VEMIC: View-aware Entropy model for Multi-view Image Compression

Susmija Jabbireddy
susmija.reddy@gmail.com
Davit Soselia
dsoselia@umd.edu
Max Ehrlich
maxehr@umd.edu
Christopher Metzler
metzler@umd.edu
Amitabh Varshney
varshney@umd.edu

The University of Maryland,
College Park, MD
United States

Abstract

With the ever-increasing amount of 3D data being captured and processed, multi-view image compression is essential to various applications, including virtual reality and 3D modeling. Despite the considerable success of learning-based compression models on single images, limited progress has been made in multi-view image compression. In this paper, we propose an efficient approach to multi-view image compression by leveraging the redundant information across different viewpoints without explicitly using warping operations or camera parameters. Our method builds upon the recent advancements in Multi-Reference Entropy Models (MEM), which were initially proposed to capture correlations within an image. We extend the MEM models to employ cross-view correlations in addition to within-image correlations. Specifically, we generate latent representations for each view independently and integrate a cross-view context module within the entropy model. The estimation of entropy parameters for each view follows an autoregressive technique, leveraging correlations with the previous views. We show that adding this view context module further enhances the compression performance when jointly trained with the autoencoder. Experimental results demonstrate superior performance compared to both traditional and learning-based multi-view compression methods.

1 Introduction

The increasing demand for multimedia content generation and consumption creates a pressing need for the development of efficient data compression techniques. Multi-view imaging systems, which capture a scene from various angles, are indispensable for virtual and augmented reality (VR and AR), 3D reconstruction, and surveillance systems. These systems present considerable challenges in the storage and transmission of very large amounts of

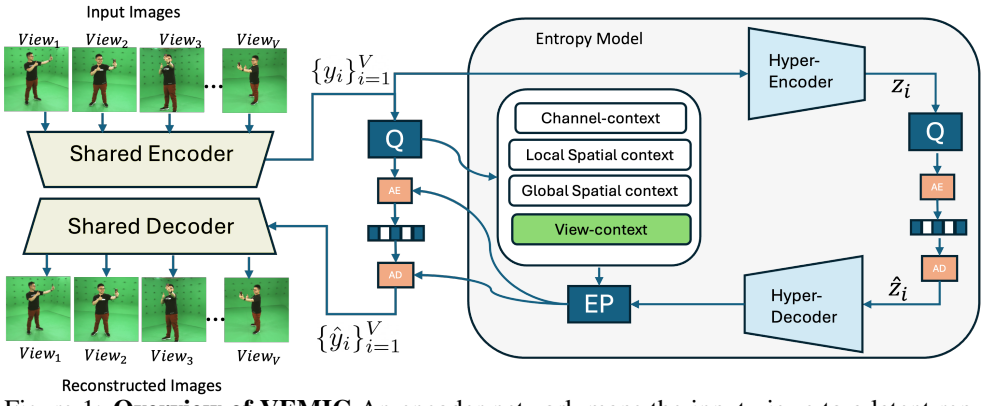

Figure 1: **Overview of VEMIC** An encoder network maps the input views to a latent representation y_k, which is then quantized and entropy-coded. The context module consists of channel context and spatial context blocks similar to MLIC++ [14] and a novel view-context block (shown in green). The view-context module learns the cross-view correlations, which are then used in entropy coding to achieve improved bitrate reductions.

data. Thus, improving compression techniques is crucial for efficiently handling, storing, and transmitting multi-view image data.

In this paper, we present a learning-based multi-view compression model for a general multi-camera setup. While significant progress has been made on learning-based single-view compression [2, 3, 8, 21], multi-view compression has often been overlooked. One of the main challenges in training multi-view compression models is the unavailability of large-scale multi-camera datasets. In this work, we mitigate this problem by taking a fine-tuning approach in which single-view compression models trained on large-scale single-view datasets are fine-tuned on multi-view data.

