Video Quality Assessment based on Quality Aggregation Networks

Yaosi Hu, Zhenzhong Chen
Wuhan University

Dec. 2022
VQEG Meeting

Introduction

As video has become one of the most significant parts of media in our daily life, service providers have made a lot of efforts to improve the quality of their service under limited bandwidth and storage. As a consequence, effective video quality assessment (VQA) models that can predict video quality automatically and accurately are required.

Compared to image quality assessment, VQA is more challenging due to the complexity of modeling perceived quality characteristics in both spatial and temporal domain, effected by both practical distortion and human perceptual effects.

- Practical distortion
  - error signal
- **Human perceptual effects**
  - short-term effect: *visual masking effect* of human visual system
  - long-term effect: *memory effect*

G. E. Legge and J. M. Foley, "Contrast masking in human vision," Journal of the Optical Society of America, vol. 70, no. 12, pp. 1458–1471, Dec. 1980.
J. W. Suchow and G. A. Alvarez, "Motion silences awareness of visual change," Current Biology, vol. 21, no. 2, pp. 140–143, Jan. 2011.

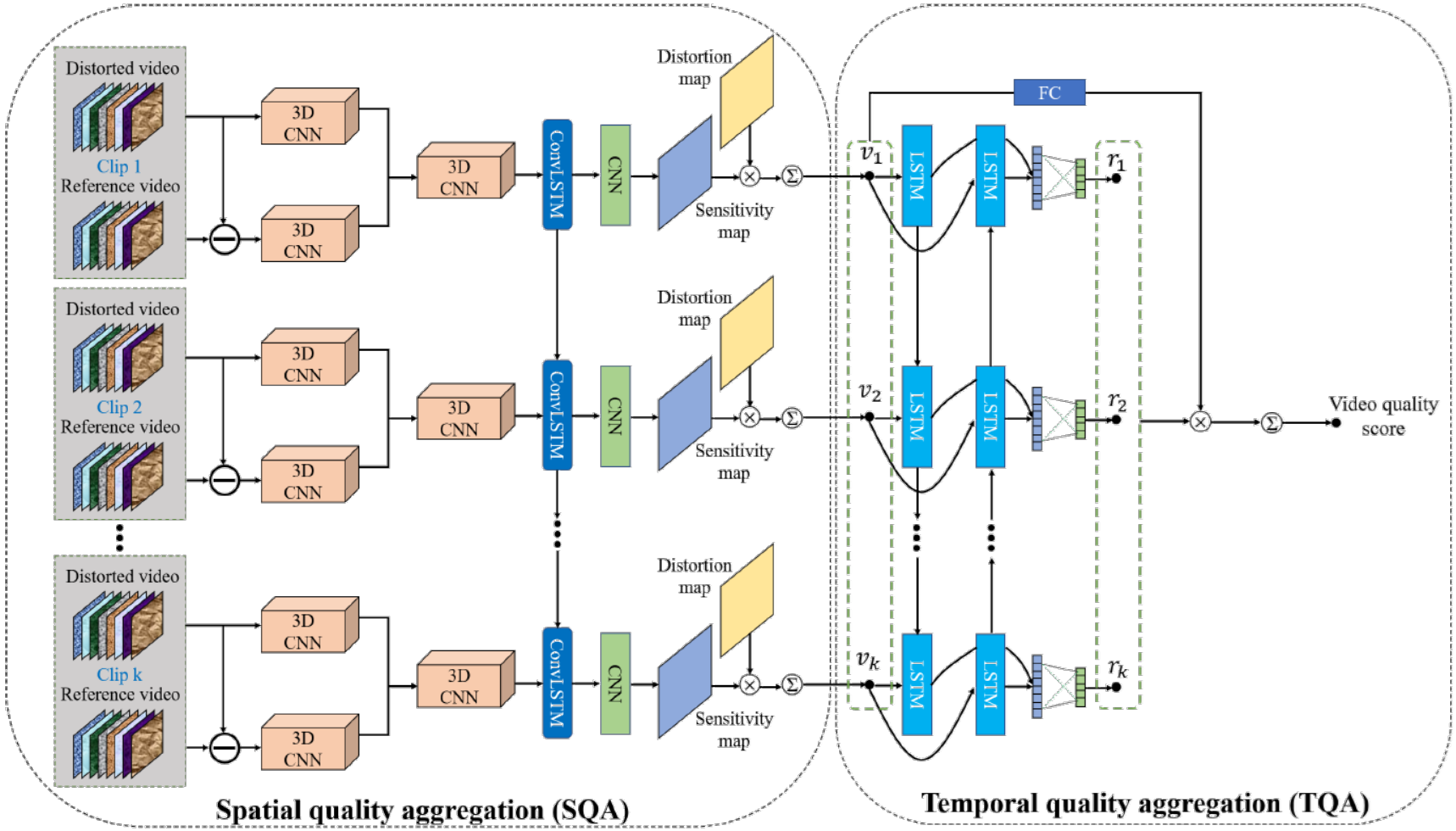
Introduction

- Nowadays, convolution neural networks has demonstrated outstanding performance in various computer vision tasks, and thus been applied in VQA to better model the perceptual quality.
- Although deep learning based methods provide promising correlation with visual perception in VQA, there still exists a dilemma where the limited understanding on human visual perception affects the performance.

