北京 2019

邀游“视”界 做你所想
Explore World, Do What You Want

Quality Assessment of In-the-Wild Videos

Tingting Jiang

Peking University
August 2019

出品: LiveVideoStack 音视频技术社区 CSDN

深圳
2019

遨游“视”界 做你所想
Explore World, Do What You Want

LiveVideoStackCon 2019 深圳

2019.12.13-14

出品: LiveVideoStack 音视频技术社区

成为讲师：speaker@livevideostack.com

成为志愿者：volunteer@livevideostack.com

赞助、商务合作：kathy@livevideostack.com

Outline

- Background
- Motivation
- Method
- Experiments
- Conclusion and Future Work

Outline

- Background
- Motivation
- Method
- Experiments
- Conclusion and Future Work

Distortions

Videos captured **in the wild** may contain **annoying distortions** due to out of focus, object motion, camera shake, or under/over exposure.

In-the-Wild vs. Synthetically-Distorted

- More content diversity
- More complex distortions that are temporally heterogeneous
- Current video quality assessment (VQA) methods (e.g., VBLIINDS and VIIDEO) validated on traditional synthetic VQA databases fail in predicting the quality of in-the-wild videos.

Quality Assessment of In-the-Wild Videos

- Helps identifying and cull low-quality videos, preventing their occurrence, or repairing/enhancing them.

- Requires no-reference general-purpose (distortion-unaware) quality assessment
  - The reference videos are not available and the shooting distortions are unknown.

Quality assessment of in-the-wild videos
is
challenging but in urgent need!

Outline

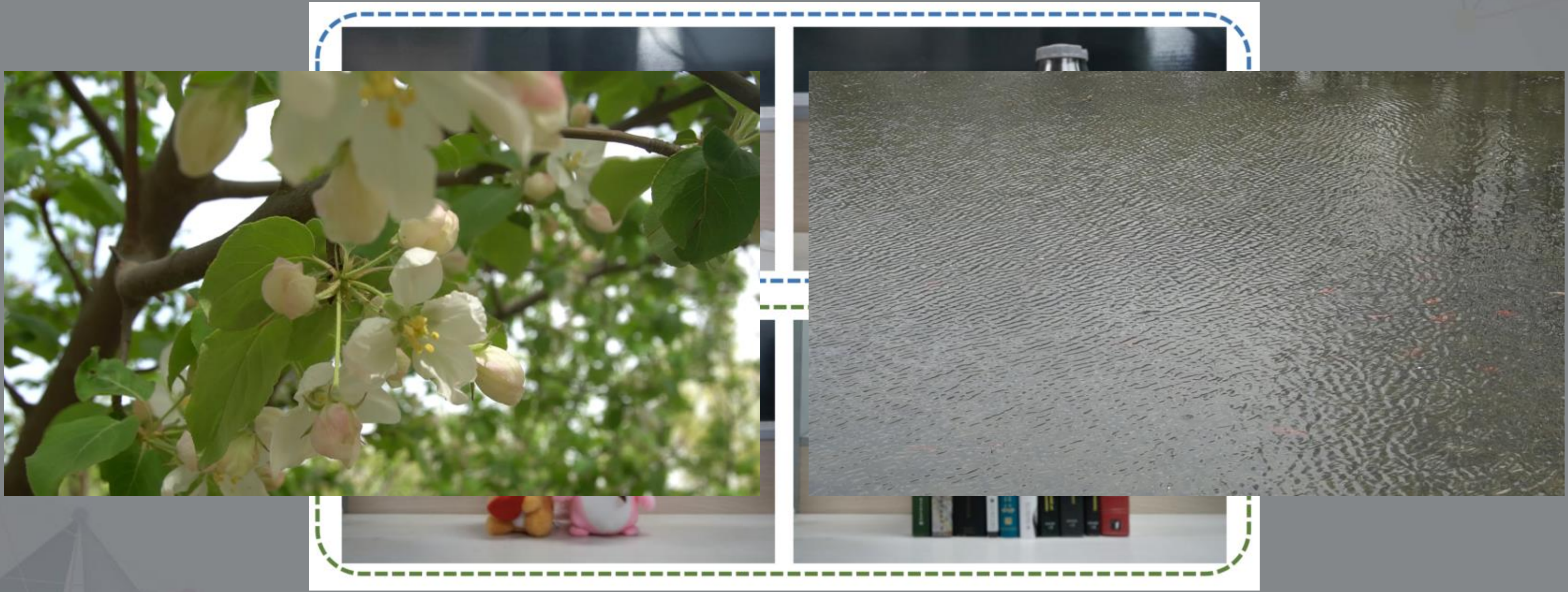
- Background
- Motivation
- Method
- Experiments
- Conclusion and Future Work

Motivation

- Human judgments of visual image/video quality depend on *content*
- Human judgments of video quality are affected by their *temporal memory*

Content-Dependency Effects

Every two images/videos in a pair are taken in the same shooting condition, and they only differ in image content.



User study shows that humans consistently prefer the left ones.

Temporal-Memory Effects

Human judgments of current frame rely on the current frame and information from previous frames.

- Long-term dependencies exist in the VQA problem.
- Temporal hysteresis effects in the frame-quality aspect
  - humans remember poor quality frames in the past and lower the perceived quality scores for following frames, even when the frame quality has returned to acceptable levels [1].