While simple fine-tuning can yield good performance gains, these models still operate on each image independently without leveraging any cross-view information. Since multi-view images exhibit significant correlations, we postulate that harnessing these correlations can further enhance the compression performance. We aim to capture these correlations within the entropy models, which work alongside the autoencoder in conventional learning-based compression models. The autoencoder encodes the images into latent representations and reconstructs the images back from these latents. The entropy model enhances compression efficiency through accurate prediction of the probability distribution of the quantized latent representations. Advanced entropy models often employ various context modules that capture the dependencies within the data to estimate the entropy parameters. These context modules generally focus on identifying and utilizing dependencies found within individual images. In this work, we focus on improving the context modules for multi-view compression task by utilizing the information from cross-view features.

We choose the single-view compression model from MLIC++ (Multi-Reference Entropy Model for Learned Image Compression) [14] as our pretraining network. MLIC++ utilizes three types of context modules – channel context, local spatial context, and global spatial context. These modules independently extract the channel and spatial context features from the latent representations of each image. In addition to these modules, for multi-view compression, we incorporate a view context module that can gather information from different

views. We illustrate the overall pipeline in Figure 1. Our context module uses an attention-based framework to autoregressively generate view-based information from the previously seen views. Specifically, the view context for the current view is generated by attending to the feature maps from the previous views. The newly added cross-view context module parameters are learned from scratch, while all previous layers are initialized from MLIC++.

We empirically show that our approach, the View-aware Entropy model for Multi-view image compression (VEMIC), can yield significant performance gains compared to both traditional and learning-based multi-view compression techniques. Fine-tuning single-view compression models can be greatly beneficial, especially in scenarios with limited data availability, which is common in multi-view image datasets. Furthermore, incorporating multi-view context modules can further boost the compression performance. In summary, our contributions are:

- We present a fine-tuning-based strategy for adapting the single-view compression models for multi-view compression problem.
- To model the multi-view correlations in the context module, we present an autoregressive view-channel context block based on the attention mechanism.
- We empirically show that our approach can yield significant gains in terms of rate-distortion performance compared to competing approaches on various multi-view datasets.

2 Related Work

2.1 Single Image Compression

Traditional codecs Classical image compression standards, such as JPEG [28] and JPEG 2000 [22], use handcrafted algorithms to reduce the size while preserving the original visual quality. These methods typically have a three-stage compression pipeline. First, the input image is transformed to a compact representation, often using discrete cosine (DCT) [1] or wavelet transforms [32]. The transformed input is then quantized. Finally, the discrete quantized output is encoded into a bitstream using entropy coders such as Huffman or Arithmetic coding in a lossless manner. Despite their effectiveness, these techniques have many limitations. First, the input image is divided into small blocks, which are independently transformed, thereby introducing block artifacts. Second, each stage is locally optimized, limiting global performance. Finally, they do not adapt to the specific characteristics of the input data, often resulting in suboptimal performance.

Learned image codecs Learning-based image compression algorithms have achieved impressive performance [11, 14, 27], often surpassing conventional codecs. Early research in learned image compression used autoencoder architectures [13] with three basic modules – encoder, quantizer, and decoder. The encoder network transforms the image into a latent representation. This representation is converted into a discrete representation using a quantizer function, and finally, a decoder reconstructs the image from the discrete latent representation. All three modules are jointly optimized to minimize the total rate-distortion cost. The latent representations are further compressed using entropy coding, where frequently occurring patterns are represented with fewer bits and the rare patterns with more bits. Toderici *et al.* [27] presented an end-to-end learned image compression method using recurrent neural

networks that outperforms JPEG. Since then, several improvements have been made to the encoder/decoder transforms to reduce the correlations in the latent representation.

2.2 Multi-view Image Compression

Conventional multi-view image compression standards [25, 26] are extensions of the video codecs and are developed to improve compression performance by exploiting the redundancies between different views. These methods utilize motion estimation and disparity estimation techniques to harness the view correlations. However, they only support the YUV420 format and are still developing. These codecs do not match the performance of the single image codecs that support YUV444 planar or RGB format. On the other hand, most of the existing learning-based multi-view approaches [9, 17, 19, 30, 31] mainly focus on stereo images, and extending the stereo-based methods to multi-view images is not trivial.