Contribution:

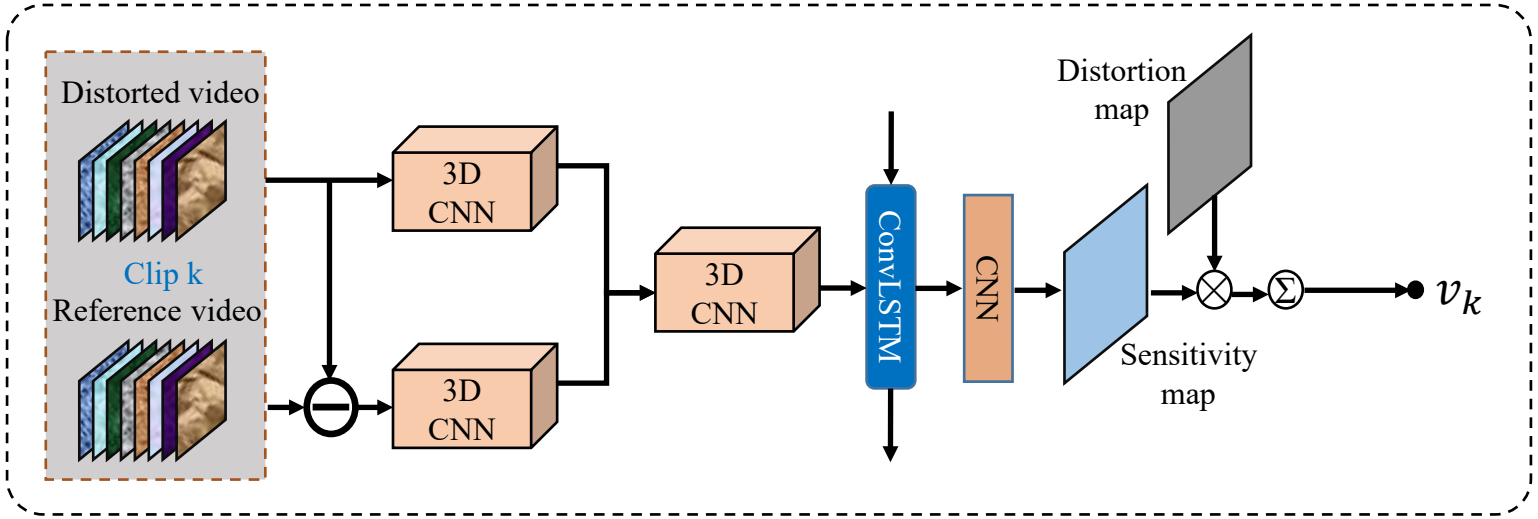
- A Quality Aggregation Network (QAN) approach for full-reference VQA is proposed.
  - Spatial Quality Aggregation (SQA) network considers visual masking effect in frame-level.
  - Temporal Quality Aggregation (TQA) network model memory effect in video-level.

Quality Aggregation Network (QAN)



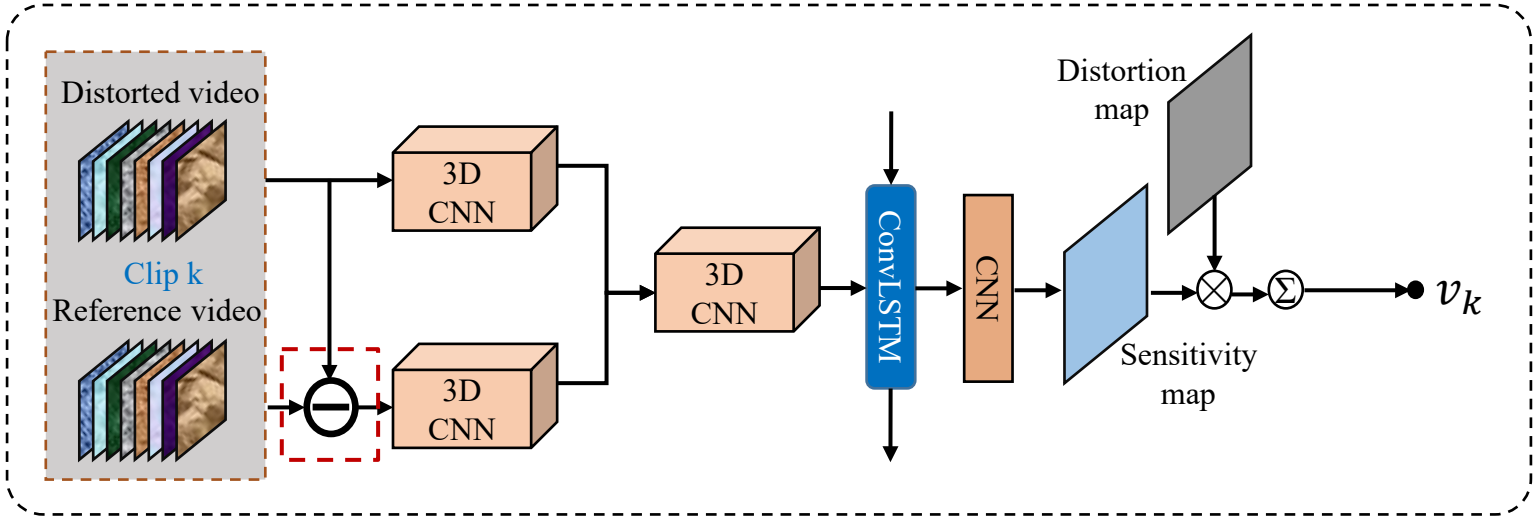
Quality Aggregation Network (QAN)

