Supplementary Materials for CLIP-Fusion: A Spatio-Temporal Quality Metric for Frame Interpolation

Göksel Mert Çökmez[1], Yang Zhang[2], Christopher Schroers[2], Tunç Ozan Aydın[2]

[1]ETH Zürich, [2]Disney Research | Studios

mert.coekmez@alumni.ethz.ch,{yang.zhang,christopher.schroers,tunc.aydin}@disneyresearch.com

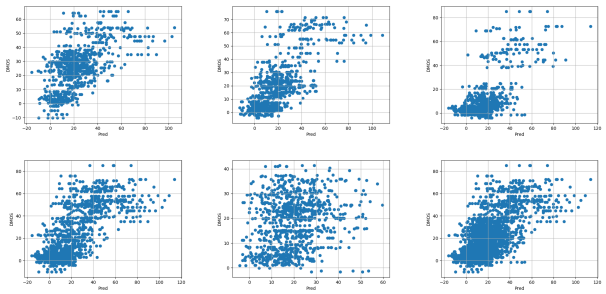

Figure 1. Scatter plot detailing the relation between the ground truth DMOS and our model's prediction when trained on cross-validation of BVI-VFI dataset in a **full-reference** setting. First row from left to right: 30fps, 60fps, 120fps. Second row from left to right: DL, non-DL and Overall.

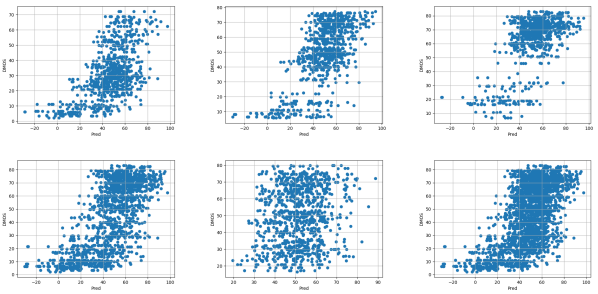

Figure 2. Scatter plot detailing the relation between the ground truth MOS and our model's prediction when trained on cross-validation of BVI-VFI dataset in a **no-reference** setting. First row from left to right: 30fps, 60fps, 120fps. Second row from left to right: DL, non-DL and Overall.

1. Additional details on experiments

1.1. Experiment results revisited

In our paper, we included cross-validation results in the BVI-VFI dataset for our model in full- and no-reference settings. This includes the PLCC and SRCC values on *30fps*, *60fps*, *120fps*, *non-DL*, *DL* and *Overall* categories. As a supplement to these results, we also provide scatter plots for each of these categories to better visualize the correlation between our model's predictions and the DMOS values provided by human participants. The linear trend can be observed in Fig. 1 and Fig. 2, the scatter plots for the full- and no-reference settings respectively.

1.2. Cross-dataset evaluation on VFIPS and BVI-VFI

In addition to cross-validating the candidate models on the BVI-VFI dataset, the BVI-VFI paper also performs a cross-dataset evaluation for each candidate model, meaning that the models are not trained on any subset of the BVI-VFI dataset. For our additions to the list in Tab. 1, namely *LPVPS* [2], we use the publicly available pre-trained LPVPS weights for LPVPS results. Namely *Ours*, we also train our model on the VFIPS dataset [2] to utilize the same settings as LPVPS, except for the training epochs.

The experiment is conducted in both full- and no-reference settings. Full-reference results displayed in Tab. 1 indicate that deep learning-based models such as our model and ST-GREED [4] yield a mediocre performance when they are not trained specifically for the task at hand. Following this trend, all models, including ours, perform poorly in a no-reference setting as seen in Tab. 2. Our model underperformed that, unlike the performance when cross-validating on the BVI-VFI dataset.

Our interpretation of these findings (in Tab. 1 and Tab. 2) is that the datasets contain significant domain gaps due to differing methodologies that have been utilized during subjective data collection. Specifically, the BVI-VFI dataset employs the Double Stimulus Continuous Quality Scale (DSCQS) methodology, while the VFIPS dataset uses Two Alternative Forced Choice (2AFC) scores. This difference in data collection methods (domain gap) between the two datasets hinders the investigation of the models' generalizability.

