LiDAR-Guided Learned Image Compression for Machine Vision

Chia-Hao Kao, *Member, IEEE*, Alessandro Gnutti, *Member, IEEE*,
Wen-Hsiao Peng, *Fellow, IEEE*, and Riccardo Leonardi, *Fellow, IEEE*

***Abstract*—Learned image compression is increasingly being used not only for human viewing but also to support machine understanding. This has led to the development of Coding for Machines (CfM), which focuses on compressing visual data for downstream tasks. While some methods improve CfM by leveraging semantic cues or recognition outputs at the encoder side, they often assume access to high-level information that is not available in practice and require costly computation to extract. In this paper, we propose a learned image coding framework for machine vision that uses LiDAR depth as a low-cost and widely available auxiliary signal. Instead of re-training the entire codec, we introduce lightweight adapters at the encoder side, keeping the base model fixed. We also avoid full image reconstruction by transforming the latent representation through a bridge module that connects directly to pruned recognition networks. Experiments across various tasks, including standard recognition models and Multimodal Large Language Models (MLLMs), show that our method achieves strong rate-accuracy performance, while lowering computational cost at the decoder side.**

***Index Terms*—Learned Image Compression, Coding for Machines, LiDAR Depth map, RGB-D Recognition.**

I. Introduction

GIVEN the ever-growing volume of visual data being exchanged and stored, the need for efficient image compression has never been more critical. Learned image compression has made significant progress over the last decade, continuously pushing the frontier of coding efficiency. From early hyperprior-based methods [1] to recent advancements [2], [3] that have outperformed the intra-coding of VVC [4], the latest video coding standard, learned image compression has demonstrated its potential to surpass traditional handcrafted codecs.

At the same time, with the rapid advancement of computer vision networks and the development of large-scale models, compressed visual data are no longer consumed exclusively by human viewers but also by machines. The surge of applications in autonomous driving, robotic perception, remote sensing, and surveillance necessitates compression techniques that aim primarily at machine-driven tasks. However, mainstream image compression techniques have traditionally focused on optimizing reconstruction quality for human perception, using metrics such as PSNR and MS-SSIM, which may not be necessary even for advanced machine vision tasks. As a result, regular image compression systems are not ideal for machine-driven tasks [5], [6], especially at low bitrates.

This study was carried out within the LICAM project – funded by European Union – Next Generation EU within the PRIN 2022 program (D.D. 104 - 02/02/2022 Ministero dell'Università e della Ricerca).

Chia-Hao Kao and Alessandro Gnutti are with the Department of Information Engineering, CNIT – University of Brescia, 25123 Brescia, Italy.

Wen-Hsiao Peng is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan.

Riccardo Leonardi is with the Department of Information Engineering, CNIT – University of Brescia, 25123 Brescia, Italy, and also with the Department of Electronics Engineering, University of Rome Tor Vergata, 00133 Roma, Italy.

Corresponding author: Alessandro Gnutti, e-mail: alessandro.gnutti@unibs.it.

A. Coding for Machines

Coding for Machines (CfM) is a paradigm that prioritizes the compression of visual data to facilitate downstream recognition and analysis tasks rather than human perception. Earlier attempts at CfM primarily focused on task-specific approaches [7], [8], simply optimizing the designed codec for a single predetermined downstream task network. These methods, while showing improved performance, have high computational requirements due to the need of re-training for each downstream task and lack practicality. More recently, some works [5], [6] propose to transfer an existing off-the-shelf codec from targeting human perception to machine vision, with only the added lightweight modules trainable, leading to a more efficient and practical solution for CfM.

Within the field of CfM, several methods have attempted to leverage additional information to improve coding efficiency. A common approach [9]–[11] is region-of-interest (ROI) coding, where an ROI map indicating objects' locations guides the encoder to allocate more bits to foreground objects while reducing the bitrate for irrelevant background regions. Another approach [12], [13] utilizes pre-trained large-scale models to generate auxiliary side information (e.g., segmentation maps) from the RGB input images before compression. However, both approaches imply that some form of semantic recognition must be performed at the encoder side, which (i) limits the practical deployment of CfM as it was initially conceived to offload heavy compute from edge devices (hosting image encoders) to the cloud (hosting image decoders) and (ii) significantly increases computational overhead if using a neural network to extract semantic information before encoding. An alternative approach proposed by [14] employs generated segmentation maps during training to guide the codec toward preserving only the object boundaries. While this may benefit certain tasks, its reliance on edge-focused representations limits generalizability, as different downstream tasks often demand more visual information beyond object boundaries.