2.3 Entropy modeling

An entropy model estimates the probability distribution of the quantized latent representations. To improve the coding efficiency, entropy models learn a prior on the latent representation, which is used with the entropy coding algorithms to yield a compressed bit stream. Early works used element-wise independent entropy models to estimate the probability distribution of the latent representations and independently encoded each element with an arithmetic coder. Later works used more sophisticated models to explicitly estimate the entropy using hyperpriors and other parametric models [3, 8, 21]

Balle *et al*. [3] introduced a hyperprior architecture to improve the entropy model. They use a conditional Gaussian model parameterized by scale to estimate the entropy parameters from the hyper-latent representation. The compressed hyper-latent is added as side information to the bitstream, allowing the decoder to use the conditional entropy model. The model is trained end-to-end to jointly optimize the autoencoder, the quantized hyper-latent representation, and the conditional entropy model. The Gaussian scale entropy model reduces the spatial dependencies within the latent representation, thereby improving the compression performance. Minnen *et al*. [21] extended the scale model to a Gaussian mixture model, where both mean and scale parameters are estimated from the entropy model. They demonstrated the use of conditional means to further reduce the spatial dependencies in the latent representation. In addition, augmenting the hyperprior model with an autoregressive model that predicts the latent from their causal context further improved the rate-distortion performance. Though the auto-regressive model is effective, it requires the model to sequentially decode each symbol, which can result in longer decoding times. Minnen *et al*. [20] proposed the channel-conditioned auto-regressive models where the latent representation is divided into slices along the challenge dimension, and the entropy parameters for each slice are conditioned on the latents of the previous slices. Given the effectiveness of the context-adaptive models, several improvements to the entropy models have been proposed to exploit the context along the spatial dimension [10, 18, 34] and channel dimensions [11, 16]. Most of the current works exploit the context modules captured within an image. In our work, we exploit the effect of context captured across multiple viewpoint images of the same scene.

3 Method

3.1 Overall architecture

The goal of multi-view image compression is to compress a set of images $X = \{x_1, x_2, \cdots x_V\}$ of a scene captured from V different viewpoints while preserving their visual quality. In contrast to single-view image compression networks [3, 11, 14] that operate on each image independently, we use the multi-view correlations to improve the rate-distortion tradeoff.

In this work, we utilize an autoencoder-based framework for the task of multi-view compression. Specifically, an encoder network transforms each image set $\{x_i\}_{i=1}^{V}$ into a set of latent representations $\{y_i\}_{i=1}^{V}$. These latent representations are quantized to $\{\hat{y}_i\}_{i=1}^{V}$ in a differentiable manner, which is then entropy-coded into a bitstream that is transmitted to the decoder. The decoder reconstructs the images $\{\hat{x}_i\}_{i=1}^{V}$ from this code. The model is trained in an end-to-end manner to jointly optimize for quality and bitrate reduction.

We build upon the work of MLIC++ [14], which is the state-of-the-art model for single image compression. The encoder and the decoder networks are composed of several convolution-based residual blocks with progressive downsampling/upsampling layers. In addition, a context-based entropy model, called Multi-reference Entropy Model (MEM), is learned to estimate the distributions of the latent representations, which are then used in entropy coding. The MEM module captures the local, global, and channel-wise correlations in the latents. These correlations are then combined with the hyper-prior side information. Using the estimated parameters of the latents, entropy coding techniques such as arithmetic coding [23] can perform a lossless compression of the quantized latent representations. An outline of this framework is illustrated in Figure 1.

In particular, the MEM module employs three types of contexts – channel context, local spatial context, and global spatial context. The channel-wise context module divides the latent representations $\hat{y}_i$ into several slices $\{\hat{y}_i^1, \hat{y}_i^2, \hat{y}_i^3, \ldots\}$ along the channel dimension. The channel context for each slice s is computed using all previous slices $\hat{y}_i^{<s}$ with a shallow network. Next, for each slice, spatial contexts are computed by partitioning the latents $\hat{y}_i$ into two halves – anchor ($\hat{y}^s_{i,anchor}$) and non-anchor ($\hat{y}^s_{i,non-anchor}$), following a checkerboard pattern. The anchor part is context-free, while the context for the non-anchor part $\hat{y}^s_{i,non-anchor}$ is computed from the anchor part $\hat{y}^s_{i,anchor}$ using an attention block. Finally, the global spatial context module uses an efficient linear attention layer [24] to capture the global correlations between the non-anchor part $\hat{y}^s_{i,non-anchor}$, and the anchor part of the current slice $\hat{y}^s_{i,anchor}$ and the previous slice $\hat{y}_i^{s-1}$.

3.2 View Context module

As MLIC++ [14] is designed for single-view image compression, all the context modules exploit the correlations within the latent representations of a single image. In a multi-view image framework, additional correlations are present across the latent representations from different views. We hypothesize that augmenting the view-based correlations in the entropy model can further reduce the bit rate. To achieve this, we propose a cross-view context module to capture the dependencies across different viewpoint representations. We use an autoregressive framework, where the entropy parameters for view v are estimated using the latent representations of previous views using an attention mechanism. Specifically, the view context for a slice s of view v ($\hat{y}_v^s$) uses an attention layer where the queries are the previous slices of the same view $\hat{y}_v^{<s}$, the keys are the previous slices of the previous views $\hat{y}_{<v}^{<s}$, and

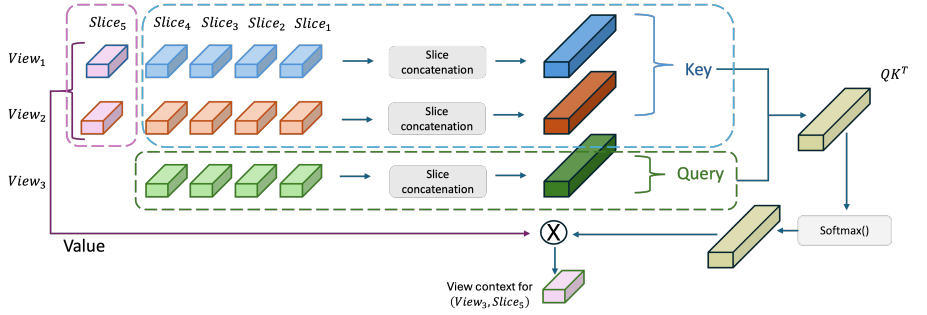

Figure 2: **View context module.** The channel context for a new slice of a current view is estimated using an attention mechanism that calculates the correlations between the current view with all previously decoded views. The queries are the features of all previous slices of the current view (shown in green), the keys are the features of all previous slices from all previous views (shown in blue and orange), and the values are the features of the current slice from previous views (shown in pink).

the values are taken from the previous views of the same slice $\hat{y}^s_{<v}$. This is illustrated in Figure 2.

Let $\hat{y}^s_v \in \mathbb{R}^{h\times w\times c}$, where h and w represent the spatial dimension and c is the number of channels. The resultant key matrix K will be of dimension $\mathbb{R}^{(v-1)hw\times(s-1)c}$, the query matrix Q will be $\mathbb{R}^{hw\times(s-1)c}$ and the value matrix V will be of dimension $\mathbb{R}^{(v-1)hw\times c}$. We also add view-based positional embeddings to the features before computing the attention matrix. The attention map represents the similarity between each spatial location of the current view and the spatial locations of all previous views. Using this attention map, the view context is obtained as a weighted combination of all previous view features.

3.3 Fine tuning

Training multi-view compression methods are often challenging due to the unavailability of large-scale multi-camera datasets. To mitigate this, we use pre-trained single-view compression models as initialization and then add our newly introduced view context modules to the entropy model. We use the publicly available checkpoints from MLIC++ [14] as our pre-trained model. The newly added view-context modules are randomly initialized, and the entire model is fine-tuned end-to-end on a multi-view image dataset.

3.4 Training objective

Our model optimizes the rate-distortion objective on multi-view images as:

$$\mathcal{L} = \sum_{k=1}^{K}\sum_{v=1}^{V} \lambda D(x_{k,v},\hat{x}_{k,v}) + R(\hat{y}_{k,v}) + R(\hat{z}_{k,v})$$

The distortion D term represents the image quality metric, such as mean-squared error or structural similarity term, between the original image x and the reconstructed image $\hat{x}$. The rate term R represents the estimated code length or the number of bits used to encode the latent representation $\hat{y}$ and the corresponding hyper-latent representation $\hat{z}$. K represents the number of multi-view images, and V represents the number of views in each image. λ controls the trade-off between the compression rate and the image quality.