Table 1. The performance of **cross-dataset** evaluation (full-reference quality metrics). Our model and LPVPS are trained on the VFIPS dataset and tested over all the DMOS values in the BVI-VFI dataset. No subset of the BVI-VFI dataset is used during training. Best models and second best models are marked accordingly.

	30fps		60fps		120fps		non-DL		DL		Overall	
Model	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
FAST	0.54	0.49	0.65	0.73	0.61	0.72	0.47	0.45	0.76	0.77	0.63	0.70
PSNR	0.52	0.47	0.60	0.67	0.55	0.63	0.49	0.44	0.68	0.70	0.58	0.65
FRQM	0.50	0.44	0.60	0.64	0.57	0.62	0.82	0.80	0.47	0.49	0.50	0.58
FovVideoVDP	0.45	0.42	0.55	0.64	0.51	0.61	0.58	0.54	0.61	0.66	0.56	0.64
FloLPIPS	0.49	0.47	0.57	0.59	0.58	0.61	0.46	0.43	0.63	0.67	0.58	0.61
GMSD	0.52	0.49	0.58	0.65	0.53	0.63	0.47	0.40	0.66	0.68	0.57	0.63
C3DVQA	0.34	0.25	0.45	0.57	0.42	0.66	0.41	0.37	0.49	0.60	0.43	0.54
SpEED	0.40	0.48	0.51	0.67	0.53	0.63	0.32	0.43	0.59	0.70	0.57	0.64
ST-GREED	0.16	0.11	0.33	0.14	0.27	0.03	0.11	0.06	0.30	0.08	0.26	0.06
LPVPS	0.19	0.19	0.28	0.34	0.27	0.28	0.23	0.25	0.29	0.30	0.26	0.29
Ours	0.56	0.53	0.66	0.65	0.68	0.61	0.32	0.32	0.69	0.69	0.63	0.62

Table 2. The performance of **cross-dataset** evaluation (no-reference quality metrics). Our model is trained on the VFIPS dataset and tested over all the MOS values in the BVI-VFI dataset. No subset of the BVI-VFI dataset is used during training. Best models and second best models are marked accordingly.

	30fps		60fps		120fps		non-DL		DL		Overall	
Model	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BRISQUE	0.16	0.01	0.15	0.00	0.15	0.08	0.14	0.13	0.17	0.06	0.12	0.00
ChipQA	0.11	0.03	0.26	0.03	0.16	0.06	0.05	0.01	0.12	0.05	0.12	0.03
VIDEVAL	0.11	0.05	0.10	0.04	0.08	0.05	0.23	0.19	0.05	0.03	0.06	0.04
deepIQA_NR	0.18	0.14	0.11	0.11	0.07	0.04	0.06	0.02	0.12	0.11	0.08	0.06
NIQE	0.19	0.05	0.23	0.12	0.18	0.03	0.22	0.19	0.15	0.03	0.15	0.08
VIIDEO	0.12	0.12	0.28	0.12	0.30	0.22	0.15	0.04	0.31	0.29	0.23	0.19
FastVQA	0.33	0.13	0.21	0.29	0.31	0.28	0.38	0.38	0.17	0.19	0.22	0.25
Ours (no-ref)	0.13	0.13	0.17	0.19	0.15	0.09	0.09	0.1	0.17	0.2	0.14	0.17

1.3. Cross-validation details

In our experiments, where we compare our model with other general-purpose video and image quality assessment metrics, our results are obtained as a result of repeated experiments on 20 splits of cross-validation on the BVI-VFI dataset [1]. To recreate the experiment settings used in the BVI-VFI paper, 80% of the 36 source videos in the BVI-VFI dataset are used for training and the remaining 20% are used for testing. This process is repeated 20 times instead of 1000 times in the BVI-VFI paper, to keep our training and testing times within a reasonable amount. Due to time constraints, this process is repeated 16 times instead for our cross-validation experiments regarding our model with optical flow in Tab. 3.

To maintain reproducibility, all 20 splits are saved and reused for all experiments in which training is performed over BVI-VFI. Fig. 3 shows the number of occurrences for each subject in training and testing splits. On average, every subject appears in the training set 15.56 times with a standard deviation of 1.95 compared to an average of 4.44 times in the test set with a standard deviation of 1.95. This is in line with the expected number of occurrences of 16 times in the training set and 4 times in the test set.

