Learned Video Compression for Attribute images in Video-based Point Cloud Compression

Seungmin Noh[a], Yongheon Kim[a], Kaeun Lee[a], and Haechul Choi[a][‡]

Abstract

3D point clouds are widely used in applications such as autonomous driving and augmented reality, and Video-based Point Cloud Compression (V-PCC) has been the primary approach for compressing such data. V-PCC projects 3D data onto two-dimensional (2D) images and compresses them using traditional video coding standards. In this study, we propose replacing conventional codecs with a neural network‑based model, Deep Contextual Video Compression (DCVC), to compress 2D attribute images of point cloud. To adapt DCVC, originally trained on natural images, to attribute images, the proposed method employs an N-stage cascaded training strategy for fine-tuning, Experimental results on MPEG-I Common Test Condition sequences show that the proposed model achieves an average BD-rate gain of 28.00% over the baseline DCVC model and provides superior reconstruction quality, even at low bitrates. These findings demonstrate the feasibility of deploying learned video codecs for point cloud compression.

Keyword : Point Cloud Compression, Video-based Point Cloud Compression, Deep Contextual Video Compression, N-stage cascaded training strategy

Ⅰ. Introduction

A point cloud is a data format used to represent objects, surfaces, or environments in three-dimensional (3D) space, consisting of a set of points distributed within that space. Recently, point cloud services have expanded beyond traditional graphics and augmented reality to location-based applications such as autonomous driving[1], further emphasizing their importance. However, point cloud data contains a vast number of points, resulting in large data sizes, and efficient storage and transmission have become critical challenges. To address this issue, point cloud compression techniques are essential[2].

Reflecting this demand, the Moving Picture Experts Group (MPEG) under ISO/IEC JTC 1 developed the Video-based Point Cloud Compression (V-PCC) standard[3]. V-PCC projects 3D point clouds into two-dimensional (2D) images, generating attribute images that represent color and texture information, geometry images that encode depth information, and occupancy maps that in-

a) Hanbat National University

‡ Corresponding Author : Haechul Choi
E-mail: choihc@hanbat.ac.kr
Tel: +82-42-821-1196
ORCID: https://orcid.org/0000-0002-7594-0828

dicate the spatial occupancy of points[4]. These 2D images are then compressed into bitstreams using traditional video codecs for transmission or storage.

V-PCC is codec-agnostic, allowing for the application of various video codecs[5]. This flexibility highlights the potential of incorporating recently emerging learned video codecs into V-PCC[6]. However, existing learned video codecs are primarily trained on natural images and therefore fail to fully capture the irregular and nonuniform characteristics of attribute images in point clouds. Considering this limitation, enabling learned codecs to adapt to the properties of the projected images of point cloud data would significantly enhance their contribution to V-PCC compression.

In this study, we propose an effective method of applying Deep Contextual Video Compression (DCVC)[7], a learned video codec, to the attribute image compression in V-PCC. DCVC leverages convolutional neural networks (CNNs) to capture spatial and temporal contexts in detail and jointly optimizes the encoder and decoder in an end-to-end manner, thereby achieving high compression efficiency and reconstruction quality. These characteristics allow it to preserve essential details in point cloud attribute images while effectively reducing redundant information during compression.

To address the limitations of existing DCVC models trained on natural images, this study fine-tunes DCVC on point cloud attribute images containing color and texture information, leading to improved compression efficiency optimized for attribute images. Moreover, by introducing an optimized training strategy, this study demonstrates the feasibility of applying learned video codecs to point cloud compression.

Ⅱ. Proposed method

1. V-PCC with a Learned Video Codec

In this study, we propose a method to efficiently compress attribute images in V-PCC by projecting 3D point clouds into two-dimensional (2D) attribute images and employing DCVC, a learned video codec. Conventional V-PCC adopts signal processing‐based video codecs as the standard for data compression, with High Efficiency Video Coding (HEVC, H.265)[8] being the most representative example. However, because V-PCC is not tied to a specific codec, it has the flexibility to incorporate different codecs. This property highlights the potential of integrating recently emerging learned video compression methods into V-PCC.
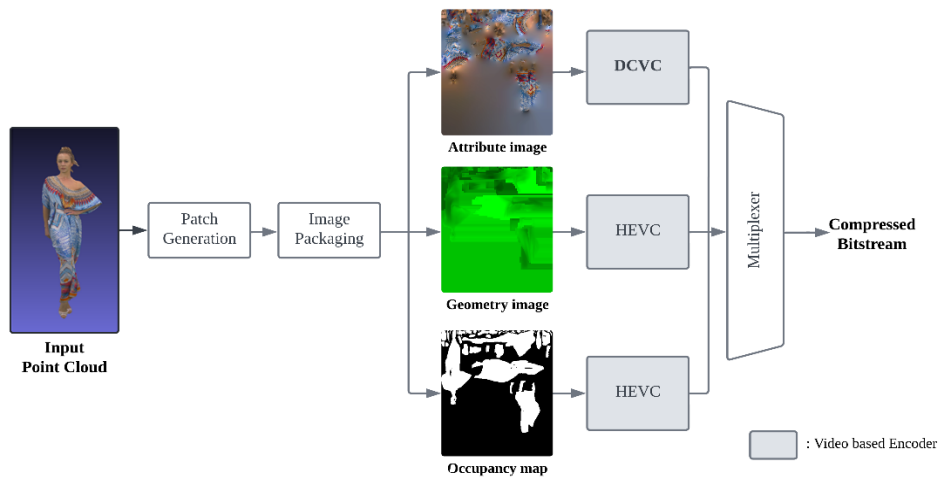

Fig. 1. V-PCC Encoder structure using Learned Video compression for 2D attribute image of point cloud

Attribute images, in particular, are characterized by irregular boundaries and nonuniform object distributions, making it difficult for codecs trained on natural images to capture these features adequately. By training the codec specifically on attribute images, these limitations can be overcome, enabling more efficient compression. To this end, we employ DCVC and train it to adapt to the unique characteristics of attribute images, thereby improving both compression efficiency and reconstruction quality.

Figure 1 illustrates the overall point cloud compression framework proposed in this study. The input point cloud is projected into 2D images consisting of attribute images, geometry images, and occupancy maps. In this study, DCVC is first applied to attribute images as a representative case to verify the feasibility of learned video compression within the V-PCC framework, while geometry and occupancy images are encoded using HEVC following the conventional pipeline. The outputs are then multiplexed to generate the final compressed bitstream.

2. Training of Deep Contextual Video Compression

DCVC is trained end-to-end using a rate‑distortion objective, where different quality levels are obtained by varying the parameter λ during training. Because each λ produces a distinct operating point, multiple DCVC models are typically trained independently to cover various bitrate-quality levels. In addition, DCVC training is computation-ally demanding, as each model is optimized from scratch using a large-scale video dataset. These characteristics highlight the need for more efficient training procedures. The three strategies proposed in this work are designed to reduce training cost and improve flexibility while maintaining the baseline DCVC framework.

In this study, we present three training strategies to improve the attribute image compression performance of DCVC. The first strategy is to train the learned video codec specifically for 2D attribute images. DCVC is trained on the color and texture information of 2D attribute images generated from 3D point clouds, thereby reinforcing its ability to learn the spatial and temporal characteristics of attribute images. Unlike natural images, attribute images exhibit irregular and unique patterns; thus, customized training data are constructed to optimize the DCVC model.

DCVC has a strong ability to capture spatial and temporal correlations across frames, and by applying this to attribute images, it can efficiently compress the fine details of projected point clouds, reduce information loss, and improve reconstruction quality. The training data for the DCVC model consist of 2D attribute images derived from the 3D point cloud datasets provided by MPEG-I[9]. As shown in Figure 2, these images are generated by projecting point cloud attributes onto a 2D plane using a V-PCC encoder, with each pixel containing 3D attribute information such as color and texture. This setup provides an effective environment for DCVC to learn point cloud data.

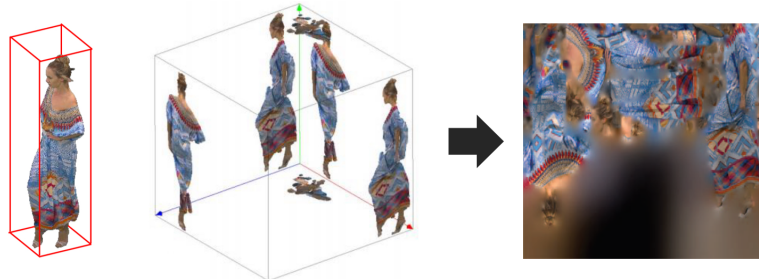The dataset used for training is described in detail in

Fig. 2. Example of 2D attribute images that project 3D attribute information

Table 1 and is designed to reflect diverse data characteristics. To improve the generalization performance of the model and prevent overfitting, data augmentation[10] was applied. Specifically, random cropping to a resolution of 256×256, vertical flipping, and horizontal flipping were employed. Using the augmented dataset, the DCVC model is trained to enhance learning efficiency and represent point cloud data more accurately.

Table 1. Point cloud train dataset

Dataset name	frames	fps	Number of Points
Levi	150	30	~1,040,000
Thomas	300	30	~1,050,000
Football	300	30	~1,110,000
Mitch	300	30	~1,280,000
Exercise	300	30	~608,000
Model	300	30	~810,000
Thaidancer	300	30	~3,100,000

DCVC requires separate training for multiple quality levels, which leads to significant computational cost and long training time. In addition, independently training each model can cause inconsistent convergence behaviors across different operating points. To address these limitations, we introduce two complementary ideas: the second strategy reduces training time through a cascaded training scheme, and the third strategy adjusts the set of quality levels to avoid unnecessary model training while retaining performance at relevant bitrates.

The second strategy aims to shorten training time by replacing the auto-regressive training scheme of the original DCVC with the "N-Stage-Cascaded" approach introduced in Temporal Context Mining for Learned Video Compression (DCVC-TCM)[11]. While the conventional auto-regressive model achieves strong compression performance through context modeling, it suffers from lengthy training time because the prediction of the current context from the previous one is performed sequentially. To address this limitation, the training process is divided into multiple stages, and the number of frames processed in

each stage is gradually increased. Each stage consists of a training mode, a learning rate[12], and the size of the decoded picture buffer (DPB), which are applied independently. The training modes are organized into four stages: (1) minimizing the distance between predicted and ground-truth motion vectors to improve motion estimation accuracy[13]; (2) optimizing the compression efficiency of motion vectors; (3) training with consideration of both similarity between reconstructed and original frames and the number of bits consumed; and (4) integrating all previous steps to jointly optimize reconstruction quality and compression efficiency. These training modes are sequentially applied throughout the overall training process.

The third strategy involves controlling the quality level of DCVC for efficient compression. In video compression, the quality level determines the trade-off between bitrate and reconstruction quality. To balance computational efficiency and model specialization at different quality levels, the training process is divided into two phases. In the initial phase, the model is trained at high quality levels to preserve fine details, while in the fine-tuning phase, the model is adapted to lower quality levels. The two phases differ in terms of stage configuration and training mode application. Specifically, in the initial phase, greater emphasis is placed on training modes 1 to 3 to adequately learn motion vector estimation and reconstruction quality. In contrast, the fine-tuning phase builds upon the high-quality model from the initial phase and applies only the fourth training mode to prevent unnecessary overfitting. Furthermore, the DPB size is increased in the fine-tuning phase to allow reference to a wider temporal range. Different learning rates are also applied: a uniform learning rate across all stages in the initial phase, and a lower rate in the fine-tuning phase for more precise model optimization.

The loss function for training DCVC follows the principle of Rate-Distortion Optimization (RDO)[14], defined as:

$$L = \lambda D + R \tag{1}$$

where λ is a hyperparameter that balances the distortion (D) between the original and reconstructed attribute images and the bitrate cost(R) predicted by the entropy model. DCVC is trained using five quality levels, with λ values set to {128, 256, 512, 1024, 2048} during training. The proposed training strategies are expected to optimize the performance of the learned video codec by adapting to the characteristics of 2D attribute images generated in V-PCC, thereby improving both compression efficiency and reconstruction quality.