[1] Seshadrinathan and Bovik, Temporal hysteresis model of time varying subjective video quality, ICASSP 2011.

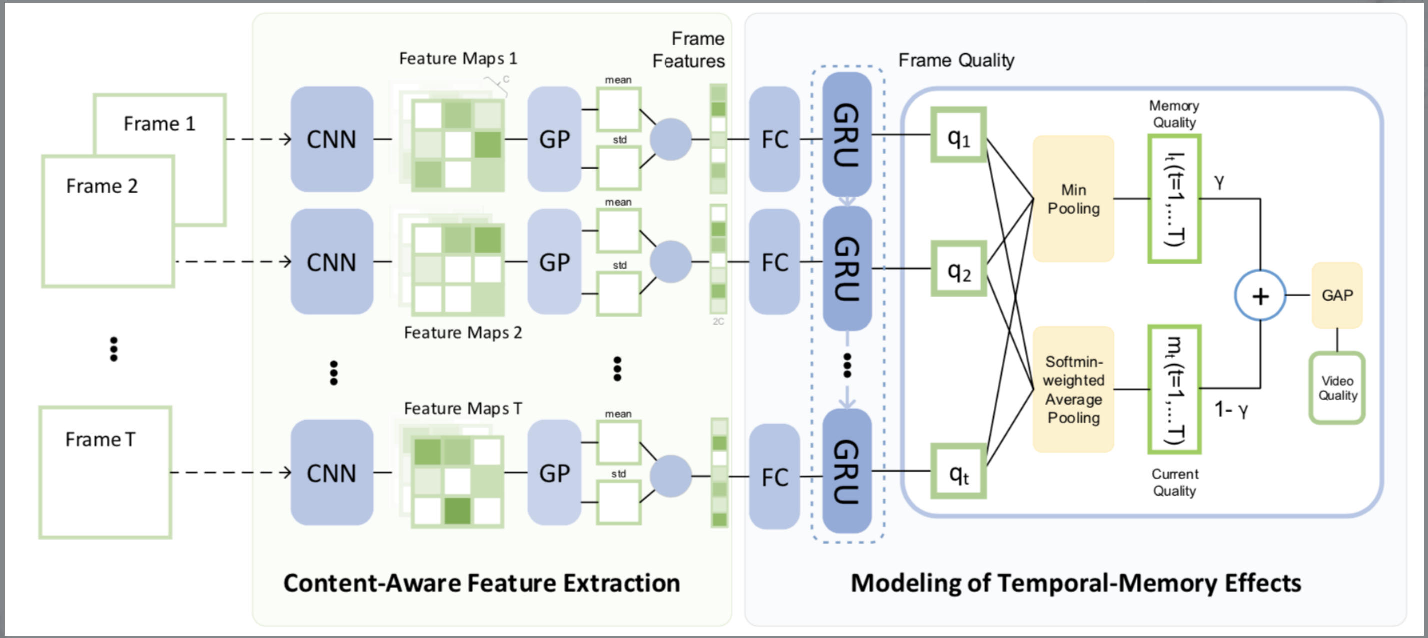
Motivation

- A deep neural network integrates the two effects
- Video quality depends on both the distortion and the content
  - Extract content-aware perceptual features from pre-trained image classification CNN models
- Temporal-memory effects exists in the VQA problem
  - In the feature integration aspect, GRU captures long-term dependencies.
  - In the quality pooling aspect, a differentiable subjectively-inspired temporal pooling layer accounts for the temporal hysteresis effects.

