Efficient video super resolution

高效视频超分辨率

主讲人：凌海涛　　时间：2025.3.30

1924-2024
中山大學 世纪华诞
100th ANNIVERSARY
SUN YAT-SEN UNIVERSITY

目录

CONTENTS

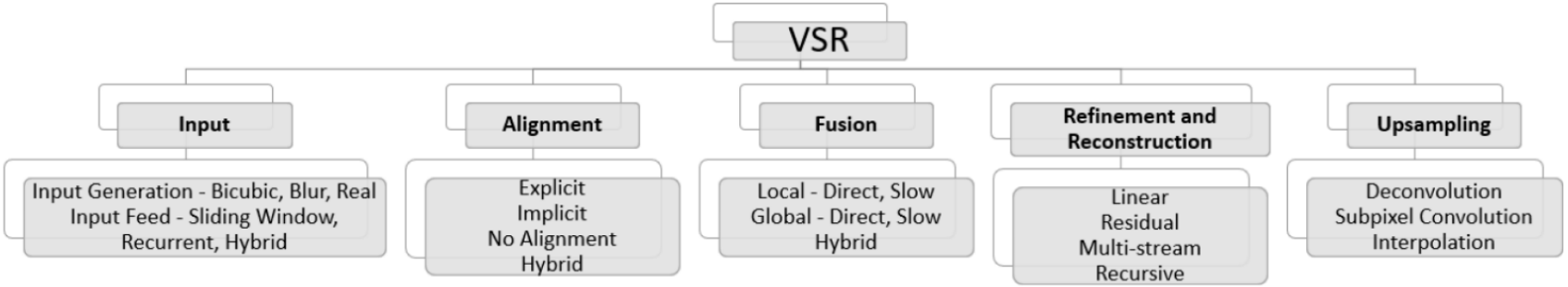
01

1924-2024
中山大學 世纪华诞
100th ANNIVERSARY
SUN YAT-SEN UNIVERSITY

VSR的五个基本组成部分

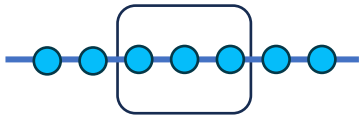
VSR与SR的区别

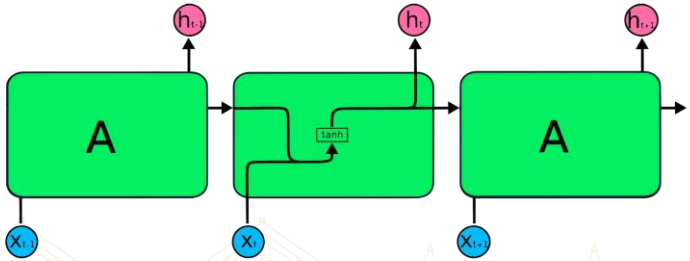
- 与单图超分（SISR）和多图超分（MISR）不同，视频超分（VSR）引入了时间维度的信息。
- 时间维度的引入增加了VSR的复杂性，要求模型对多个时间分散的帧进行融合和对齐。
- 下面从五个方面详细介绍VSR和ISR的区别。



VSR的五个基本组成部分——输入

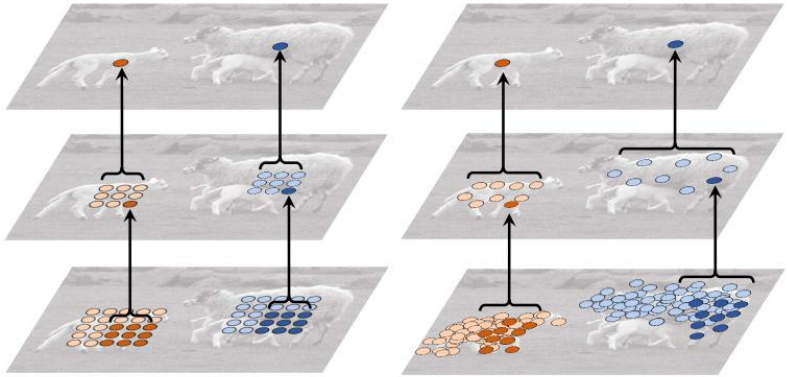
- 输入生成：与ISR类似，常采用Bicubic、Gaussian下采样的方式获得HR-LR对。
- 输入方式：因为要考虑时间维度，主要包括滑动窗口和递归的方式
  - 滑动窗口（Sliding Window）：每次处理窗口内的固定数量的帧。比如，每次使用3帧进行处理，窗口会在时间轴上滑动。
  - 递归（Recurrent）：与RNN类似，当前帧的超分辨率预测不仅依赖于当前帧，还依赖于之前的所有帧。





VSR的五个基本组成部分——对齐

- 对齐（Alignment）是指对齐低分辨率相邻帧及对应目标物体的过程，确保不同帧或图像的特征能够正确地对应到同一位置上，提高时序一致性。
- 显式对齐：
  - 如基于光流的运动估计和补偿。
- 隐式对齐：
  - 如可变性卷积。



(a) standard convolution　　(b) deformable convolution

VSR的五个基本组成部分——融合、细化、重建与上采样

- 融合（Fusion）指的是将不同来源的特征或信息进行特征融合，通过一定的机制将多个图像或视频帧的细节与信息整合到一起，从而生成高质量的输出。
- 旨在提升超分后的纹理细节。
- 常见的方式包括：特征拼接、注意力机制、加权求和等。

- 细化与重建（Refinement and Reconstruction）指的是模型的骨架网络。
- 包括残差连接、多支路网络、递归神经网络、3D卷积神经网络等。

- 上采样（Upsampling）与ISR类似
- 常见方式包括反卷积、PixelShuffle、双线性插值、Bicubic插值等。

02

1924-2024
中山大學 世纪华诞
100th ANNIVERSARY
SUN YAT-SEN UNIVERSITY

高效视频超分论文分享

用于实时渲染的高效超分

Efficient Video Super-Resolution for Real-time Rendering with Decoupled G-buffer Guidance

Mingjun Zheng* Long Sun* Jiangxin Dong Jinshan Pan†
School of Computer Science and Engineering, Nanjing University of Science and Technology
{mingjunzheng, cs.longsun, jxdong, jspan}@njust.edu.cn

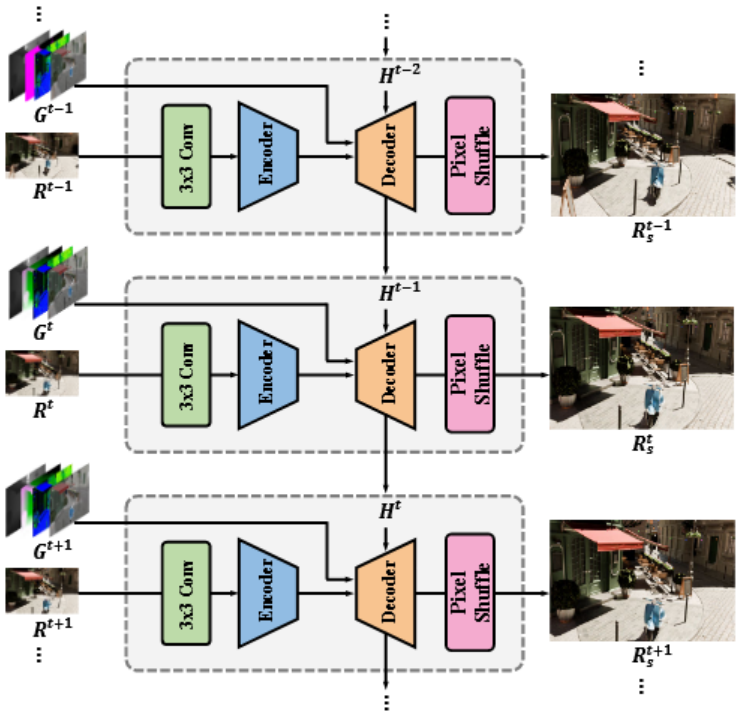
- 南京理工发表在CVPR 2025的一篇工作。
- 通过解耦的G-Buffer引导实现了用于实时渲染的高效视频超分。
- 270*480的视频在3090上做4倍超分成1080P，帧率可达126FPS。

