SeeClear: Semantic Distillation Enhances Pixel Condensation for Video Super-Resolution

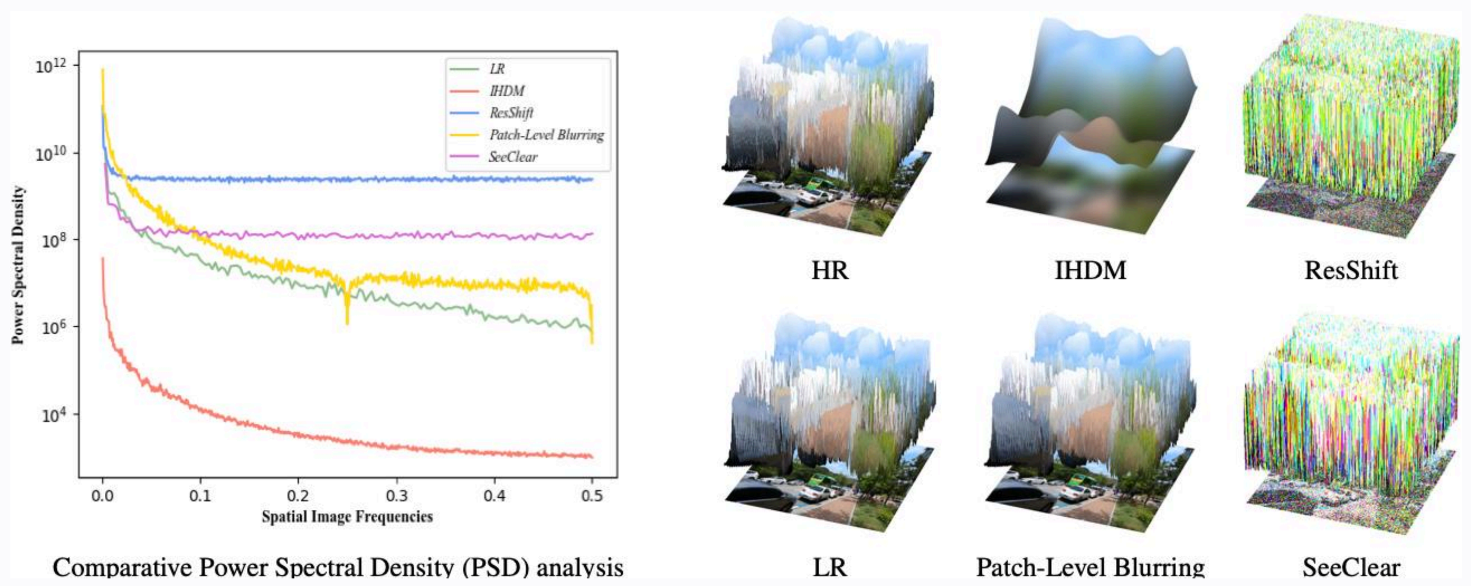
Qi Tang[1], Yao Zhao[1], Meiqin Liu[1], Chao Yao[2]

[1]Beijing Jiaotong University

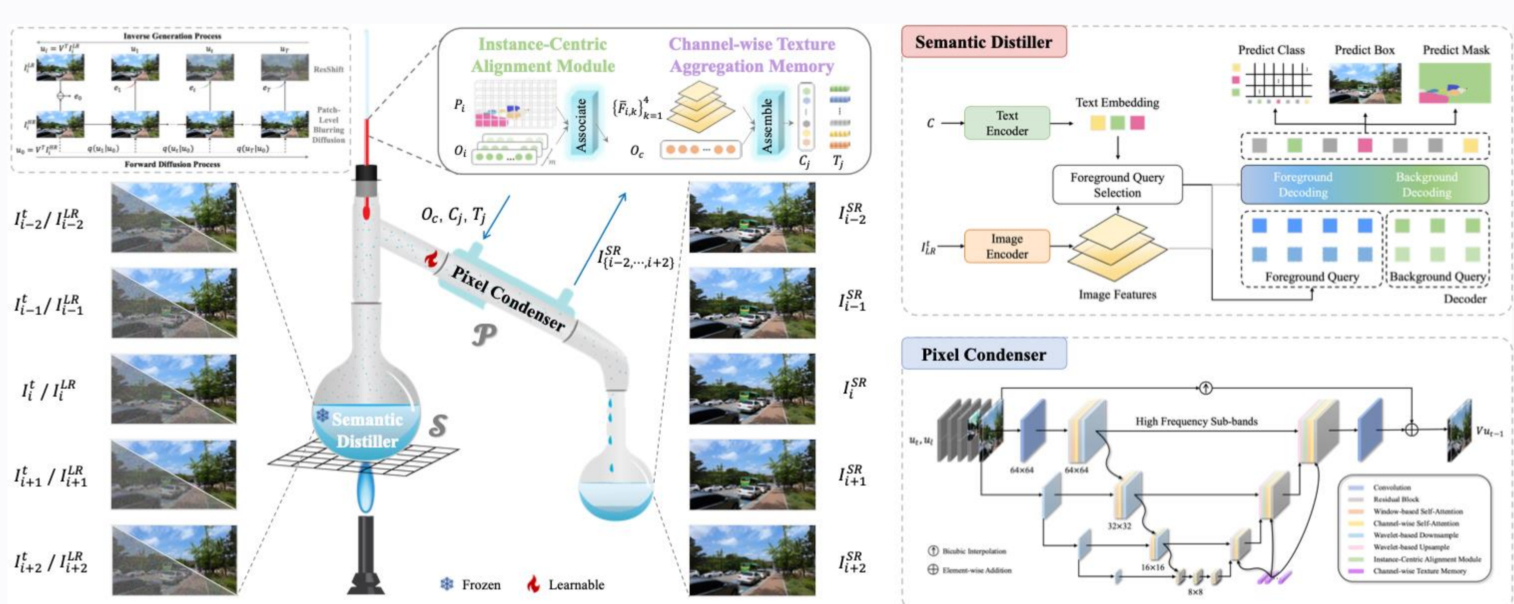
[2]University of Science and Technology Beijing

Motivation

- Diffusion-based Video Super-Resolution (VSR) is renowned for generating perceptually realistic videos, yet it grapples with maintaining detail consistency across frames due to stochastic fluctuations.
- The traditional approach of pixel-level alignment is ineffective for diffusion-processed frames because of iterative disruptions.
- Solely disrupting frames with additive noise is inadequate to depict the degradation of high-resolution videos.



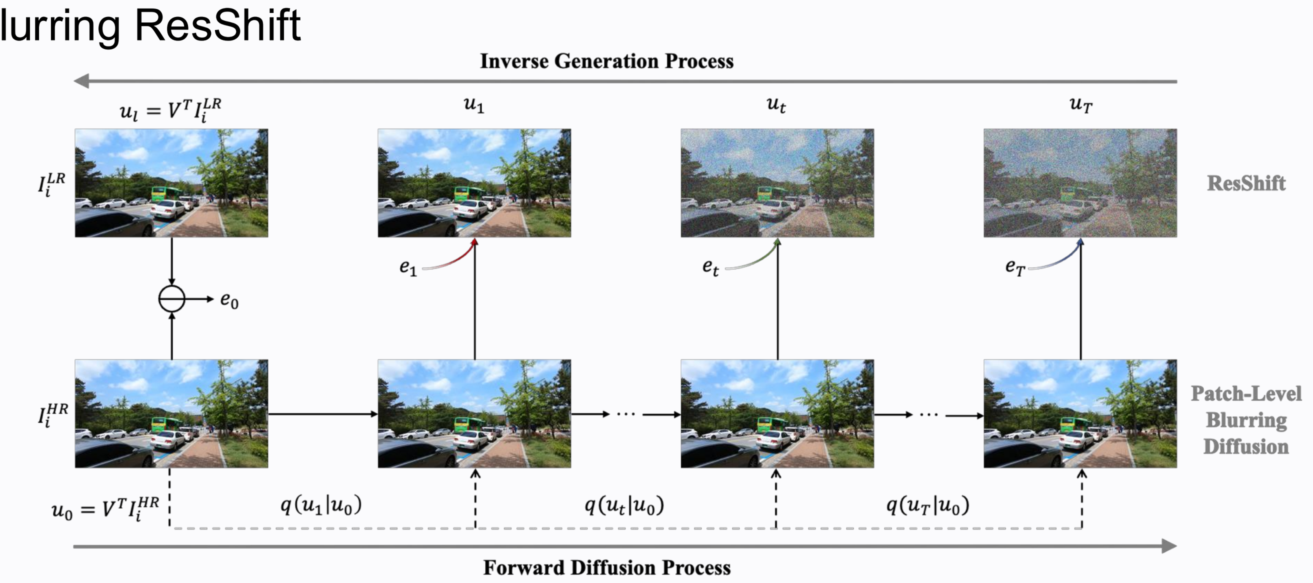
Overall Architecture



The illustration of **SeeClear**. It comprises the diffusion process incorporating patch-level blurring and residual shift mechanism and a reverse process. During the reverse process, Semantic Distiller for semantic embedding extraction and U-shaped Pixel Condenser are employed for iterative denoising. The devised InCAM and CaTeGory are inserted into the U-Net to utilize the diverse semantics for inter-frame alignment in the diffusion-based VSR framework.