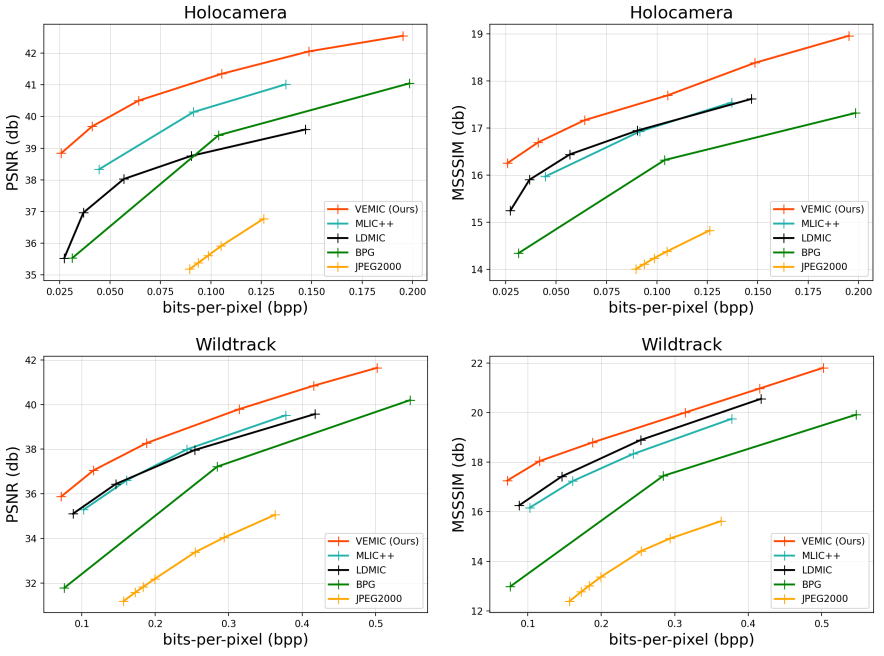

Figure 3: **Rate-distortion curve on Holocamera and Wildtrack datasets** The plot on the left shows the rate-distortion curve with the PSNR metric, while the one on the right shows the curve with the MS-SSIM metric. Our approach, VEMIC, outperforms all prior approaches, achieving state-of-the-art results.

4 Experiments

Datasets We evaluate our approach on two publicly available multi-view image datasets – Holocamera [10] and Wildtrack [7]. The Holocamera dataset contains 4032×3040 resolution volumetric captures of 30 static scenes acquired from 300 different viewpoints. We use the 240 viewpoints along the four sidewalls of the volumetric capture studio and ignore the 60 viewpoints from the ceiling. We encode six images at a time, corresponding to the six camera locations placed vertically, one below the other. This provides structured data and establishes a spatial relationship across the viewpoints. The Wildtrack dataset consists of pedestrian surveillance videos of 1920×1080 resolution captured from seven randomly placed cameras. Both these multi-view datasets have overlapping field-of-view across viewpoints, which leads to view redundancies.

Implementation details Our implementation is based on the publicly available CompressAI [4] library. We initialize our model using the pre-trained checkpoints corresponding to different λ values provided by MLIC++ [14]. Following the settings from MLIC++ and CompressAI, we use $\lambda \in \{67, 130, 250, 483\} \times 10^{-4}$ and MSE distortion loss. We train each model using Adam optimizer [15] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a batch size of 4 and start with a learning rate of 10^{-4} for both models and on both datasets. On the Holocamera dataset, the learning rate reduces to 10^{-5} at 100 epochs, 10^{-6} at 180 epochs, and 10^{-7} at 220 epochs, and the network is trained for 250 epochs. On the Wildtrack dataset,

Method	Holocamera		Wildtrack	
	PSNR	SSIM	PSNR	SSIM
JPEG2000 [22]	201.86 %	228.15 %	122.39 %	127.22 %
LDMIC [33]	-14.59 %	-48.76 %	-30.91 %	-49.37 %
MLIC++ [14]	-21.26 %	-30.56 %	-31.04 %	-41.25 %
VEMIC (Ours)	**-62.79 %**	**-68.21 %**	**-50.51 %**	**-62.06** %

Table 1: **BD-rate comparison.** The table shows the BD-rate values with BPG [5] as the baseline. VEMIC outperforms prior approaches on both Holocamera and Wildtrack datasets.