用于实时渲染的高效超分

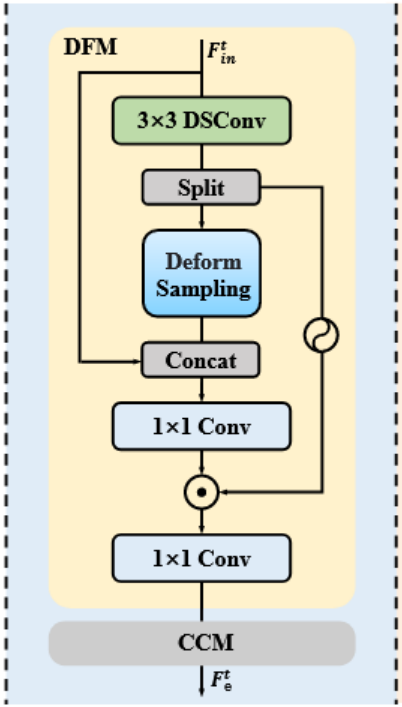
- 现有研究的不足
  - 目前的VSR模型通常依赖于双端特征传播，即需要当前帧和相邻帧的信息来进行恢复，与实时渲染任务相冲突，因为未来帧的信息在实时渲染中是未知的。
  - 在时序上进行特征融合常常采用光流估计的方法，这个方法的计算成本会随着分辨率的提升而大大增加，显著限制了模型的推理效率。
  - 目前的VSR模型主要基于Transformer架构，计算成本高昂而难以实时渲染。

- 什么是G-Buffer?
  - 这是一种储存场景几何信息的缓存结构，主要包含了图像的深度信息（Depth）、法线信息（Normal）、运动向量（Motion Vector）和双端反射分布函数（Bidirectional Reflectance Distribution Function, BRDF）

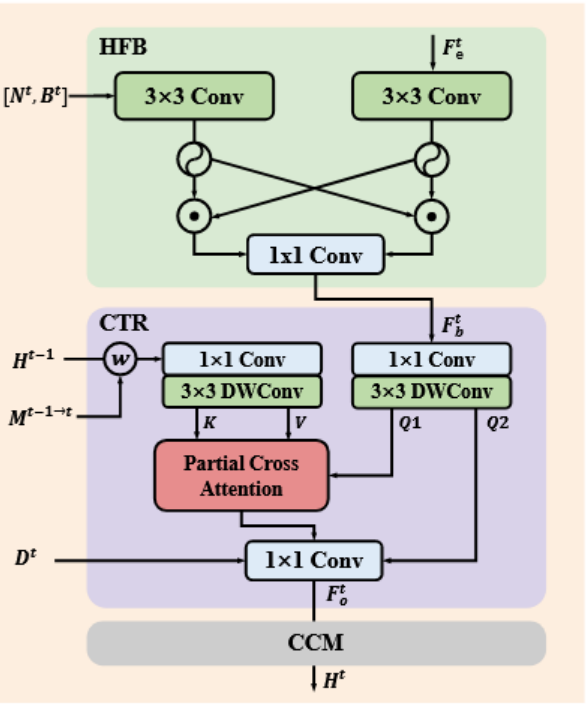
用于实时渲染的高效超分——整体框架



(a) Overall Architecture

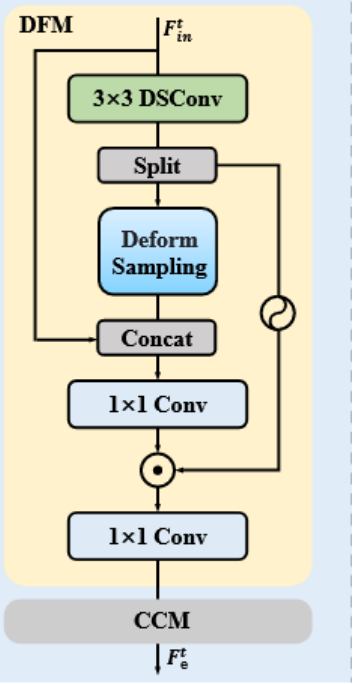


(b) Encoder Block

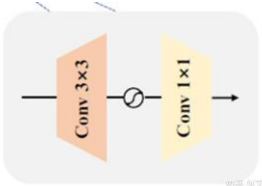


(c) Decoder Block

用于实时渲染的高效超分——编码器



(b) Encoder Block



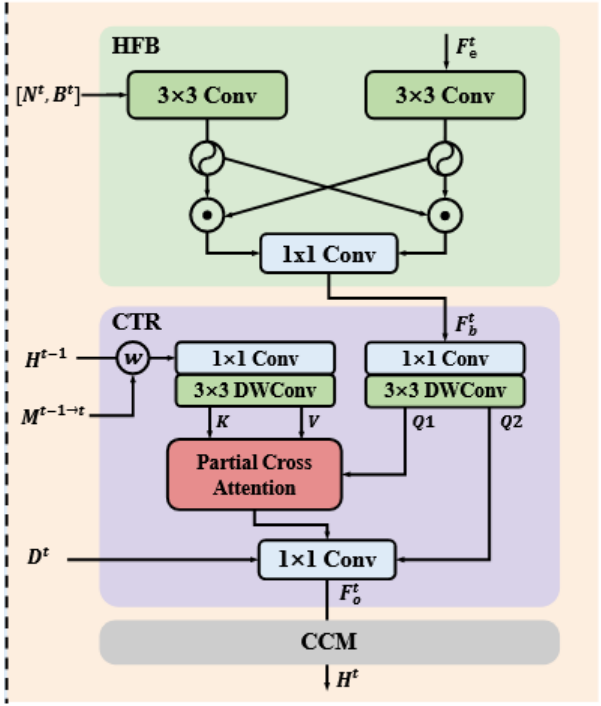
- 传统提取非局部信息的方法主要依赖于池化操作和大核深度卷积。但都存在缺陷，池化操作会平滑空间信息，而大核深度卷积无法捕捉跨通道的特征。
- 通过可变形采样的策略，更好地建模跨通道的非局部特征，具体来说，给定采样的大小S_h和S_w：

$$\hat{X}_{[i,j,:]} = \sum_{c=0}^{C_{in}} w_{[c,:]} \cdot X_{[i+\Delta_i(c),j+\Delta_j(c),c]} + b, \tag{1}$$

