LAVIB: A Large-scale Video Interpolation Benchmark

Alexandros Stergiou

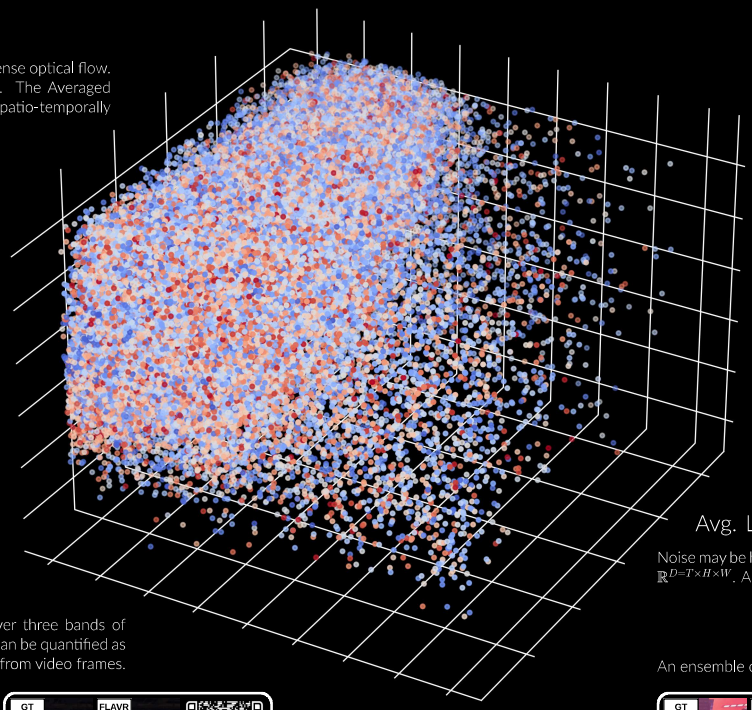
UNIVERSITY OF TWENTE.

NEURAL INFORMATION PROCESSING SYSTEMS

4K 17.2K videos | 283K video segments | 77.6 hours | 17M frames

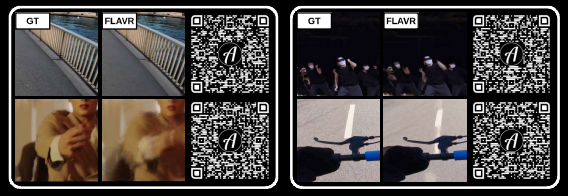
Avg. Flow Magnitude

Motion magnitudes are quantified with dense optical flow. FlowFormer is used on each frame pair. The Averaged Flow Magnitude (AFM) is defined as the spatio-temporally averaged optical flow.



Avg. Relevant Luminance

The perception of luminance is done over three bands of color. Average Relative Luminance (ARL) can be quantified as the weighted sum for each color channel from video frames.



Symbol cheat sheet

low → high	high → low
(AFM)	(AFM)
(ALV)	(ALV)
(ARMS)	(ARMS)
(ARL)	(ARL)

Avg. RMS Contrast

Contrast relates to differences in raw pixel values. *Average Root Mean Square* (ARMS) is used to calculate the difference between each pixel from each frame of $\mathbf{V}$ and the corresponding pixel in the channel-averaged $\overline{\mathbf{V}}$.

$$\text{ARMS}(\mathbf{V}) = \frac{1}{T}\sum_{t\in\mathbb{R}^T}\sqrt{\frac{1}{HW}\sum_{s\in\mathbb{R}^{HW}}(\mathbf{V}_{t,s}-\overline{\mathbf{V}}_{t,s})}$$

Avg. Laplacian Variance

Noise may be highlighted or suppressed by different sensors capturing video $\mathbf{V}$ of dimensions $\mathbb{R}^{D=T\times H\times W}$. ALV is formulated by applying Laplacian of Gaussians (LoG) and averaging:

$$\text{ALV}(\mathbf{V},\sigma,K) = \frac{1}{D}\sum_{r\in\mathbb{R}^D}\sum_{i=1}^{K}\sum_{j=1}^{K}\underbrace{-1\frac{1}{\pi\sigma^4}(1-\frac{i^2+j^2}{2\sigma^2})e^{-\frac{i^2+j^2}{2\sigma^2}}}_{\text{LoG}(i,j)}\mathbf{V}_{r-[i,j]}$$

An ensemble of kernels is used to calculate the final value $\frac{1}{|K|}\sum_{K\in\mathcal{N}}\text{ALV}(\mathbf{V},\sigma,K)$ with $\sigma = 1.4$.