Spatial Quality Aggregation (SQA)



Quality Aggregation Network (QAN)

Spatial Quality Aggregation (SQA)

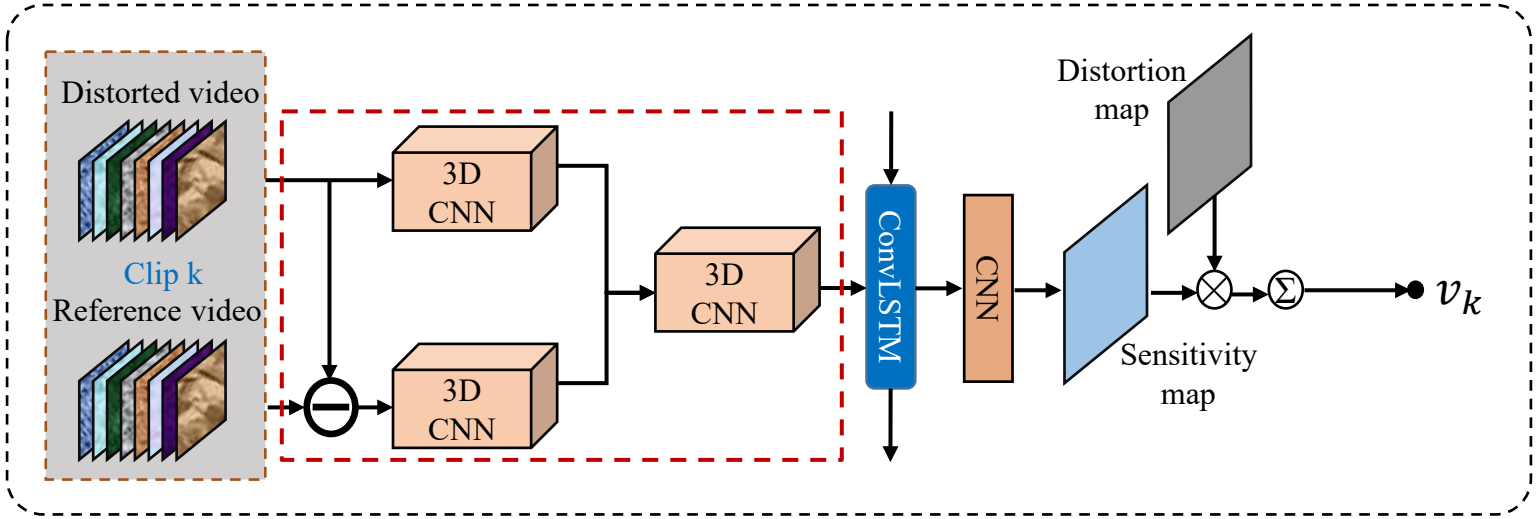


Error map[1]: $$X_t^e = \frac{\log\left(1/\left(\left(X_t^r - X_t^d\right)^2 + \epsilon/255^2\right)\right)}{\log\left(255^2/\epsilon\right)}$$

[1] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1676–1684.

Quality Aggregation Network (QAN)

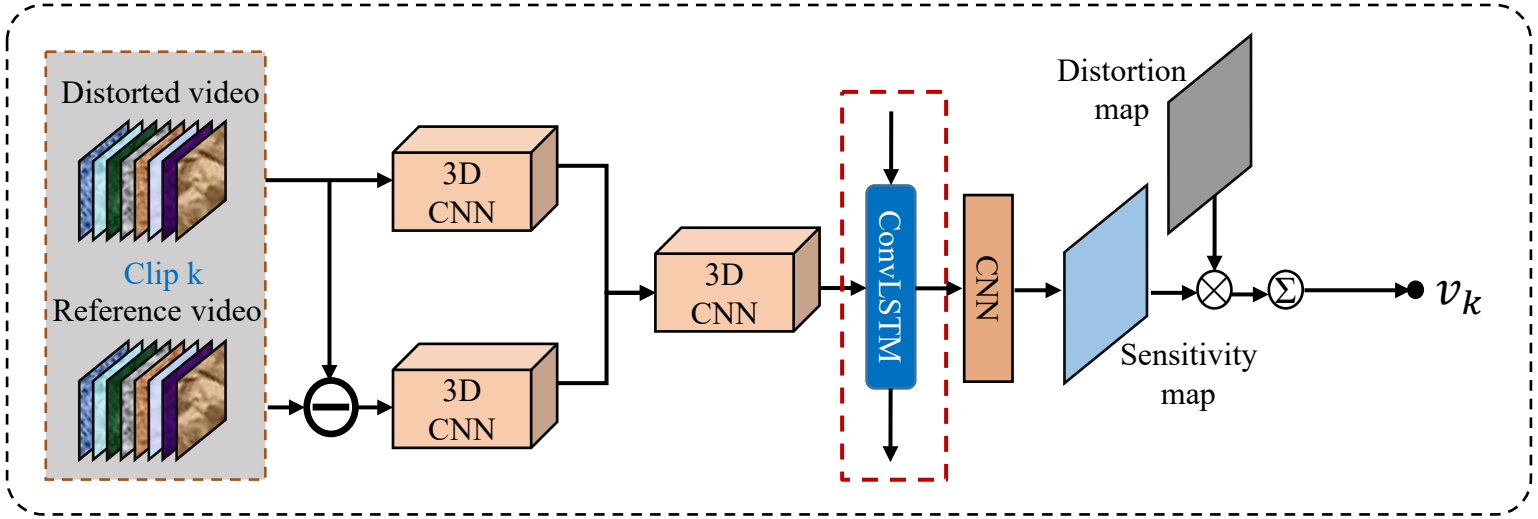
Spatial Quality Aggregation (SQA)



