AHG16: Lightweight Multi-resolution CFTE Model and Color Calibration Post-processing Algorithm for Generative Face Video Compression

JVET-AM0058

Zihan Zhang, Shanzhi Yin, Shiqi Wang, Bolin Chen, Ru-Ling Liao, Jie Chen, Yan Ye

City University of Hong Kong, Alibaba Group

Jul. 2025

1 Proposal Review

香港城市大學 City University of Hong Kong

達摩院 DAMO ACADEMY

Review:

- The current AHG16 GFVC software employs **a single-resolution lightweight CFTE model**, optimized through methods from JVET-AG0139 and JVET-AH0109.
  - JVET-AG0139 proposed the integration of **depthwise separable convolution** into **the dense motion module** of the CFTE method.
  - JVET-AH0109 proposed to **reduce the number of channels** and selectively replace some of the 3x3 convolutions within **the generation module** with **depthwise separable convolution layers**.
- JVET-AL0156 proposed to utilize **channel-wise distribution calibration** as an **optional post-processing step** in GFVC decoding to adjust the color distribution of the generated frames to match that of the original frames.

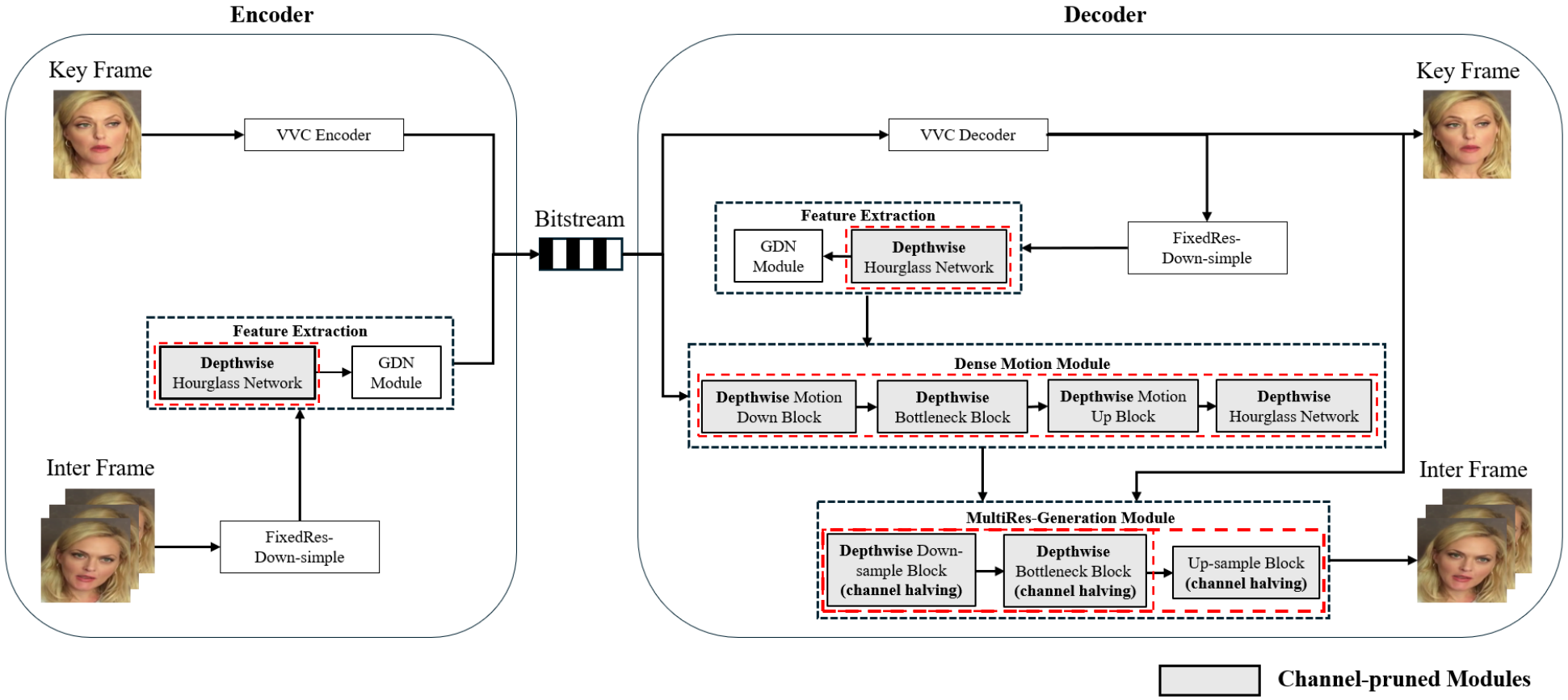
1 Proposal Review

- **Problem**:
  - The lightweight methods in the current AHG16 software were implemented based on the **single-resolution CFTE model**.
  - **The hourglass networks** in the feature extraction and dense motion modules still contain **excessive parameters**.
  - The generated facial frames occasionally exhibit **color shift issues**.