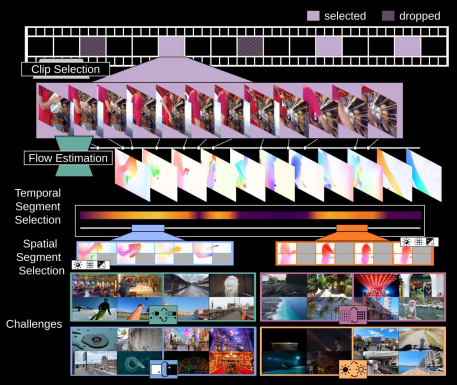


Figure 1. **LAVIB segment selection and challenges pipeline.** Tubelets are made from clips based on embeddings, AFM, ARL, ALV, and ARMS.

Table 1. **Multi-frame interpolation scores** over triplets, and septuplets. Video duration increase is denoted with $\{\times 2, \times 3, \times 4\}$.

Model	triplet			septuplets		
	×2	×3	×4	×2	×3	×4
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Baseline	32.10/0.963	31.58/0.952	30.42/0.937	32.69/0.976	32.10/0.972	31.95/0.918
FLAVR	33.44/0.981	33.07/0.975	32.86/0.968	33.62/0.985	33.41/0.980	33.28/0.962

Table 2. **PSNR and SSIM scores on OOD challenges.** Flow-based challenges are denoted with / , / for blur, / for contrast, and / for luminance.

(a) AFM

Model	→		→	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
RIFE	25.34	0.832	28.75	0.926
EMA-VFI	30.21	0.936	34.89	0.929
FLAVR	**30.67**	**0.959**	**35.66**	**0.991**

(b) ALV

Model	→		→	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
RIFE	26.52	0.873	29.31	0.906
EMA-VFI	31.26	0.948	34.30	0.972
FLAVR	**31.78**	**0.962**	**34.67**	**0.975**

(c) ARMS

Model	→		→	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
RIFE	26.28	0.836	25.42	0.855
EMA-VFI	32.79	0.964	30.65	0.951
FLAVR	**33.02**	**0.982**	**31.11**	**0.977**

(d) ARL

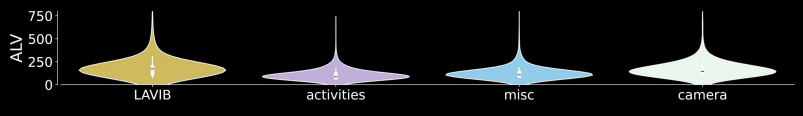
Model	→		→	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
RIFE	26.83	0.872	26.95	0.865
EMA-VFI	33.55	0.974	33.41	0.968
FLAVR	**33.97**	**0.980**	**34.20**	**0.976**

Datasets. Compared to prior efforts, LAVIB provides a large-scale general-purpose dataset of standardized 4K 60 fps videos. It features a significant variance across Average Flow Magnitude (AFM), Average Relevant Luminance (ARL), and Average Laplacian Variances (ALV) in videos.

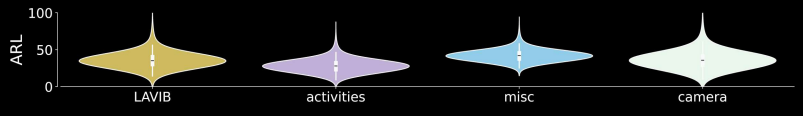
Dataset	Dataset statistics		Video statistics		Average video metrics		
	Year	Tot. Mins	Tot. Vids	Src Res.	FPS	AFM	ARL	ALV
UCF101	2012	1,600	13,320	240p	25	2.43 ± 1.85	53.37 ± 13.42	53.99 ± 18.37
Xiph	2020	4	19	**2160p**	60	26.21 ± 25.19	60.64 ± 10.77	95.24 ± 62.32
Inter4K	2021	83	1,000	**2160p**	60	56.38 ± 14.34	56.79 ± 14.48	25.05 ± 24.05
X4K1KFPS	2021	191	4,423	**2160p**	**960**	266.87 ± 178.72	53.95 ± 12.07	135.67 ± 78.19
Vimeo90K	2017	356	91,701	720p	30	49.63 ± 18.32	59.68 ± 20.89	26.26 ± 29.25
LAVIB (ours)	2024	**4,660**	**283,484**	**2160p**	60	63.10 ± 58.41	38.34 ± 28.69	199.78 ± 197.79

Vocabulary Three core components are used for creating search terms from the vocabulary to source 4K YouTube videos; locations, activities, or specific objects/settings relevant to videos. Locations and activities include two levels of hierarchies. The structure of search terms changes based on the selected sub-group.

ALV distributions for all LAVIB and videos from activities, misc, and camera queries.



ARL distributions for all LAVIB and videos from activities, misc, and camera queries.



Project page: