A Perceptual Quality Metric for Video Frame Interpolation

Portland State University

Qiqi Hou, Abhijay Ghildyal, and Feng Liu

ECCV TEL AVIV 2022

Overview

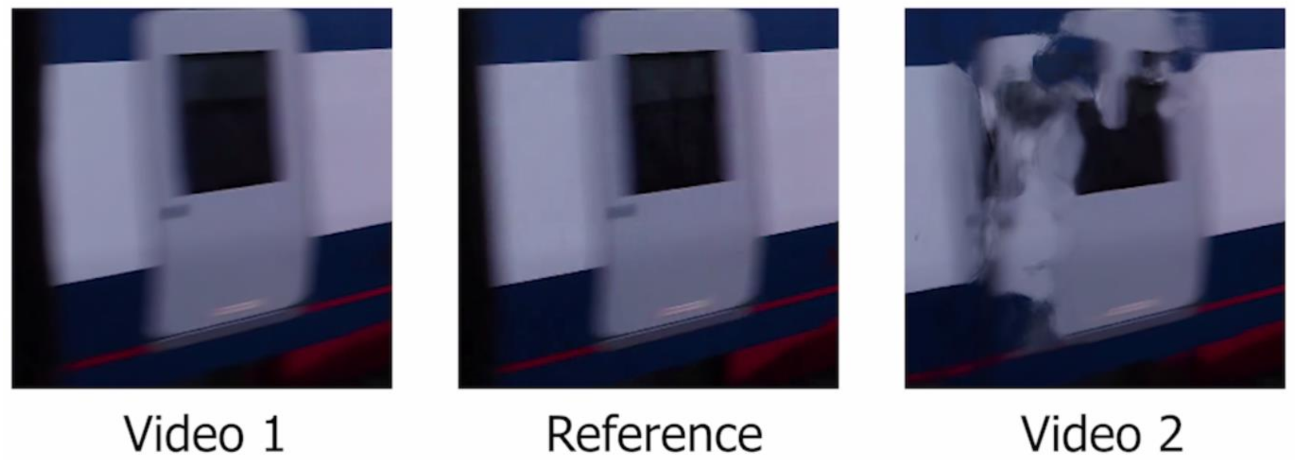
As video frame interpolation results often exhibit unique artifacts, existing quality metrics sometimes are not consistent with human perception.

Contributions
1) Provide the first video perceptual similarity metric dedicated to video frame interpolation,
2) Design a novel neural network architecture for video perceptual quality assessment based on the Swin Transformers,
3) Build a large video frame interpolation perceptual similarity dataset.

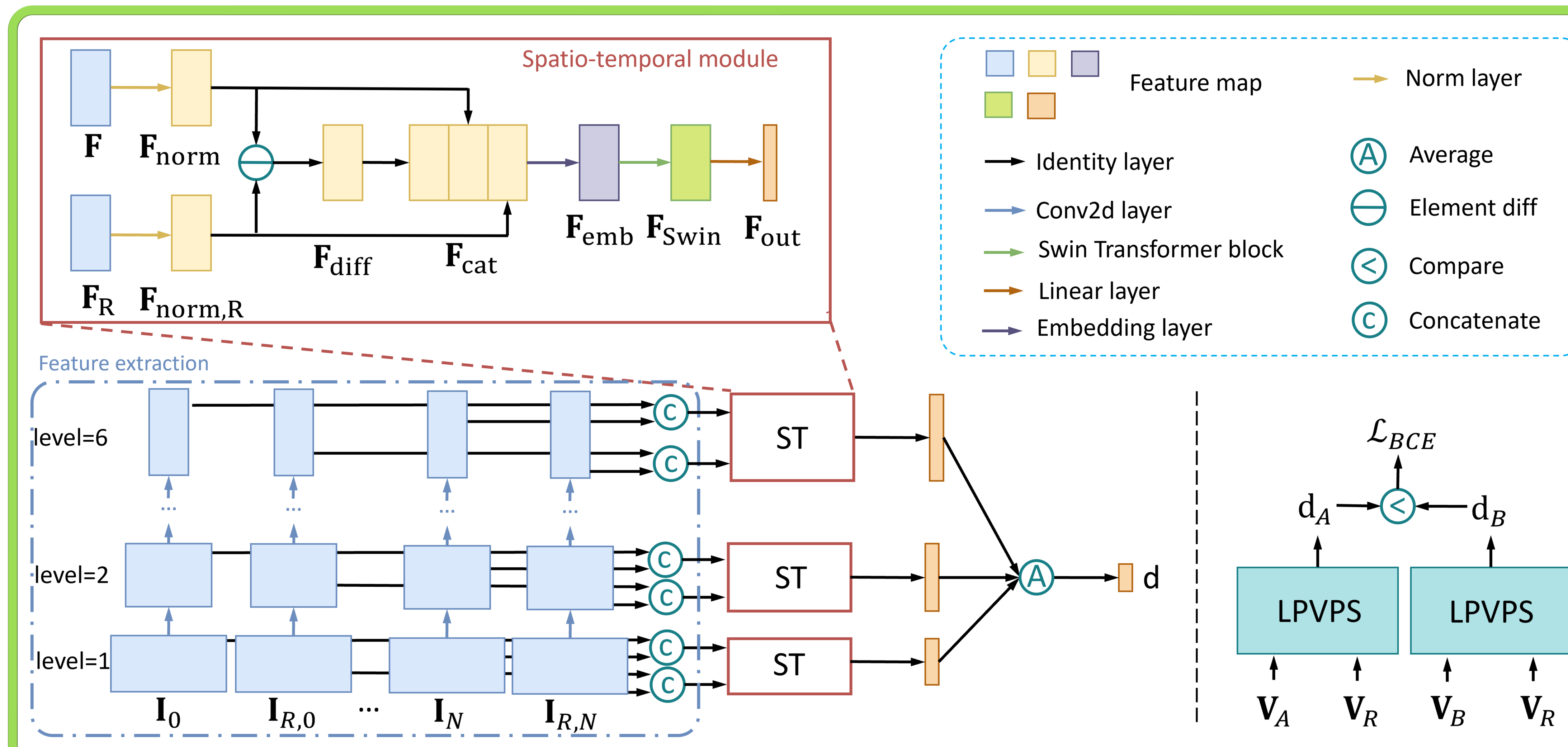
Unique distortions in video frame interpolation results.