2. Integrating optical flow information

Due to the success of FAST [7], which is an optical flow-based video quality assessment metric, in all our full-reference experiments; we introduce optical flow into our model as well. Due to the multi-scale nature of our model, we decided to employ SPyNET [6], as it provides optical flows in multiple resolutions.

As shown in Fig. 4, this requires a slight modification to our feature extraction network. In addition to extracting CLIP [5] features using the modified CLIP visual backbone, we also extract the optical flow of reference and distorted input videos. The extracted features are then concatenated with the CLIP features in the channel dimension. The rest of the network is not modified, as the only difference is that

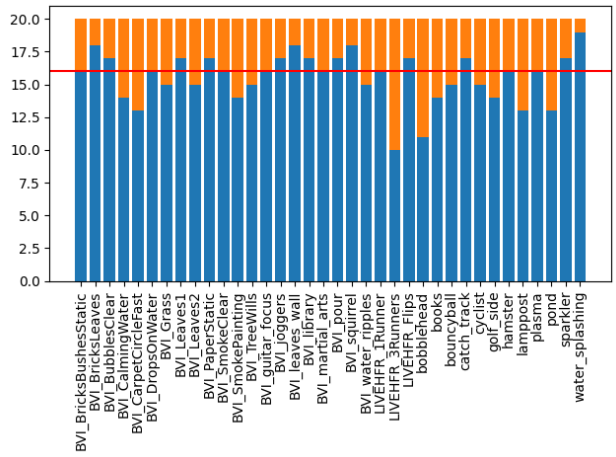

Figure 3. Number of occurrences in training and testing datasets for each subject. Blue bars indicate occurrences in the training set, and orange bars indicate occurrences in the testing set. Horizontal red line marks the 80% threshold.

our feature tensor in each level simply has two additional channels. The element-wise absolute difference, the subsequent concatenation, and the Video Swin Transformers [3] function in an identical manner.

As it can be observed from the cross-validation results in Tab. 3, our model with optical flow is nonetheless outperformed by our model without optical flow in all categories except in *non-DL*. Although we choose to employ SPyNET due to its multi-resolution output which pairs nicely with our multi-scale architecture, it should nevertheless be noted that the SPyNET was developed in 2017. Therefore, an interesting topic for potential future research would encompass the implementation of a more recent optical flow algorithm in a multi-resolution setting, to truly assess the contribution of optical flow for task-specific video quality assessment.

3. Computational cost and performance

Due to possessing a considerably larger feature extraction network, our model exhibits a 55% slower per-frame inference time during evaluation compared to LPVPS, without speed optimization. However, it is worth noting that we achieve the reported performance with only 5 epochs of training, each taking $\sim$ 20 minutes, compared to 20 epochs of $\sim$ 8 minutes for LPVPS. This reduces the total training time from $\sim$ 160 minutes for LPVPS to $\sim$ 100 minutes for our model.

Moreover, although our full-reference performance remains comparable to other methods, such as FAST [41], we can observe this in Tab. 6 of the main paper that our model offers 3% to 36% performance increase compared to the second-best method in all categories except *non-DL* in no-reference settings. We believe that the flexible nature and retention of no-reference performance of our model should also be considered when evaluating its overall performance.

References

[1] Duolikun Danier, Fan Zhang, and David R. Bull. BVI-VFI: A video quality database for video frame interpolation. *IEEE Trans. Image Process.*, 32:6004–6019, 2023. 2

[2] Qiqi Hou, Abhijay Ghildyal, and Feng Liu. A perceptual quality metric for video frame interpolation. In *European Conference on Computer Vision*, pages 234–253. Springer, 2022. 1

[3] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3192–3201. IEEE, 2022. 3

[4] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. ST-GREED: space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Trans. Image Process.*, 30:7446–7457, 2021. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[6] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[7] Jinjian Wu, Yongxu Liu, Weisheng Dong, Guangming Shi, and Weisi Lin. Quality assessment for video with degradation along salient trajectories. *IEEE Trans. Multim.*, 21(11):2738–2749, 2019. 2