- 3D CNN Component: effectively extracting spatial-temporal features and fuse multi-frame masking effects over the target frame.
  - slow-fusion approach is adopted in early spatio-temporal feature extraction stage before the concatenation

Quality Aggregation Network (QAN)

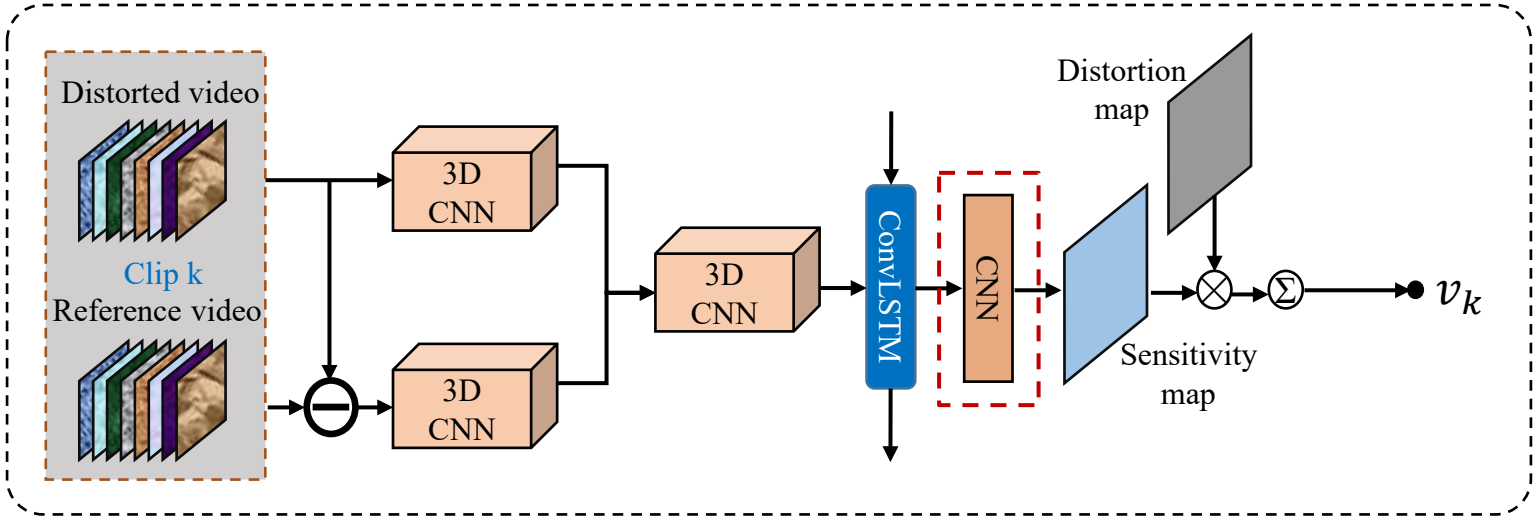
Spatial Quality Aggregation (SQA)



- ConvLSTM Component: preserving spatial information, while also modeling the influence of previous motion on current short-term representation.

Quality Aggregation Network (QAN)

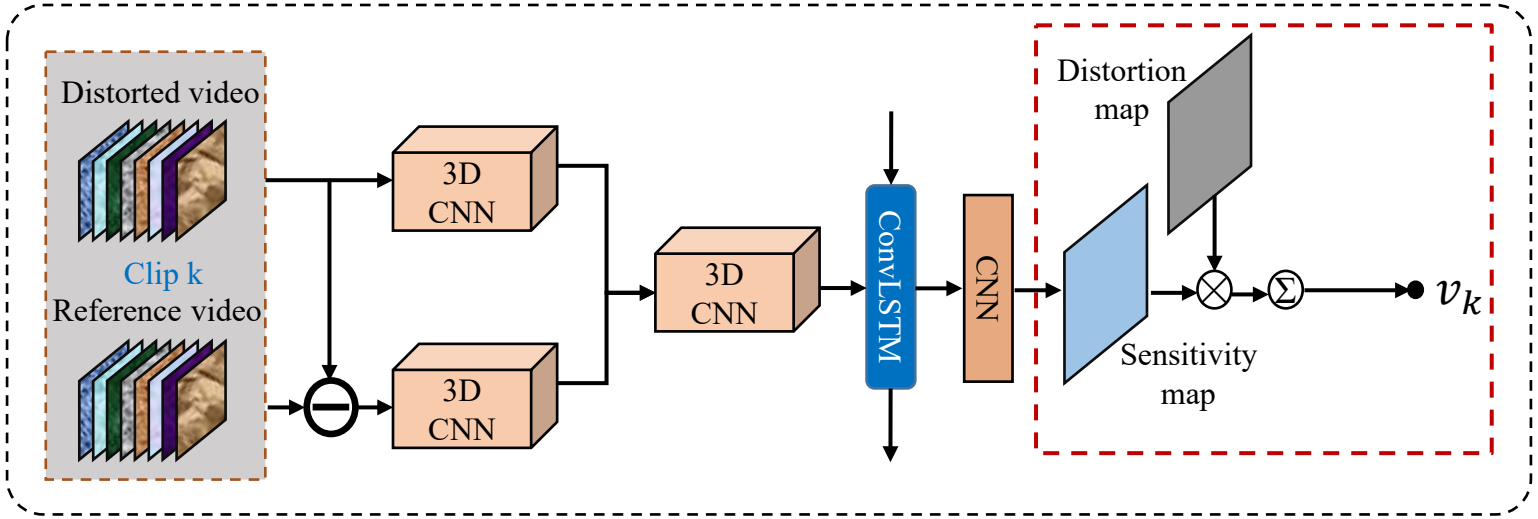
Spatial Quality Aggregation (SQA)



