



Tencent
腾讯

WEVSR: Video Diffusion Generators for Real-World Video Super-Resolution with Wavelet-Enhanced VAE Encoder

Yuying Chen, Zhirui Liu, Linyan Jiang, Qifan Gao, Xianguo Zhang, Jianhou Gan, Wenqi Ren



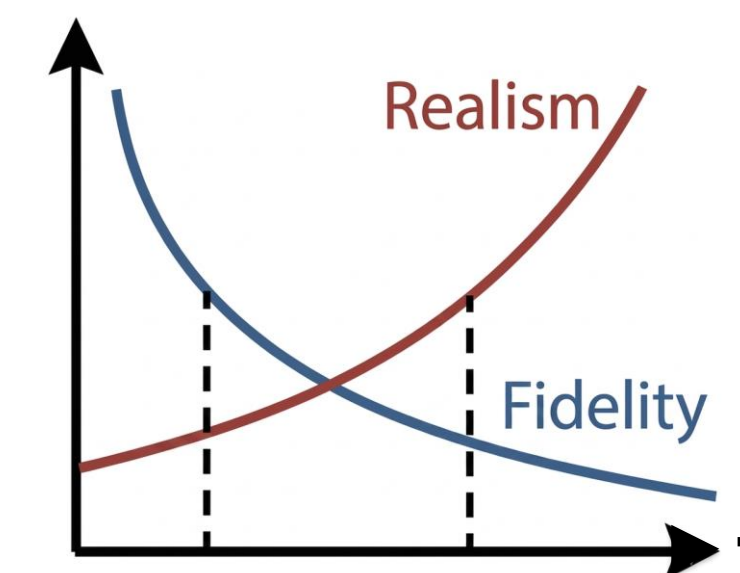
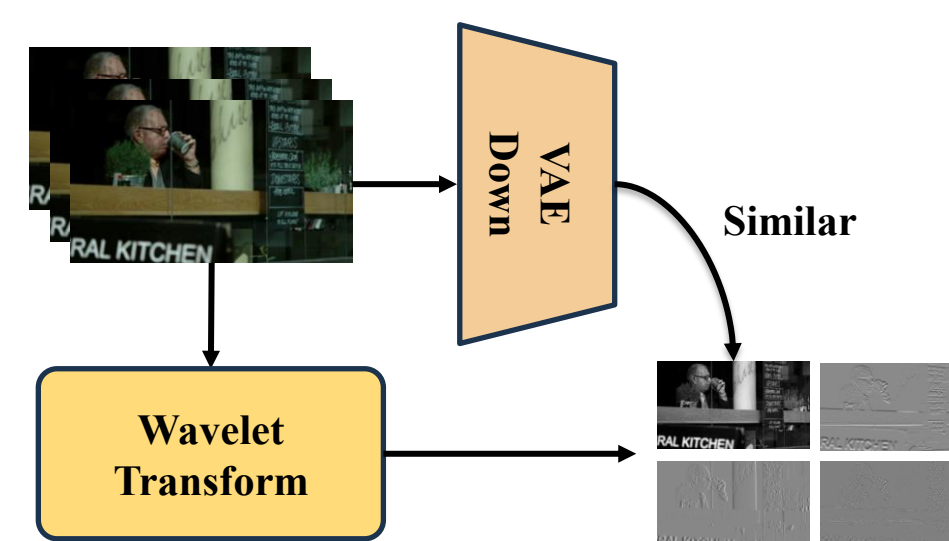
ICML
International Conference
On Machine Learning

