NEURAL INFORMATION PROCESSING SYSTEMS

Hierarchical Integration Diffusion Model for Realistic Image Deblurring

Zheng Chen[1], Yulun Zhang[2*], Ding Liu[3], Bin Xia[4], Jinjin Gu[5,6], Linghe Kong[1*], Xin Yuan[7]

[1]Shanghai Jiao Tong University, [2]ETH Zürich, [3]Bytedance Inc, [4]Tsinghua University,
[5]Shanghai AI Laboratory, [6]The University of Sydney, [7]Westlake University

Introduction

Motivation

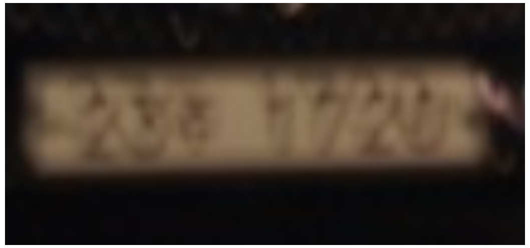
- **Regression-based** methods: Show remarkable success, especially in terms of distortion-based metrics (e.g., PSNR). But recover images with fewer details.
- **Generative models**: Generate more perceptually plausible results. But produce undesired artifacts not present in the original images.
- **Deblurring**: non-uniform blur in real scenarios.

Blurry | GT PSNR↑/LPIPS↓ | MPRNet **30.96**/0.114 | DvSR 29.77/**0.089**

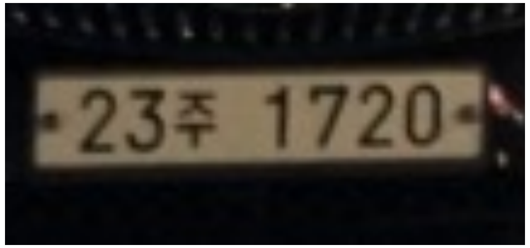
Introduction

Motivation

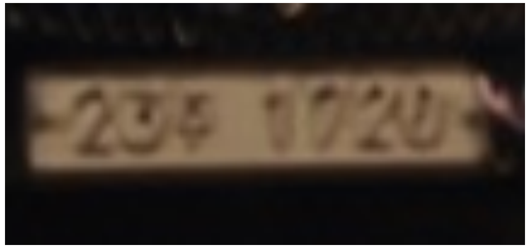
- Therefore, we design HI-Diff, which hierarchically integrates Transformer (Regression-based) and Diffusion model (Generative) for realistic image deblurring.
- HI-Diff leverages the power of diffusion models to generate prior in compact latent space.
- The generate prior is applied to guide the regression-based deblurring process from multiple scales with the hierarchical integration module.