- **Contribution**: Building on the current AHG16 GFVC software,
  - we propose to optimize the multi-resolution CFTE model through two primary strategies: **architectural (i.e., depthwise separable convolution) and operational (i.e., pruning) optimizations**.
  - we propose incorporating the **color calibration method** as a **post-processing step** into the AHG16 GFVC software to address the occasional color shift issues observed in some generated video frames.

2 Proposed Lightweight Multi-resolution CFTE Method

■ The framework of lightweight multi-resolution CFTE

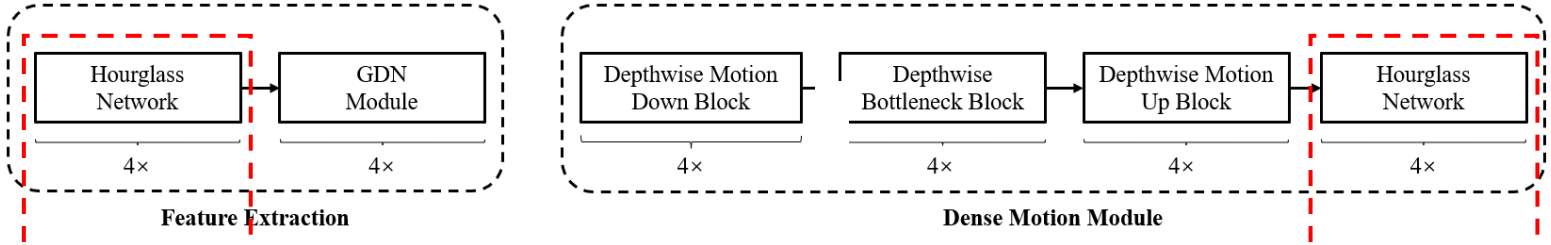


- A subset of 3×3 convolutions was replaced with **depthwise separable convolutions**.
- The number of channels was **halved** in the **generation modules**.
- A **pruning strategy** was applied to eliminate redundant channels.

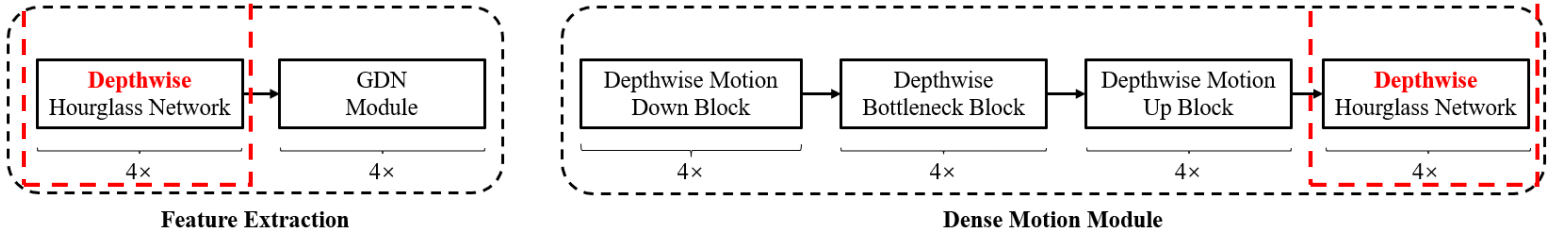
2 Architectural Optimization

- The **hourglass networks** account for the majority of parameters in both the feature extraction and dense motion modules.

