BEV-MMC: Bird's-Eye-View-Based Multimodal Compression for Enhanced Visual Recognition

Zhiwei Dong
Department of Computer Science and Engineering
Santa Clara University
Santa Clara, CA 95053 USA
zdong@scu.edu

Ying Liu†
Department of Computer Science and Engineering
Santa Clara University
Santa Clara, CA 95053 USA
yliu15@scu.edu

***Abstract*—Today, visual data compression is essential not only for human perception but also for machine analysis. With the growing volumes of visual data, challenges in data storage and transmission demand advanced compression techniques for downstream tasks. Most current works focus on either single-modal compression or multimodal compression without unifying the modalities effectively. In this study, we introduce a bird's-eye-view (BEV)-based multimodal compression framework that jointly compresses camera images and LiDAR point clouds to enhance the accuracy of downstream tasks. A core innovation of our approach is the inception-triggered soft element-wise mask, which leverages multi-scale feature extraction to effectively capture richer information and reduces spatial-channel redundancy. Compared to the existing state-of-the-art methods, our proposed framework has achieved BD-rate gains of 88.12% for 3D object detection and 98.71% for BEV map segmentation compared to the referenced BEV-based compression framework and 95.43% for 3D object detection compared to the referenced multimodal compression framework.**

***Index Terms*—3D object detection, bird's-eye-view map segmentation, coding for machines, image compression, multimodal learning**

I. Introduction

In recent years, the rapid development of computer vision has significantly advanced real-world visual recognition systems, such as those in autonomous vehicles. Camera-based machine vision systems [1]–[3] provide rich semantic information, which is advantageous for 2D vision tasks. However, they lack the geometric information that is critical in 3D settings. In contrast, LiDAR-based machine vision systems [4]–[6] excel in 3D vision tasks, as the LiDAR point cloud captures geometric details that spatially describe environments and precisely locate objects. To take advantage of both camera and LiDAR systems, recently multimodal machine vision systems [7]–[9] have been a promising solution for visual recognition. These systems integrate data from camera and LiDAR sensors into a unified feature space, which then serve for specific downstream tasks such as classification, detection, and segmentation. For example, BEVFusion [7] unifies camera and LiDAR data into a shared bird's-eye-view (BEV) representation space and demonstrates that BEV-based multimodal frameworks outperform BEV-based single-modal frameworks in both 3D object detection and BEV map segmentation tasks.

However, in real-world applications, sensor data is captured locally in vehicles and then transmitted to remote servers for analysis. Therefore, visual recognition systems face challenges due to the large volume of data transmission. To optimize bandwidth during data transmission while ensuring high accuracy in downstream vision tasks, efficient compression for machine recognition has become an urgent necessity. Current compression methods for machine vision involved in 3D settings are mainly implemented in the 3D space [10]–[12]. Some studies have focused on compression of LiDAR-transformed BEV features for machine vision tasks [13]. In recent years, multimodal compression has been a new direction for machine visual recognition. Existing methods either leave multimodal data in separated domains [14], [15] or unify them in 3D space [16].

In this paper, we propose a novel BEV-based multimodal compression framework BEV-MMC for 3D object detection and BEV map segmentation. Our motivation is, BEVFusion [7] has demonstrated BEV-based multimodal perception systems that integrate multimodal data in a unified BEV domain are more powerful than single-modal systems and less complex than multimodal systems that unify features in other domains. Our proposed framework extracts BEV features from camera images and LiDAR point clouds separately and fuses them into a shared BEV domain then compresses the fused features for the downstream 3D object detection and BEV map segmentation tasks. Additionally, we introduce an inception-triggered soft element-wise mask module that provides an adaptive solution to the variation of sparsity in fused BEV features. This mask is element-wise multiplied to fused BEV representation, significantly reducing bit rates during compression. To our knowledge, BEV-MMC is the first BEV-based multimodal compression framework for 3D object detection and BEV map segmentation tasks. The key contributions of this paper are as follows:

- For the first time in the literature, we propose a BEV-based multimodal (camera and LiDAR) compression framework for 3D object detection and BEV map segmentation.
- We propose a fusion module that comprises a CNN-based

† Corresponding author.
This work is supported in part by the National Science Foundation under Grant ECCS-2138635 and the NVIDIA Academic Hardware Grant.

fuser and an inception-triggered soft element-wise mask, which adaptively selects informative multimodal BEV features for compression.
- We demonstrate that joint optimization of the feature fusion module and the compression model can achieve high coding efficiency for the downstream machine vision tasks.

