Improving Reconstruction Fidelity in Generative Face Video Coding using High-Frequency Shuttling

Goluck Konuko, Giuseppe Valenzise, Anthony Trioux

Improving Reconstruction Fidelity in Generative Face Video Coding using High-Frequency Shuttling

Goluck Konuko[1], Giuseppe Valenzise[1], Anthony Trioux[2]
[1]*Univ. Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, France*
[2]*Xidian University, School of Telecommunications Engineering, Xi'an, China*
[1]{goluck.konuko, giuseppe.valenzise}@centralesupelec.fr, [2]anthony_trioux@xidian.edu.cn

***Abstract*—Generative face video coding (GFVC) schemes applied to talking head videos have recently demonstrated significant coding gains compared to traditional coding frameworks, particularly at ultra-low bitrates. Despite advancements in the field, these methods still face challenges in handling large pose and facial expression changes, as well as (dis-)occlusions. Recently, a hybrid approach (HDAC+) that combines a low-quality video coded with a conventional codec and animation-based coding has been proposed for standardization and shown to partially address these issues. Although HDAC+ shows promising results, it still struggles with generating accurate images. In this paper, we propose HDAC-HF, an improvement to the reconstruction process in HDAC+. Based on empirical observations that some pose and expression details are lost during animation, we introduce a high-frequency (HF) shuttling mechanism to enhance reconstruction fidelity at the decoder side, inspired by recent advancements in video super-resolution. By enhancing the flow of high-frequency details in the feature domain, we improve the reconstruction of facial expressions and poses. Qualitative and quantitative experiments confirm that the proposed method improves reconstruction without any additional bitstream or signaling cost compared to the baseline HDAC+ codec.**

***Index Terms*—Video Compression, Generative Face Video Coding, Frequency Shuttling, Video Conferencing**

I. Introduction

Achieving low-bitrate compression with high-quality reconstruction remains a significant challenge in video coding. Although end-to-end learned image [1] and video compression [2]–[4] frameworks have shown competitive performance against traditional codecs like HEVC [5] and VVC [6], they are generally tailored for broad applications and fall short in addressing the specific needs of video conferencing. Deep image animation offers a promising approach to compress talking head videos, that are characteristic of video conferencing, at ultra-low bitrates, *i.e.,* below 10kbps. Inspired by the First Order Animation model [7], GFVC frameworks such as [8]–[13] demonstrated viable video reconstruction with high perceptual quality at very low bitrates. This is largely due to the compactness of the bitstream information required to decode very long video sequences. The underlying principle is video self-reenactment. That is, given a reference frame (typically the first frame in a sequence), the subsequent frames can be reconstructed using a generative network given the motion predictors between the reference frame and the following target frames. The effectiveness of the GFVC methods is shown on their ability to reproduce motion from a small set (hence low bitrate) of motion keypoints/landmarks on a speaker's face. However, the initial GFVC methods were demonstrably limited in the complexity of motions they could reproduce and had limited perceptual fidelity when there are large pose changes between the reference and the target frames. Occlusions in the foreground and dis-occlusion of background details also limit the effectiveness of these methods. Further, the generative nature of these frameworks limits their performance in reconstructing videos with satisfactory pixel fidelity.

The observed limitations of purely animation-based GFVC frameworks [8]–[11] have been primarily addressed through hybrid [12] or predictive [13] animation frameworks. In particular, the Hybrid Deep Animation Codec (HDAC) [12] functions similarly to a layered multiple description scheme: one layer is encoded at low quality using a conventional codec such as HEVC or VVC, while an additional stream contains animation information, such as keypoints and intra-coded frames of a Group of Pictures (GoP). Both streams can be decoded independently, but HDAC decodes them jointly, using the conventionally coded layer to obtain a structural and content prior that enhances the perceptual and pixel fidelity of the animated video. Recently, this approach has attracted attention in GFVC standardization, since the animation bitstream can be conveniently transmitted through SEI (Supplemental Enhancement Information) messages to the decoder. Therefore, an extended version of HDAC that supports multiple quality levels for the conventionally coded layer through a single model has been proposed for standardization in the JVET GFVC Adhoc group [14]. In this paper, we employ this version of HDAC as the baseline for our method. To avoid confusion with the initial version in [12], we denote it as HDAC+ in the following.