B. LiDAR Depth Information

To overcome these limitations, we explore the possibility of leveraging a lightweight, readily available, and low-cost source of auxiliary information at the encoder side for CfM, thereby removing the need for additional semantic processing. In this context, it is observed that the increasing integration

of depth sensors in consumer devices, such as smartphones and autonomous vehicles, has made depth maps a commonly available modality, often captured alongside RGB images when a user takes a picture. Although this depth information is originally used for applications such as depth-of-field rendering and augmented reality, they have also been shown to be highly informative for various machine vision tasks. The additional geometric cues provided by depth maps are able to enhance feature extraction, improve robustness and enable better understanding and recognition performance in tasks such as semantic segmentation [15]–[17] and object detection [17], [18].

Among the various types of depth sensors, LiDAR (Light Detection and Ranging) is a depth-sensing technology that uses laser pulses to measure distances, enabling the construction of high-resolution 3D maps of the environment [19]. Apple integrated LiDAR into its consumer devices starting with the iPad Pro in March 2020, followed by the iPhone 12 Pro series and subsequent models [20]. The LiDAR scanner emits laser beams and calculates the return time to estimate depth, generating accurate spatial representations via dedicated processing algorithms.

The inclusion of LiDAR in Apple's devices has expanded the potential of mobile applications. For instance, in augmented reality (AR), it enhances scene understanding and improves the realism and stability of digital overlays. This has led to more immersive ARKit-based applications and games. In photography, LiDAR assists with faster and more accurate autofocus, particularly in low-light conditions, and improves depth estimation for portrait mode and AR photography. Additionally, the depth data supports advanced scene and object recognition, contributing to accessibility features that help visually impaired users better navigate their surroundings.

C. Contributions and Organization

A novel learned image compression system that incorporates LiDAR depth map to enhance coding efficiency for machine vision tasks is proposed in this work. Although LiDAR is commonly used in high-level computer vision tasks, the proposed system aims to exploit its information to effectively guide image compression for machine vision applications, leveraging the strong correlation between depth cues and object structures. Note that the present work focuses on coding for machine vision, where the compressed representation is optimized directly for downstream task performance rather than for reconstruction quality. Prior research [21], instead, explored the use of LiDAR depth maps for learned image compression intended for human consumption, explicitly optimizing the codec for conventional image fidelity metrics. In a similar spirit, several recent works have investigated guided or side-information-assisted image compression frameworks aimed at improving pixel-level or perceptual reconstruction quality [22]–[24]. However, the potential of readily available LiDAR depth maps for CfM remains largely unexplored, since previous CfM works focus on adopting costly auxiliary information.

The proposed framework is illustrated in Fig. 1. It builds on an existing learned image codec optimized for human perception and proposes three key components to adapt it to machine vision: (1) LiDAR-augmented encoding incorporates depth information into the encoding process via a dedicated **depth branch** module, effectively guiding it to retain semantic information most relevant to downstream machine vision tasks. (2) Instead of re-training the entire encoder for each downstream task, the proposed approach freezes the pretrained image codec and introduces additional, trainable modules (i.e., **adapters**) between encoding blocks in the encoder. These add-on, trainable modules are designed to be lightweight, and their removal reverts the codec to its original role of performing reconstruction for human vision, which can be regarded as one of our target task. This offers a practical and efficient solution to CfM considering the large number of existing tasks and networks, while preserving the flexibility of decoding images for human viewing at the same time. (3) Latent **bridge** decoding enables the direct adaptation of latent representations for downstream task networks instead of using reconstructed images, removing the computational cost needed for full image decoding and some of the early recognition process. Unlike [21], which requires the explicit transmission of the LiDAR depth map, our approach uses depth information exclusively at the encoder, avoiding additional bandwidth usage.

With a generalized design, our proposed method is shown to be applicable to a wide variety of downstream machine vision tasks and different codec architectures, demonstrating effectiveness over applications including traditional RGB or RGB-D recognition tasks, and emerging Multimodal Large Language Models (MLLMs), which process visual data alongside textual inputs. Experiments validate the advantages of integrat-

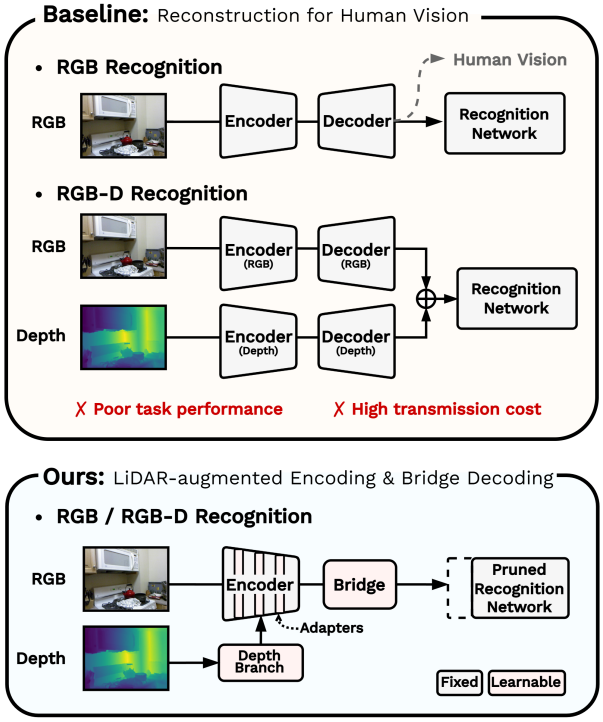


Fig. 1. High-level comparison of our proposed method and the baseline using reconstructed images for human vision. Our proposed method leverages device captured depth information for image compression for machine vision.