Module	Feature Extraction			Dense motion Module		
Network Structure	**Hourglass Network**	**Others**	**Total**	**Hourglass Network**	**Others**	**Total**
Lightweight CFTE (Single-resolution)	**13.2M**	0.9M	14.1M	**9.3M**	3.0M	12.3M



(a) Feature extraction and dense motion module in the current AHG16 GFVC software

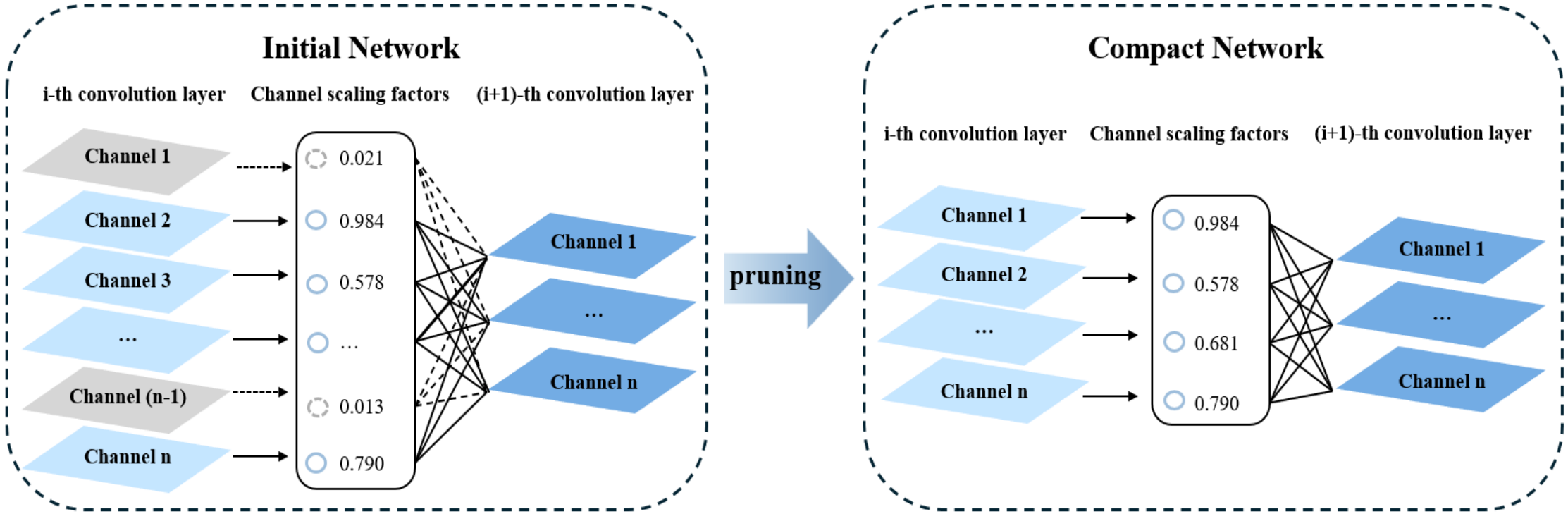


(b) Feature extraction and dense motion module in this contribution

- Compared to the feature extraction and dense motion modules in the current AHG16 software, we replace all 3×3 convolution layers **in the hourglass network** of both modules with **depthwise separable convolutions**.

2 Operational Optimization

■ Channel pruning mechanism



- We use the magnitude of **scaling factors in BatchNorm layers** to **identify and remove** less important channels in the corresponding convolutional layers.
- Then we **fine-tune** the **pruned lightweight model** to compensate for loss caused by channel removal.

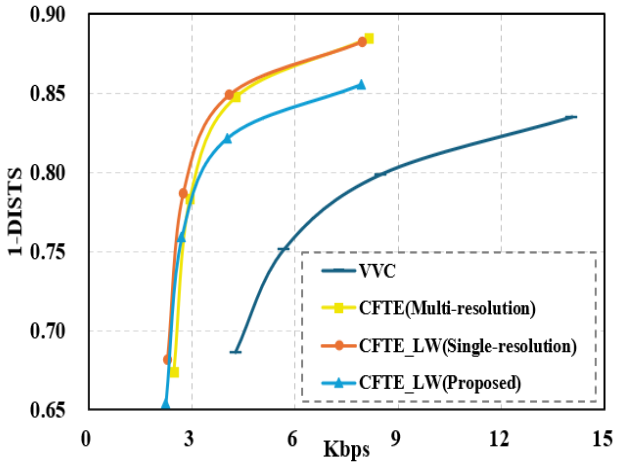
3 Lightweight Results: Complexity Reduction

- The number of parameters and kMACs/pixel of the multi-resolution CFTE, the single-resolution lightweight CFTE, and the proposed multi-resolution lightweight CFTE.