II. Related Work

A. 3D Visual Recognition

Based on the input data, 3D visual recognition can be categorized into camera-based, LiDAR-based, and multimodal approaches, with the latter garnering significant interests in recent years. Camera-based methods primarily focus on compensating for the lack of depth information, which is crucial for 3D object detection [17], [18]. Depth prediction methods assist research in transforming camera images from a perspective view into a bird's-eye-view (BEV) by using a view transformer [19]–[22]. LiDAR-based methods typically either directly extract global point cloud features [4], [23], or voxelize the original point clouds to a lower resolution then extract BEV features [24]–[26]. Since LiDAR point clouds provide geometric information while camera images offer semantic information, methods that extract features from both modalities and fuse them into a unified feature space such as the BEV space have shown superior performance compared to approaches that rely on a single-modal [7], [9], [27].

B. Compression for Machine Visual Recognition

Compression for machine vision tasks can be categorized as single-modal or multimodal compression. One research trend of single-modal methods compresses LiDAR 3D point cloud in the original 3D space. For instance, SPCGC [12] separates geometric and semantic information, compressing them with distinct modules then concatenating the two information for downstream analysis. In contrast, some studies [10], [28] compress point cloud's geometry and attributes together and perform downstream tasks based on the reconstructed point cloud. PCHM-Net [11] and HM-PCGC [29] focus on balancing compression for human and machine vision tasks. PCHM-Net introduces two parallel compression branches for human and machine vision, utilizing encoder-decoder modules that share identical architectures and weights across branches. HM-PCGC employs a framework where geometric information is extracted via pretrained geometry encoders, while semantic features are captured using a transformer-based module. Additionally, HM-PCGC integrates a multi-objective loss constraint, combining geometry and semantic constraints to balance performance for human and machine vision tasks effectively. A newer research direction in this stream involves implementing compression process in a transformed feature space, rather than in the original 3D space. One such method, PC4M [13] first compresses extracted LiDAR-only BEV features and then performs 3D object detection and BEV map segmentation based on the reconstructed BEV features.

As to multimodal methods, one stream of research keeps multimodal data in their original domains. FUTR3D-based [30] methods first separately extract features from camera and LiDAR data then fuse them with a series of anchors to query useful features from each source. Finally queried informative features are compressed for downstream tasks [14]. A multi-object tracking study [15] also separately extracts features from LiDAR and camera data without integrating them into a unified domain. It generates two attention-based soft masks, which are applied to each feature source through element-wise multiplication. After that, the two masked features are fused for compression and downstream analysis. Another study [31] extracts depth information from camera image to assist the compression of LiDAR point cloud geometric information, rather than to perform machine tasks. On the other hand, some works have explored to unify multimodal data in 3D space [16]. It first transforms camera image into point cloud by using PENet [32], then combines the transformed camera data with the LiDAR point cloud. The combined point cloud is subsequently voxelized for feature extraction, and the extracted features are losslessly compressed for downstream tasks.

III. Proposed Method

Fig. 1 shows our proposed BEV-based multimodal compression framework BEV-MMC. It has four main components: BEV feature extractors, a fusion module, a compression model, and a task head. We adopt the pretrained feature extractors and task head from BEVFusion [7], which has pioneered the BEV domain integration of multimodal inputs for visual recognition. The compression model we employ is MSRB [33] due to its balance of complexity and effectiveness.

A. The Overall Framework

The feature extractors $\mathbf{FE}_c$ and $\mathbf{FE}_l$ first extract BEV features $\mathbf{b}_c$ and $\mathbf{b}_l$ from the raw camera images $\mathbf{x}_c$ captured by 6 cameras and LiDAR point cloud $\mathbf{x}_l$, respectively. Then, $\mathbf{b}_c$ and $\mathbf{b}_l$ are stacked channel-wise to form $\mathbf{b}_s$, which is processed by a CNN-based fuser $\mathbf{FS}$ to produce the fused BEV feature $\mathbf{b}_f$. An inception module $\mathbf{I}$ and a mask module $\mathbf{M}$ further extract and select features from $\mathbf{b}_f$ to generate a soft mask $\mathbf{m}$ which is then element-wise multiplied with $\mathbf{b}_f$, resulting in the masked feature $\mathbf{x}_m$. Then a compression model $\mathbf{C}$ compresses $\mathbf{x}_m$ into a bit stream and reconstructs the masked feature $\widehat{\mathbf{x}}_m$. Finally, $\widehat{\mathbf{x}}_m$ is fed into the task head $\mathbf{H}$ to output the visual recognition results $\mathbf{p}$.

B. Baseline Model

We define a baseline model BEV-MMC-c that consists of feature extractors $\mathbf{FE}_c$ and $\mathbf{FE}_l$, the fuser $\mathbf{FS}$, the compression model $\mathbf{C}$ and the task head $\mathbf{H}$. We train the weights of $\mathbf{C}$, while the weights of $\mathbf{FE}_c$, $\mathbf{FE}_l$, $\mathbf{FS}$ and $\mathbf{H}$ are from the pretrained BEVFusion model [7] and are frozen. We adopt the

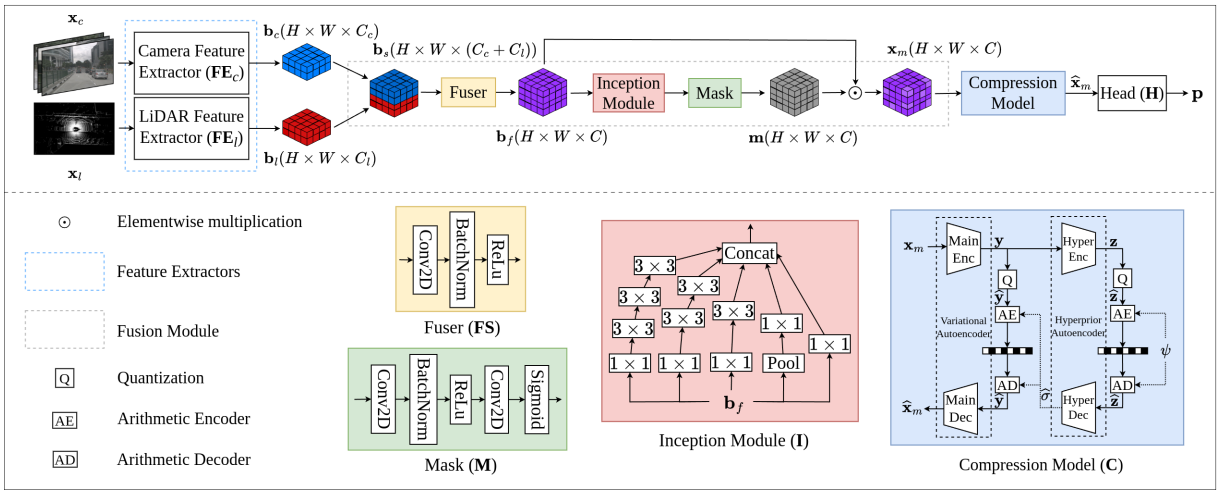

Fig. 1. The proposed BEV-MMC framework.

joint optimization [34] to the following loss function during training:

$$\begin{aligned} \mathrm{L} &= \mathrm{R} + \alpha \mathrm{D} + \mathbf{w}\mathbf{L}_{task} \\ &= \mathbb{E}[-\log_2(p_{\hat{\mathbf{y}}|\hat{\mathbf{z}}}(\hat{\mathbf{y}}|\hat{\mathbf{z}}))] + \mathbb{E}[-\log_2(p_{\hat{\mathbf{z}}|\psi}(\hat{\mathbf{z}}|\psi))] \\ &+ \alpha \cdot \mathrm{D}(\mathbf{x}_m, \hat{\mathbf{x}}_m) + \mathbf{w}\mathbf{L}_{task}, \end{aligned} \tag{1}$$

where R is the bit rate loss of the quantized latent representations $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$, generated by the variational autoencoder and the hyperprior autoencoder, respectively. D is the distortion loss measured by the mean squared error (MSE) between input BEV feature $\mathbf{x}_m$ and reconstructed BEV feature $\hat{\mathbf{x}}_m$, $\mathbf{w}$ are the weights of the loss terms and $\mathbf{L}_{task}$ are the task-specific losses. Specifically, for 3D object detection, $\mathbf{w} = \begin{bmatrix} \mathrm{w}_c & \mathrm{w}_h & \mathrm{w}_r \end{bmatrix}$, $\mathbf{L}_{task} = \begin{bmatrix} \mathrm{L}_{cls} & \mathrm{L}_{heatmap} & \mathrm{L}_{reg} \end{bmatrix}^T$, where L_{cls} is the cross-entropy loss of the predicted object class, $\mathrm{L}_{heatmap}$ is the class-specific center heatmap loss proposed by [35] and L_{reg} is the regression loss of the predicted 3D object bounding box coordinates. For BEV map segmentation, $\mathbf{w} = \mathrm{w}_m$, $\mathbf{L}_{task} = \mathrm{L}_{map}$, which is the focal loss between the ground-truth and predicted segmentation maps.

C. Joint Optimization of the Fuser and Compression Model

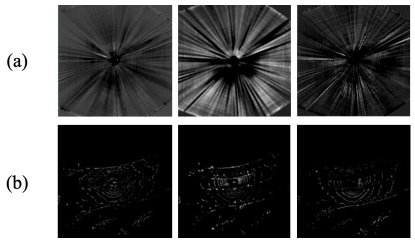

Fig. 2. BEV features: (a) camera BEV features, and (b) LiDAR BEV features.

Upon visualizing the camera and LiDAR BEV features produced by the baseline model in Fig. 2 (a) and (b), we observe significant redundancies in the camera BEV features. This occurs because the pretrained **FS** in the BEVFusion model is not optimized with an entropy loss, limiting its ability to reduce bit rate consumption. To address this problem, we define our second model BEV-MMC-f, which jointly trains **FS** and **C** with the same network architecture and loss function as the baseline model.

D. Inception-Triggered Soft Element-Wise Mask

To more effectively reduce redundancy, we propose an inception-triggered soft element-wise mask to adaptively select informative features from $\mathbf{b}_f$ for compression. We start with a simple CNN-based mask module **M**, which is placed between **FS** and **C**. It generates a soft mask $\mathbf{m} \in [0,1]^{H \times W \times C}$. The element-wise multiplication is then performed between $\mathbf{b}_f$ and $\mathbf{m}$ to produce the filtered feature $\mathbf{x}_m$ for compression. This model is referred to as BEV-MMC-m. Further, to enhance the effectiveness of mask-based information selection, we propose an inception-like module **I**. It has five branches with different kernel sizes: 7×7, 5×5, 3×3, 1×1, and global pooling, as shown in Fig. 1. **I** and **M** process the fused BEV feature $\mathbf{b}_f$ to produce the soft mask $\mathbf{m}$, followed by the element-wise multiplication and compression. This enhanced model, named BEV-MMC-i, is able to capture richer information by utilizing multi-scale feature extraction.

IV. Experimental Studies

A. Dataset and Settings

The NuScenes dataset [36] is widely utilized in 3D perception research due to its comprehensive and robust multi-sensor data. Each sample in the NuScenes dataset includes six images captured by cameras at different positions, along with a point cloud generated by a 32-beam LiDAR scan. We train our models for 3D object detection and BEV map segmentation tasks on the NuScenes training set and evaluate them on the NuScenes validation and test sets. We employ the AdamW optimizer [37]. The training process begins with a learning rate of 1e-4 for the first eight epochs, followed by a decay

of 0.1 for two additional epochs. A batch size of 8 is used throughout all training processes.

To achieve various bit rates, we apply the following settings to the loss function hyper parameters to train our models BEV-MMC-f, BEV-MMC-m and BEV-MMC-i. For 3D object detection, we set α to $\{16, 32, 64, 128, 256\}$, set w_c and w_h to $\{1, 1, 2, 4, 8\}$, and set w_r to $\{0.25, 0.25, 0.5, 1, 2\}$. For BEV map segmentation, we set α to $\{16, 128, 256, 512, 1024\}$ and set w_m to $\{0.5, 4, 8, 16, 32\}$.

The baseline model BEV-MMC-c is trained with different loss function hyper parameters because its compression capability solely relies on $\mathbf{C}$. For 3D object detection, we set α to $\{1, 2, 4, 8, 16\}$, set w_c and w_h to $\{0.5, 1, 2, 4, 8\}$, and set w_r to $\{0.125, 0.25, 0.5, 1, 2\}$. For BEV map segmentation, we set α to $\{8, 32, 64, 128, 256\}$ and set w_m to $\{0.25, 1, 2, 4, 16\}$.

B. Evaluation Metrics

We use bits per pixel (BPP) to measure coding efficiency. Since this study involves multimodal inputs, our definition of pixels refers to BEV-domain pixels instead of RGB-domain pixels in traditional 2D visual compression. The evaluation metric used for 3D object detection in this study is the mean average precision (mAP) [36], [38], calculated across ten object classes: car, truck, construction vehicle, bus, trailer, barrier, motorcycle, bicycle, pedestrian, and traffic cone. For BEV map segmentation, we use the mean intersection-over-union (mIoU) [39], measured across six zones: drivable area, pedestrian crossing, walkway, stop line, car parking area, and lane divider. To measure model complexity, we adopt the number of trainable parameters measured in millions (M) and the amount of calculation measured in giga floating point operations (GFLOPs).

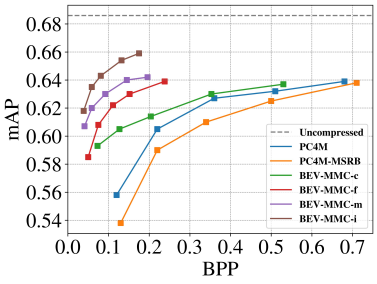

Fig. 3. The rate-accuracy performance of 3D object detection on the nuScenes validation set.

C. Quantitative Results

Fig. 3 presents the rate-mAP curves for 3D object detection on the nuScenes validation set, comparing our proposed BEV-MMC frameworks (BEV-MMC-c, BEV-MMC-f, BEV-MMC-m, and BEV-MMC-i) with PC4M-MSRB and PC4M [13]. Both PC4M-MSRB and PC4M represent state-of-the-art BEV-based LiDAR point cloud compression frameworks for machine vision tasks. These frameworks utilize the pretrained BEVFusion-L [7] as the visual recognition backbone, incorporating a BEV feature codec between the LiDAR feature extraction module and the task-specific head. The key distinction

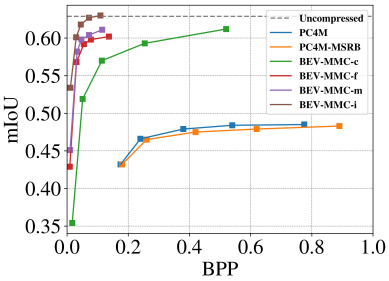

Fig. 4. The rate-accuracy performance of BEV map segmentation on the nuScenes validation set.