TABLE I
COMPARISON OF RELATED CODING FOR MACHINES PRIOR WORKS AND OUR PROPOSED METHOD.

	Bypass Image Reconstruction	Auxiliary Information	Target Tasks	Re-trained Modules for a New Task
Single-task Optimization [7], [8]	✗	None	RGB Recog.	Full model
Universal Codec [25], [26]	✗	None	RGB Recog.	None
Compressed Domain Processing [8], [27]–[31]	✓	None	RGB Recog. / MLLM	Full model / Bridge
Task-specific Adaptation [5], [6]	✗	None	RGB Recog.	Adapters
Auxiliary Data Utilization [10]–[13], [32]	✗	ROI map / Semantic Info.	RGB Recog.	Full model / None
Proposed Method	✓	(LiDAR) Depth map	RGB/RGB-D Recog. & MLLM	Adapters & Bridge

ing LiDAR depth into the compression system, showcasing improved recognition accuracy, reduced bitrate, and increased efficiency for machine vision applications. In comparison to existing CfM frameworks, our proposed system combines compressed domain recognition with low-cost LiDAR depth information, enabling efficient and effective compression for a wide range of machine vision tasks. In particular for RGB-D recognition tasks, most existing CfM methods and human-optimized codec require additionally transmission of depth map to enable decoder-side recognition, while our proposed method implicitly encodes depth information in latent representation. The contributions of this work are summarized as follows:

- To our best knowledge, this work represents the first-ever learned image coding for machine system that utilizes the low-cost LiDAR depth map as auxiliary information.
- Our framework is lightweight and practical as it adapts a fixed pre-trained codec with LiDAR-guided encoder modulation and a latent bridge, enabling task-specific inference directly from compressed latents without full image reconstruction.
- It generalizes across diverse tasks, from conventional recognition networks (RGB and RGB-D) to multimodal large language models, and consistently achieves superior rate–accuracy performance compared to baselines.

The remainder of the paper is organized as follows. Sec. II provides an overview of related works, with a focus on learned image compression for machine vision and depth map-assisted computer vision. Sec. III details the proposed method, while Sec. IV presents a comprehensive experimental evaluation. Ablation studies and extended analysis are reported in Sec. V. Finally, Sec. VI concludes this work.

II. RELATED WORK

This section provides an overview of related works on learned image compression methods for machine vision (Sec.II-A) and depth map assisted computer vision approaches (Sec.II-B). The comparison between related works and our proposed method is summarized in Table I.

A. Learned Image Compression for Machine Vision

With the increasing prevalence of computer vision applications, the learned image compression methods tailored for machine vision has recently gained significant attention. Early attempts at coding for machines focused primarily on optimizing the compression system for a specific downstream task [7], [8]. While end-to-end training with a task-specific loss achieves strong rate-accuracy performance, it requires re-training for each task and model, making it computationally prohibitive. In contrast, more recent methods [25], [26] aim to develop general compression frameworks suitable for multiple downstream target tasks simultaneously. Feng et al. [25] utilizes contrastive learning to extract semantically rich features suitable for various downstream tasks, while Tian et al. [26] leverages large vision models to guide the feature learning process for ensuring broad applicability. While this line of methods provides an efficient and universal solution, the performance may be suboptimal for individual tasks due to their general design.

A different strategy has emerged in recent works [5], [6], which aims to adapt a pre-trained base codec for different machine tasks rather than training an entirely new system in an end-to-end manner. These methods introduce task-specific, plug-and-play adaptation modules into an off-the-shelf, fixed image codec, allowing efficient transfer to machine vision without modifying the base codec. By training only the added lightweight modules, this approach avoids the expensive full model re-training while preserving task-specific performance.

On the other hand, many studies have been exploring the approach of directly performing machine vision recognition in the compressed domain rather than reconstructed images. Early attempts [27], [28] directly feed the compressed image latents into a partial recognition model for inference, and jointly train the entire system, including downstream network, in an end-to-end fashion. Mei et al. [8] propose to employ an additional bridge network connecting learned image codec and object detection network. More recently, the authors in [30], [33] propose a compressed domain face detector using the latent code present in the JPEG AI architecture [34]. Several methods have also explored the scalable approach, which only select a portion of the latent representation for downstream computer vision task. Liu et al. [7] first select several channels from the compressed latents and utilize a simple network with upsampling to bridge the pruned task network for inference. Furthermore, in [29], the authors propose a gate module to dynamically select suitable subset of the latent representation. This line of methods enables to skip image reconstruction, thus saving computational complexity, while maintaining downstream performance for the task being considered.