Video Frame Interpolation Quality Dataset

We collected a Video Frame Interpolation Perceptual Similarity (VFIPS) dataset. It contains 25,887 samples.

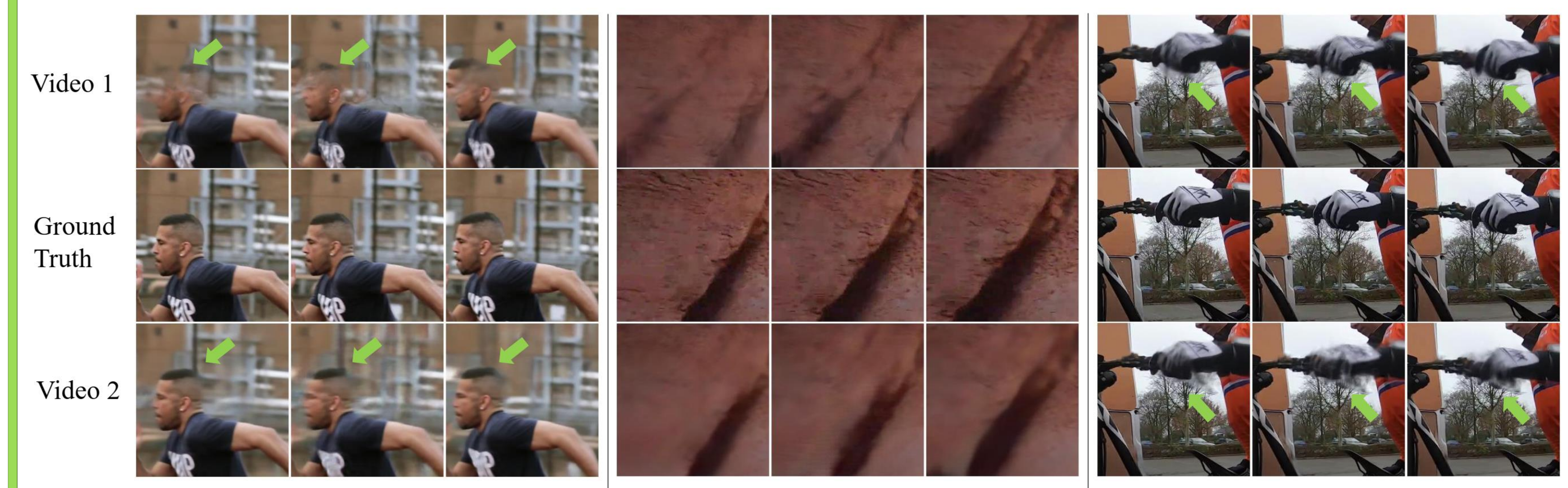


Each sample consists of two videos synthesized from different interpolation methods, the reference video, and its perceptual judgments.