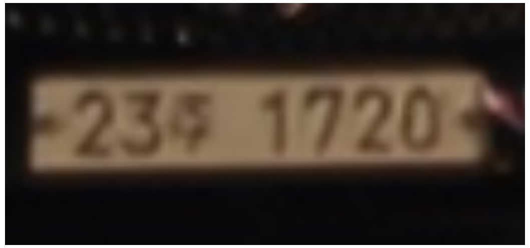
Blurry



GT

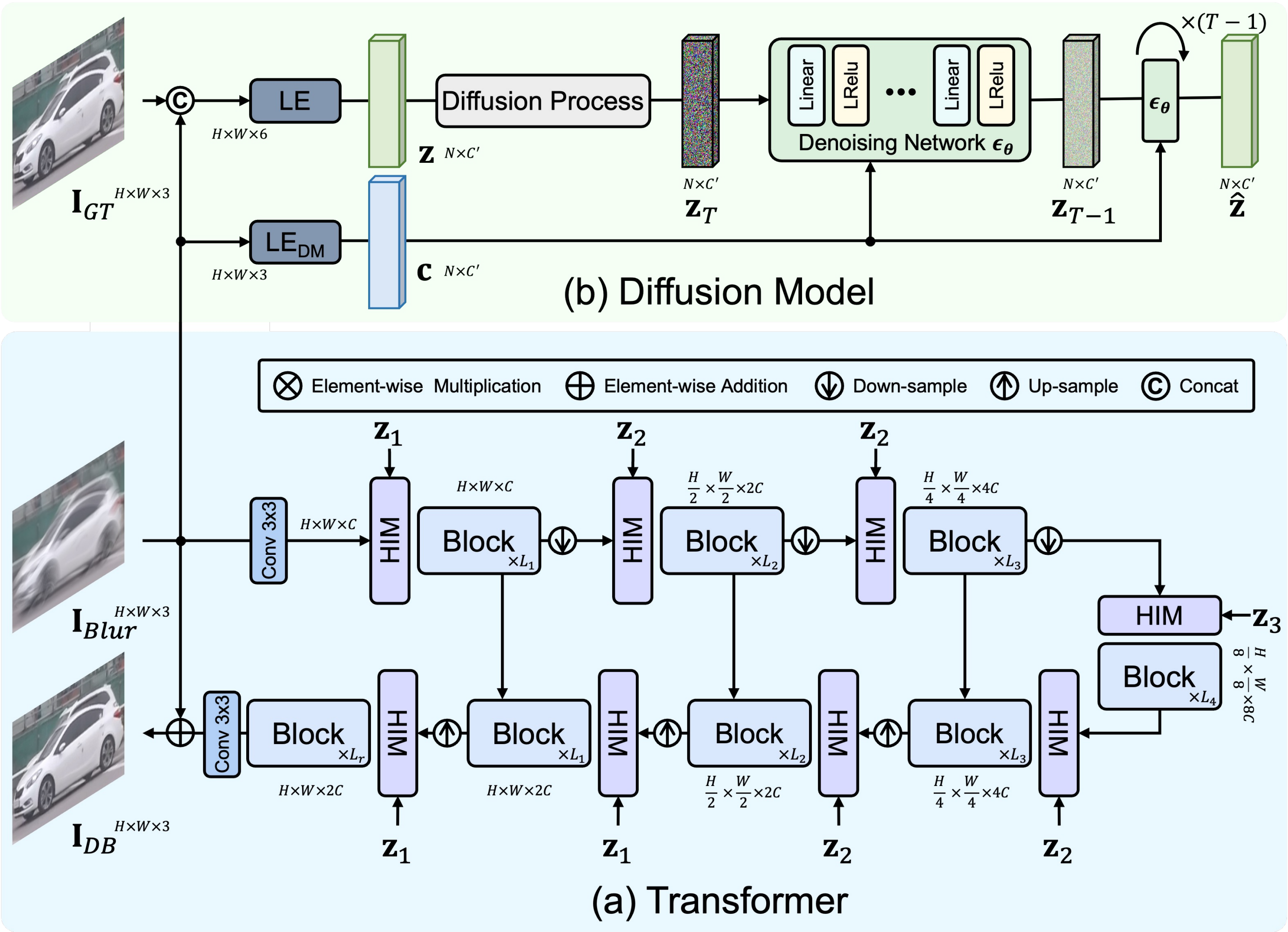


Restormer (SOTA)



HI-Diff (ours)

Method



Framework

- Compositions: Transformer and diffusion model.

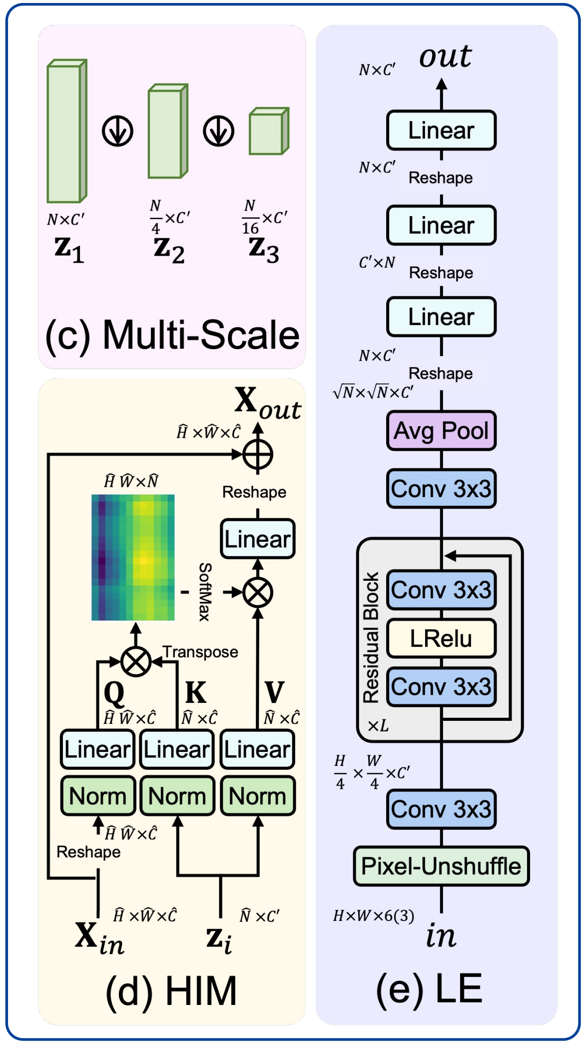
Transformer

- Hierarchical encoder-decoder architecture, guided by prior (z).

Diffusion model

- Perform in the highly compacted latent space to generate prior.

Method



Latent Encoder (LE)

- Compress the image into a compact latent representation for DM.

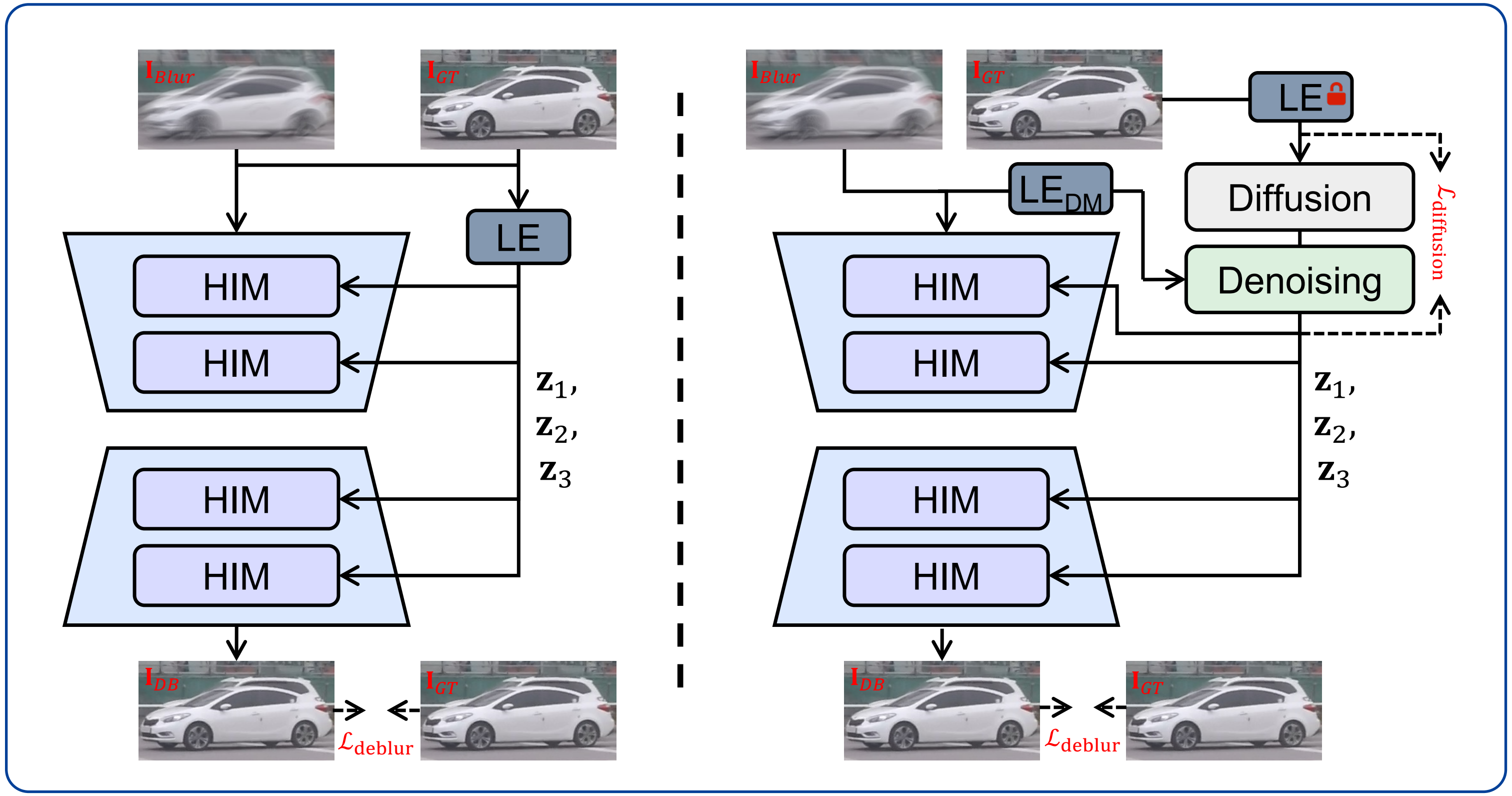
Hierarchical Integration Module (HIM)