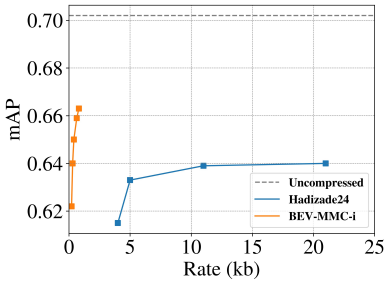

Fig. 5. The rate-accuracy performance of 3D object detection on the nuScenes test set.

lies in their codec designs: PC4M-MSRB employs the MSRB model as the BEV feature codec, while PC4M enhances it with a fusion block and an attention block. Similarly, our BEV-MMC framework leverages the pretrained BEVFusion model but stands out by extracting features from both camera and LiDAR data and integrating them through a fusion module. From the results, we observe that to achieve an mAP of 0.64, the baseline model BEV-MMC-c requires more than 0.50 BPP, BEV-MMC-f takes about 0.24 BPP, BEV-MMC-m lowers the BPP to less than 0.15, and the best-performing model, BEV-MMC-i, which incorporates the inception-based mask for multi-scale feature selection, achieves the same performance with less than 0.08 BPP. In contrast, PC4M demands about 0.70 BPP.

For BEV map segmentation, as illustrated in Fig. 4, at 0.10 BPP, BEV-MMC-c achieves approximately 0.57 mIoU, BEV-MMC-f reaches 0.60 mIoU, BEV-MMC-m further improves the mIoU to 0.61, and BEV-MMC-i achieves the highest performance with 0.63 mIoU. In comparison, PC4M stays at about 0.48 mIoU even increasing BPP to almost 0.80.

Fig. 5 shows the rate-mAP curves for 3D object detection on the nuScenes test set, comparing BEV-MMC-i with Haizade24 [14], the state-of-the-art multimodal compression framework for 3D object detection. To compare our results with Haizade24, we adopt the same method for measuring coding efficiency. Specifically, we use the size of the compressed bitstream (in kilobytes) that encodes six camera images and one LiDAR point cloud. To achieve 0.64 mAP, Haizade24 takes about 15 kilobytes, whereas BEV-MMC-i only needs 0.3 kilobytes, which demonstrates our framework's superior

TABLE I
BD-RATE GAIN WITH RESPECT TO DETECTION MAP AND SEGMENTATION MIOU (PC4M IS THE ANCHOR) AND MODEL COMPLEXITIES

Model	Train **C**	Train **FS**	Train **H**	Mask	Inception	BD-rate		Parameters (/M)		GFLOPs	
						Detection	Segmentation	Detection	Segmentation	Detection	Segmentation
BEV-MMC-c	✓					-31.22 %	-87.93 %	62.22	64.54	537.65	699.18
BEV-MMC-f[1]	✓	✓				-66.89 %	-95.87 %	62.22	64.54	537.65	699.18
BEV-MMC-m	✓	✓	✓	✓		-79.36 %	-96.25 %	63.40	65.71	575.91	737.44
BEV-MMC-i	✓	✓	✓	✓	✓	**-88.12 %**	**-98.71 %**	65.04	67.34	628.92	790.46

[1] Our experiments show that jointly training **H** doesn't improve the prediction accuracy of BEV-MMC-f.