Ⅲ. Experiments

For performance evaluation, we used sequences from the Common Test Conditions (CTC)[15] provided by MPEG-I, specifically Queen, Redandblack, Loot, and Soldier from Category A, and Longdress from Category B. To quantitatively assess the effectiveness of fine-tuning on attribute images in V-PCC, two models were built: a natural image‑based DCVC (Base Model) and an attribute image‑based DCVC (Ours Model). The Base Model was fine-tuned on the Vimeo-90k dataset[16] consisting of natural images, while the Ours Model was fine-tuned on a dataset of 2D attribute images derived from point clouds. Both models share the same network architecture and training settings, with the only difference being the training dataset.

To validate the efficiency of the proposed training strategy, we compared the coding performance of the Base Model, the Ours Model, and the traditional codec HEVC with two presets (x265 veryslow and x265 veryfast)[17], using reconstruction quality in the attribute image domain‑that is, the 2D packed representations of point clouds‑as the evaluation criterion. The rate‑distortion (RD) curves showing the relationship between Peak Signal-to-Noise Ratio (PSNR)[18] and bits per pixel (BPP) for the three models are presented in Figure 3.
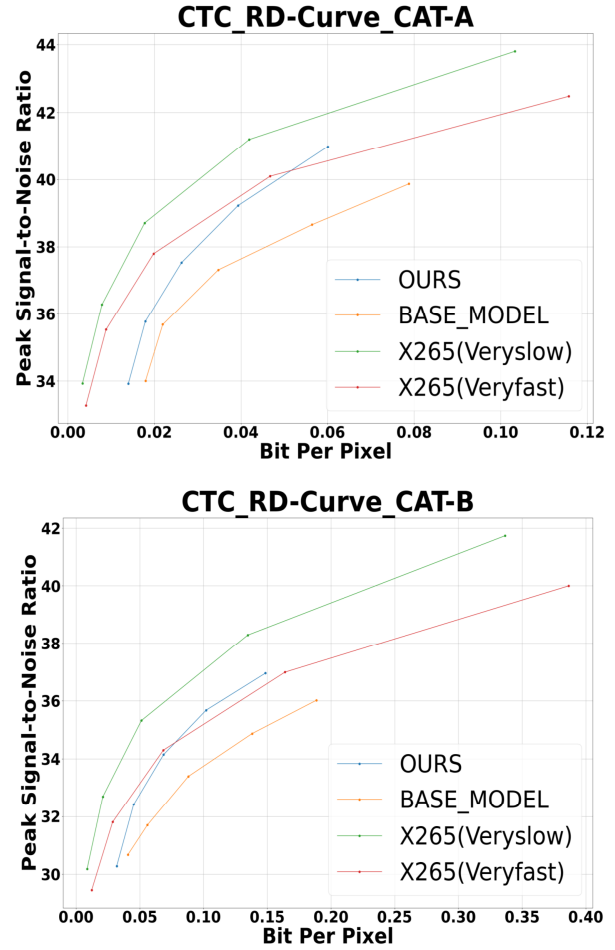

Fig. 3. RD-Curve Comparison for Cat-A and Cat-B: Base Model, Ours Model, and HEVC (H.265) Presets

To enable a fair comparison between the learned models and the signal processing-based codec, the rate points are selected according to the native contral parameters of each compression framework. For the learned models (Base Model and Ours), rate points are generated using the five predefined DCVC quality levels corresponding to λ values of {128, 256, 512, 1024, 2048}. In contrast, the HEVC codec is evaluated using two x265 presets (veryslow and veryfast), where QP values of {22, 27, 32, 37, 42} are selected to span a bitrate range comparable to that of the learned models. Although the two systems rely on different

