	56.67	27.55	19.06	57.72	31.27
DCDM	**18.13**	**50.89**	**22.94**	**18.91**	**50.90**	**25.49**	**18.87**	**56.89**	**29.34**	**19.23**	**58.12**	**31.94**
HR image	-	-	48.07	-	-	45.14	-	-	48.07	-	-	45.14

• Quantitative Ablation Study on Textzoom Dataset.

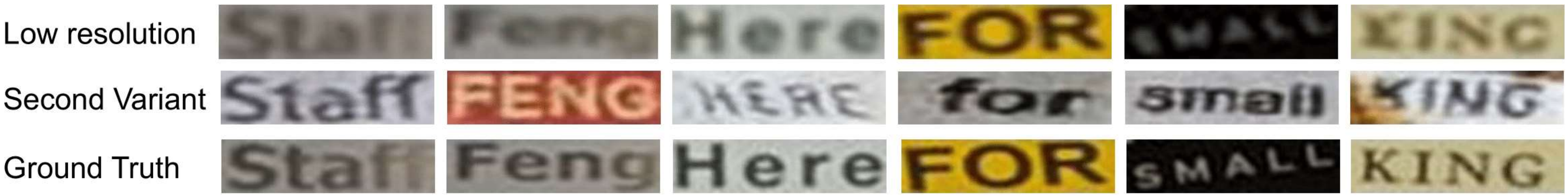
Method	Comp.		Accuracy of CRNN			Accuracy of MORAN			Accuracy of ASTER		
	TD	ID	*Easy*	*Medium*	*Hard*	*Easy*	*Medium*	*Hard*	*Easy*	*Medium*	*Hard*
DCDM w/o text	✗	✓	65.2%	56.5%	41.3%	78.1%	61.7%	44.5%	80.4%	64.5%	46.7%
DCDM	✓	✓	**65.7%**	**57.3%**	**41.4%**	**78.4%**	**63.5%**	**45.3%**	**81.8%**	**65.1%**	**47.4%**
			(↑0.5%)	(↑0.8%)	(↑0.1%)	(↑0.3%)	(↑1.8%)	(↑0.8%)	(↑1.4%)	(↑0.6%)	(↑0.7%)

Experimental Results

• Qualitative Result on Textzoom



• Qualitative Results of Ablation Study



Conclusion

• **Novel Approach for STISR:** The proposed Diffusion-Conditioned Diffusion Model (DCDM) offers a unique method for scene text image super-resolution by incorporating two distinct diffusion modules for text and image denoising.

• **Text Prior and Conditioning:** The latent diffusion module generates text priors by mapping noise to character embeddings, using low-resolution image encodings, and leveraging the CLIP-based character-level model (CL-CLIP).

• **Comparison to Conventional Methods:** Unlike traditional STISR approaches that rely on a text recognizer, DCDM eliminates the need for one, posing the question of whether it's necessary during inference.

• **Improved Performance:** Experimental results on TextZoom and Real-CE datasets showed that DCDM improves upon state-of-the-art methods quantitatively and qualitatively, with generated images maintaining high realism and fidelity.