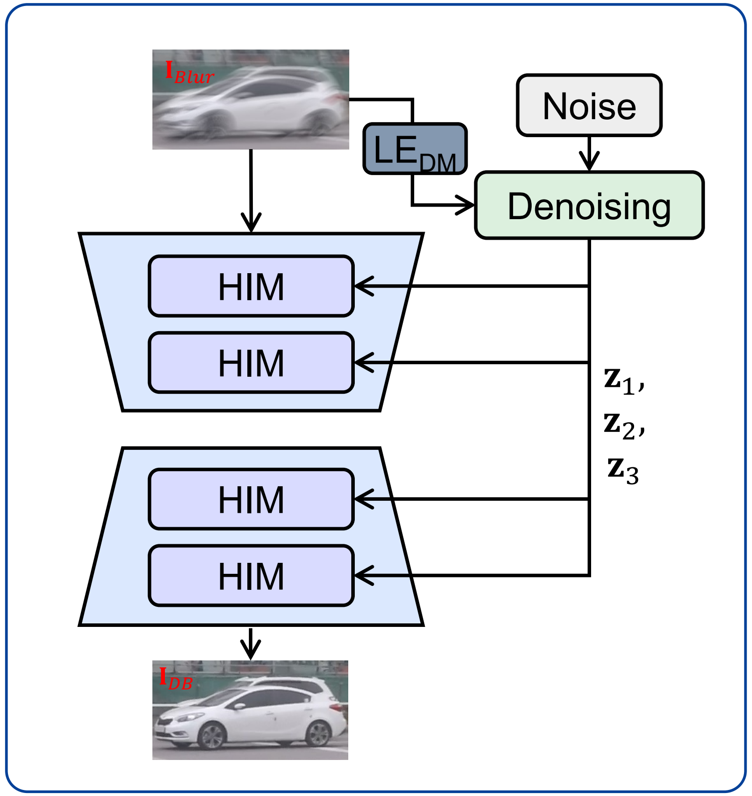
- Effectively integrate the prior feature and intermediate feature of Transformer.
- The multiple-scale prior (z_1, z_2, z_3) adapts to different scale intermediate features with cross-attention:

$$\mathbf{Q} = \mathbf{W}_Q\mathbf{X}_r, \mathbf{K} = \mathbf{W}_K\mathbf{z}_i, \mathbf{V} = \mathbf{W}_V\mathbf{z}_i,$$

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{SoftMax}(\mathbf{Q}\mathbf{K}^T/\sqrt{\hat{C}})\cdot\mathbf{V}$$

Method





Training

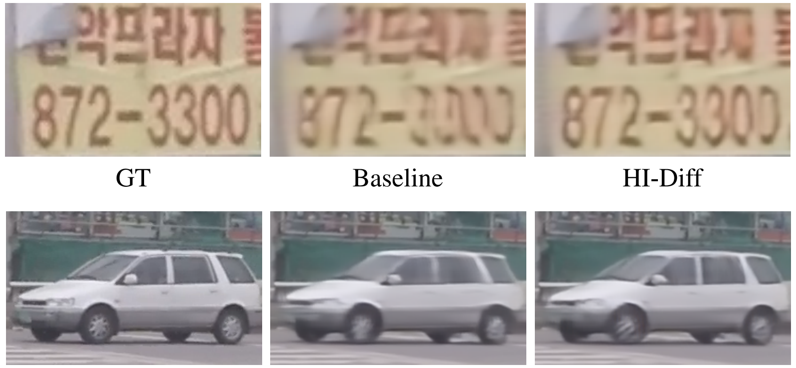
- Stage one, Loss: $\mathcal{L}_{\text{deblur}}$
- Stage two (joint training), Loss:$\mathcal{L}_{\text{deblur}} + \mathcal{L}_{\text{diffusion}}$

Inference

- Input: blurry input image, $\mathbf{I}_{Blur}$.