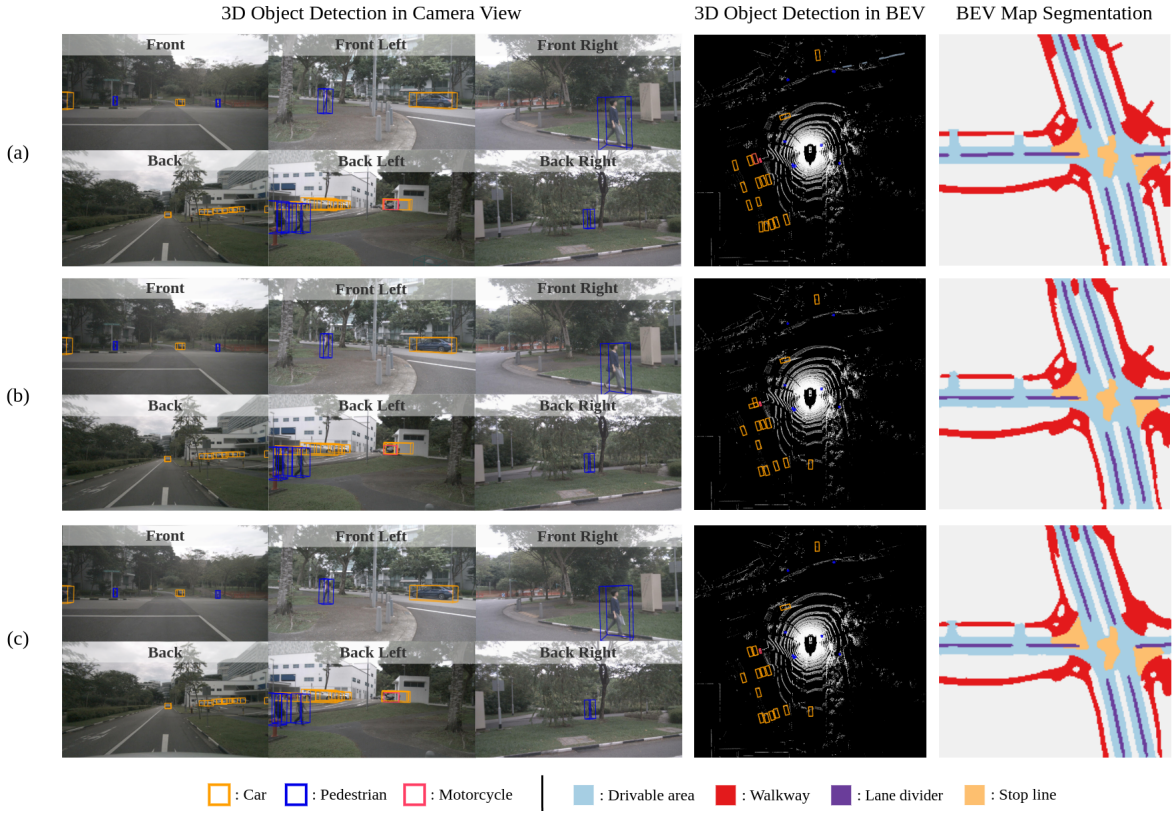

Fig. 6. The visualization of 3D object detection and BEV map segmentation: (a) ground truth, (b) BEVFusion results with uncompressed inputs, and (c) BEV-MMC-i results.

efficiency. The BD-rate gain of BEV-MMC-i compared to Haizade24 is 95.43%.

Table I summarizes the BD-rate gains with PC4M as the anchor, along with model complexities. Compared to PC4M, our best model, BEV-MMC-i, achieves BD-rate gains of 88.12% for 3D object detection and 98.71% for BEV segmentation. Despite incorporating the inception-triggered element-wise soft mask, BEV-MMC-i increases trainable parameters by only 3% and 5%, and GFLOPs by 13% and 17%, for 3D object detection and BEV map segmentation respectively.

D. Visual Results

Fig. 6 shows the visual results, including ground truth, predictions from the BEVFusion model with uncompressed data, and those from BEV-MMC-i at 0.17 BPP. The visualization provides six camera images of different views and one point cloud of BEV for the 3D object detection task, as well as one BEV segmentation map for the BEV map segmentation task. From these results, we observe that our proposed BEV-MMC-i can locate the same objects as accurately as the BEVFusion model and produce segmentation maps that closely align with the ground truth, exhibiting remarkable precision.

V. CONCLUSIONS

In this work, we present a BEV-based multimodal compression framework for enhanced visual recognition. Our proposed inception-triggered soft element-wise mask enables the fusion module to effectively integrate BEV features extracted from cameras and LiDAR and adaptively select salient information for the subsequent compression process. Additionally, we jointly train the fusion module, compression model, and task-specific prediction head to balance coding efficiency and prediction accuracy. Experimental results demonstrate that our BEV-MMC framework outperforms referenced state-of-

the-art methods. For future research, we plan to incorporate transformer structures into the fusion module to better capture and utilize global information. We also aim to design a more efficient feature codec specifically tailored for BEV features.

REFERENCES

[1] Y. Wang, V. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," 2021.

[2] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," 2022.

[3] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13750–13759, 2022.

[4] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: deep hierarchical feature learning on point sets in a metric space," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 5105–5114, Curran Associates Inc., 2017.

[5] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–779, 2019.