This paper aims to enhance the effectiveness of HDAC+ by introducing a high-frequency shuttling mechanism to improve the reconstruction fidelity of HDAC-generated images. Although HDAC+ has significant quality gains compared to animation-only decoding approaches, we have observed that the reconstructed video can still exhibit distortions, such as improperly reconstructed facial expressions. We hypothesize that these reconstruction errors are partially due to a loss of feature details throughout the reconstruction process. Recent research on video super-resolution has demonstrated that high-frequency shuttling *i.e.* filtered skip connections, can

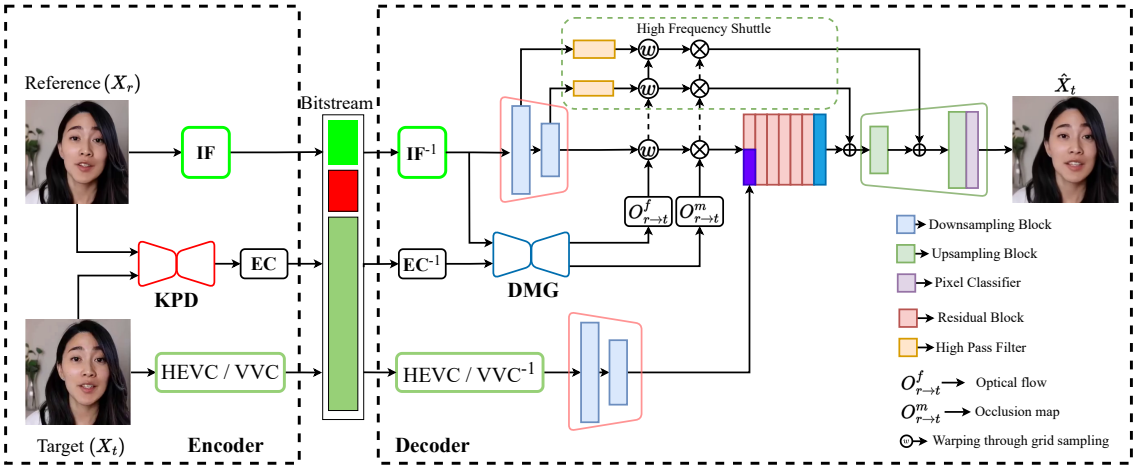

Fig. 1: The proposed HDAC-HF architecture comprising two compressed bitstreams: the animation-based GFVC bistream and a HEVC/VVC bistream used as a conditional base layer during the frame reconstruction process. Our method proposes a high-frequency shuttling mechanism that enables improved reconstruction fidelity for facial expressions and poses without additional signaling cost compared to the baseline HDAC+. IF: Intra Frame coding; KPD: Keypoint Detector; EC: Entropy Coding; DMG: Dense Motion Generation.

significantly improve reconstruction quality in deep video generation [15]. This is in-turn inspired by the need to minimize aliasing in autoencoder frameworks [16]. While in both cases, the objective is anti-aliasing the generated pictures, we observe that this architecture can be used to transfer fine-scale information such as edge details that are typically lost in the generative framework such as HDAC [12]. As a result, we propose a novel motion and occlusion-aware high frequency shuttling mechanism that is applicable to the animation framework. Our experiments on the extension to the HDAC+ model, i.e., HDAC-HF, confirm that high-frequency shuttling is also beneficial for facial animation, leading to better preservation of pixel and perceptual detail.

II. Preliminaries

Generative Face Compression frameworks [8]–[10] using animation relies on a compressed reference frame X_r and sparse motion keypoints extracted from the reference frame K_r and each target frame K_t. A generative autoencoder uses the reference frame features ε_r and keypoints to predict the optical flow between the reference and the target ($O^f_{r\to t}$) and an occlusion mask ($O^m_{r\to t}$) that guides the generator in-painting process. The reference features are subsequently transformed as follows:

$$\hat{\varepsilon}_r = O^m_{r\to t} \cdot w(O^f_{r\to t}, \varepsilon_r) \tag{1}$$

where $w(\cdot,\cdot)$ denotes feature warping through the optical flow, using grid sampling. However, this process relies on the generalization ability of the motion predictor and close alignment between the reference and target frames in the feature space. Unfortunately, the latter can be inaccurate, especially in the presence of large pose changes as well as facial occlusions.

The hybrid generative face compression framework (HDAC) [12] improves the capabilities of the generative face compression (GFVC) process by incorporating a low-quality version of the target frame, X^b_t, which is encoded and transmitted using a standard codec such as HEVC [5] or VVC [6]. During the animation process, the features (ε^b_t) of this low-quality frame are used as a conditioning input to the animation generator.

Improvements to the optimization process of the HDAC framework were included in the JVET-AH0114 standardization effort. We refer to the standardized HDAC model as HDAC+ [14]. HDAC+ enhanced reconstruction fidelity of HDAC by including an additional MSE term in the loss function at training time. The MSE loss is weighted differently depending on the quantization level of the VVC layer. Furthermore, HDAC+ has been trained using multiple VVC quantization parameters, resulting in models that generalize better across different test conditions [14]. In addition to using the standardized model HDAC+ as a baseline in this paper, we also employ the test conditions defined in [14], which enables comparing performance fairly and reproduce the results. We use this architecture to show the potential of high-frequency shuttling in improving pixel fidelity for hybrid animation frameworks.

III. Proposed Method

The proposed coding framework (HDAC-HF), shown in Fig. 1 offers improvements over the original HDAC architecture [12]. We present the novel high-frequency shuttling mechanism for HDAC in Section III-A, with a focus on details coming from feature maps, the choice of high-pass filter, and the intuition behind its effectiveness.

A. High Frequency Shuttling

The hybrid animation framework achieves notably higher pixel fidelity in the reconstructed output compared to animation-only GFVC frameworks. However, previous works have demonstrated its limitations in recovering expression and pose details in animated videos. In this paper, we hypothesize that this loss of detail is due to the generation process, where some relevant feature details are lost during the animation and upsampling/reconstruction blocks. Nonetheless, high-frequency details necessary for accurately reconstructing facial expressions and poses are still available in the Intra frames, which are encoded using a conventional codec with higher quality than the VVC coded layer. The basic idea of HDAC-HF is to enhance the animation and generation of target frames by *transferring details from the Intra frame features to the generator decoder*.

To this end, we draw inspiration from the concept of high-frequency shuttling [15] used for high-fidelity video super-resolution. In [15], high-frequency details are obtained by subtracting a low-pass filtered version of the features and then transferred through the network as skip connections. This method has been shown to improve edge detail reconstruction in super-resolved videos. Our approach follows a similar strategy by directly filtering the features through a high-pass filter. Additionally, since the features are used for target frame reconstruction rather than reference frame reconstruction, we apply warping using dense reconstructed optical flow and an occlusion mask directly learned within the animation generator. This process, similar to the conventional generation method of HDAC+ illustrated in Figure 1, maximizes the effectiveness of the skip features.