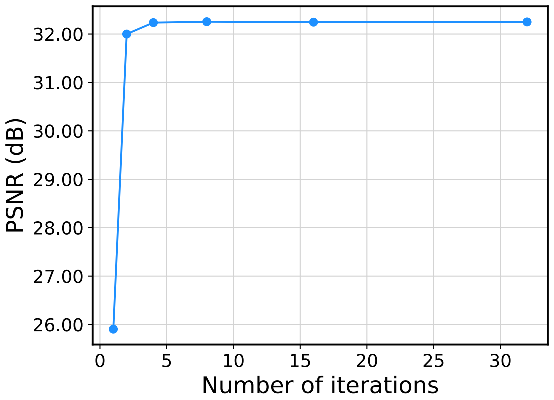
Experiments

Method	Prior	Multi-Scale	Joint-training	Params (M)	FLOPs (G)	PSNR (dB)	SSIM
Basline	✗	✗	✗	19.13	117.25	31.96	0.9528
Single-Guide	✓	✗	✓	21.98	125.39	32.00	0.9534
Split-Training	✓	✓	✗	23.99	125.47	30.73	0.9434
HI-Diff (ours)	✓	✓	✓	23.99	125.47	32.24	0.9558



GT Baseline HI-Diff

GT Single HI-Diff



Ablation

- Prior improves performance.
- Hierarchical Integration is effective.
- DM is efficient in highly compact latent space.

Experiments

Method	GoPro [28]		HIDE [39]		RealBlur-R [34]		RealBlur-J [34]	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DeblurGAN [22]	28.70	0.858	24.51	0.871	33.79	0.903	27.97	0.834
DeepDeblur [28]	29.08	0.914	25.73	0.874	32.51	0.841	27.87	0.827
DeblurGAN-v2 [23]	29.55	0.934	26.61	0.875	35.26	0.944	28.70	0.866
SRN [43]	30.26	0.934	28.36	0.915	35.66	0.947	28.56	0.867
DBGAN [56]	31.10	0.942	28.94	0.915	33.78	0.909	24.93	0.745
MT-RNN [30]	31.15	0.945	29.15	0.918	35.79	0.951	28.44	0.862
DMPHN [55]	31.20	0.940	29.09	0.924	35.70	0.948	28.42	0.860
SAPHN [42]	31.85	0.948	29.98	0.930	N/A	N/A	N/A	N/A
SPAIR [32]	32.06	0.953	30.29	0.931	N/A	N/A	28.81	0.875
MIMO-UNet+ [5]	32.45	0.957	29.99	0.930	35.54	0.947	27.63	0.837
TTFA [4]	32.50	0.958	30.55	0.935	N/A	N/A	N/A	N/A
MPRNet [54]	32.66	0.959	30.96	0.939	35.99	0.952	28.70	0.873
HINet [2]	32.71	0.959	30.32	0.932	35.75	0.949	28.17	0.849
Restormer [53]	32.92	0.961	31.22	0.942	36.19	0.957	28.96	0.879
Stripformer [44]	33.08	0.962	31.03	0.940	36.08	0.954	28.82	0.876
HI-Diff (ours)	33.33	0.964	31.46	0.945	36.28	0.958	29.15	0.890