[6] X. Zhu, H. Zhou, T. Wang, F. Hong, W. Li, Y. Ma, H. Li, R. Yang, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar-based perception," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6807–6822, 2022.

[7] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2774–2781, 2023.

[8] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: a simple and robust lidar-camera fusion framework," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, (Red Hook, NY, USA), Curran Associates Inc., 2024.

[9] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1526–1535, 2021.

[10] B. Liu, S. Li, X. Sheng, L. Li, and D. Liu, "Joint optimized point cloud compression for 3d object detection," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1185–1189, 2023.

[11] L. Liu, Z. Hu, and J. Zhang, "Pchm-net: A new point cloud compression framework for both human vision and machine vision," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1997–2002, 2023.

[12] L. Xie, W. Gao, H. Zheng, and G. Li, "Spcgc: Scalable point cloud geometry compression for machine vision," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 17272–17278, May 2024.

[13] X. Gao, Z. Zhu, and L. Yu, "Bird's-eye-view-based lidar point cloud coding for machines," in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5, 2023.

[14] H. Hadizadeh and I. V. Bajić, "Learned multimodal compression for autonomous driving," in *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2024.

[15] X. Li, O. A. Hanna, C. Fragouli, S. Diggavi, G. Verma, and J. Bhattacharyya, "Feature compression for multimodal multi-object tracking," in *MILCOM 2023 - 2023 IEEE Military Communications Conference (MILCOM)*, pp. 139–143, 2023.

[16] C. Tian, Z. Li, H. Yuan, R. Hamzaoui, L. Shen, and S. Kwong, "Feature compression for cloud-edge multimodal 3d object detection," 2024.

[17] T. Wang, X. Zhu, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," *CoRR*, vol. abs/2107.14160, 2021.

[18] H. Chen, W. Tian, P. Wang, F. Wang, L. Xiong, and H. Li, "Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2024.

[19] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, p. 4867–4873, July 2020.

[20] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV*, (Berlin, Heidelberg), p. 194–210, Springer-Verlag, 2020.

[21] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11135–11144, 2020.

[22] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," in *British Machine Vision Conference*, 2018.

[23] B. Graham, M. Engelcke, and L. v. d. Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9224–9232, 2018.

[24] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2018.

[25] M. Nie, Y. Xue, C. Wang, C. Ye, H. Xu, X. Zhu, Q. Huang, M. B. Mi, X. Wang, and L. Zhang, "Partner: Level up the polar representation for lidar 3d object detection," 2023.

[26] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection," 2022.

[27] J. Fu, C. Gao, Z. Wang, L. Yang, X. Wang, B. Mu, and S. Liu, "Eliminating cross-modal conflicts in bev space for lidar-camera 3d object detection," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16381–16387, 2024.

[28] M. Ulhaq and I. V. Baji'c, "Learned point cloud compression for classification," *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2023.

[29] X. Ma, Y. Xu, X. Zhang, L. Tang, K. Zhang, and L. Zhang, "Hm-pcgc: A human-machine balanced point cloud geometry compression scheme," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2265–2269, 2023.

[30] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 172–181, 2023.

[31] Y. Lin, T. Xu, Z. Zhu, Y. Li, Z. Wang, and Y. Wang, "Your camera improves your point cloud compression," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[32] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, "Penet: Towards precise and efficient image guided depth completion," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13656–13662, 2021.

[33] H. Fu, F. Liang, J. Liang, B. Li, G. Zhang, and J. Han, "Asymmetric learned image compression with multi-scale residual block, importance scaling, and post-quantization filtering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4309–4321, 2023.

[34] C. Gao, Z. Li, L. Li, D. Liu, and F. Wu, "Rethinking the joint optimization in video coding for machines: A case study," in *2024 Data Compression Conference (DCC)*, pp. 556–556, 2024.

[35] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3d object detection and tracking," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11779–11788, 2021.

[36] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628, 2020.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[38] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[39] W. K. Fong, R. Mohan, J. V. Hurtado, L. Zhou, H. Caesar, O. Beijbom, and A. Valada, "Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3795–3802, 2022.