Another line of research has explored incorporating auxiliary information into the compression pipeline to enhance

coding efficiency for machine vision. Several methods [9]–[11], [32] utilize Region-of-Interest (ROI) maps, oftentimes generated by a neural network, that indicates the location of objects in an image for encoding. By allocating more bits on the object regions, or preserving all objects, these methods aim to reduce encoded bits while keeping relevant semantic information. More recent works [12], [13] further leverage large foundation models, e.g. Grounded-SAM [35] and MLLMs [36], to extract high-level semantic information before encoding. While these approaches demonstrate performance improvement given auxiliary information, generating ROI maps or semantic annotations requires substantial computational resources, making them impractical. Moreover, performing recognition at the encoder side is inappropriate with the fundamental premise of coding for machines, where the encoder is typically assumed to be resource-constrained. Although the prior work [14] attempts to mitigate this issue by leveraging auxiliary information only during training, its generalization remains uncertain. Specifically, the method uses SAM-generated segmentation maps [37] to guide the codec toward preserving object edges while discarding most texture and high-frequency details. However, the assumption that object shapes alone suffice for downstream tasks does not always hold, and reliably identifying and preserving object boundaries across diverse images is inherently challenging and requiring a computational-heavy base codec.

Beyond traditional recognition tasks, recent efforts have also explored compression methods tailored for emerging machine vision applications, such as MLLMs, which exhibit remarkable reasoning capabilities. However, given their large model size up to billion-scale, common coding for machine methods cannot be directly applied due to training infeasibility. A recent work [31] addresses this challenge by introducing a surrogate loss that is backpropagated only through the visual encoder of the MLLM, circumventing the training issue.

In this work, we propose a learned image compression system that leverages the LiDAR depth maps to enhance image coding for machine vision tasks. By integrating task-specific adaptation modules into a fixed base codec, the method ensures efficient compression while maintaining strong downstream performance for a wide range of computer vision tasks.

B. Depth Map Assisted Computer Vision

Depth information has been widely recognized as beneficial across various computer vision tasks, spanning both high-level and low-level vision applications. Unlike traditional RGB images, which primarily capture texture and color information, depth maps additionally provide geometric and structural cues that are useful for various applications. Many prior methods have been proposed for developing the neural network integrating depth information alongside RGB images for tasks such as semantic segmentation [16], [17], object detection [17], [18]. Specifically, Cao et al. [15] has shown that incorporating depth information improves semantic segmentation performance by 2–4% compared to using RGB images as input alone. In terms of low-level computer vision, Gnutti et al. [21] first proposed a LiDAR depth map-assisted compression system optimized for human perception. Such approach employs a prompt-based architecture, inspired by [38], to incorporate depth maps into both the encoder and decoder. The additional modality assists the compression in improving RD performance even when considering the added bitrate required to transmit depth map explicitly. Even though the advantages of depth information in machine vision and compression have been demonstrated, there has not been any attempt on leveraging LiDAR depth map in image coding for machines.

III. Proposed Method

This section introduces the proposed framework. Sec. III-A provides technical background on learned image compression. The high-level system architecture is outlined in Sec. III-B, followed by a description of the LiDAR-augmented encoding and latent decoding via the bridge module in Sec. III-C and III-D, respectively. Finally, the overall training procedure is detailed in Sec. III-E.

A. Preliminaries

Most modern VAE-based learned image compression systems follow the similar framework of hyperprior [1], which consists of the analysis transform g_a, the synthesis transform g_s, and the hyperprior autoencoder h_a, h_s. Given an RGB image $x \in \mathbb{R}^{3\times H\times W}$ of spatial resolution $H\times W$, g_a encodes the image into a compact latent representation $y \in \mathbb{R}^{C_y\times \frac{H}{16}\times \frac{W}{16}}$ of a smaller resolution. The latents y are quantized and entropy encoded for transmission to the decoder side, generating $\hat{y}$, which is decoded through g_s to produce the reconstructed image $\hat{x} \in \mathbb{R}^{3\times H\times W}$. To efficiently encode the quantized latents $\hat{y}$, a side information $z \in \mathbb{R}^{C_z\times \frac{H}{64}\times \frac{W}{64}}$ is derived through the hyper analysis transform h_a. The quantized side information $\hat{z}$ then goes through the hyper synthesis transform h_s to generate the probability estimates of the image latents $\hat{y}$ for entropy coding.