- 2D CNN Component: compressing the spatio-temporal features into sensitivity map M_{t_k}, which can be seen as the ***mask*** in visual masking effect.

Quality Aggregation Network (QAN)

Spatial Quality Aggregation (SQA)

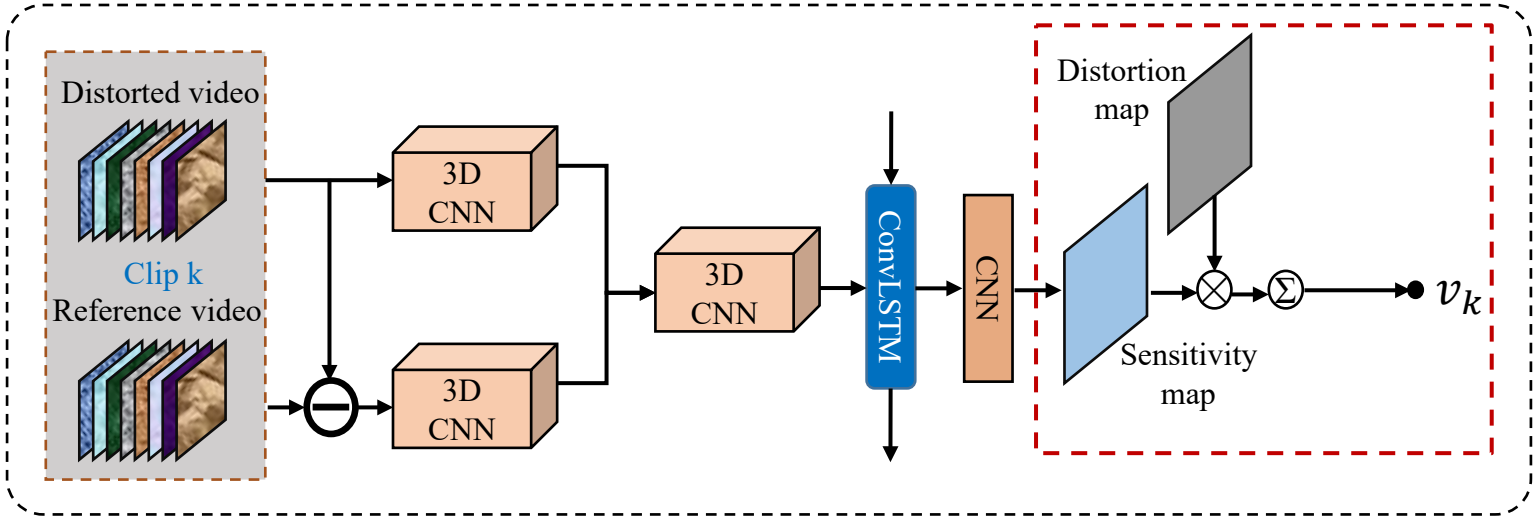


- Regression Component: aggregating spatial distortion information and regressing features into the predicted scores.
  - Since zero-padding is used before each convolution operation, values tend to be zero near the borders around the sensitivity map. p_n pixels near the borders around the map are excluded.

$$\nu_k = \frac{\sum_{(i,j)\in\Omega}\left(M_{t_k}\cdot X^e_{t_k}\right)}{(H-p_n\cdot 2)\cdot(W-p_n\cdot 2)}$$

Quality Aggregation Network (QAN)

Spatial Quality Aggregation (SQA)



- **Loss Function**:

$$\mathcal{L}_{\mathrm{I}} = \lambda_1 \left\| \mathbf{u}_{\mathrm{I}} - \mathbf{u}_{\mathrm{g}} \right\|_2^2 + \lambda_2 TV + \lambda_3 \left\| \xi \right\|_2^2$$

  - ***MSE loss***: optimizing towards better prediction quality.
  - ***Total variation (TV) regularization***[1]: relieving high-frequency noise in the perceptual sensitivity map.