$$\Delta_i(c) = (c \bmod S_h) - 1,$$

$$\Delta_j(c) = (\lfloor \frac{c}{S_h} \rfloor \bmod S_w) - 1.$$

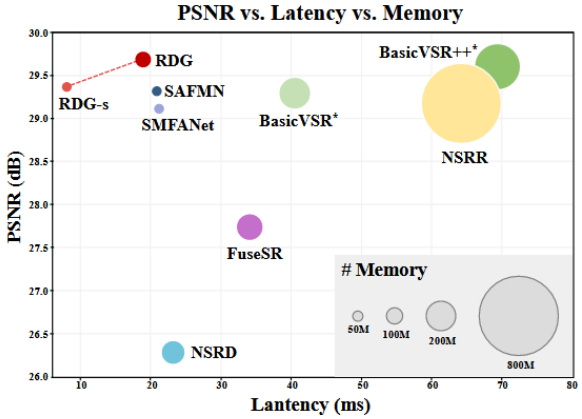
用于实时渲染的高效超分——解码器



(c) Decoder Block

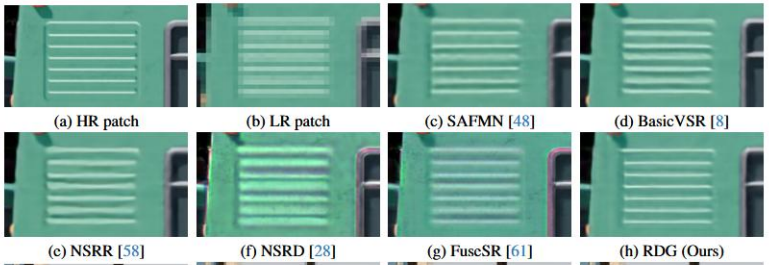
- 回顾：G-Buffer包含深度信息（Depth）、法线信息（Normal）、运动向量（Motion Vector）和双端反射分布函数（Bidirectional Reflectance Distribution Function, BRDF）
- 高频特征增强器（High-frequency Feature Booster，HFB）
  - 主要由交叉门控机制将编码器特征和G-Buffer中的高频信息（法线、双端反射分布函数）进行融合。
- 帧间时序精炼器（Cross-frame Temporal Refiner，CTR）
  - 单凭光流估计的运动向量来传播物体运动信息，当存在物体遮挡或长距离位移时，扭曲后的帧可能会出现错位错误。
  - CTR通过结合前一帧的隐藏状态、运动向量和深度信息，来提升视频恢复的时序一致性

用于实时渲染的高效超分——结果



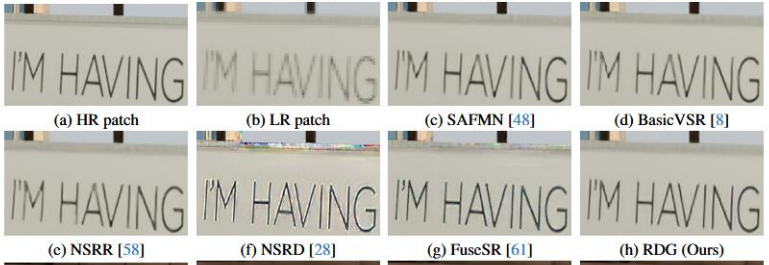
Methods	Efficiency Metrics #Params	#Memory	#Latency	Testing Data (PSNR/SSIM/VMAF) Bar	Bistro	Forest	NYC	Square	Squre-night	Tenshu	ZeroDay	Average
SAFMN [48]	0.240M	49.46M	20.88ms	30.85	25.92	25.77	29.21	26.49	33.01	31.46	31.91	29.33
				0.8192	0.8104	0.6439	0.906	0.8332	0.9048	0.8913	0.901	0.8387
				90.31	79.33	61.32	93.48	90.64	82.49	73.78	81.63	81.62
SMFANet [60]	**0.197M**	56.72M	21.23ms	30.78	25.93	25.74	28.27	26.39	32.81	31.26	31.76	29.12
				0.8164	0.8052	0.6421	0.8848	0.8264	0.9001	0.8860	0.9025	0.8329
				92.27	81.50	62.92	91.04	92.80	87.70	72.58	82.75	82.95
BasicVSR* [8]	1.760M	249.68M	40.34ms	30.84	25.92	25.62	29.32	26.50	32.92	31.35	31.90	29.30
				0.8182	0.8106	0.6356	0.9064	0.8335	0.9025	0.8881	0.9041	0.8374
				93.64	79.62	63.19	93.82	90.29	86.54	75.34	84.64	83.38
BasicVSR++* [9]	2.408M	404.60M	67.12ms	30.96	26.10	25.76	29.87	26.70	33.17	31.64	32.07	29.53
				0.8220	0.8159	0.6462	0.9163	0.8399	0.9084	0.8933	0.9053	0.8716
				93.84	79.66	64.31	96.97	91.04	87.28	76.36	85.65	84.39
NSRR [58]	0.535M	766.97M	66.86ms	30.86	25.93	25.79	29.24	26.48	33.14	31.36	31.94	29.34
				0.8218	0.8125	0.6487	0.9075	0.8333	0.9058	0.8898	0.8985	0.8397
				91.93	77.80	63.70	92.36	88.51	85.64	75.14	83.76	82.35
NSRD [28]	1.61M	166.50M	23.06ms	27.75	22.33	21.33	28.75	23.08	31.50	27.13	28.42	26.29
				0.8287	0.7066	0.5276	0.8133	0.7658	0.8880	0.8881	0.8648	0.7854
				71.29	64.73	54.66	77.32	69.78	65.03	63.99	72.98	67.47
FuseSR [61]	2.247M	198.25M	33.99ms	28.09	24.33	21.94	29.44	24.58	32.88	28.88	31.79	27.74
				0.8228	0.7550	0.5705	0.9100	0.8161	0.8907	0.8830	0.8978	0.8182
				79.59	69.45	54.58	75.58	77.23	78.49	68.34	71.42	71.83
RDG-s (Ours)	0.304M	61.45M	**7.93ms**	**31.03**	26.18	25.74	28.93	26.65	33.19	31.39	31.83	29.37
				0.8274	0.8159	0.6597	0.8989	0.8326	0.9073	0.8891	0.9061	0.8421
				96.29	83.94	70.14	98.36	95.09	89.40	78.75	85.82	87.22
RDG (Ours)	1.474M	161.76M	18.78ms	30.96	**26.32**	**25.93**	**30.20**	**27.02**	**33.21**	**31.70**	**32.16**	**29.69**
				0.8299	**0.8231**	**0.6601**	**0.9204**	**0.8425**	**0.9112**	**0.8937**	**0.9141**	**0.8494**
				98.62	**86.04**	**71.52**	**99.12**	**96.65**	**92.85**	**79.84**	**87.06**	**88.96**

frame 163 from Square



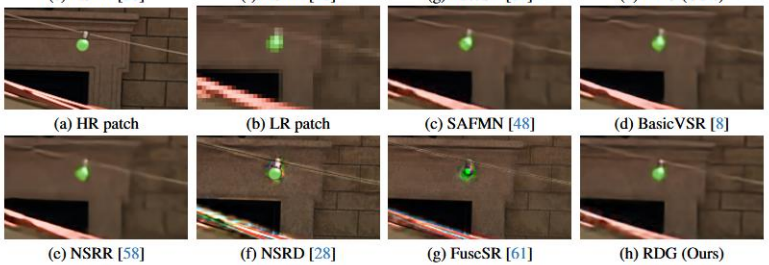
(a) HR patch (b) LR patch (c) SAFMN [48] (d) BasicVSR [8]
(c) NSRR [58] (f) NSRD [28] (g) FuseSR [61] (h) RDG (Ours)

frame 080 from NYC



(a) HR patch (b) LR patch (c) SAFMN [48] (d) BasicVSR [8]
(c) NSRR [58] (f) NSRD [28] (g) FuseSR [61] (h) RDG (Ours)

frame 103 from Bistro



(a) HR patch (b) LR patch (c) SAFMN [48] (d) BasicVSR [8]
(c) NSRR [58] (f) NSRD [28] (g) FuseSR [61] (h) RDG (Ours)