B. System Overview

This paper presents a novel learned image compression system that integrates two key components: (i) LiDAR-augmented encoding, which incorporates depth information for encoding, and (ii) latent decoding via a bridge component, which enables the adaptation of the latent representation to downstream tasks without requiring full image reconstruction. The overall architecture is illustrated in Fig. 2. Given an RGB image x and a LiDAR depth map d of the same resolution, the encoder g_a is modulated using depth features extracted from d via the proposed depth branch d_a and LiDAR-augmented Spatial-Frequency Modulation Adapters (L-SFMA), derived from [6]. This approach aims to preserve relevant depth information in the latents, ensuring the downstream task performance while minimizing bitrate. The quantized latent representation $\hat{y}$ is then directly adapted for inference in a machine vision network. The proposed latent bridge aligns the compressed latents with the intermediate features of the task model, skipping both image decoding and some early feature extraction of the downstream network. The net effect is reduced computational complexity.

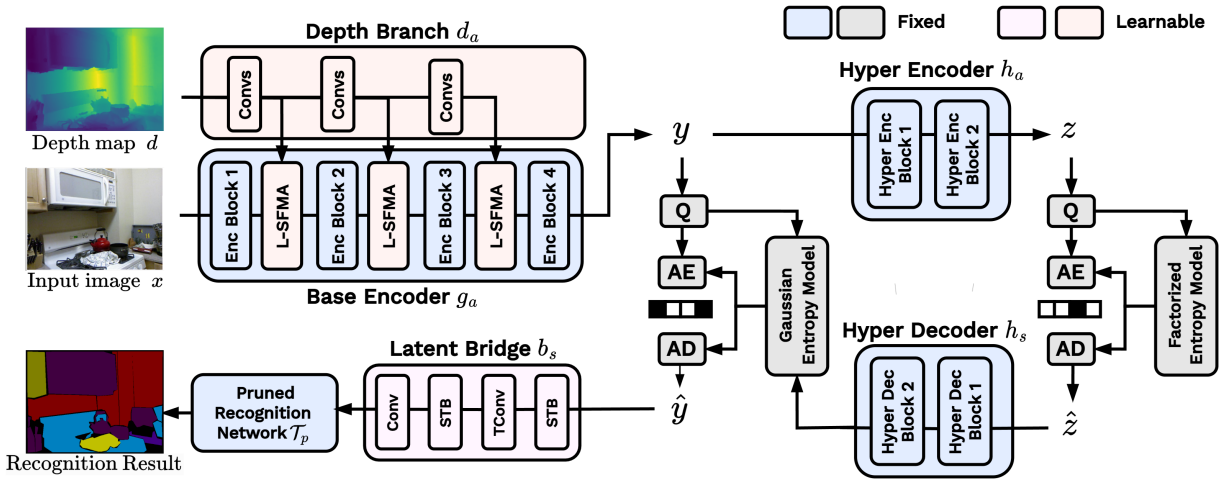


Fig. 2. Overall architecture of the proposed method.

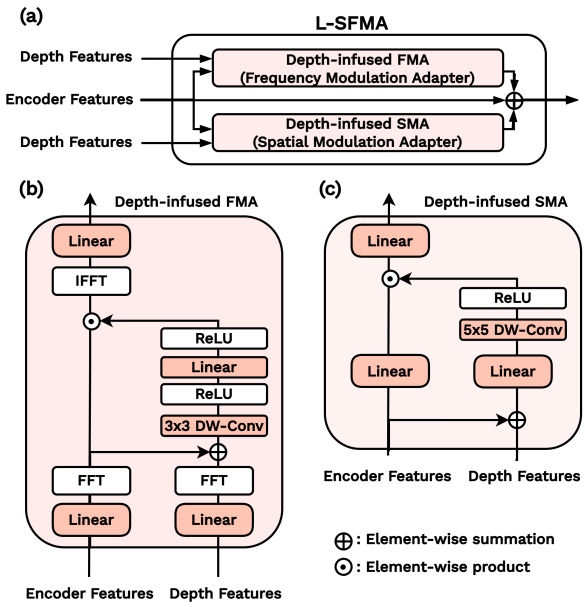


Fig. 3. Detailed architecture of (a) LiDAR-augmented SFMA (L-SFMA), which consists two modulation adapters: (b) depth-infused FMA for frequency domain and (c) depth-infused SMA for spatial domain.

Following the prior work [6], our base encoder g_a is pre-trained for human perception and remains fixed throughout the training process. The same applies to the hyperprior autoencoder h_a, h_s and entropy models. This design ensures an efficient coding-for-machines framework that maintains compatibility with human-viewing applications while enabling task-specific adaptation to task performance. Since only a portion of the system is trainable for each downstream task, this also avoids the expensive full re-training given the large number of possible downstream tasks. For this work, it is assumed that the LiDAR depth maps are captured by devices and available at the encoder alongside the RGB images. Considering certain scenarios where this is not true, we also explore the effectiveness of our method with model-generated depth information (Sec. IV).