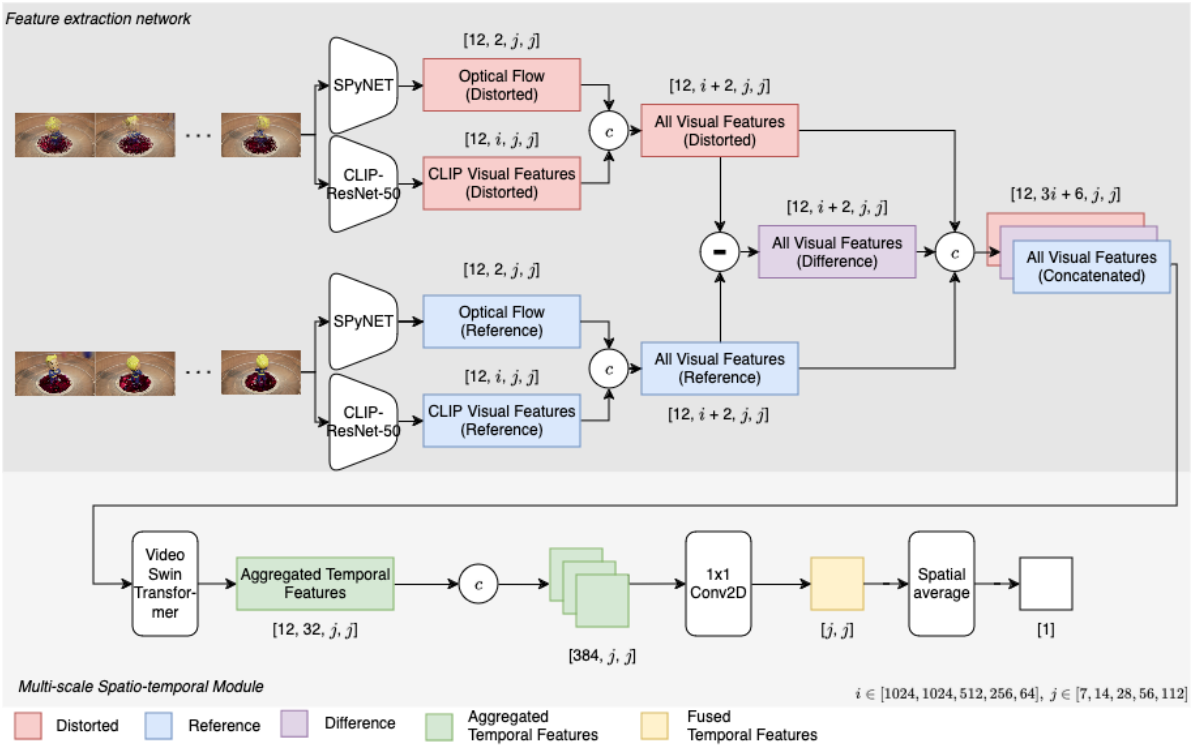

Figure 4. Modified architecture with optical flow.

Table 3. Cross validation performance of our model with optical flow over all the DMOS values in BVI-VFI dataset. Best models and second best models are marked accordingly.

Model	30fps		60fps		120fps		non-DL		DL		Overall	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
LPVPS	0.45 (0.12)	0.61 (0.13)	0.63 (0.11)	0.65 (0.09)	0.65 (0.14)	0.58 (0.11)	0.32 (0.12)	0.36 (0.11)	0.61 (0.11)	0.66 (0.09)	0.53 (0.11)	0.61 (0.09)
Ours w/ optical flow	0.65 (0.11)	0.59 (0.10)	0.70 (0.07)	0.74 (0.06)	0.70 (0.09)	0.68 (0.11)	0.32 (0.08)	0.40 (0.08)	0.73 (0.08)	0.75 (0.08)	0.69 (0.09)	0.72 (0.08)
Ours w/o optical flow	0.67 (0.11)	0.63 (0.10)	0.76 (0.08)	0.76 (0.07)	0.75 (0.07)	0.67 (0.12)	0.31 (0.09)	0.34 (0.10)	0.77 (0.07)	0.76 (0.07)	0.73 (0.08)	0.72 (0.07)