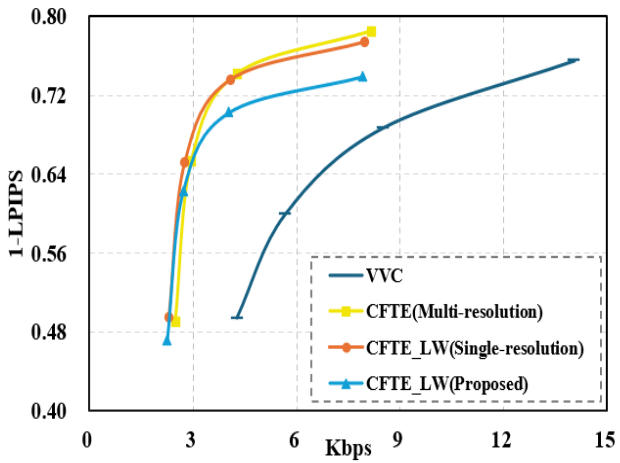
	Parameters	kMACs/pixel@256px	kMACs/pixel@512px
Multi-resolution CFTE	58.2M	1072.8	392.9
Single-resolution Lightweight CFTE	28.3M	209.3	-
Proposed Multi-resolution Lightweight CFTE	6.0M	180.9	79.9

3 Lightweight Results: Rate-distortion Performance

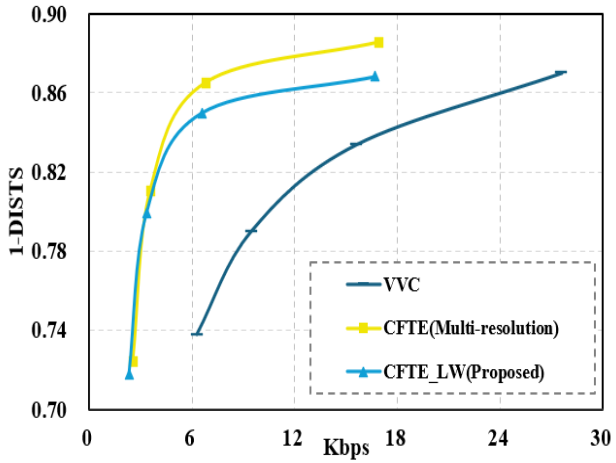
- The Rate-Distortion performance of multi-resolution CFTE, single-resolution lightweight CFTE, and proposed lightweight CFTE as well as VVC anchor in terms of rate-DISTS and rate-LPIPS.



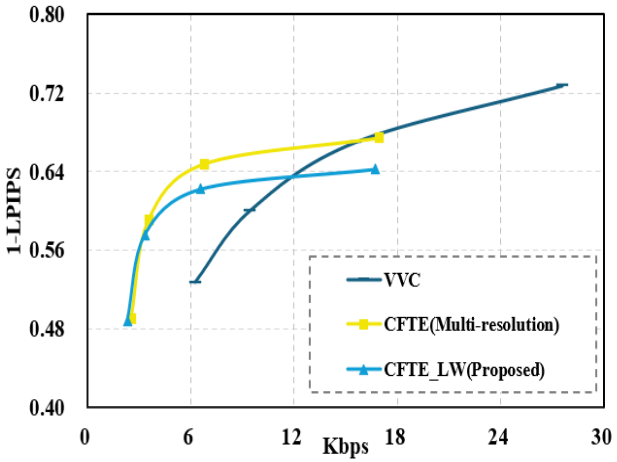
(a) 256×256, Rate-DISTS



(b) 256×256, Rate-LPIPS



(c) 512×512, Rate-DISTS



(d) 512×512, Rate-LPIPS

Sequences	Over VTM-22.2	
	DISTS	LPIPS
Class A(256px)	-60.21%	-59.66%
Class B(256px)	-46.34%	-38.94%
Class C(512px)	-66.65%	-52.61%
Class D(512px)	-59.24%	-43.00%
Overall	-57.63%	-47.86%

3 Lightweight Results: Subjective Results

■ Class A Seq 001

Original

VVC@7.10k

CFTE@5.95k

CFTE-LW(Single-resolution)@5.75k

CFTE-LW(Proposed)@5.62k

■ Class B Seq 005

Original

VVC@7.08k

CFTE@6.08k

CFTE-LW(Single-resolution)@5.84k

CFTE-LW(Proposed)@5.87k

3 Lightweight Results: Subjective Results

■ Class C Seq 013

Original

VVC@5.25k

CFTE@4.43k

CFTE-LW(Proposed) @4.22k

■ Class D Seq 004

Original

VVC@10.38k

CFTE@9.12k

CFTE-LW(Proposed) @8.84k

4 Color Shift in Generated Facial Frames

- The generated facial frames in the proposed lightweight CFTE occasionally exhibit **color shift issues**.