C. LiDAR-augmented Encoding

To adapt a pre-trained image codec to a machine vision task with the aid of depth information, we introduce a dedicated depth branch d_a to extract features from the input depth map d on top of the base encoder g_a. These features are incorporated into the encoding process through the proposed LiDAR-augmented SFMAs (L-SFMAs), which are inserted between the encoding blocks of g_a. Specifically, three L-SFMAs are placed at different stages of the encoder, each processing the corresponding intermediate features from both d_a and g_a at the same spatial resolution.

Taking inspirations from the Spatial Feature Transform (SFT) [39] and its extension [6], the proposed L-SFMA aims to transform the input features for specific downstream tasks by incorporating depth information. Specifically, similar to [6], the modulations are performed in both spatial and frequency domains accounting for the distinct feature characteristics for different tasks in each domain. To achieve this, L-SFMA introduces two components: the Spatial Modulation Adapter (SMA) and the Frequency Modulation Adapter (FMA), both of which are designed to infuse LiDAR depth information. Fig. 3 illustrates the detailed architecture of L-SFMA and its components. Unlike prior methods that solely rely on input features for modulation, the proposed L-SFMA leverages depth features as an additional guidance signal. The output of both adapters are aggregated with the original encoder features through element-wise summation (see Fig. 3a).

The Frequency Modulation Adapter (FMA) adjusts the encoder features in the frequency domain by leveraging LiDAR depth information (see Fig. 3b). First, both input features are transformed with Fast Fourier Transform (FFT) to extract their spectra. The transformed encoder features are then refined through an element-wise product with a modulation matrix. The modulation matrix is computed with the summation of two transformed input features via a lightweight processing pipeline consisting of a convolutional layer, a linear transformation, and ReLU activations. At the end, an Inverse FFT (IFFT) is applied to revert the output features back to the spatial domain.

In contrast, the Spatial Modulation Adapter (SMA) operates

TABLE II
VARIOUS DOWNSTREAM MACHINE VISION TASKS AND CORRESPONDING NETWORKS, DATASETS, AND EVALUATION METRICS.

Task	Network	Dataset	Metrics
	Traditional RGB Recognition Network		
Classification	ResNet50 [40]	ImageNet	Top-1 Accuracy
Object Detection	Faster R-CNN [41] (ResNet101)	COCO2017	mAP
Instance Segmentation	Mask R-CNN [42] (ResNet101)	ARKitScenes [43]	mAP
Semantic Segmentation	DeeplabV3 [44] (ResNet101)	NYU-v2 [45]	Pixel Accuracy
	Traditional RGB-D Recognition Network		
Semantic Segmentation	ShapeConv [15] (ResNeXt101)	NYU-v2	Pixel Accuracy mIoU

directly in the spatial domain, without any spectral transformation (see Fig. 3c). It follows a similar modulation pipeline to the FMA, employing convolutional layers to adaptively refine encoder features. By simultaneously modulating in both the spatial and frequency domains with additional LiDAR depth information, the proposed approach effectively preserves essential task-relevant features without changing its compression capabilities.

D. Latent Decoding Via Bridge

Instead of employing a conventional image decoder, a latent bridge b_s is introduced for transforming directly the quantized image latents $\hat{y}$ into a form suitable for the downstream task. Consisting of two convolutional layers and two Swin Transformer Blocks (STB), the latent bridge ensures that the compressed latent information, which retains sufficient semantic information, aligns with the intermediate features of a recognition network. By integrating the latent representation into the target task network, the computational overhead is significantly reduced while maintaining high task accuracy. Note that the depth map information is incorporated solely in the encoder, eliminating the need for explicitly signaling the depth map d in the bitstream.

Given that the latent representation already captures rich low-level features through the encoding network, it becomes feasible to prune the recognition model by removing certain early stages that are primarily responsible for extracting these low-level features. Notably, this further reduces computational complexity on the decoder side.

E. Training Objective

The proposed system is trained using a Rate-Distortion (RD) objective, following common learned compression approaches. However, unlike conventional compression methods optimized for human perception, which typically adopt the Mean Squared Error (MSE) between the original and reconstructed images as the distortion metric, our approach defines distortion based on the downstream task network to better capture task-specific degradation.

Accordingly, the overall loss function is formulated as

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

where λ is the hyperparameter balancing between R and D. R represents the estimated bits required to transmit both $\hat{y}$ and $\hat{z}$:

$$R = -\log p(\hat{z}) - \log p(\hat{y}|\hat{z}), \tag{2}$$

and D denotes the machine vision-based distortion that incorporates both a task loss $\mathcal{L}_{\text{Task}}$ and a perceptual loss $\mathcal{L}_{\text{Perc}}$ [5], [6], and is expressed as

$$D = \mathcal{L}_{\text{Task}}\left(\mathcal{T}_{\text{P}}\left(b_s(\hat{y})\right), \text{GT}\right) + \mathcal{L}_{\text{Perc}}\left(x, b_s(\hat{y}); \mathcal{T}, \mathcal{T}_{\text{P}}\right). \tag{3}$$