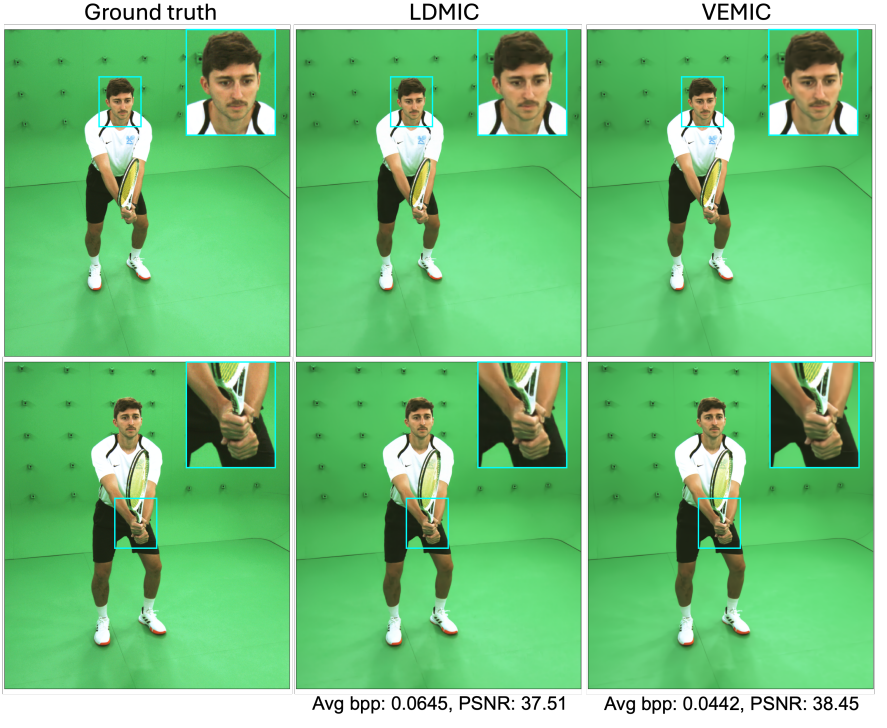

Figure 4: **Qualitative results on Holocamera dataset.** The two rows represent the scene captured from two different viewpoints. The left column shows the uncompressed ground-truth image, the middle column shows the reconstructions from LDMIC, and the right shows the reconstructions from our VEMIC. Our approach achieves very high reconstruction quality while requiring much lower bits per pixel than LDMIC [33].

the learning rate reduces to 10^{-5} at 60 epochs, 10^{-6} at 84 epochs, and 10^{-7} at 110 epochs, and the network is trained for 140 epochs. The Holocamera images are downsampled to 2716×2048 to accommodate the GPU memory limitations. During training, we use image patches of size 512×512. The image patches correspond to identical regions across all the viewpoints. We perform testing on full-size images.

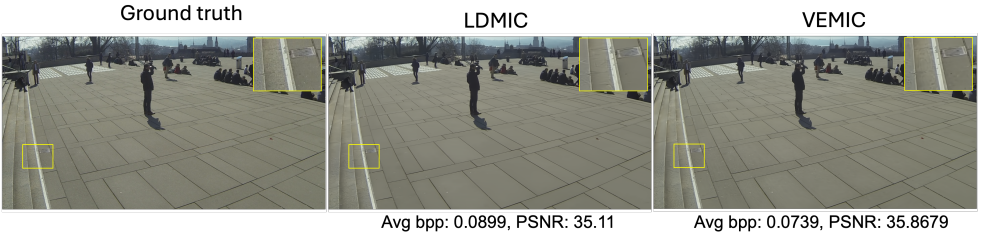

Figure 5: **Qualitative results on Wildtrack dataset.** The panel in the middle shows the reconstructions of LDMIC, while the panel on the right shows the reconstructions from VEMIC. Our approach achieves high reconstruction quality with lower bpp than LDMIC.