Original **CFTE-LW(Proposed) @5.87k**


Original **CFTE-LW(Proposed) @5.62k**

- Additionally, the generated facial frames in other GFVC models in the AhG16 software occasionally exhibit **color shift issues**, taking CFTE as an example.

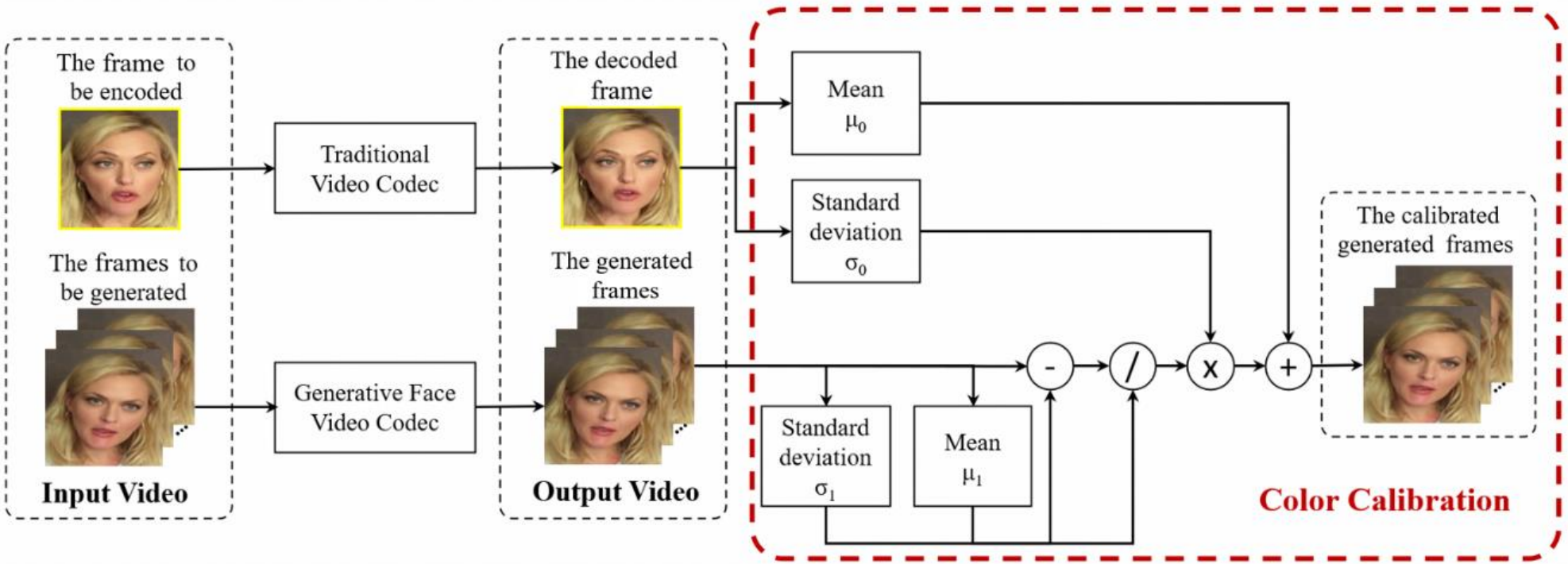

Original **CFTE@4.43k**


Original **CFTE@9.12k**

- We propose incorporating the **color calibration method** as a **post-processing step** into the AHG16 GFVC software.