Here, $\mathcal{L}_{\text{Task}}$ corresponds to the specific training loss associated with the target downstream task (e.g. the cross-entropy loss for classification), with $\mathcal{T}_{\text{P}}$ as the pruned recognition network and GT as the ground truth. The perceptual loss $\mathcal{L}_{\text{Perc}}$, on the other hand, measures the distance between the intermediate features of the recognition network $\mathcal{T}$ when using the uncompressed image x as input, and those of the pruned recognition network $\mathcal{T}_{\text{P}}$ when using the transformed latents $b_s(\hat{y})$ as input. Through the two loss terms, our training objective ensures that the semantic information essential to the task is preserved and the task performance is maintained. It is important to note that throughout training, only depth branch d_a, L-SFMAs, and latent bridge b_s are updated, while the base codec, i.e., g_a, h_a, h_s, and the pruned recognition network remain frozen.

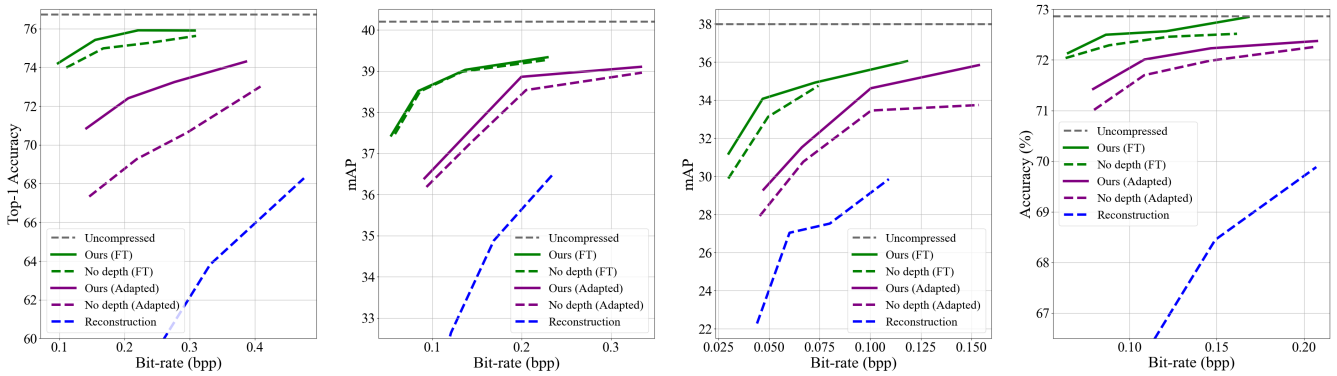
IV. EXPERIMENTAL RESULTS

The experimental setting is first described in Sec. IV-A. To demonstrate the versatility of the proposed method, which is designed to be general and adaptable across a broad spectrum of downstream machine vision tasks, we evaluate it on RGB-D recognition networks besides the common RGB machine vision tasks in Sec. IV-B. Furthermore, we compare our proposed LiDAR-guided framework with existing CfM methods in Sec. IV-C to demonstrate its effectiveness. Additional results of the proposed approach when applied to multimodal large language models (MLLMs) are presented in Sec. IV-D, with the aim of further assessing and extending its generalizability. Finally, computational complexity aspects are discussed in Sec. IV-E.

A. Experimental Settings

The specific tasks, recognition networks, datasets, and evaluation metrics are reported in Table II. It is important to note that all the employed downstream networks are pre-trained and fixed. Additionally, the proposed method is applicable to different base codecs, as demonstrated in the experiments using TIC [46], a Transformer-based codec for RGB/RGB-D tasks, and ELIC [3], a CNN-based codec for MLLM tasks. For each task, a separate system (i.e., depth branch, adapters and bridge) is trained at four different λ values to control the bitrate and explore the rate-performance trade-off.

We consider two configurations under our framework: **Adapted** and **FT**, which stands for *Full Fine-tuning*. Both share the same architecture, combining LiDAR-augmented encoding with latent bridge decoding. In the **Adapted** setting, the base codec is frozen and only the lightweight adapters and bridge are trained, providing an efficient and modular



(a) Image Classification (b) Object Detection (c) Instance Segmentation (d) Semantic Segmentation

Fig. 4. The rate-accuracy comparison on traditional RGB recognition tasks.

TABLE III
BD-RATE AND BD-METRIC COMPARISON (RECONSTRUCTION AS THE ANCHOR).