More specifically, we construct a feature pyramid with high-frequency details from the encoder network of the animation generator, where each feature set is computed as follows:

$$\varepsilon^i_{hf} = f_{hpf}(\varepsilon^i_r), \forall i \in (1, 2) \tag{2}$$

where ε^i_{hf} are high frequency features, f_{hpf} is a high pass filter and ε^i_r are the reference frame features at level i of the downsampling encoder. The high-frequency features are then aligned with the transformed features used in the image generation process and added at each level of the upsampling decoder. To align the high-frequency features, the optical flow and occlusion mask predicted between the reference and target frames are applied as follows:

$$\hat{\varepsilon}^i_{hf} = O^m_{r\to t} \cdot w(O^f_{r\to t}, \varepsilon^i_{hf}). \tag{3}$$

In the decoder upsampling network, the high frequency features are added to the transformed reference features before each upsampling block as follows:

$$\hat{\varepsilon}^1_t = d^{up}_1(\hat{\varepsilon}^1_r + \hat{\varepsilon}^2_{hf}) \tag{4}$$

$$\hat{\varepsilon}^2_t = d^{up}_2(\hat{\varepsilon}^1_t + \hat{\varepsilon}^1_{hf}), \tag{5}$$

where d^{up}_i are the decoder upsampling blocks and $\hat{\varepsilon}^2_t$ are the final features maps used for target frame prediction.

-1	-1	-1
-1	8	-1
-1	-1	-1

Filter 1

0	1	0
1	-4	1
0	1	0

Filter 2

Fig. 2: Approximations of the Laplacian filter kernel.

To validate the effectiveness of high-frequency shuttling, we consider two variants of a discrete Laplacian filters shown in Figure. 2 : a simple high-pass filter, denoted as *Filter 1* filter and a second order high pass filter denoted as *Filter 2*.

RESULTS AND DISCUSSION

B. Datasets, Training and Evaluation

We reproduce the standardized model (HDAC+) [14] and implement the proposed high-frequency shuttling mechanism. The training dataset follows GFVC conditions and consists of 18k videos from the VoxCeleb Dataset with a resolution of 256×256 pixels. The training videos are encoded with VVC (VTM 12) and HEVC (HM-16), with QP values of 50,48,46 and 44. At training time, the base layer frame for each target frame in a batch is sampled randomly between those QP values. The loss function is proportionally adjusted to account for the quality of the base layer frame. For HDAC+ [14] we train two models, *i.e.,* one for each base layer codec. Our proposed method (HDAC-HF) is trained instead with VVC base layer only. Despite this, at inference time, we observe that the models maintain robust performance even when the base layer is changed to HEVC, while for HDAC+ this is not the case and a color shift has been observed when a different codec is used at inference. We follow the JVET-AH0114 [14] test protocol using the standardized settings for all GFVC frameworks as well as HEVC and VVC anchor methods.

C. Rate-Distortion Performance Evaluation

Table I presents the relative bitrate savings of our proposed coding framework, HDAC-HF with Laplacian filter, compared to the baseline HDAC+, a residual GFVC coding method called RDAC [13], and conventional HEVC and VVC codecs. Five quality metrics are used to evaluate performance: FSIM, MS-SSIM, LPIPS, msVGG, and DISTS.

As shown in the table, HDAC-HF achieves significant gains compared to conventional codecs and the residual-based approach RDAC. RDAC's residual coding consumes significant bitrate, reducing the overall efficiency of that codec. The comparison with HDAC+ is more nuanced. The BD-BR gains are favorable but more limited, with an increase in the average rate for similar DISTS quality. We attribute this loss in DISTS, which focuses on texture preservation, to the loss of detail in some textures not necessarily associated with facial expressions or pose details, as shown in the visual examples in the next section. These features might be somewhat lost in the feature encoder of the intra-frame. Further study on this phenomenon is ongoing.