Methods

➢ Blurring ResShift



- Transition Kernel:

$$q(\boldsymbol{u}_t \mid \boldsymbol{u}_0) = \mathcal{N}(\boldsymbol{u}_t \mid \boldsymbol{D}_t \boldsymbol{u}_0, \eta_t \boldsymbol{E}), \quad t \in \{1, \cdots, T\},$$
$$\boldsymbol{u}_0 = \boldsymbol{V}^{\mathrm{T}} I_i^{HR},$$

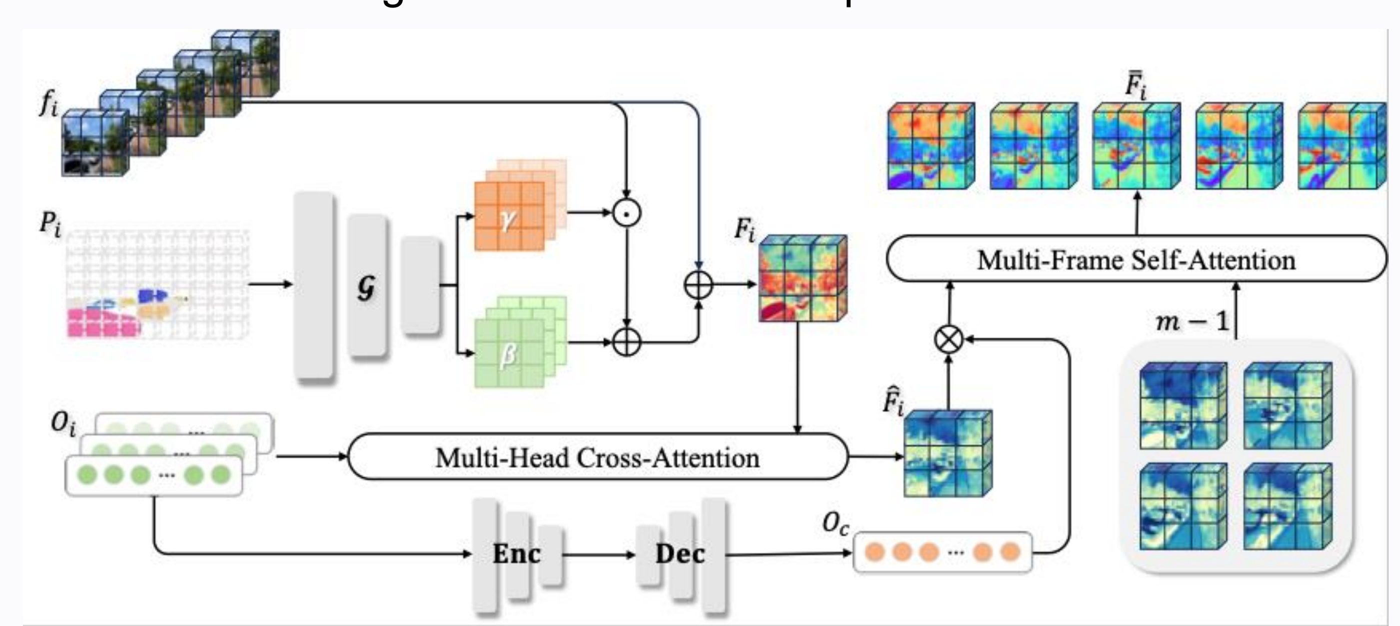
- Forward Diffusion Process:

$$q(\boldsymbol{u}_t \mid \boldsymbol{u}_0, \boldsymbol{u}_l) = \mathcal{N}(\boldsymbol{u}_t \mid \boldsymbol{D}_t \boldsymbol{u}_0 + \eta_t \boldsymbol{e}_t, \kappa^2 \eta_t \boldsymbol{E}), \quad t \in \{1, \cdots, T\},$$
$$\boldsymbol{e}_t = \boldsymbol{u}_l - \boldsymbol{D}_t \boldsymbol{u}_0$$