Metrics We use the peak signal-to-noise ratio (PSNR) and multi-scale structural similarity index (MS-SSIM) [29] to measure the reconstructed image quality compared to the ground-truth image. We plot the rate-distortion curves to compare different methods. In addition, we also calculate the Bjøntegaard Delta-Rate (BD-Rate) [6] to indicate the average bitrate savings at the same level of distortion.

Benchmarks We compare our approach, VEMIC, with some of the popular traditional codecs and the more recent learning-based codecs. For the traditional codecs, we use JPEG-2000 [27] and BPG [5], which are widely used for single-view image compression. For learned compression models, we perform comparisons with MLIC++ [14] and LDMIC [33], the state-of-the-art models for single-view and multi-view compression, respectively. For MLIC++, we perform comparisons with zero-shot evaluation, where the pre-trained model is evaluated directly on the multi-view datasets.

Results The rate-distortion curves of our approach in comparison with all baselines are reported in Figure 3. We observe that our method, VEMIC, achieves better compression performance on both datasets, outperforming prior approaches by a significant margin. We notice PSNR and MS-SSIM improvements consistently across all λ values. The test set for both datasets consists of scenes that were not seen during training, demonstrating that our method effectively generalizes to new scenes within each dataset.

To quantify the average bit rate savings, we also show the BD-rate [6] values for VEMIC and other methods in Table 1 with BPG [5] method as the baseline. We observe that VEMIC achieves state-of-the-art results on Holocamera and Wildtrack datasets on both PSNR and MS-SSIM metrics. Our VEMIC achieves a BD-rate improvement of 72.91% over the prior state-of-the-art multi-view image compression method, LDMIC, on the Holocamera dataset.

We also visualize the qualitative results in Figure 4. We observe that VEMIC can reconstruct the input images with very high fidelity. While the visual quality of both LDMIC and VEMIC are comparable, our approach achieves a much lower number of bits per pixel.

5 Conclusion

In this work, we address the problem of learning-based multi-view image compression using a view-aware entropy model. Our idea is to inject the view information into the channel context module, which can then be used in the entropy coding to improve the bitrate of the

latent codes. We design the view-aware context module using an autoregressive attention-based model that learns the correlations between the current views and all previous views. We then show how these view context modules can be added to the single-view pre-trained networks and finetuned end-to-end on multi-view datasets. Through experimental results on various multi-view image datasets, we show that our approach can outperform prior baselines and establish a new state-of-the-art for the multi-view compression problem.

Limitations and Future Work While our approach focuses on leveraging the cross-view context in the entropy model alone, there may be advantages in incorporating view information from various perspectives into both the encoder and decoder. This integration could facilitate enhanced information exchange. However, we defer this aspect to be explored in future research endeavors. Our current method, though effective, is limited to a relatively small number of viewpoints. As the number of viewpoints increases, the computational demands of self-attention become a constraint. Future work could address this by adopting memory-efficient attention mechanisms. Additionally, expanding this approach to multi-view video compression offers an exciting opportunity to exploit temporal correlations and redundancies, further advancing the state of the art.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful feedback to improve the paper. Additionally, we would also like to thank Yogesh Balaji and Sharmila Duppala for their helpful discussions. This work has been supported by NSF Grants 18-23321, 22-35050, and the State of Maryland's MPower initiative. C.M. was supported in part by AFOSR YIP award no. FA9550-22-1-0208. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the research sponsors.

References

[1] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete Cosine Transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. doi: 10.1109/T-C.1974.223784.

[2] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2016.

[3] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkcQFMZRb.

[4] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. CompressAI: a PyTorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

[5] Fabrice Bellard. BPG image format. 2014. URL https://bellard.org/bpg/.

[6] Gisle Bjøntegaard. Calculation of average PSNR differences between RD-curves. 2001. URL https://api.semanticscholar.org/CorpusID:61598325.

[7] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. WILDTRACK: A multi-camera HD dataset for dense unscripted pedestrian detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018. doi: 10.1109/CVPR.2018.00528.

[8] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized Gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] Xin Deng, Wenzhe Yang, Ren Yang, Mai Xu, Enpeng Liu, Qianhan Feng, and Radu Timofte. Deep homography for efficient stereo image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1492–1501, 2021.

[10] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14771–14780, 2021.

[11] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5718–5727, 2022.