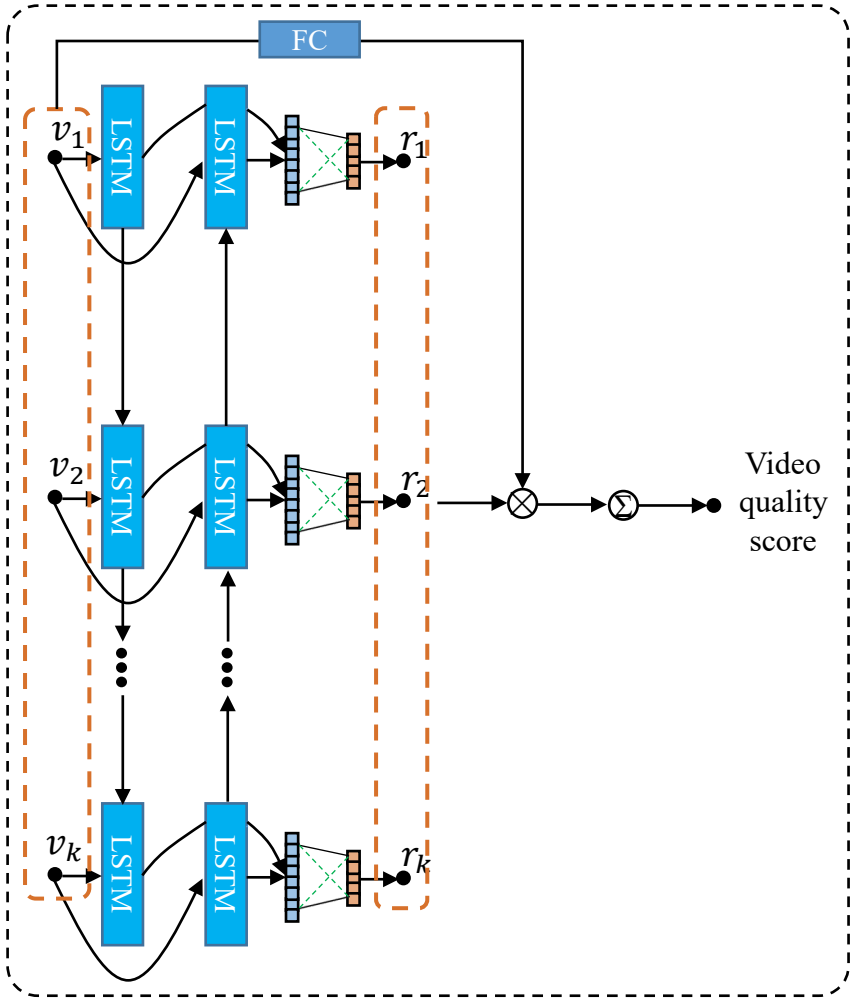
$$TV(X_w) = \frac{1}{H \cdot W} \sum_{(x,y)} \left(\mathrm{sobel}_h(M)^2 + \mathrm{sobel}_v(M)^2 \right)^{\frac{3}{2}}$$

  - ***L2 regularization***: applied to parameters to avoid the overfitting issue.

[1] A. Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. Int. J. Comput. Vis., 2016.

Quality Aggregation Network (QAN)

Temporal Quality Aggregation (TQA)



Based on SQA, TQA aims at predicting the retrospective quality score and exploring the temporal pattern.

- Bidirectional LSTM: weighing the importance of the frame-level quality scores.

$$r_k = \text{LeakyReLU}(W^r h_k + b^r)$$

- Score aggregation:

$$u_{\text{II}} = \sum_{k=1}^{K} r_k FC(\nu_k)$$

- Loss function:

$$\mathcal{L}_{\text{II}} = \|\mathbf{u}_{\text{II}} - \mathbf{u}_{\text{g}}\|_2^2$$

Experiments

Setup

- Datasets:
  - LIVE Video Quality Database
  - CSIQ Video database

- The training of the proposed QAN consists of two stages:
  - At Stage I, SQA is trained to predict the frame-level quality scores.
  - At Stage II, TQA is trained to predict retrospective quality score using the pre-trained SQA.

- Patch-based training approach is adopted to obtain more video clips as training samples, of which the shape is **112×112** in spatial dimension and **8** frames in temporal dimension.
- Using leave-2-fold-out cross-validation strategy for fair comparison.