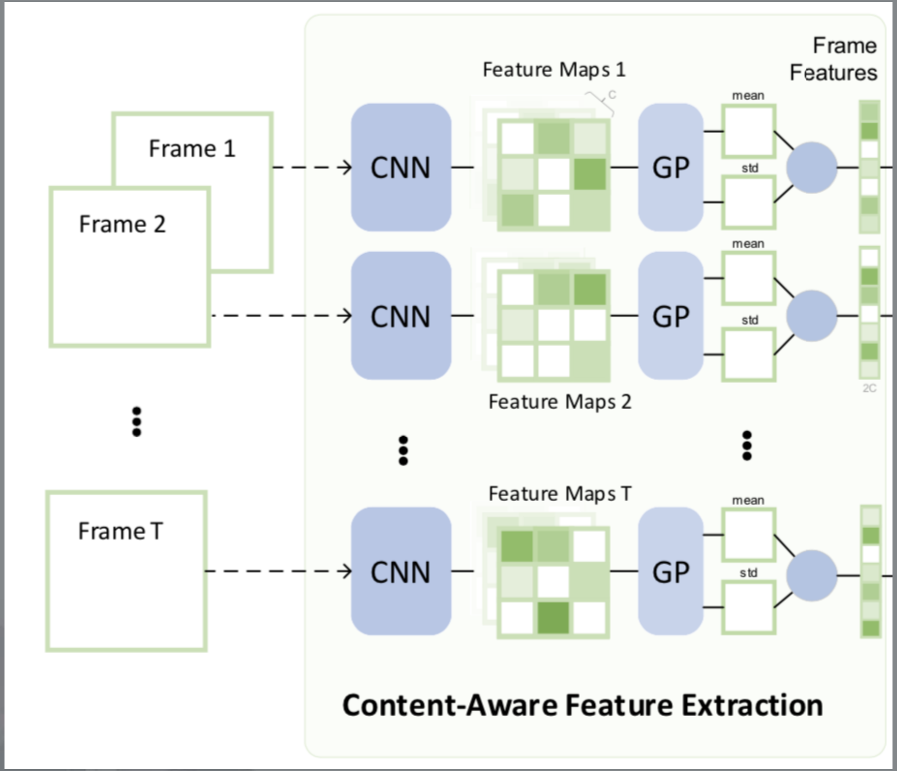
Outline

- Background
- Motivation
- Method
- Experiments
- Conclusion and Future Work

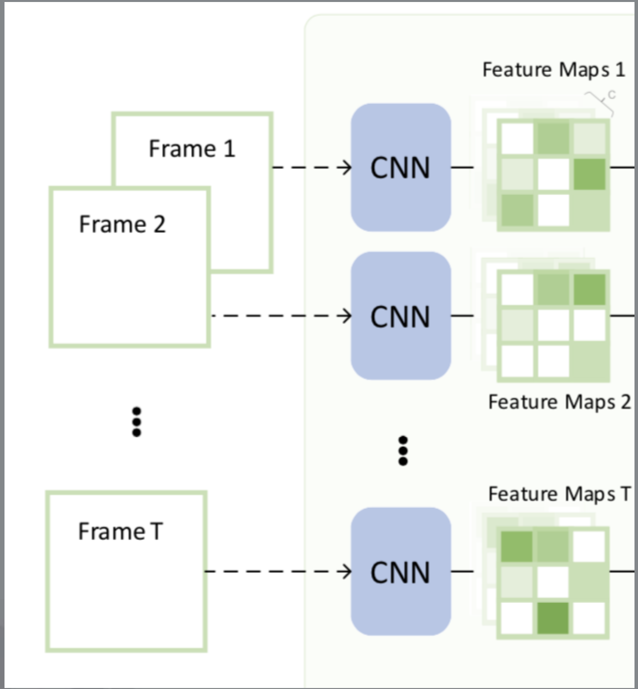
Overall Framework



Content-Aware Feature Extraction

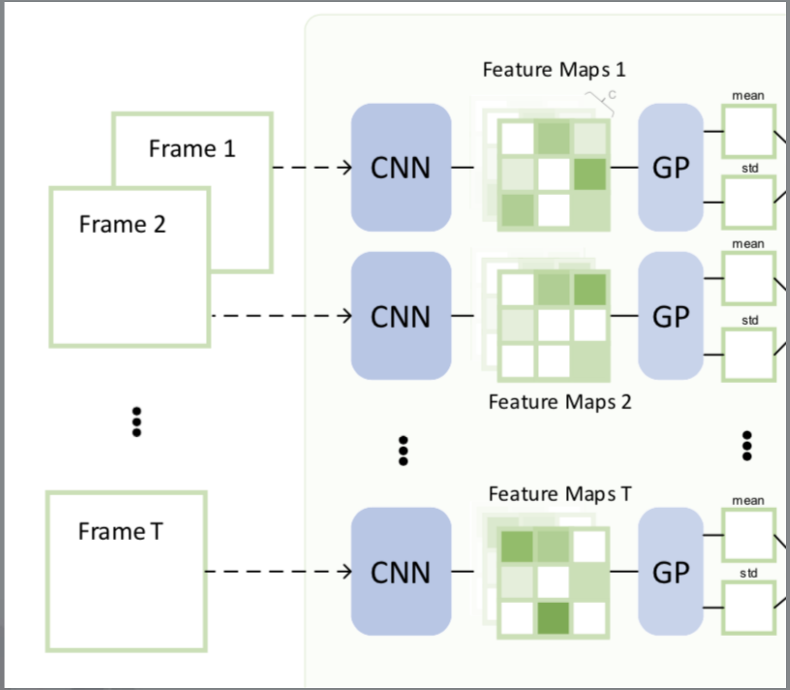


Content-Aware Feature Extraction



$$\mathbf{M}_t = \mathrm{CNN}(\mathbf{I}_t)$$

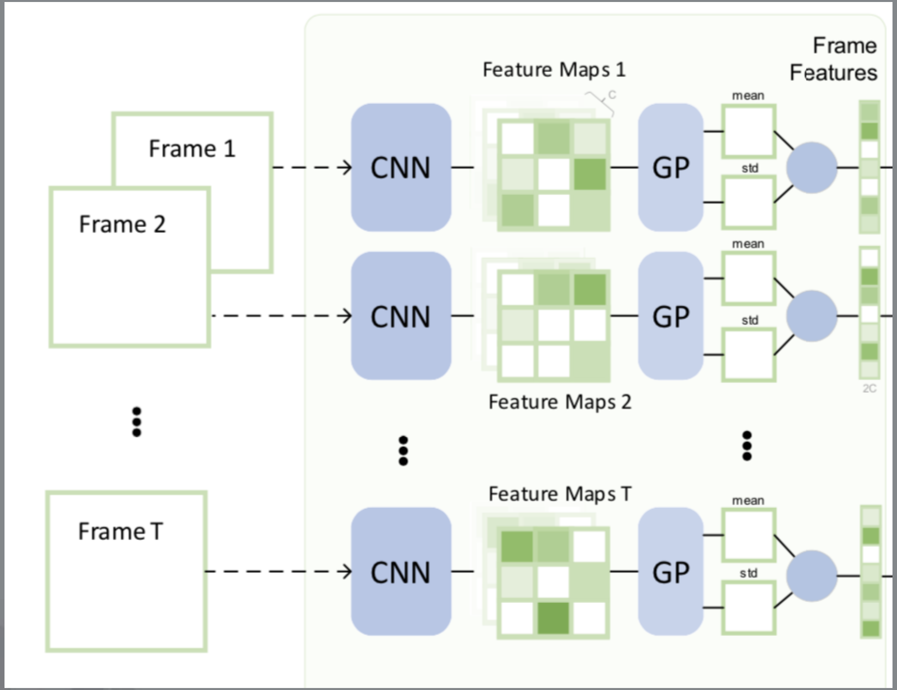
Content-Aware Feature Extraction



$$\mathbf{M}_t = \mathrm{CNN}(\mathbf{I}_t)$$

$$\mathbf{f}_t^{\mathrm{mean}} = \mathrm{GP}_{\mathrm{mean}}(\mathbf{M}_t),$$
$$\mathbf{f}_t^{\mathrm{std}} = \mathrm{GP}_{\mathrm{std}}(\mathbf{M}_t).$$

Content-Aware Feature Extraction