Table 2. PSNR and BPP of the RD-curves for the Base Model and Ours Model

	Cat-A				Cat-B			
	PSNR		BPP		PSNR		BPP	
	Base Model	Ours	Base Model	Ours	Base Model	Ours	Base Model	Ours
Rate 1	33.999	33.916	0.018	0.014	30.671	30.278	0.041	0.032
Rate 2	35.675	35.767	0.022	0.018	31.695	32.414	0.056	0.045
Rate 3	37.313	37.529	0.035	0.026	33.392	34.143	0.088	0.069
Rate 4	38.656	39.244	0.056	0.039	34.870	35.677	0.138	0.102
Rate 5	39.868	40.981	0.079	0.060	36.017	36.952	0.189	0.149

rate-control mechanisms, aligning their output bitrates in this manner allows for a fair and meaningful comparison of rate‐distortion performance.

Under the same bitrate conditions, the Ours Model generally achieved higher PSNR across most bitrate ranges compared to the Base Model, demonstrating superior reconstruction quality. In addition, the results confirmed that performance remained stable across different quality levels. The quantitative performance metrics of the two models are

presented in Table 2, and the Bjøntegaard Delta-rate (BD-rate)[19] analysis results are summarized in Table 3. The Ours Model achieved an average BD-rate gain of 28.00% over the Base Model, demonstrating its capability for more efficient compression.

Figure 4 presents a visual comparison of the reconstructed attribute images, showing that the Ours Model reconstructs color and texture details more accurately than the Base Model. This observation is consistent with the quantitative analysis results and further supports the superior reconstruction quality achieved by the proposed method.

Meanwhile, the proposed model consistently improved PSNR across all BPP ranges compared with the Base Model, but its performance was comparable to the veryfast preset of HEVC (H.265) and somewhat lower than the ve-

Table 3. BD-Rate of the RD Curves for the Base Model and Ours

	Cat-A				Cat-B
	Loot	Redandblack	Soldier	Queen	Longdress
BD-rate [%]	-20.65%	-56.54%	-19.49%	-7.73%	-35.61%
	-26.20%				
	-28.00%				

Fig. 4. Reconstruction Results for Each Trained Model: Original Video, Base Model's Reconstruction, and Ours Reconstruction (from left to right)

ryslow preset. These results indicate that the proposed model still has certain limitations in comparison with the traditional signal processing‐based codec HEVC. Nevertheless, research on learned video codecs has been rapidly advancing, with continuous improvements in performance. Therefore, if a learned video codec with higher performance than DCVC were adopted as the base model, further performance gains could be expected.

Table 4 presents the End-to-End BD-AttrRate evaluation results comparing the attribute compression performance of the Ours Model with the Base Model. This metric quantitatively evaluates attribute compression performance by reflecting the correlation between bitrate and reconstruction quality in 3D point clouds. The experiments were conducted under the 32-frame Random Access (RA) configuration of the V-PCC standard[20].

Table 4. End-to-End BD-AttrRate Results for 3D Point Cloud Sequences

Category	Sequence	End-to-End BD-AttrRate [%]		
		Luma	Chroma Cb	Chroma Cr
Cat-A	Loot	-0.2%	79.3%	3.0%
	Redandblack	-40.4%	-62.3%	-36.5%
	Soldier	-9.4%	-5.7%	32.1%
	Queen	42.9%	87.7%	197.7%
Cat-B	Longdress	-20.6%	-29.9%	-23.8%
Cat-A average		-1.8%	24.7%	49.1%
Cat-B average		-20.6%	-29.9%	-23.8%
Overall average		-5.5%	13.8%	34.5%

The results show that, while the Ours Model exhibited average performance degradation of 13.8% and 34.5% in Chroma Cb and Cr, respectively, it achieved a 5.5% improvement in Luma. Considering the relative importance of Luma and Chroma, the Ours Model can be regarded as showing an overall trend of meaningful performance improvement compared with the Base Model.

On the other hand, the results in Table 4 indicate a smaller gain than those observed in the 2D reconstruction-based evaluation. This difference can be attributed to the evaluation criteria: whereas 2D evaluation measures quality across the entire image area, 3D evaluation considers only the quality within the occupancy region of 2D attribute images. For instance, in the Queen sequence, even though the non-occupancy regions were reconstructed with high visual quality, those areas do not correspond to the point cloud and are thus excluded from the 3D quality assessment, leading to lower overall performance metrics.

To overcome this limitation, improvements in network architecture and optimization strategies that explicitly incorporate occupancy information during training are required. Incorporating occupancy-aware training would allow the model to better capture the characteristics of V-PCC data and, ultimately, lead to enhanced attribute compression performance for 3D point clouds.

Ⅳ. Conclusion

In this study, we demonstrate the feasibility of deploying learned video codecs for point cloud compression and proposed an approach to enhance the compression performance of both 2D attribute images and 3D point cloud data by improving the training strategy of the DCVC model. Experimental results demonstrated that the proposed strategy achieved superior efficiency in 2D video coding, with overall PSNR improvements over the Base Model as shown in the RD-curve. In addition, BD-rate analysis confirmed its effectiveness by reporting an average gain of 28.00%.

In comparison with HEVC (H.265), the proposed model achieved performance comparable to the veryfast preset but somewhat lower than the veryslow preset. These results suggest that the proposed training strategy holds strong potential for further improvements if applied to more advanced learned video codecs. They also provide positive implications regarding the applicability and scalability of learning-based video compression techniques.

In terms of 3D point cloud attribute quality, the improvements were relatively modest compared with those observed in 2D packed image quality. This is because only the occupancy regions of 2D attribute images determine the quality of 3D point clouds. To overcome this limitation, future work should focus on network architectures and optimization strategies that explicitly incorporate occupancy information during training. In addition, while this study employs the original DCVC as the baseline, applying the proposed training strategy to more advanced variants－such as DCVC-TCM[11], Neural Video Compression with Diverse Contexts (DCVC-DC)[21]－is an important direction for future research. Exploring these models may further improve both 2D attribute image compression efficiency and 3D point cloud reconstruction quality.

References

[1] D. Graziosi, O. Nakagami, S. Kuma, A. Zaghetto, T. Suzuki, and A. Tabatabai, "An overview of ongoing point cloud compression standardization activities: video-based (V-PCC) and geometry-based (G-PCC)," *APSIPA Transactions on Signal and Information Processing*, vol. 9, e13, pp.1-17, 2020.
doi: https://doi.org/10.1017/ATSIP.2020.12

[2] C. Cao, M. Preda, and T. Zaharia, "3D point cloud compression: A survey," in Proc. 24th Int. Conf. on 3D Web Technology (Web3D), pp. 1－9, 2019.
doi: https://doi.org/10.1145/3329714.3338130

[3] ISO/IEC 23090-5:2023, *Information Technology - Coded Representation of Immersive Media, Part 5: Visual Volumetric Video-Based Coding (V3C) and Video-Based Point Cloud Compression (V-PCC)*, ISO, 2023.

[4] ISO/IEC JTC 1/SC 29/WG 7, *V-PCC Codec Description*, ISO/IEC JTC 1/SC 29/WG 7 N00100, 2021.

[5] J. Kim, J. Im, S. Rhyu, and K. Kim, "3D motion estimation and compensation method for video-based point cloud compression," *IEEE Access*, vol. 8, pp. 83538－83547, May 2020.
doi: https://doi.org/10.1109/ACCESS.2020.2991478

[6] Z. Liang, J. Liu, M. Dasari, and F. Wang, "Fumos: Neural compression and progressive refinement for continuous point cloud video streaming," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 6, pp. 2849－2859, June 2024.
doi: https://doi.org/10.1109/TVCG.2024.3372096

[7] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 18114－18125, 2021.

[8] B. Bross, W.-J. Han, G. J. Sullivan, J.-R. Ohm, and T. Wiegand, "High efficiency video coding (HEVC) text specification draft 8," ITU-T/ISO/IEC Joint Collaborative Team on Video Coding (JCT-VC), JCTVC-J1003, July 2012.