Experiments

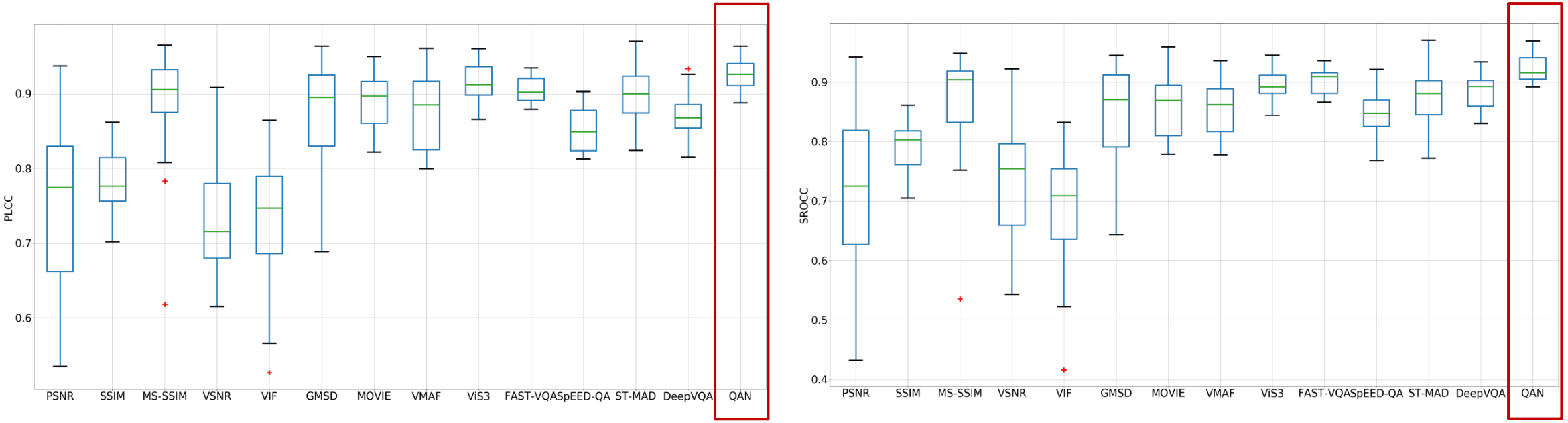
Performance Comparison

Performance comparison on LIVE VQA database.

Metrics	PLCC					SROCC				
	Wireless	IP	H.264	MPEG-2	**ALL**	Wireless	IP	H.264	MPEG-2	**ALL**
PSNR	0.8358	0.8200	0.8452	0.8217	0.7539	0.7550	0.7244	0.8058	0.7912	0.7151
SSIM [1]	0.8509	0.8477	0.8392	0.7480	0.7835	0.7604	0.7738	0.7988	0.6824	0.7915
MS-SSIM [20]	0.8677	0.9009	0.8662	0.8489	0.8852	0.7746	0.8057	0.8286	0.7999	0.8645
VSNR [21]	0.7792	0.9059	0.9034	0.8791	0.7380	0.7217	0.8171	0.8995	0.8643	0.7372
VIF [22]	0.8011	0.8482	0.8566	0.8321	0.7286	0.7651	0.7854	0.8302	0.7825	0.6890
GMSD [23]	0.8752	0.8991	0.8388	0.8460	0.8719	0.7984	0.7613	0.8148	0.8027	0.8476
MOVIE [24]	0.9070	0.8643	0.8888	0.8900	0.8888	0.8302	0.7702	0.8434	0.8437	0.8611
VMAF [25]	0.9075	0.8707	0.8759	0.8561	0.8749	0.8392	0.7841	0.8392	0.8118	0.8563
ViS_3 [18]	0.9052	0.9312	0.8972	0.8757	0.9157	0.8069	0.8540	0.8455	0.8282	0.8943
FAST-VQA [26]	**0.9403**	0.9472	0.9392	0.9175	0.9060	0.8508	0.8400	0.9116	0.8791	0.9026
SpEED-QA [27]	0.8949	0.9230	0.8931	0.8410	0.8512	0.8344	0.8476	0.8529	0.7986	0.8531
ST-MAD [28]	0.9070	0.9002	**0.9502**	0.9195	0.8986	0.8307	0.8235	0.9222	0.8870	0.8745
DeepVQA [11]	0.8877	0.8119	0.8869	0.9076	0.8700	0.8631	0.7143	0.9095	0.9225	0.8845
QAN	0.9397	**0.9703**	0.9454	**0.9693**	**0.9272**	**0.8988**	**0.8943**	**0.9333**	**0.9548**	**0.9232**

Experiments

Performance Comparison



Box plot of PLCC/SROCC distributions of the VQA metrics for 20 iterations on the LIVE VQA database.

Experiments

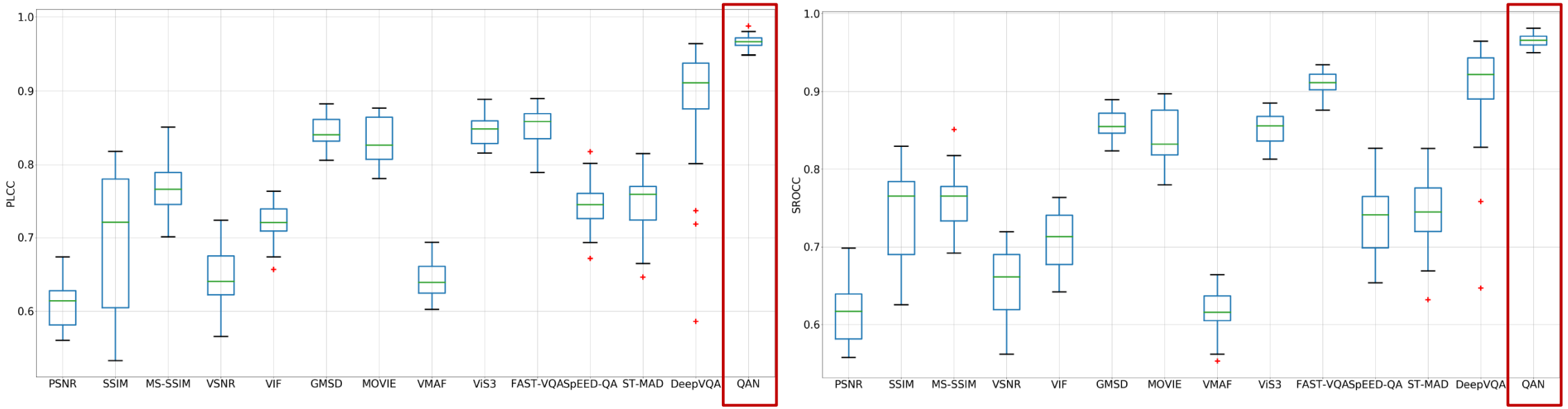
Performance Comparison

Performance comparison on CSIQ database.