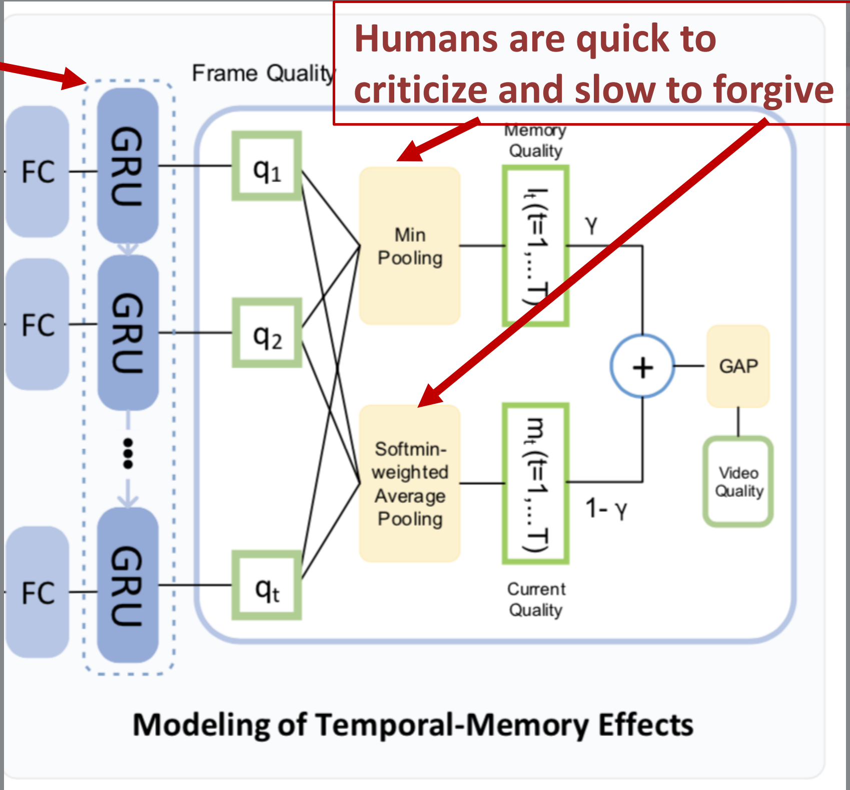
$$\mathbf{M}_t = \mathrm{CNN}(\mathbf{I}_t)$$

$$\mathbf{f}_t^{\text{mean}} = \mathrm{GP}_{\text{mean}}(\mathbf{M}_t),$$
$$\mathbf{f}_t^{\text{std}} = \mathrm{GP}_{\text{std}}(\mathbf{M}_t).$$

$$\mathbf{f}_t = \mathbf{f}_t^{\text{mean}} \oplus \mathbf{f}_t^{\text{std}},$$

Modeling of Temporal-Memory Effects

Long-term dependencies



Humans are quick to criticize and slow to forgive

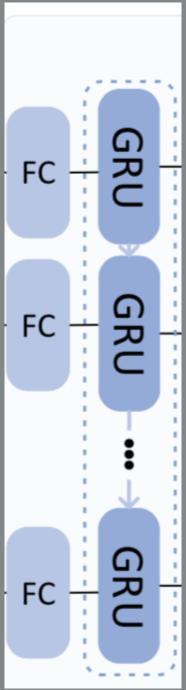
Modeling of Temporal-Memory Effects

- Long-term dependencies modeling

$$\mathbf{x}_t = \mathbf{W}_{fx}\mathbf{f}_t + \mathbf{b}_{fx}$$

$$\mathbf{h}_t = \mathrm{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1})$$