使用可变形注意力金字塔实现快速在线视频超分

Fast Online Video Super-Resolution with Deformable Attention Pyramid

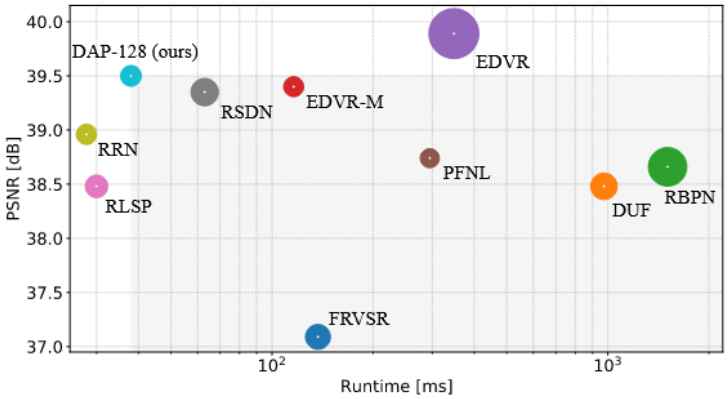
Dario Fuoli[1] Martin Danelljan[1] Radu Timofte[1,2] Luc Van Gool[1,3]

[1]Computer Vision Lab, ETH Zürich, Switzerland [2]CAIDAS, University of Würzburg, Germany [3]KU Leuven, Belgium

{dario.fuoli, martin.danelljan, vangool}@vision.ee.ethz.ch, radu.timofte@uni-wuerzburg.de

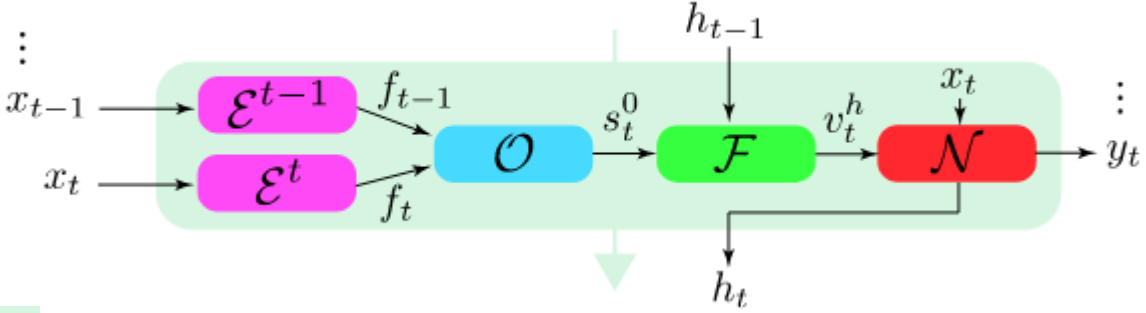
- 发表在WACV 2023的一篇工作。
- 提出了一个可变形的注意力金字塔，解决了传统VSR中帧对齐和信息融合计算成本高的问题，实现了在线视频超分。
- 320*180的视频在2080ti上做4倍超分成720P，帧率可达26.3FPS

使用可变形注意力金字塔实现快速在线视频超分



- 这篇工作的动机：注意力机制能够比卷积更好地捕捉全局信息，但二次复杂度对于在线视频超分的任务来说难以忍受。
- 作者受光流估计的启发，认为只有发生了明显位移的像素区域需要计算注意力。

使用可变形注意力金字塔实现快速在线视频超分



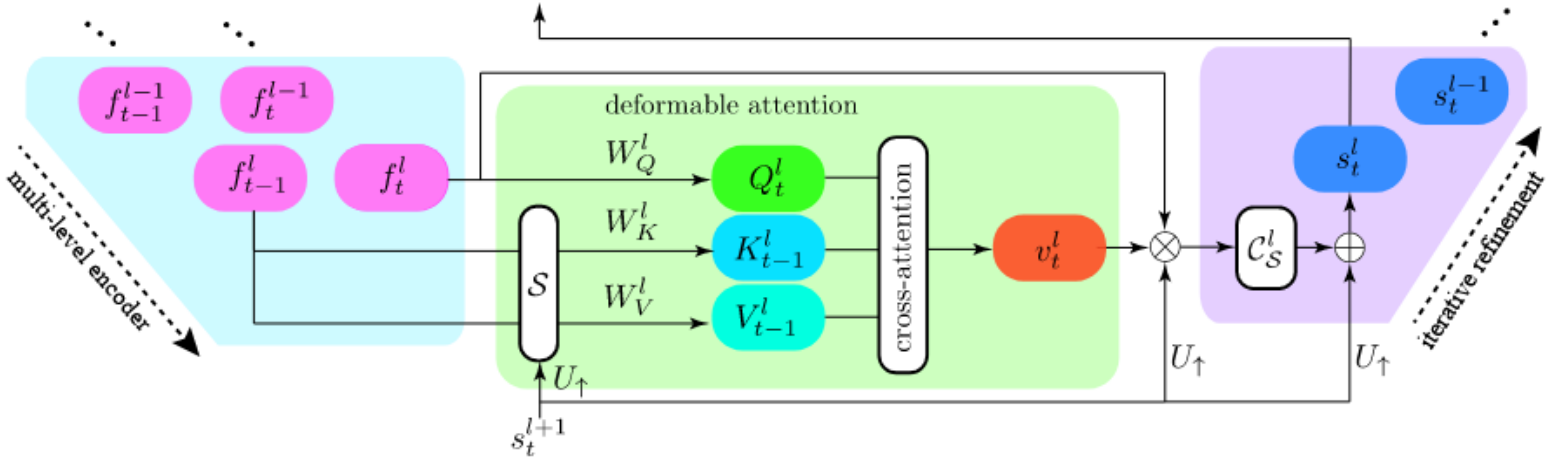
$\mathcal{E}^{t-1}$ UNet编码器。提取前一帧的空间信息。

$\mathcal{O}$ 可变形偏移量预测模块。用于预测相邻两帧间的偏移量s_t^0。

$\mathcal{F}$ 信息融合模块。根据帧间的偏移量和历史状态，通过交叉注意力完成特征融合。

$\mathcal{N}$ 残差卷积处理模块。它用于对融合后的特征进行进一步处理，最终生成超分辨率的输出帧 y_t，并更新当前的隐藏状态 h_t 为下一个时间步 t+1做准备。