[9] ISO/IEC JTC 1/SC 29/WG 7, *Common Test Conditions for G-PCC*, ISO/IEC JTC 1/SC 29/WG 7 document N800, Jan. 2024.

[10] C. Shorten and T. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, pp. 1－48, July 2019.
doi: https://doi.org/10.1186/s40537-019-0197-0

[11] X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, "Temporal context mining for learned video compression," *IEEE Transactions on Multimedia*, early access, Feb. 2022.
doi: https://doi.org/10.1109/TMM.2022.3149876

[12] A. Defazio, A. Cutkosky, H. Mehta, and K. Mishchenko, "Optimal linear decay learning rate schedules and further refinements," arXiv preprint arXiv:2302.06675, 2023. Available: https://arxiv.org/abs/2302.06675

[13] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 10998－11007, June 2019.
doi: https://doi.org/10.1109/CVPR.2019.01126.

[14] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74－90, Nov. 1998.

[15] ISO/IEC JTC 1/SC 29/WG 11, *Common Test Conditions for PCC*, ISO/IEC JTC 1/SC 29/WG 11 N19329, Jan. 2020.

[16] T. Xue, X. Huang, J. Wang, and Z. Wang, "Vimeo-90K: Toward high-quality video frame interpolation," arXiv preprint arXiv: 1711.09078v3, Nov. 2017. Available: https://arxiv.org/abs/1711.09078

[17] x265 Documentation, *x265 HEVC Encoder Software Manual*, [Online]. Available: https://x265.readthedocs.io/en/master/. (accessed Jan. 20, 2025).

[18] G. Bjontegaard, *Calculation of average PSNR differences between RD-curves*, VCEG-M33, ITU-T Video Coding Experts Group, Apr. 2001.

[19] J. Ryu, S. Lee, and D. Shim, "No-reference PSNR estimation of H.264/AVC video," in Proc. Summer Conf. Korean Inst. of Broadcast and Media Engineers (KIBME), pp. 277－278, July 2010.

[20] ISO/IEC JTC 1/SC 29/WG 11, *Common Test Conditions for V3C and V-PCC*, ISO/IEC JTC 1/SC 29/WG 11 document N19518, July 2020.

[21] J. Li, B. Li, and Y. Lu, "Neural video compression with diverse contexts," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 22616-22626, 2023.
doi: https://doi.org/10.1109/CVPR52729.2023.02166

Introduction Authors

Seungmin Noh

- 2025 : Bachelor of Science, Department of Information and Communication Engineering, Hanbat National University
- 2025 ~ Present : Pursuing Master of Science, Department of Intelligence Media Engineering, Hanbat National University
- ORCID : https://orcid.org/0009-0002-2216-0817
- Research interests : Video Coding, Computer Vision, Point Cloud Compression, Deep Learning

Yongheon Kim

- 2025 : Bachelor of Science, Department of Information and Communication Engineering, Hanbat National University
- 2025 ~ Present : Pursuing Master of Science, Department of Intelligence Media Engineering, Hanbat National University
- ORCID : https://orcid.org/0009-0005-4152-8252
- Research interests : Computer Vision, Video Coding, Deep Learning

Kaeun Lee

- 2024 : Bachelor of Science, Department of Information and Communication Engineering, Hanbat National University
- 2024 ~ Present : Pursuing Master of Science, Department of Intelligence Media Engineering, Hanbat National University
- ORCID : https://orcid.org/0009-0008-2142-7034
- Research interests : Video Coding, Computer Vision, Deep Learning, 3D Point Cloud, Point Cloud Compression

Haechul Choi

- 1997 : Bachelor of Science in Electronics Engineering, Kyungpook National University
- 1999 : Master of Science in Electrical Engineering, Korea Advanced Institute of Science and Technology
- 2004 : Ph.D in Electrical Engineering, Korea Advanced Institute of Science and Technology
- 2010 ~ Present : Professor, Department of Intelligence Media Engineering, Hanbat National University
- ORCID : https://orcid.org/0000-0002-7594-0828
- Research interests : Image Processing, Video Coding, Computer Vision