$$q_t = \mathbf{W}_{hq}\mathbf{h}_t + \mathbf{b}_{hq}$$



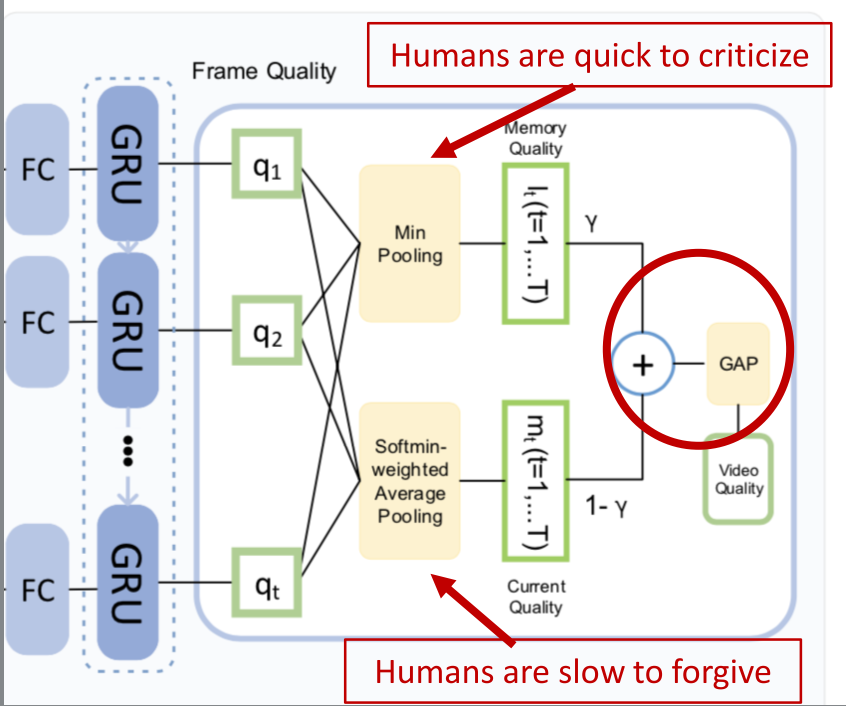
Modeling of Temporal-Memory Effects

- Subjectively-inspired temporal pooling

$$l_t = q_t, \quad \text{for } t = 1,$$
$$l_t = \min_{k \in V_{prev}} q_k, \quad \text{for } t > 1,$$

$$m_t = \sum_{k \in V_{next}} q_k w_t^k,$$
$$w_t^k = \frac{e^{-q_k}}{\sum_{j \in V_{next}} e^{-q_j}}, k \in V_{next},$$

$$q_t' = \gamma l_t + (1-\gamma) m_t,$$
$$Q = \frac{1}{T} \sum_{t=1}^{T} q_t',$$



Implementation Details

- Content-aware feature extraction module: ResNet-50 pre-trained on ImageNet, res5c layer
- Long-term dependencies part: a single FC layer that reduces the feature dimension from 4096 to 128, followed by a single-layer GRU network whose hidden size is set as 32
- Subjectively-inspired temporal pooling layer: τ and γ are set as 12 and 0.5, respectively.
- Training: L1 loss, Adam with an initial learning rate 0.00001 and training batch size 16 (PyTorch implementation: https://github.com/lidq92/VSFA)

Outline

- Background
- Motivation
- Method
- Experiments
- Conclusion and Future Work

Experimental Settings

- Databases
  - KoNViD-1k: 1200 videos, 960x540, 8s with 24/25/30fps
  - LIVE-Qualcomm: 208 videos, 1920x1080, 15s with 30 fps
  - CVD2014: 234 videos, 640×480 or 1280×720, 10-25s with 11-31fps
- Compared methods
  - NR-VQA: VBLIINDS, VIIDEO
  - NR-IQA: BRISQUE, NIQE, CORNIA
- Basic evaluation criteria
  - prediction monotonicity: SROCC, KROCC
  - prediction accuracy: PLCC, RMSE

Performance Comparison

Method	Overall Performance SROCC↑	KROCC↑	PLCC↑	RMSE↓	LIVE-Qualcomm [10] SROCC↑	p-value (<0.05)	KROCC↑	PLCC↑	RMSE↓
BRISQUE [27]	0.643 (± 0.059)	0.465 (± 0.047)	0.625 (± 0.053)	3.895 (± 0.380)	0.504 (± 0.147)	1.21E-04	0.365 (± 0.111)	0.516 (± 0.127)	10.731 (± 1.335)
NIQE [29]	0.526 (± 0.055)	0.369 (± 0.041)	0.542 (± 0.054)	4.214 (± 0.323)	0.463 (± 0.105)	5.28E-07	0.328 (± 0.088)	0.464 (± 0.136)	10.858 (± 1.013)
CORNIA [51]	0.591 (± 0.052)	0.423 (± 0.043)	0.595 (± 0.051)	4.139 (± 0.300)	0.460 (± 0.130)	4.98E-06	0.324 (± 0.104)	0.494 (± 0.133)	10.759 (± 0.939)
VIIDEO [28]	0.237 (± 0.073)	0.164 (± 0.050)	0.218 (± 0.070)	5.115 (± 0.285)	0.127 (± 0.137)	9.77E-11	0.082 (± 0.099)	-0.001 (± 0.106)	12.308 (± 0.881)
VBLIINDS [35]	0.686 (± 0.035)	0.503 (± 0.032)	0.660 (± 0.037)	3.753 (± 0.365)	0.566 (± 0.078)	1.02E-05	0.405 (± 0.074)	0.568 (± 0.089)	10.760 (± 1.231)
Ours	**0.771** (± 0.028)	**0.582** (± 0.029)	**0.762** (± 0.031)	**3.074** (± 0.448)	**0.737** (± 0.045)	-	**0.552** (± 0.047)	**0.732** (± 0.0360)	**8.863** (± 1.042)