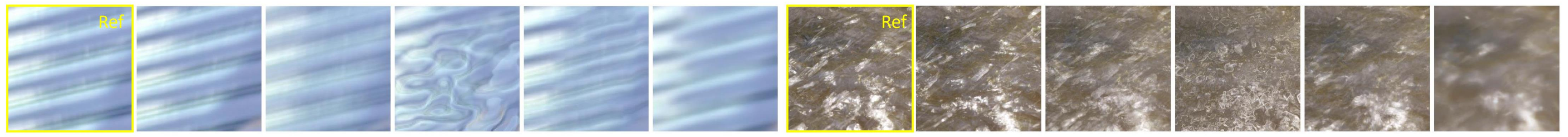
Our network architecture. Learned Perceptual Video Patch Similarity (LPVPS) takes a video V and its reference V_R as input and predicts their perceptual similarity d.

Visual Comparison



	Human	LPIPS	DISTS	PIM1	VMAF	Ours	Human	LPIPS	DISTS	PIM1	VMAF	Ours	Human	LPIPS	DISTS	PIM1	VMAF	Ours
Video 1		✓	✓	✓	✓			✓	✓	✓	✓		✓					✓
Video 2	✓					✓	✓					✓		✓	✓	✓	✓	

Visual examples on the VFIPS dataset. Green arrows are used to label the area with noticeable difference. We mark the preference of each method using ✓. Compared to other methods, our method is consistent with humans.



Human	4th	3rd	5th	1st	2nd	Human	4th	1st	5th	3rd	2nd
VMAF	4th	3rd	5th	2nd	1st	VMAF	1st	2nd	3rd	4th	5th
LPIPS	3rd	4th	5th	1st	2nd	LPIPS	2nd	1st	5th	3rd	4th
Ours	4th	3rd	5th	1st	2nd	Ours	4th	1st	5th	3rd	2nd

Visual examples on the BVI-VFI dataset. Yellow rectangles are used to show the reference video. We report the rank for the distorted videos.

Experiments

Table 1: Comparison with state-of-the-art methods.

	Method	VFIPS (val.) 2AFC	BVI-VFI [19] (test) SROCC	PLCC	KROCC
Image	PSNR	0.763	0.742	0.722	0.656
	SSIM [79]	0.784	0.739	0.746	0.639
	MS-SSIM [80]	0.794	0.772	0.789	0.667
	LPIPS (VGG) [97]	0.808	0.628	0.796	0.517
	DISTS [21]	0.801	0.597	0.763	0.517
	PIM-1 [6]	0.787	0.492	0.668	0.428
	Watson-DFT [17]	0.800	0.628	0.706	0.538
Video	STRRED [7]	0.777	0.614	0.682	0.539
	VMAF [69]	0.805	0.583	0.614	0.483
	DeepVQA [77]	0.588	0.369	0.271	0.300
	VSFA [41]	0.660	0.108	0.486	0.050
	Ours	**0.830**	**0.794**	**0.870**	**0.700**

Our method outperforms the state-of-the-art methods by a large margin in the VFIPS dataset and the BVI-VFI dataset.

Table 2: Comparison on the X-TEST(4K) dataset [73].

Method	PSNR	SSIM [79]	MSSSIM [80]	LPIPS [97]	STRRED [7]	VMAF [69]	Ours
2AFC	0.752	0.637	0.737	0.748	0.722	0.735	**0.789**

Our method outperforms the state-of-the-art methods on the high-resolution-large-motion X-TEST(4K) dataset.

Ablation Study

We examine our feature extractor, our spatio-temporal module, and the LPIPS-annotated training examples.

Effectiveness of the feature extractor

Extractor	SROCC	PLCC	KROCC	Param(M)	Runtime(ms)
AlexNet [37]	0.761	0.832	0.650	14.5	12.8
I3D [10]	0.659	0.758	0.550	20.3	33.2
Ours-3D	0.728	0.738	0.639	8.6	13.6
Ours-2D	**0.794**	**0.870**	**0.700**	**4.6**	**10.4**

Effectiveness of the ST module

ST Module	SROCC	PLCC	KROCC
None	0.617	0.663	0.539
Conv3D	0.761	0.819	0.661
Original Swin	0.728	0.766	0.639
Ours-Swin w. LN	0.724	0.746	0.611
Ours-Swin	**0.794**	**0.870**	**0.700**

Effectiveness of the Annotations

Annotations	SROCC	PLCC	KROCC
Human	0.719	0.753	0.611
Automatic	0.653	0.687	0.567
All	**0.794**	**0.870**	**0.700**

Our code and models can be downloaded at: https://github.com/hqqxyy/VFIPS