Introduction

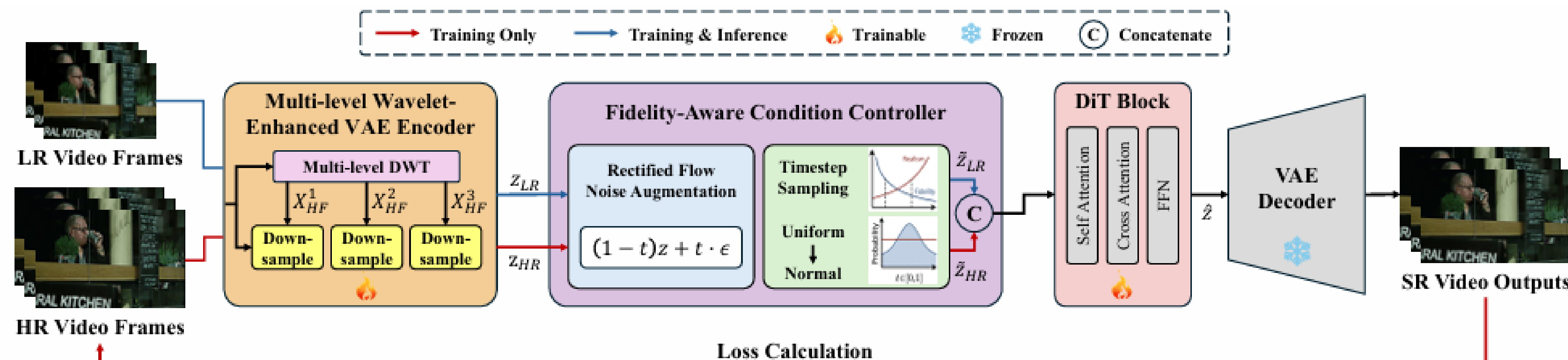
- **Large video diffusion models** demonstrates great generative capability but the image condition controlling remains challenging
- We introduce the multi-level DWT, flow-based noise augmentation and fidelity-aware timestep sampling to balance the **generation realism** and **fidelity** for video super-resolution
- **Two-stage training pipeline**
 - Finetune the multi-level wavelet-enhanced VAE encoder with the pretrained decoder being frozen
 - Finetune the pretrained DiT backbone with the wavelet-enhanced VAE encoder and fidelity-aware controller

Key Motivation

- ❑ **Multi-level Wavelet-Enhanced VAE Encoder:** Similar to wavelet transform, the VAE encoder produce a half scale latent by cutting out high-freq in each downsample layer
- ❑ **Fidelity-Aware Condition Controller:** The trade off between realism and fidelity of the restored video frames can be adjusted by the sampling timestep



Overall Pipeline



Experimental Results

Numerical performance on VSR benchmarks

Dataset	Metric	RealViFormer	Upscale-A-Video	MGLD-VSR	STAR	SeedVR-7B	SeedVR2-7B	Ours
SPMCS	PSNR↑	24.185	22.12	23.41	22.58	20.78	22.90	23.82
	SSIM↑	0.663	0.562	0.633	0.609	0.575	0.638	0.673
	LPIPS↓	0.378	0.405	0.369	0.420	0.395	0.322	0.294
	DISTS↓	0.186	0.198	0.166	0.229	0.166	0.134	0.127
UDM10	PSNR↑	26.70	23.29	26.11	24.66	24.29	26.26	28.10
	SSIM↑	0.796	0.717	0.772	0.747	0.731	0.798	0.845
	LPIPS↓	0.285	0.301	0.273	0.359	0.264	0.203	0.130
	DISTS↓	0.166	0.175	0.144	0.195	0.124	0.101	0.070
YouHQ40	PSNR↑	23.26	22.41	22.62	22.15	20.60	22.46	23.67
	SSIM↑	0.606	0.565	0.576	0.575	0.546	0.600	0.612
	LPIPS↓	0.362	0.375	0.356	0.451	0.323	0.274	0.237
	DISTS↓	0.193	0.186	0.166	0.213	0.134	0.110	0.104
VideoLQ	NIQE↓	4.153	3.676	3.864	5.915	4.933	4.948	4.412
	MUSIQ↑	54.65	51.45	53.49	40.50	48.35	45.76	54.13
	CLIP-IQA↑	0.411	0.327	0.333	0.243	0.258	0.257	0.356
	DOVER↑	7.035	6.935	8.109	6.891	7.416	7.236	8.234
RealVSR	NIQE↓	4.056	3.849	3.627	3.927	3.851	3.877	3.792
	MUSIQ↑	47.542	72.858	73.241	73.912	73.825	73.540	75.690
	CLIP-IQA↑	0.458	0.540	0.553	0.537	0.566	0.562	0.568
	DOVER↑	10.39	12.58	12.23	12.46	13.29	12.34	14.07
AIGC30	NIQE↓	4.618	3.924	3.917	4.449	3.809	4.554	3.325
	MUSIQ↑	49.890	54.318	53.863	52.170	56.921	50.055	60.092
	CLIP-IQA↑	0.298	0.382	0.377	0.326	0.411	0.300	0.414
	DOVER↑	7.318	8.176	8.128	7.492	8.202	7.485	8.731

Visual performance on Real VSR benchmarks