Method	KoNViD-1k [12] SROCC↑	p-value	KROCC↑	PLCC↑	RMSE↓	CVD2014 [31] SROCC↑	p-value	KROCC↑	PLCC↑	RMSE↓
BRISQUE [27]	0.654 (± 0.042)	6.00E-06	0.473 (± 0.034)	0.626 (± 0.041)	0.507 (± 0.031)	0.709 (± 0.067)	7.03E-07	0.518 (± 0.060)	0.715 (± 0.048)	15.197 (± 1.325)
NIQE [29]	0.544 (± 0.040)	7.31E-11	0.379 (± 0.029)	0.546 (± 0.038)	0.536 (± 0.010)	0.489 (± 0.091)	1.73E-10	0.358 (± 0.064)	0.593 (± 0.065)	17.168 (± 1.318)
CORNIA [51]	0.610 (± 0.034)	6.77E-09	0.436 (± 0.029)	0.608 (± 0.032)	0.509 (± 0.014)	0.614 (± 0.075)	5.69E-09	0.441 (± 0.058)	0.618 (± 0.079)	16.871 (± 1.200)
VIIDEO [28]	0.298 (± 0.052)	4.22E-15	0.207 (± 0.035)	0.303 (± 0.049)	0.610 (± 0.012)	0.023 (± 0.122)	3.02E-14	0.021 (± 0.081)	-0.025 (± 0.144)	21.822 (± 1.152)
VBLIINDS [35]	0.695 (± 0.024)	6.75E-05	0.509 (± 0.020)	0.658 (± 0.025)	0.483 (± 0.011)	0.746 (± 0.056)	2.94E-06	0.562 (± 0.0570)	0.753 (± 0.053)	14.292 (± 1.413)
Ours	**0.755** (± 0.025)	-	**0.562** (± 0.022)	**0.744** (± 0.029)	**0.469** (± 0.054)	**0.880** (± 0.030)	-	**0.705** (± 0.044)	**0.885** (± 0.031)	**11.287** (± 1.943)

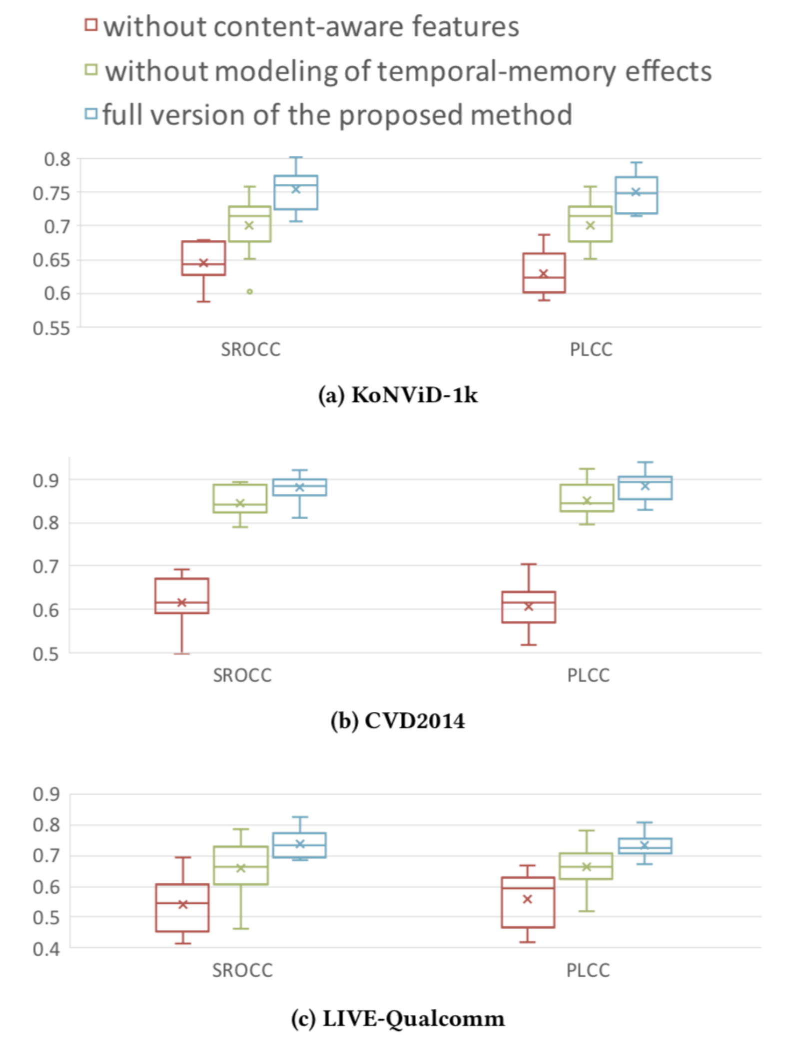
Significantly outperforms other methods by large margins

Ablation Study

On KoNViD-1k, CVD2014 and LIVE-Qualcomm,

- The removal of the content-aware features causes 14.57%, 30.00%, 26.87% decrease in terms of SROCC, where p-values are 1.10E-05, 1.76E-08, 2.47E-06.

- Temporal modeling provides 7.70%, 4.14%, 12.01% SROCC gains, where the p-values are 4.00E-04, 1.11E-04, and 8.49E-03.



(a) KoNViD-1k

(b) CVD2014

(c) LIVE-Qualcomm

Choice of Feature Extractor

- Pre-trained image classification model

Table 2: Performance of different pre-trained image classification models on KoNViD-1k.