使用可变形注意力金字塔实现快速在线视频超分



- 主要通过 $\mathcal{S}$ 预测相邻两帧发生偏移的位置，只计算这些位置的交叉注意力，从而大大降低计算成本。

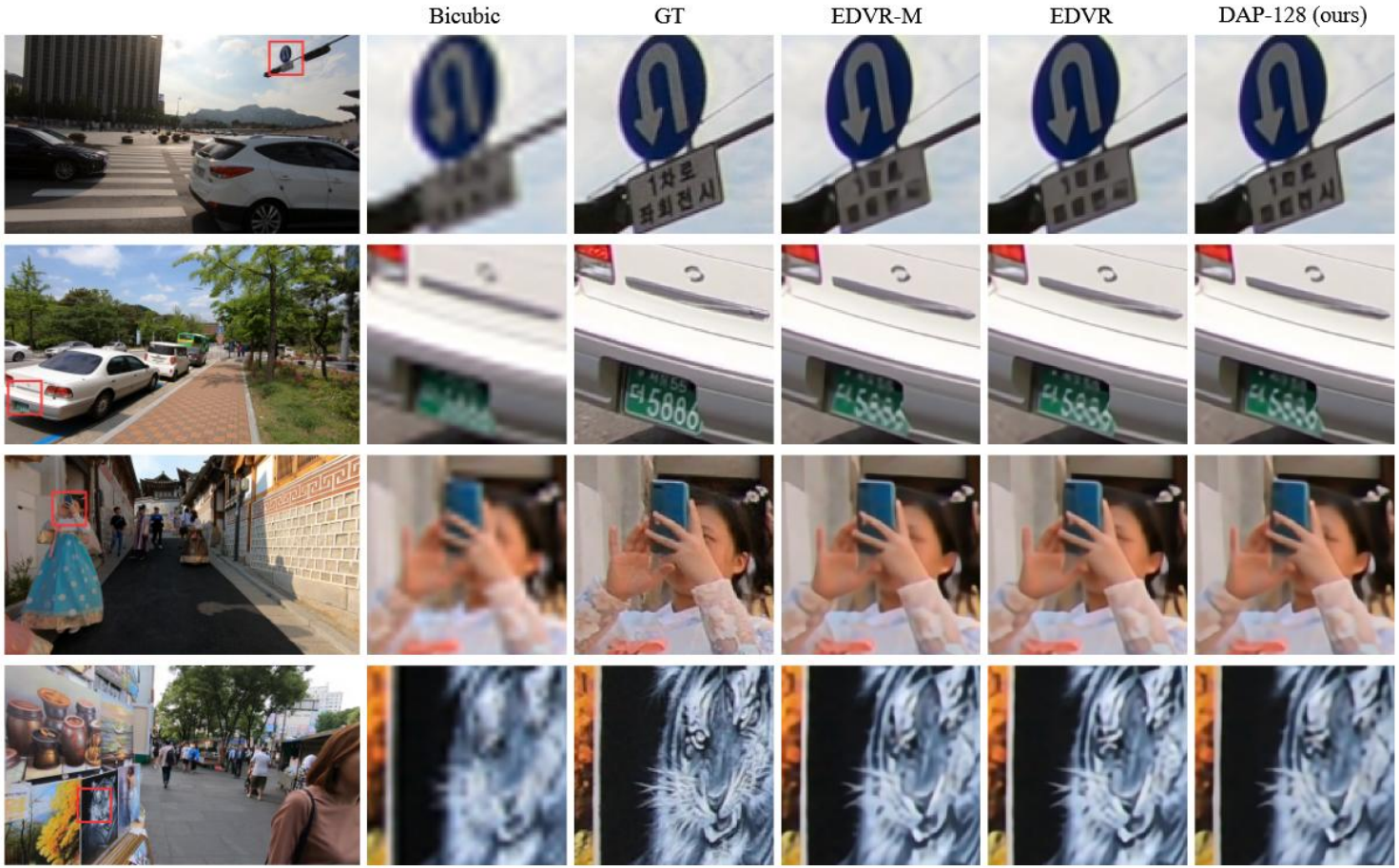
使用可变形注意力金字塔实现快速在线视频超分——消融实验

Configuration	Offsets	Pyramid	Attention	Features	REDS4val [19] (Y)	REDS4 [19] (RGB)
1				64	28.77/0.7906	28.59/0.8155
2			✓	64	28.95/0.7926	28.69/0.8184
3	✓			64	29.82/0.8194	29.50/0.8461
4	✓	✓		64	30.07/0.8264	29.66/0.8507
5	✓	✓	✓	64	30.36/0.8341	29.97/0.8571
6	✓	✓	✓	128	30.77/0.8440	30.49/0.8676

使用可变形注意力金字塔实现快速在线视频超分——结果

Method	Unid.	Onl.	R-T.	Run [ms]	fps [1/s]	FLOPs [G]	MACs [G]	REDS4[19] PSNR/SSIM	UDM10[33] PSNR/SSIM	Vimeo-90K[32] PSNR/SSIM
Bicubic	✓	✓	✓	-	-	-	-	26.14/0.7292	28.47/0.8253	31.30/0.8687
TOFlow [31]	✓	✗	✗	-	-	-	-	27.98/0.7990	36.26/0.9438	34.62/0.9212
FRVSR [22]	✓	✓	✗	*137	*7.3	-	-	-	37.09/0.9522	35.64/0.9319
DUF [15]	✓	✗	✗	*974	*1.0	-	-	28.63/0.8251	38.48/0.9605	36.87/0.9447
RBPN [8]	✓	✓	✗	*1507	*0.7	-	-	30.09/0.8590	38.66/0.9596	37.20/0.9458
PFNL [33]	✓	✗	✗	*295	*3.4	-	-	29.63/0.8502	38.74/0.9627	-
MuCAN [18]	✓	✗	✗	2'208	0.5	15'853.2	7'922.8	30.88/0.8750	-	-
EDVR-M [28]	✓	✗	✗	116	8.6	925.7	462.3	30.53/0.8699	39.40/0.9663	37.33/0.9484
EDVR [28]	✓	✗	✗	348	2.9	4'037.3	2'017.3	31.09/0.8800	39.89/0.9686	37.81/0.9523
TGA [12]	✓	✗	✗	427	2.3	-	-	-	-	37.59/0.9516
RSDN [11]	✓	✓	✗	63	15.9	713.2	356.3	-	39.35/0.9653	37.23/0.9471
RRN [13]	✓	✓	✓	28	35.7	387.5	193.6	-	38.96/0.9644	-
RLSP [6]	✓	✗	✓	30	33.3	503.7	251.8	-	38.48/0.9606	36.49/0.9403
DAP-128 (ours)	✓	✓	✓	38	26.3	330.0	164.8	30.59/0.8703	39.50/0.9664	37.29/0.9476
BasicVSR [2]	✗	✗	✗	82	12.2	754.3	376.7	31.42/0.8909	39.96/0.9694	37.53/0.9498
IconVSR [2]	✗	✗	✗	100	10.0	904.9	451.9	31.67/0.8948	40.03/0.9694	37.84/0.9524
BasicVSR++ [3]	✗	✗	✗	110	9.1	837.1	418.1	32.39/0.9069	40.72/0.9722	38.21/0.9550