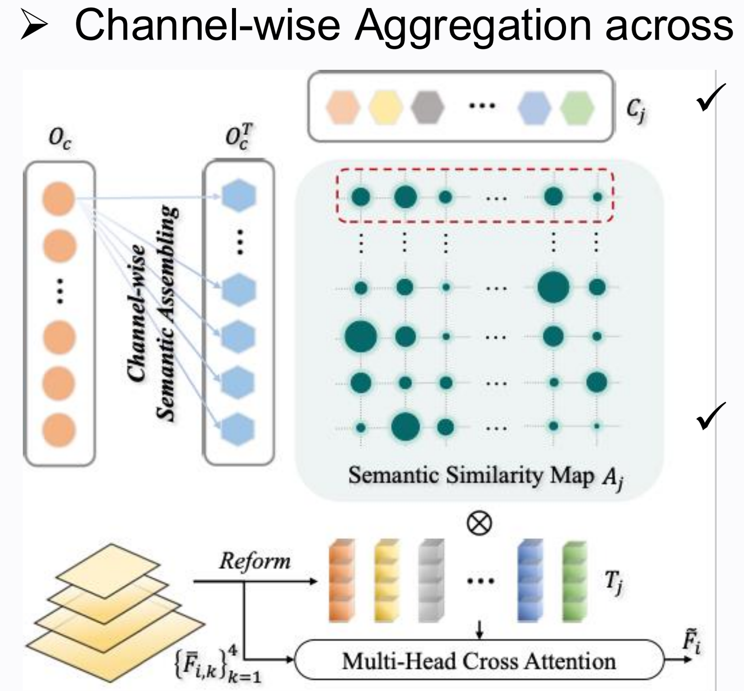
- Reverse Sampling Process:

$$p(\boldsymbol{u}_0 \mid \boldsymbol{u}_l) = \int p(\boldsymbol{u}_T \mid \boldsymbol{u}_l) \prod_{t=1} p_{\theta}(\boldsymbol{u}_{t-1} \mid \boldsymbol{u}_t, \boldsymbol{u}_l) \mathrm{d}\boldsymbol{u}_{1:T},$$
$$p(\boldsymbol{u}_T \mid \boldsymbol{u}_l) \approx \mathcal{N}(\boldsymbol{u}_T \mid \boldsymbol{u}_l, \kappa^2 \boldsymbol{E}),$$

➢ Instance-Centric Alignment within Video Clips



➢ Channel-wise Aggregation across Video Clips



✓ The **I**nstance-**C**entric **A**lignment **M**odule (InCAM) utilizes video-clip-wise tokens to dynamically relate pixels within and across frames, enhancing coherency.

✓ The **C**ha**n**nel-wise **Te**xture **A**ggregation Mem**ory** (CaTeGory) infuses extrinsic knowledge, capitalizing on long-standing semantic textures.

Results

Table 1 Performance comparisons on the REDS4 and Vid4 datasets.

Methods	Frames	REDS4 [24]			Vid4 [19]		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Bicubic	-	26.14	0.7292	0.3519	23.78	0.6347	0.3947
TOFlow [41]	7	29.98	0.7990	0.3104	25.89	0.7651	0.3386
EDVR-M [37]	5	30.53	0.8699	0.2312	27.10	0.8186	0.2898
BasicVSR [1]	15	31.42	0.8909	0.2023	27.24	0.8251	0.2811
VRT [16]	6	31.60	0.8888	0.2077	27.93	0.8425	0.2723
IconVSR [1]	15	31.67	0.8948	0.1939	27.39	0.8279	0.2739
StableSR [36]	1	24.79	0.6897	0.2412	22.18	0.5904	0.3670
ResShift [45]	1	27.76	0.8013	0.2346	24.75	0.7040	0.3166
SATeCo [6]	6	31.62	0.8932	0.1735	27.44	0.8420	0.2291
SeeClear (Ours)	5	28.92	0.8279	0.1843	25.63	0.7605	0.2573
SeeClear* (Ours)	5	31.32	0.8856	0.1548	27.80	0.8404	0.2054

Methods	Frames	REDS4 [24]			Vid4 [19]		
		DISTS ↓	NIQE ↓	CLIP-IQA ↑	DISTS ↓	NIQE ↓	CLIP-IQA ↑
Bicubic	-	0.1876	7.257	0.6045	0.2201	7.536	0.6817
TOFlow [41]	7	0.1468	6.260	0.6176	0.1776	7.229	0.7356
EDVR-M [37]	5	0.0943	4.544	0.6382	0.1468	5.528	0.7380
BasicVSR [1]	15	0.0808	4.197	0.6353	0.1442	5.340	0.7410
VRT [16]	6	0.0823	4.252	0.6379	0.1372	5.242	0.7434
IconVSR [1]	15	0.0762	4.117	0.6162	0.1406	5.392	0.7411
StableSR [36]	1	0.0755	4.116	0.6579	0.1385	5.237	0.7644
ResShift [45]	1	0.1432	6.391	0.6711	0.1716	6.868	0.7157
SATeCo [6]	6	0.0607	4.104	0.6622	0.1015	5.212	0.7451
SeeClear (Ours)	5	0.0762	4.381	0.6870	0.0947	5.305	0.7106
SeeClear* (Ours)	5	0.0641	3.757	0.6848	0.0919	4.896	0.7303