Pre-trained model	SROCC↑	KROCC↑	PLCC↑
ResNet-50	0.755 (±0.025)	0.562 (±0.022)	0.744 (±0.029)
AlexNet	0.732 (±0.040)	0.540 (±0.036)	0.731 (±0.035)
VGG16	0.745 (±0.024)	0.554 (±0.023)	0.747 (±0.022)

Choice of Feature Extractor

- Global std pooling

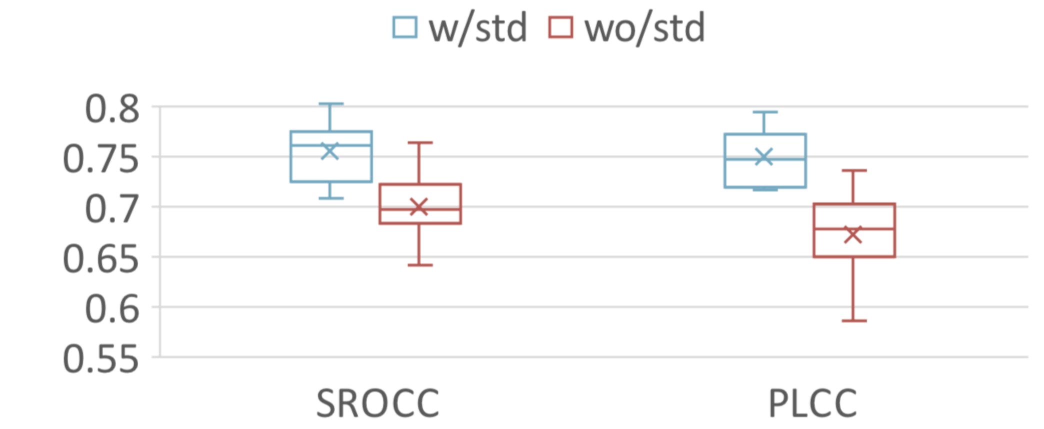


Figure 4: Effectiveness of global std pooling on KoNViD-1k.

Choices of Temporal Pooling Strategy

- Hyper-parameters in subjectively-inspired temporal pooling

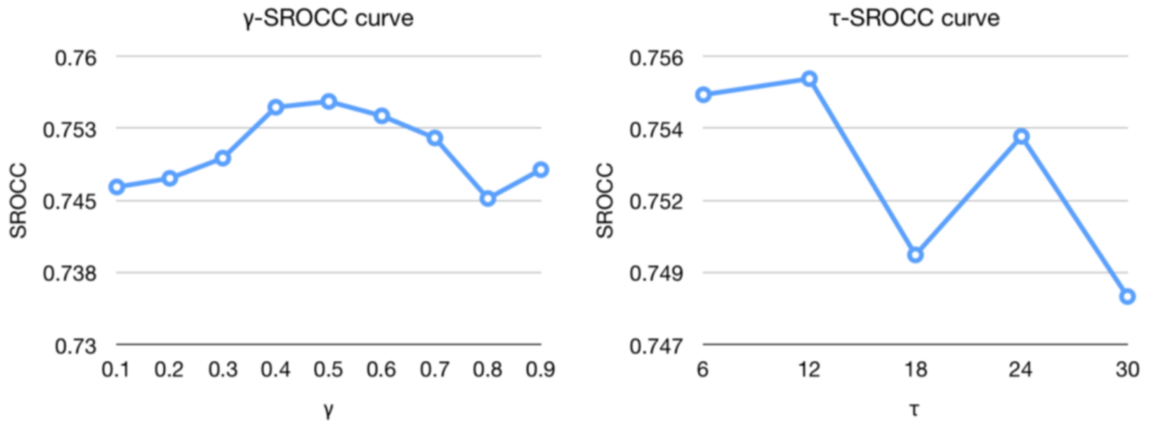


Figure 5: Performance on KoNViD-1k of different hyper-parameters in subjectively-inspired temporal pooling