使用可变形注意力金字塔实现快速在线视频超分——结果



通过T2V模型实现真实世界视频超分

STAR: Spatial-Temporal Augmentation with Text-to-Video Models for Real-World Video Super-Resolution

Rui Xie[1*], Yinhong Liu[1*], Penghao Zhou[2], Chen Zhao[1], Jun Zhou[3]
Kai Zhang[1], Zhenyu Zhang[1], Jian Yang[1], Zhenheng Yang[2], Ying Tai[1†]
[1]Nanjing University, [2]ByteDance, [3]Southwest University

https://nju-pcalab.github.io/projects/STAR

南京大学和字节的一篇工作，2025.1.6挂在Arxiv上。首篇将T2V的扩散先验集成到VSR任务中的工作，针对伪影/闪烁和保真度两方面展开的研究，针对伪影在全局注意力前加了局部注意力，针对保真度提出了一个动态频率损失。

现有挑战：

- 许多 VSR 方法仅针对特定的已知退化（如下采样），但真实视频常伴随噪声、模糊、压缩等复杂退化，导致恢复效果较差。
- 现有的扩散模型主要基于静态图像训练，仅加入时间模块并不足以确保帧间平滑过渡，会导致闪烁等问题。
- T2V 模型（如 CogVideoX-5B）虽然可以用于 VSR，但其强大的生成能力可能会损害恢复的保真度，即输出的画面过于“合成”而非真实恢复。

整体架构

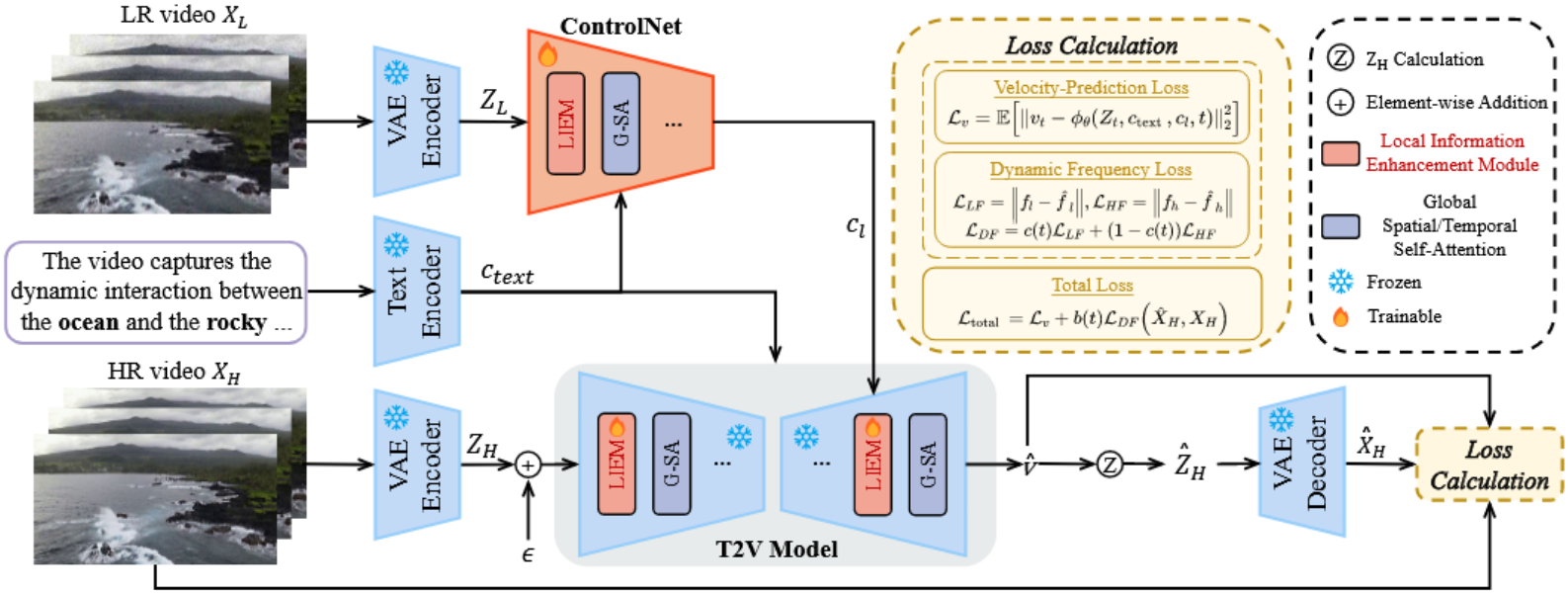
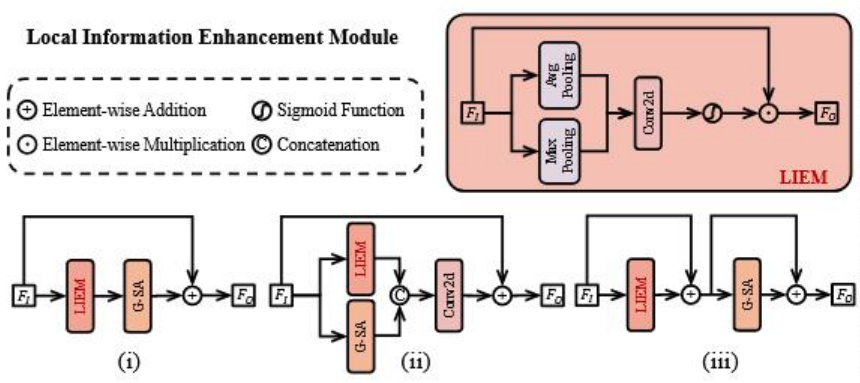


Figure 2. Overview of the proposed STAR.

训练集是OpenVid-1M，包括~200K的文本视频对，最低像素512*512，平均时长7.2s

- 模型由四部分构成：VAE、Text Encoder、ControlNet和集成了LIEM的T2V模型

局部信息增强模块(LIEM)



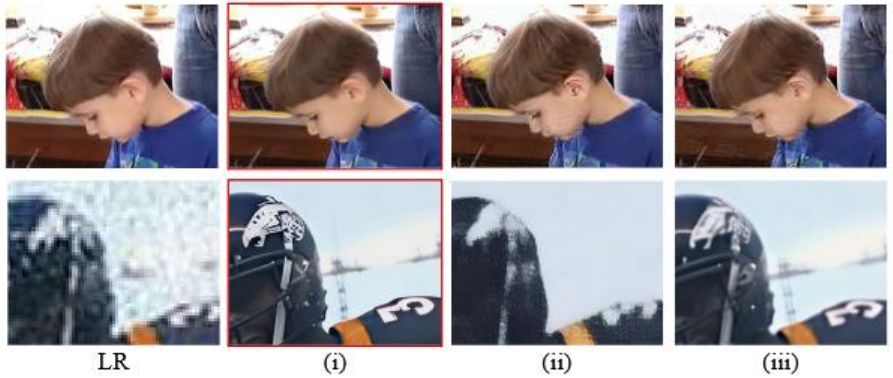
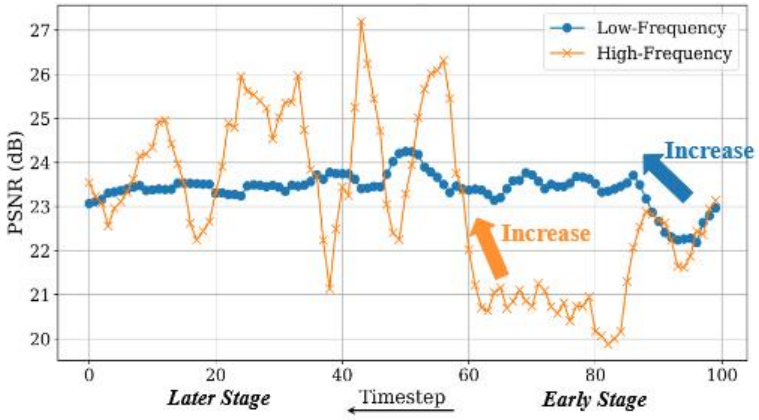