Finally, we do not compare BD-BR with other GFVC approaches such as CFTE [10] or DAC [8], as those methods

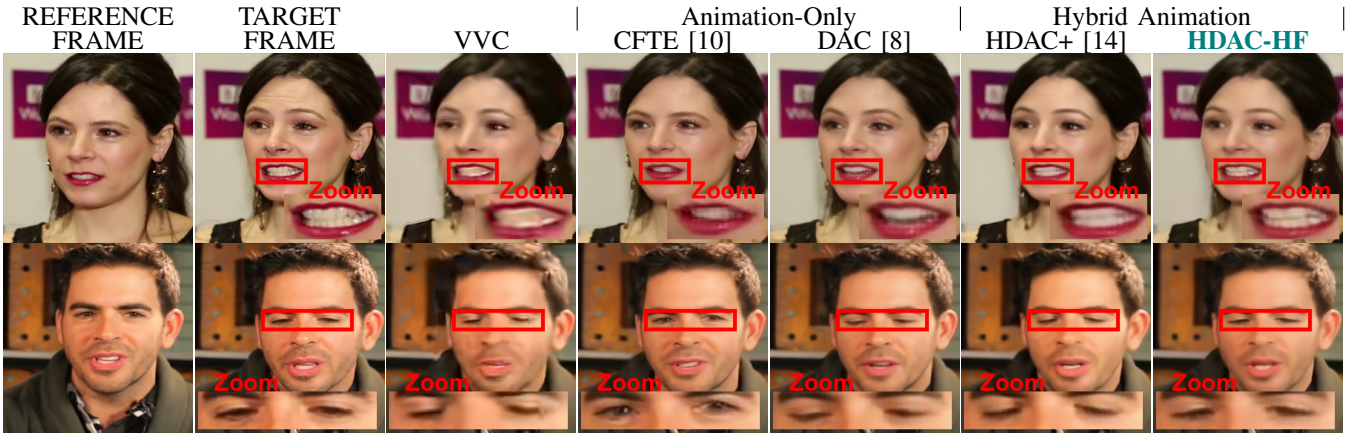

Fig. 3: **Visual Comparison:** HDAC-HF significantly outperforms conventional codecs and improves facial detail reconstruction over HDAC+, such as the teeth (**top row**) and the eyes (**bottom row**).

only target extremely low bitrates, resulting in RD curves that have little or no overlap with ours, impeding a fair BD-BR computation.

TABLE I: Bjøntegaard-Delta Bitrate (BD-BR) Performance of HDAC-HF (Using VVC Base Layer)

	PSNR-Y	MS-SSIM	LPIPS	msVGG	DISTS
RDAC [13]	-59.36	-54.42	-5.34	-44.46	-17.68
HDAC+ [14]	-6.82	-6.20	-3.23	-11.35	5.68
HEVC	-38.42	-47.12	-75.53	-63.17	-67.16
VVC	9.30	-21.66	-62.63	-47.30	-59.64

D. Visual Results

Fig. 3 shows a visual comparative analysis of the proposed framework versus the previous GFVC frameworks as well as VVC with low-delay configuration. In both examples, we observe a significant enhancement in perceptual quality when using our proposed HDAC-HF method (with Laplacian filter), as compared to VVC at low bitrates (∼12kbps). Notably, the animation-only GFVC methods, such as CFTE and DAC, demonstrate good perceptual quality but struggle with accurate facial detail reconstruction. This limitation arises from their inability to effectively predict large motions and reproduce intricate facial expressions. The previously developed hybrid coding framework, HDAC+, addresses some of these limitations by improving both perceptual and pixel fidelity over DAC and CFTE. Our proposed method, HDAC-HF, further enhances the reconstruction of facial details, such as teeth and eyes. In Example 1 (**top row**), we observe an improvement regarding the generated mouth, attributed to the HF shuttling mechanism. Similarly, in Example 2 (**bottom row**), HDAC-HF shows superior reconstruction of the eyes, achieving a closer alignment with the ground truth target frame.