5 Color Calibration for Generative Face Video Coding

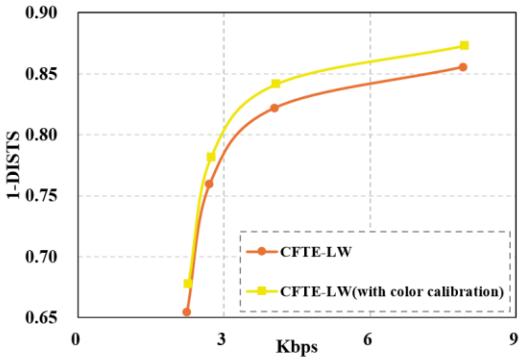
■ Color Calibration Post-processing Algorithm



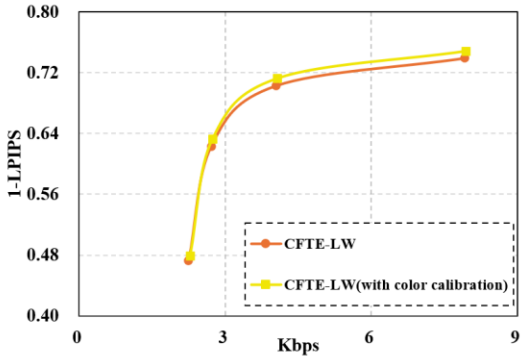
- At the decoder after frame generation, the distribution parameters of the **decoded frame and generated frame** are calculated for each color component.
- The proposed channel-wise color calibration is performed by **normalizing each channel of the generated frame** with its own distribution parameters and then shifting to the target distribution of **the corresponding VVC decoded frame**.

6 Color Calibration Results: Lightweight CFTE model

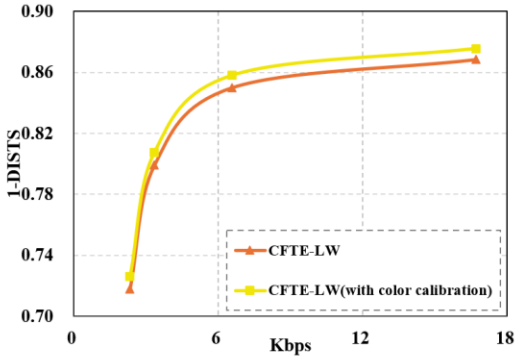
- The RD performances of proposed lightweight CFTE (with and without color calibration) at different resolutions in terms of rate-DISTS and rate-LPIPS, with visual quality comparison of VVC .



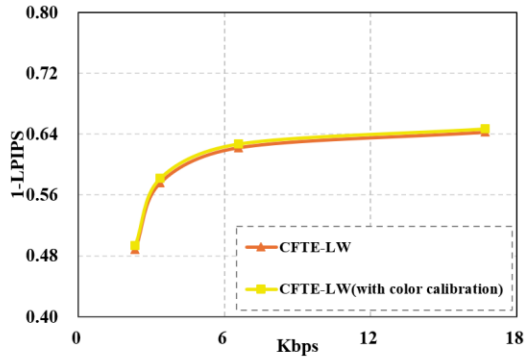
(a) 256×256, Rate-DISTS



(b) 256×256, Rate-LPIPS



(c) 512×512, Rate-DISTS



(d) 512×512, Rate-LPIPS


Original


VVC@7.10k


CFTE-LW@5.87k

Color calibration


CFTE-LW@5.87k


Original


VVC@10.38k


CFTE-LW @8.84k


CFTE-LW @8.84k

6 Color Calibration Results: GFVC models

■ **Class B Seq 005**

Original

VVC@7.10k

CFTE @6.08k

DAC@5.18k

FOMM @6.88k

FV2V @6.37k

CFTE @6.08k

DAC@5.18k

FOMM @6.88k

FV2V @6.37k

Color calibration

6 Color Calibration Results: GFVC models

■ **Class D Seq 004**

Original

VVC@10.38k

CFTE @9.12k

DAC @8.20k

FOMM @10.31k

FV2V @9.50k

CFTE @9.12k

DAC @8.20k

FOMM @10.31k

FV2V @9.50k

Color calibration

6 Conclusion

- This contribution proposes **a lightweight multi-resolution CFTE model for generative face video compression** that significantly reduces the number of parameters and computation while maintaining competitive RD performance.

- We also introduce **color calibration as an optional post-processing step** in GFVC decoding to solve the color shift issue and improve the subjective quality.

- **We recommend to 1) replace the existing lightweight CFTE model in the AHG16 GFVC software with the proposed lightweight multi-resolution CFTE model, and 2) adopt color calibration algorithm as a post-processing technique in the AhG 16 software.**