Choices of Temporal Pooling Strategy

- Pooling in subjective-inspired temporal pooling

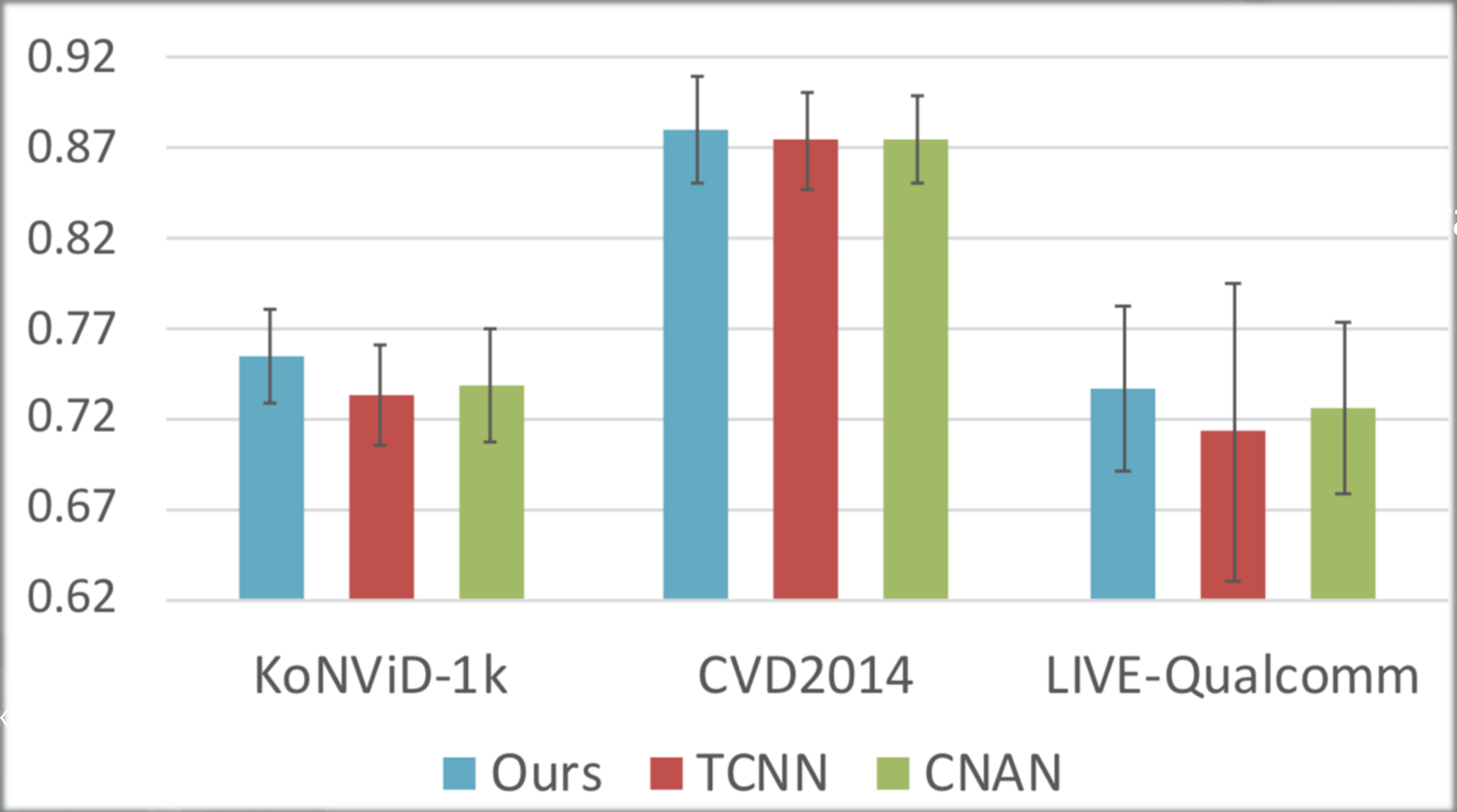
Table 3: Effectiveness of min pooling in subjective-inspired temporal pooling on KoNViD-1k.

pooling	SROCC↑	*p*-value	KROCC↑	PLCC↑
min	**0.755** (±0.025)	-	**0.562** (±0.022)	**0.744** (±0.029)
average	0.736 (±0.031)	3.04E-4	0.543 (±0.027)	0.740 (±0.027)

Choices of Temporal Pooling Strategy

•

al



[2] K

Motion information

- Motion features by optical flow statistics
  - Extract the optical low using the initialized TVNet [3]
  - Calculate optical flow statistics [4]

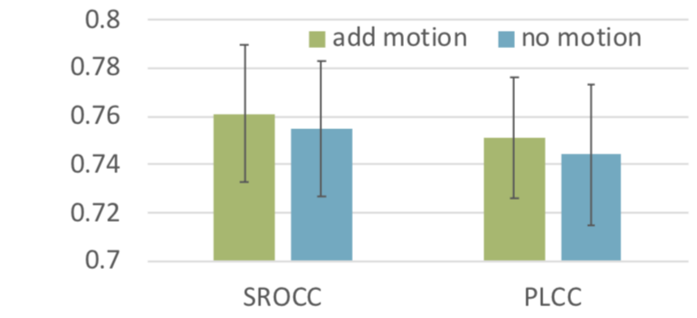


Figure 7: The performance comparison of our model with/without motion information on KoNViD-1k.

[3] Fan et al., End-to-End Learning of Motion Representation for Video Understanding, CVPR 2018
[4] Manasa and Channappayya, An optical low-based no-reference video quality assessment algorithm, ICIP 2016

Computational efficiency

Table 4: The average computation time (seconds) for four videos selected from the original databases. {xxx}frs@{yyy}p indicates the video frame length and the resolution.

Method	240frs@540p	364frs@480p	467frs@720p	450frs@1080p
BRISQUE [27]	12.6931	12.3405	41.2220	79.8119
NIQE [29]	45.6477	41.9705	155.9052	351.8327
CORNIA [51]	225.2185	325.5718	494.2449	616.4856
VIIDEO [28]	137.0538	128.0868	465.2284	1024.5400
VBLIINDS [35]	382.0657	361.3868	1390.9999	3037.2960
Ours	269.8371	249.2085	936.8452	2081.8400

Our method is faster than VBLIINDS, the method with the second best performance.

The implementation of our method can be accelerated to 30x faster or more by simply switching the CPU mode to the GPU mode.

Outline

- Background
- Motivation
- Method
- Experiments
- Conclusion and Future Work

Conclusion and Future Work

- A novel NR-VQA method for in-the-wild videos by incorporating content-dependency and temporal-memory effects.
  - Superior performance and ablation study on three in-the-wild VQA databases

- In the further study, we will consider embedding the spatio-temporal attention models into the framework
  - When and where the video is important for the VQA problem.

Thank you