Metrics	PLCC							SROCC						
	H.264	PLoss	MJPEG	Wavelet	AWGN	HEVC	**ALL**	H.264	PLoss	MJPEG	Wavelet	AWGN	HEVC	**ALL**
PSNR	0.9152	0.9300	0.8546	0.9075	0.9809	0.9056	0.6085	0.8736	0.9030	0.8277	0.8433	0.9455	0.8736	0.6150
SSIM [1]	0.9587	0.8476	0.5996	0.9419	0.8876	0.9683	0.7023	0.9415	0.8190	0.5212	0.9117	0.8509	0.9322	0.7387
MS-SSIM [20]	0.9775	0.9169	0.9573	0.9533	0.9779	0.9815	0.7674	0.9532	0.9169	0.9368	0.9126	0.9481	0.9558	0.7601
VSNR [21]	0.9655	0.9571	0.8018	0.9465	0.9683	0.9272	0.6468	0.9333	0.9264	0.7697	0.8952	0.9437	0.8978	0.6515
VIF [22]	0.9828	0.8583	0.9626	0.9605	0.9834	0.9833	0.7191	0.9550	0.8260	0.9229	0.9437	0.9532	0.9498	0.7097
GMSD [23]	0.9746	0.9410	0.9501	0.9574	0.9761	0.9830	0.8447	0.9532	0.9177	0.9169	0.9143	0.9437	0.9541	0.8581
MOVIE [24]	0.9629	0.9462	0.9536	0.9597	0.9493	0.9716	0.8322	0.9342	0.9238	0.9290	0.9333	0.9134	0.9506	0.8425
VMAF [25]	0.9787	0.8884	0.8837	0.9505	0.9492	0.9664	0.6434	0.9567	0.8900	0.8899	0.9307	0.9238	0.9394	0.6166
ViS_3 [18]	0.9580	0.9307	0.8848	0.9701	0.9818	0.9656	0.8480	0.9368	0.9056	0.8424	0.9455	0.9532	0.9429	0.8521
FAST-VQA [26]	0.9748	**0.9692**	0.9678	0.9762	0.9720	0.9801	0.8523	0.9628	**0.9567**	0.9524	0.9654	0.9558	**0.9610**	0.9099
SpEED-QA [27]	0.9867	0.9380	0.8720	0.9798	0.9686	0.9586	0.7435	**0.9697**	0.9264	0.8312	0.9567	0.9368	0.9359	0.7365
ST-MAD [28]	0.9489	0.8931	0.8547	0.9259	0.9496	0.9541	0.7446	0.9299	0.8589	0.8234	0.8944	0.9255	0.9238	0.7435
DeepVQA [11]	0.9190	0.8978	0.8603	0.8887	0.9482	0.9534	0.8798	0.8829	0.8886	0.8571	0.9086	0.9229	0.9314	0.8973
QAN	**0.9871**	0.9643	**0.9843**	**0.9837**	**0.9859**	**0.9936**	**0.9658**	0.9657	0.9457	**0.9714**	**0.9714**	**0.9686**	0.9571	**0.9654**

Experiments

Performance Comparison



Box plot of PLCC/SROCC distributions of the VQA metrics
for 20 iterations on the CSIQ database.

Experiments

Ablation Study

Ablation study on LIVE VQA database.

Metrics	PLCC				
	Wireless	IP	H.264	MPEG-2	ALL
SQA	0.8985	0.9064	0.9251	0.9510	0.9152
QAN	0.9397	0.9703	0.9454	0.9693	0.9272
Metrics	SROCC				
	Wireless	IP	H.264	MPEG-2	ALL
SQA	0.8595	0.8514	0.9274	0.9589	0.9115
QAN	0.8988	0.8943	0.9333	0.9548	0.9232

- QAN model outperforms the model with SQA, demonstrating that QAN successfully models temporal quality variation with memory effect.
- The wireless and IP distortion may lead to serious degradation on some frames, and thus leave a greater impact on the overall quality assessment compared to other frames.

Conclusion

- In summary, we propose a novel Quality Aggregation Network (QAN) model for FR-VQA task:
  - A spatial quality aggregation network is designed by adopting 3D CNN and ConvLSTM to learn spatio-temporal masking effects which greatly influences the perception of the video quality.
  - A temporal quality aggregation network is proposed based on LSTM to simulate the memory effect during video viewing and evaluation.
- Experimental results show that QAN greatly improves the overall accuracy on both LIVE VQA and CSIQ databases when compared with the SOTA models.
- Code: https://github.com/lorenzowu/QAN

Wei Wu, Yingxue Zhang, Yaosi Hu, Zhenzhong Chen, Shan Liu. "Video Quality Assessment based on Quality Aggregation Networks", IEEE Visual Communications and Image Processing Conference (VCIP), 2022.

Thanks for Listening !