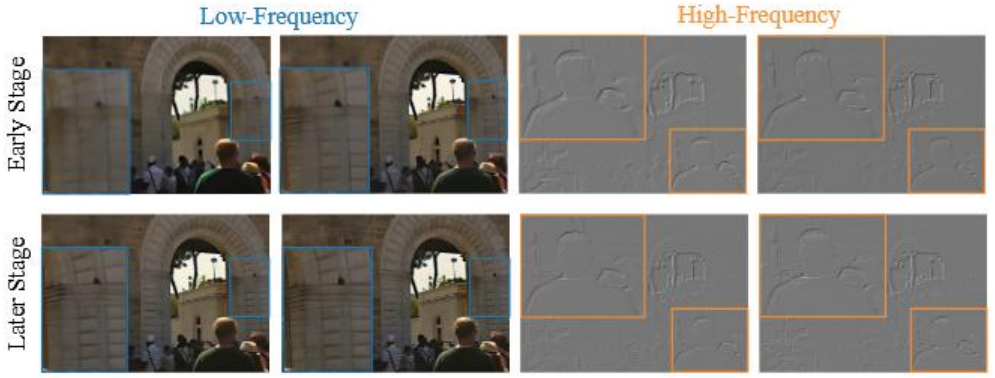
Table 3. Ablation of LIEM position.

Position	Spa-Local	Temp-Local	PSNR↑	LPIPS↓	E^*_{warp} ↓
(i)	✓		23.14	0.2015	2.83
		✓	23.61	0.2013	2.82
	✓	✓	23.65	0.1945	2.92
(ii)			23.69	**0.1943**	2.74
(iii)			23.27	0.2363	3.57
			24.51	0.2094	**1.99**

探究在不同网络层加入 LIEM的恢复效果：在全局注意力前加入 LIEM 效果最佳，既能减少降质伪影，又能增强细节。

动态频率损失(DFL)





$$\mathcal{L}_{LF} = \|f_l - \hat{f}_l\|, \mathcal{L}_{HF} = \|f_h - \hat{f}_h\|,$$

$$\mathcal{L}_{DF} = c(t)\mathcal{L}_{LF} + (1 - c(t))\mathcal{L}_{HF},$$

动机：基于观察，早期阶段模型主要恢复低频信息，后期阶段主要恢复高频信息

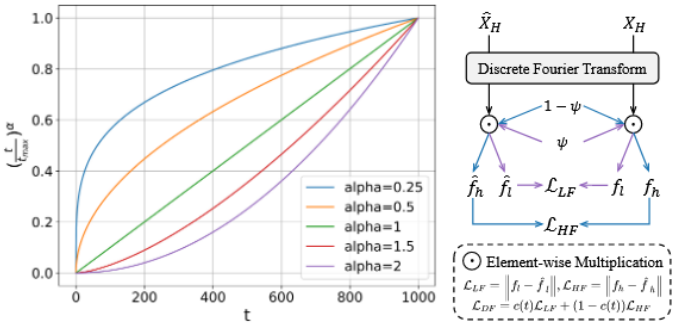


Figure 5. **Dynamic Frequency Loss. Left**: curves of weighting function $c(t)$ for different α. **Right**: details of DF loss.

Table 4. Ablation of different variants of DF loss.

Seperate	Type	PSNR↑	LPIPS↓	E^*_{warp} ↓
w/o Frequency Loss		23.69	0.1943	2.74
-	-	<u>23.72</u>	<u>0.1941</u>	<u>2.71</u>
✓	Inverse	23.67	0.1945	2.83
✓	Direct	**23.85**	**0.1903**	**2.69**

T2V 模型规模对比

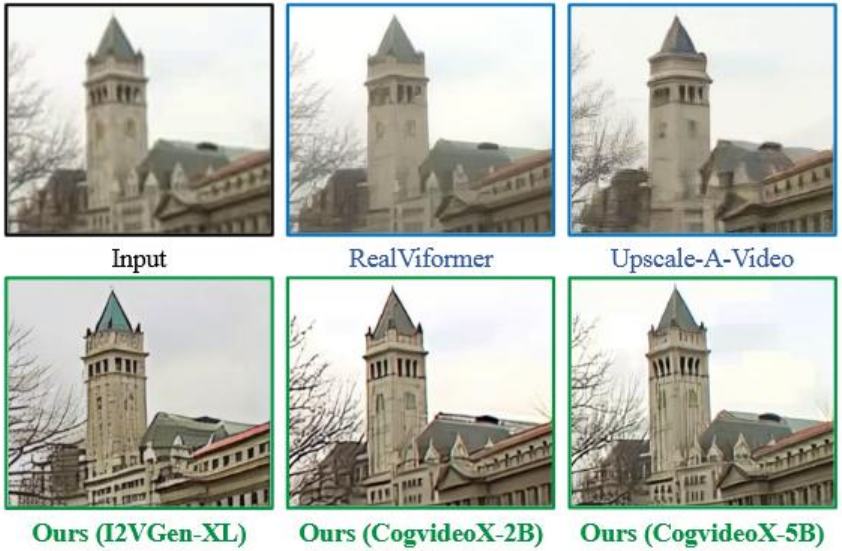


Figure 10. Illustration on scaling up with larger t2v models on a real-world low-quality video. **(Zoom-in for best view)**

Table 6. Effectiveness of T2V diffusion prior for real-world VSR.