E. Impact of the high-frequency filter

As mentioned in Section 3, the choice of the high-frequency filter has a significant impact in terms of reconstruction performance. In Fig. 4, we report the average RD curves on LPIPS. We observe that, for this task, the Laplacian filtering

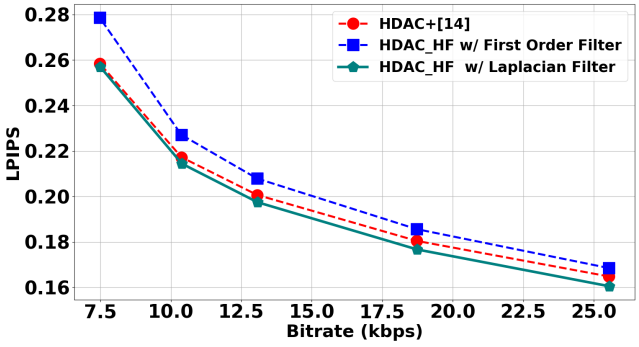

Fig. 4: Impact of the choice of the high-pass filter.

is generally better than a simple high pass filter (analogous to the residual obtained from the low pass filter proposed by [15]). However, the performance of the two kinds of filtering, as well as the relative gains over HDAC+, strongly depend on the adopted metric. For a pixel-based metric such as MS-SSIM, the gains are small but consistent, while for LPIPS, the simple high pass filter seems to deteriorate the performance of HDAC+ and is the motivation for selecting Laplacian filter with better edge detection and noise suppression characteristics. We are currently investigating the reasons for this phenomenon, and whether using higher-order filtering might further increase the performance.

IV. Conclusion

We investigate the problem of preserving facial expression and pose details in generative face video coding. Using a hybrid animation codec as our baseline, we demonstrate that transferring high-frequency details in the feature space from the reference intra-frame to the target reconstructed frames can enhance the reconstruction of facial features. The initial results are promising, further improving the performance of GFVC codecs compared to traditional coding schemes at low to very low bitrates. However, several issues and questions remain unresolved. Specifically, we hypothesize that a learnable dynamic high-pass filter similar to [17], could have significant impact performance with a trade-off on increased decoder complexity.

REFERENCES

[1] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations (ICLR)*, 2017.

[2] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 11 006–11 015.

[3] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 18 114–18 125, 2021.

[4] T. Ladune and P. Philippe, "AIVC: Artificial intelligence based video codec," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 316–320.

[5] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[6] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC)," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021.

[7] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in neural information processing systems (NeurIPS)*, vol. 32, 2019.

[8] G. Konuko, G. Valenzise, and S. Lathuilière, "Ultra-low bitrate video conferencing using deep image animation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4210–4214.

[9] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 10 039–10 049.

[10] B. Chen, Z. Wang, B. Li, R. Lin, S. Wang, and Y. Ye, "Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression," in *2022 Data Compression Conference (DCC)*. IEEE, 2022, pp. 13–22.

[11] Z. Chen, M. Lu, H. Chen, and Z. Ma, "Robust ultralow bitrate video conferencing with second order motion coherency," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2022, pp. 1–6.

[12] G. Konuko, S. Lathuilière, and G. Valenzise, "A hybrid deep animation codec for low-bitrate video conferencing," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 1–5.

[13] G. Konuko, S. Lathuilière, and G. Valenzise, "Predictive coding for animation-based video compression," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 2810–2814.

[14] B. Chen, Y. Ye, G. Konuko, G. Valenzise, S. Yin, and S. Wang, "AHG 16: Updated common software tools for generative face video compression," in *The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, doc. no. JVET-AH0114*, 2024.

[15] X. Yiran, P. Taesung, Z. Richard, Z. Yang, S. Eli, L. Feng, H. Jia-Bin, and L. Difan, "VideoGigaGAN: Towards detail-rich video super-resolution," 2024.

[16] X. Zou, F. Xiao, Z. Yu, and Y. J. Lee, "Delving deeper into anti-aliasing in convnets," *International Journal of Computer Vision*, vol. 131, pp. 67–81, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221246286

[17] S. A. Magid, Y. Zhang, D. Wei, W.-D. Jang, Z. Lin, Y. Fu, and H. Pfister, "Dynamic high-pass filtering and multi-spectral attention for image super-resolution," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4268–4277.