[12] Jonathan Heagerty, Sida Li, Eric Lee, Shuvra Bhattacharyya, Sujal Bista, Barbara Brawn, Brandon Feng, Susmija Jabbireddy, Joseph JaJa, Hernisa Kacorri, David Li, Derek Yarnell, Matthias Zwicker, and Amitabh Varshney. HoloCamera: Advanced volumetric capture for cinematic-quality VR applications. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[13] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[14] Wei Jiang and Ronggang Wang. MLIC++: Linear complexity multi-reference entropy modeling for learned image compression. In *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023. URL https://openreview.net/forum?id=hxIpcSoz2t.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[16] A. Burakhan Koyuncu, Han Gao, Atanas Boev, Georgii Gaikov, Elena Alshina, and Eckehard Steinbach. Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer*

Vision – ECCV 2022, pages 447–463, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19800-7.

[17] Jianjun Lei, Xiangrui Liu, Bo Peng, Dengchao Jin, Wanqing Li, and Jingxiao Gu. Deep stereo image compression via bi-directional coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19669–19678, 2022.

[18] Mu Li, Kai Zhang, Jinxing Li, Wangmeng Zuo, Radu Timofte, and David Zhang. Learning context-based nonlocal entropy modeling for image compression. *IEEE Transactions on Neural Networks and Learning Systems*, 34(3):1132–1145, 2023. doi: 10.1109/TNNLS.2021.3104974.

[19] Jerry Liu, Shenlong Wang, and Raquel Urtasun. DSIC: Deep Stereo Image Compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3145, 2019.

[20] David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343, 2020. doi: 10.1109/ICIP40778.2020.9190935.

[21] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018.

[22] Majid Rabbani and Rajan Joshi. An overview of the JPEG 2000 still image compression standard. *Signal Processing: Image Communication*, 17(1): 3–48, 2002. ISSN 0923-5965. doi: https://doi.org/10.1016/S0923-5965(01)00024-8. URL https://www.sciencedirect.com/science/article/pii/S0923596501000248. JPEG 2000.

[23] J. Rissanen and G. Langdon. Universal modeling and coding. *IEEE Transactions on Information Theory*, 27(1):12–23, 1981. doi: 10.1109/TIT.1981.1056282.

[24] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.

[25] Gary J Sullivan, Jill M Boyce, Ying Chen, Jens-Rainer Ohm, C Andrew Segall, and Anthony Vetro. Standardized extensions of high efficiency video coding (HEVC). *IEEE Journal of selected topics in Signal Processing*, 7(6):1001–1016, 2013.

[26] Gerhard Tech, Ying Chen, Karsten Müller, Jens-Rainer Ohm, Anthony Vetro, and Ye-Kui Wang. Overview of the multiview and 3D extensions of High Efficiency Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(1):35–49, 2016. doi: 10.1109/TCSVT.2015.2477935.

[27] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David C. Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5435–5443, 2016. URL https://api.semanticscholar.org/CorpusID:24041818.

[28] Gregory K. Wallace. The JPEG still picture compression standard. *Commun. ACM*, 34 (4):30–44, apr 1991. ISSN 0001-0782. doi: 10.1145/103085.103089. URL https://doi.org/10.1145/103085.103089.

[29] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[30] Matthias Wödlinger, Jan Kotera, Jan Xu, and Robert Sablatnig. SASIC: Stereo image compression with latent shifts and stereo attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 661–670, 2022.

[31] Matthias Wödlinger, Jan Kotera, Manuel Keglevic, Jan Xu, and Robert Sablatnig. EC-SIC: Epipolar cross attention for stereo image compression. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3436–3445, 2024.

[32] Dengsheng Zhang. *Wavelet Transform*, pages 35–44. Springer International Publishing, Cham, 2019. ISBN 978-3-030-17989-2. doi: 10.1007/978-3-030-17989-2_3. URL https://doi.org/10.1007/978-3-030-17989-2_3.

[33] Xinjie Zhang, Jiawei Shao, and Jun Zhang. LDMIC: Learning-based Distributed Multi-view Image Coding. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ILQVw4cA5F9.

[34] Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang. The devil is in the details: Window-based attention for image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17492–17501, June 2022.