Metrics	UAV	RealViformer	Ours		
			I2VGen-XL	CogX-2B	CogX-5B
PSNR↑	22.46	22.90	21.46	<u>23.18</u>	**23.60**
SSIM↑	0.6552	0.6944	0.6715	<u>0.7112</u>	**0.7400**
LPIPS↓	0.2035	0.1823	0.1779	<u>0.1571</u>	**0.1314**
DOVER↑	0.6609	0.4286	<u>0.7267</u>	0.6955	**0.7350**
E^*_{warp} ↓	5.424	4.75	5.529	**3.68**	<u>4.56</u>

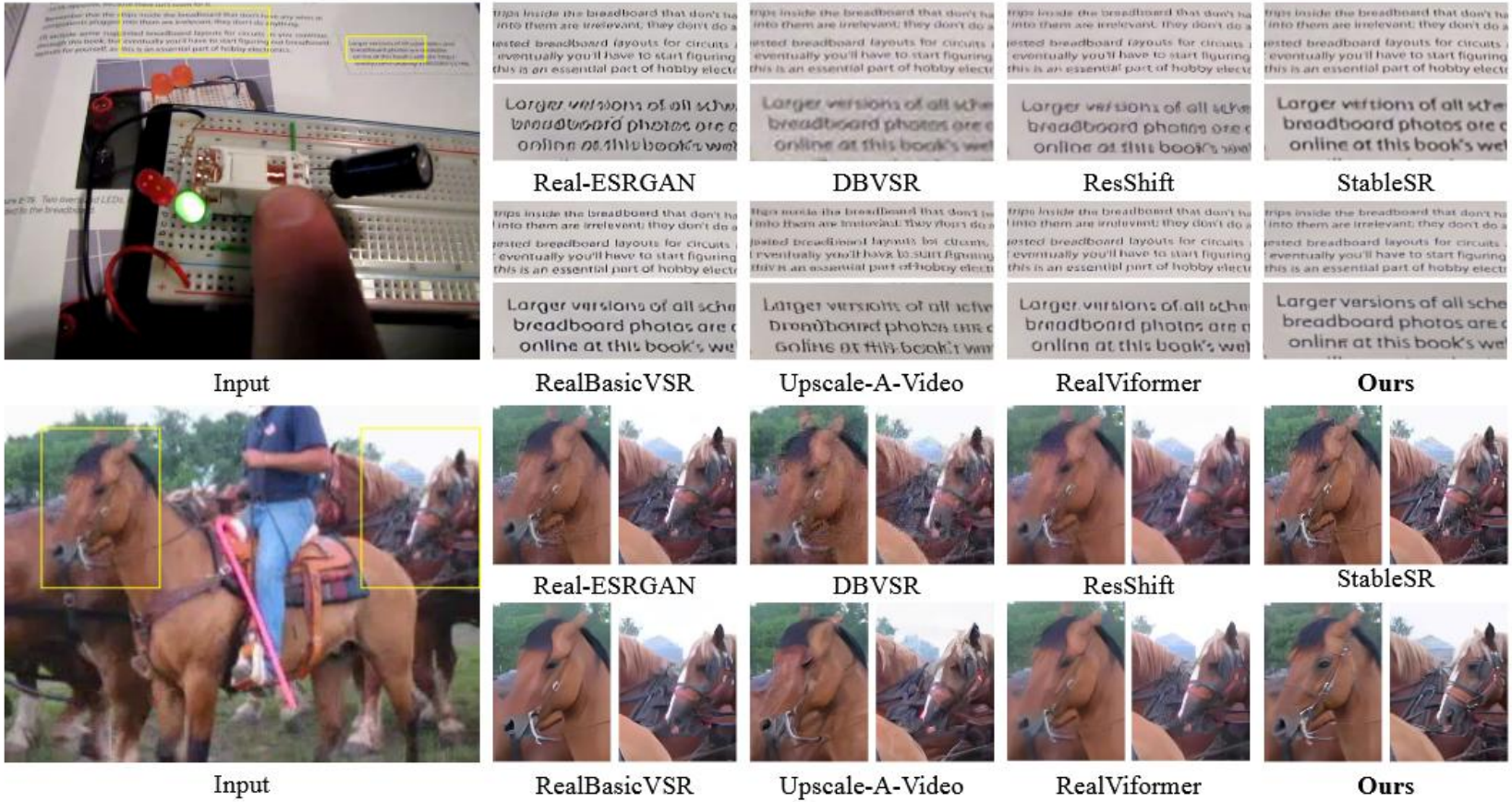
受Scaling Law启发：用更大的T2V模型能得到更好的恢复性能

结果

Table 1. Quantitative evaluations on diverse VSR benchmarks from synthetic (UDM10, REDS30, OpenVid30) and real-world (VideoLQ) sources. The best performance is highlighted in **bold**, and the second-best in underlined. E^*_{warp} refers to E_{warp} ($\times 10^{-3}$).

Datasets	Metrics	Real-ESRGAN ICCVW 2021	DBVSR ICCV 2021	RealBasicVSR CVPR 2022	RealViformer ECCV 2024	ResShift NeurIPS 2023	StableSR IJCV 2024	Upscale-A-Video CVPR 2024	MGLDVSR ECCV 2024	Ours -
UDM10	PSNR↑	22.41	19.65	23.64	**24.00**	22.90	23.50	21.29	23.74	23.91
	SSIM↑	0.6476	0.4747	0.6842	0.6896	0.5451	0.6599	0.5967	0.6826	**0.7164**
	LPIPS↓	0.2769	0.4566	0.2514	0.2325	0.4036	0.2785	0.3006	0.2195	**0.1885**
	DOVER↑	0.4831	0.0959	0.5039	0.5055	0.3252	0.3490	0.5309	0.4896	**0.5422**
	E^*_{warp} ↓	11.17	12.56	5.14	3.57	12.69	8.89	2.83	6.03	**2.68**
REDS30	PSNR↑	19.56	14.85	20.85	**20.86**	19.93	20.32	19.71	20.57	20.29
	SSIM↑	0.4862	0.2941	**0.5469**	0.5377	0.4261	0.5043	0.4315	0.5113	0.5411
	LPIPS↓	0.3376	0.5915	0.2899	0.2597	0.4422	0.3857	0.3443	**0.2240**	0.2804
	DOVER↑	0.3182	0.0600	0.3483	0.3400	0.2221	0.2519	0.2857	0.3857	**0.4017**
	E^*_{warp} ↓	19.1	18.00	8.32	**6.06**	17.40	22.14	15.65	12.28	7.30
OpenVid30	PSNR↑	24.62	21.14	24.63	**26.21**	24.29	24.91	24.41	24.73	25.30
	SSIM↑	0.7778	0.5887	0.7759	0.8080	0.6070	0.7633	0.7167	0.7686	**0.8371**
	LPIPS↓	0.1994	0.4207	0.2297	0.1881	0.3902	0.2102	0.2479	0.2074	**0.1011**
	DOVER↑	0.6992	0.1819	0.7345	0.7275	0.5435	0.6368	0.7201	0.7191	**0.7393**
	E^*_{warp} ↓	8.46	12.11	4.12	2.52	9.78	8.87	4.72	4.82	**1.82**
VideoLQ	ILNIQE↓	27.95	27.19	26.29	26.11	25.92	29.97	24.49	**23.94**	25.35
	DOVER↑	0.4967	0.3392	0.5285	0.4804	0.4113	0.4775	0.4833	0.5319	**0.5431**
	E^*_{warp} ↓	8.00	7.75	6.52	**5.10**	8.33	9.26	10.89	7.82	6.38

结果


Real-ESRGAN
DBVSR
ResShift
StableSR
Input
RealBasicVSR
Upscale-A-Video
RealViformer
Ours
Real-ESRGAN
DBVSR
ResShift
StableSR
Input
RealBasicVSR
Upscale-A-Video
RealViformer
Ours

THANKS
请各位老师同学批评指正

1924-2024
中山大學 世纪华诞
100th ANNIVERSARY
SUN YAT-SEN UNIVERSITY