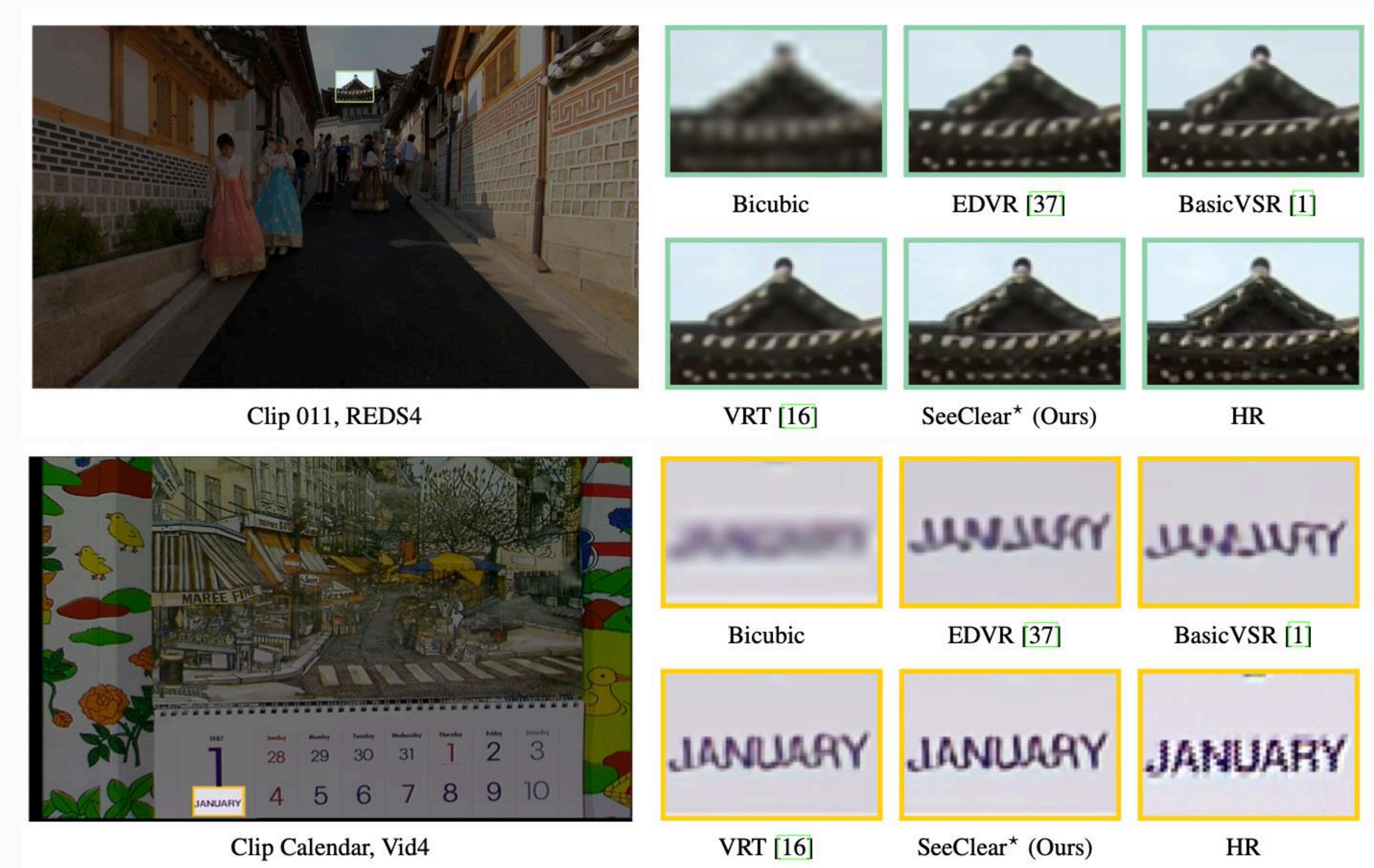


Figure 1 Qualitative results on the REDS4 and Vid4 datasets.

Analysis. SeeClear benefits from the control of dual semantics, striking a balance between superior fidelity and the generation of realistic textures while maintaining a higher temporal consistency. Besides, SeeClear is much smaller and runs faster compared to other diffusion-based method.