Synthetic

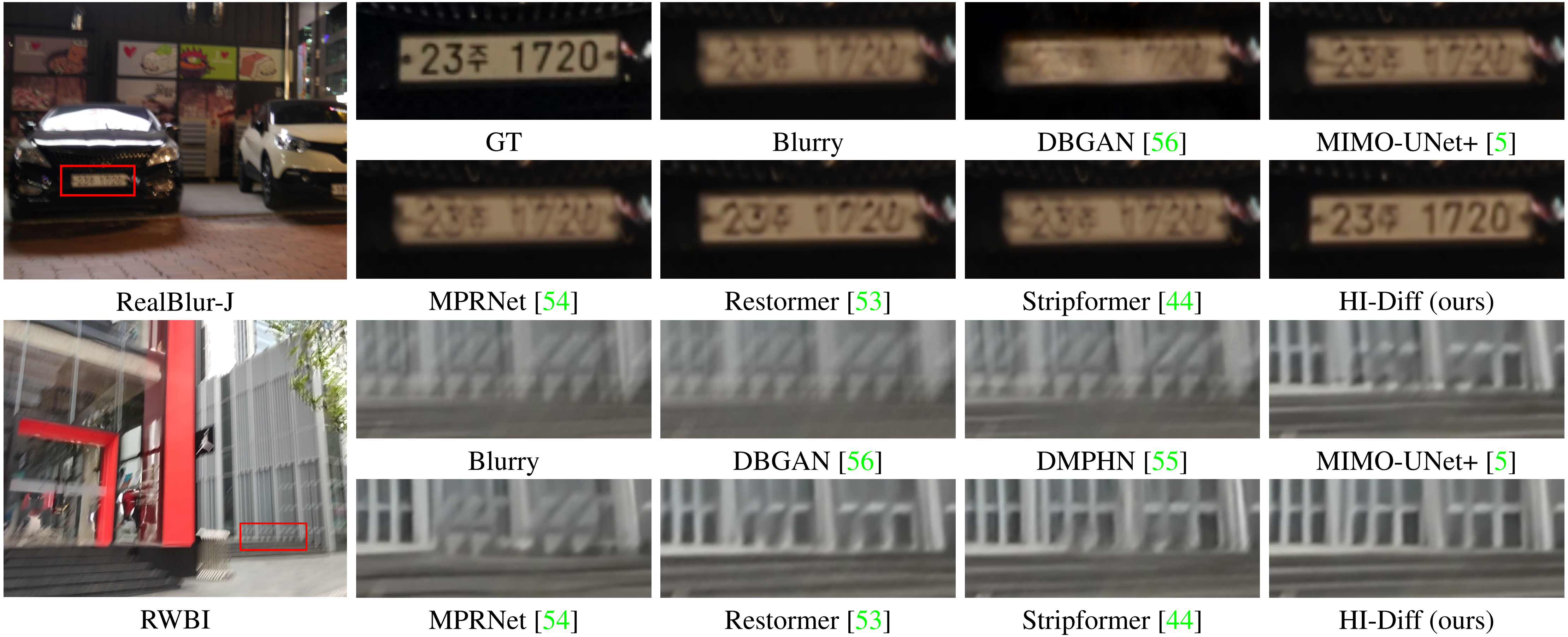
- Train only on GoPro.
- Performs well on synthetic datasets: GoPro and HIDE.
- Performs well on real-world dataset: RealBlur.
- Better generalization ability than others.

Dataset	Method	DeblurGAN-v2 [23]	SRN [43]	MIMO-UNet+ [5]	MPRNet [54]	BANet [45]	Stripformer [44]	HI-Diff (ours)
RealBlur-R [34]	PSNR ↑	36.44	38.65	N/A	39.31	39.55	39.84	41.01
	SSIM ↑	0.935	0.965	N/A	0.972	0.971	0.974	0.978
RealBlur-J [34]	PSNR ↑	29.69	31.38	31.92	31.76	32.00	32.48	33.70
	SSIM ↑	0.870	0.909	0.919	0.922	0.9230	0.929	0.941

Real-World

- Train on the RealBlur.
- Outperforms other compared methods.

Experiments



Visual comparison

- Our method reconstructs more accurate textures and sharper edges.

Experiments

Method	DMPHN [55]	MIMO-UNet+ [5]	MPRNet [54]	HINet [2]	Restormer [53]	Stripformer [44]	HI-Diff (ours)	HI-Diff-2 (ours)
Params (M)	21.70	16.11	20.13	88.67	26.13	19.71	28.49	23.99
FLOPs (G)	195.44	150.68	760.02	67.51	154.88	155.03	142.62	125.47
PSNR (dB)	31.20	32.45	32.66	32.71	32.92	33.08	33.33	33.28

Model Size

- Our method achieves a better trade-off between performance and complexity.

GT | Blurry | $\mathbf{z}_8$ (noise, ϵ) | $\mathbf{z}_4$ | $\mathbf{z}_2$ | $\mathbf{z}_0$ (output)

Diffusion

- Blur images gradually become sharp as the reverse process proceeds.

Conclusion

Contribution

- We design a novel approach called the Hierarchical Integration Diffusion Model (HI-Diff) for realistic (synthetic and real-world) image deblurring.
- Our HI-Diff leverages the power of diffusion models to generate prior and hierarchically integrates priors into the deblurring process for better generalization in complex scenarios.
- Our HI-Diff achieves superior performance on synthetic and real-world blur datasets.

Github Repo

Poster

- Time: Wed 13 Dec 5 p.m. CST - 7 p.m. CST
- Place: Great Hall & Hall B1+B2 #909

Thanks!