			No depth (Adapted)		Ours (Adapted)		No depth (FT)		Ours (FT)	
		Metric	BD-metric↑	BD-Rate↓	BD-metric↑	BD-Rate↓	BD-metric↑	BD-Rate↓	BD-metric↑	BD-Rate↓
RGB Recog.	Classification	Acc.	10.84	-65.04	14.37	-79.35	20.66	-	22.38	-
	Obj. Detection	mAP	3.55	-53.44	3.93	-57.81	5.47	-71.91	5.53	-76.02
	Instance Seg.	mAP	3.94	-29.52	4.90	-58.13	7.62	-	7.87	-
	Semantic Seg.	Acc.	4.59	-	4.86	-	5.98	-	6.02	-
RGB-D Recog.	Semantic Seg.	Acc.	-4.00	-	1.12	-37.63	-2.47	82.13	2.54	-64.19
	Semantic Seg.	mIoU	-2.98	-	0.73	-36.89	-1.71	10.54	1.87	-63.24

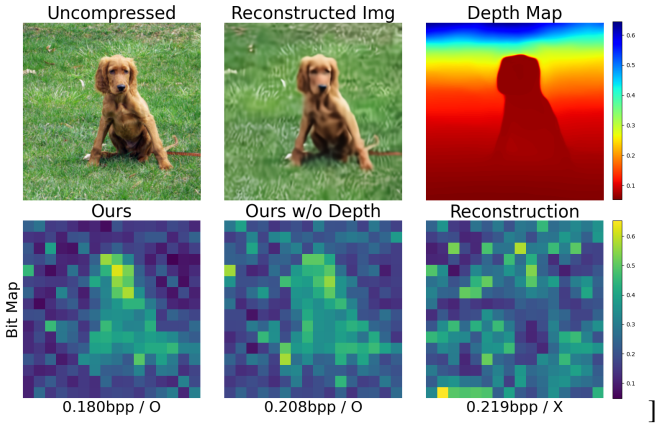


Fig. 5. Visualization of the reconstructed images and bit allocation maps of compressed latents on image classification for ours and baseline methods. The bits-per-pixel and prediction result ('O' for correct and 'X' for incorrect) are provided below each image. Our proposed method more effectively allocate bits to foreground object, achieving correct prediction with lower bitrate.

solution. In the **FT** setting, the entire system, including the base codec, is updated end-to-end. This allows the system to reach improved performance, but it results in substantially higher computational and storage costs.

As a baseline, we compare against a variant **No depth** in which LiDAR depth information is absent under two configurations. Specifically, we train the system of the architecture but using a uniform map filled with 0 as depth input to the depth branch. This is to highlight the impact of LiDAR depth as auxiliary information for machine vision tasks. Additionally, we include two reference baselines: **Uncompressed**, which directly feeds uncompressed RGB images (and depth maps, when available) into the downstream networks, serving as an upper bound; and **Reconstruction**, which uses images reconstructed by the base codec as input to downstream networks, reflecting the performance of human-centric compression.

B. Performance on Traditional Recognition Networks

For these experiments, we adopt the pre-trained Transformer-based Image Compression (TIC) [46] as the base codec. Among the datasets used, ARKitScenes [43] and NYU Depth V2 [45] both provide ground-truth depth maps. Particularly, the former is the first RGB-D dataset captured with the Apple LiDAR scanner, offering a setting that closely mirrors real-world application scenarios. Since ground truth annotations for RGB recognition tasks are not available in ARKitScenes, we employ a more powerful pre-trained model, YOLOv11 [47], to generate pseudo ground truths (i.e., object mask annotations) for instance segmentation. For the other computer vision datasets that do not include depth information (i.e., ImageNet and COCO2017), we generate depth maps using a pre-trained depth estimation network [48]. This also serves to validate the robustness of the proposed method across various scenarios, including cases where LiDAR depth information is not readily available. Training is conducted on the predefined training split of each dataset, with evaluation performed on its respective validation split.

We utilize pre-trained recognition networks for downstream tasks, but modify their architectures by removing the first few layers, since we consider that the image encoding process has already produced proper features, as mentioned in Sec. III-D. Specifically, for the recognition tasks, the first three processing

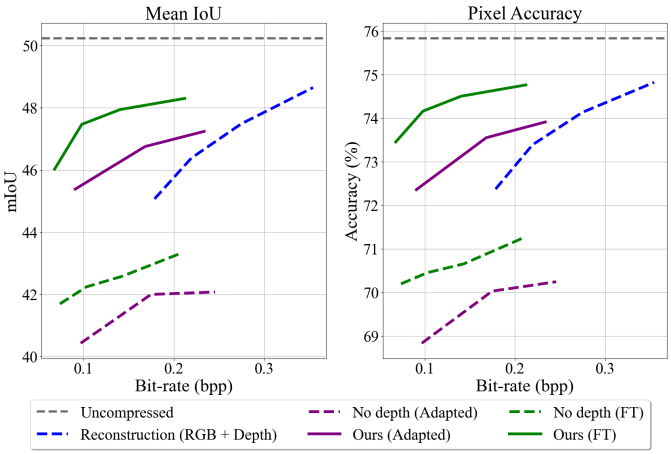


Fig. 6. The rate-accuracy comparison on traditional RGB-D recognition tasks.

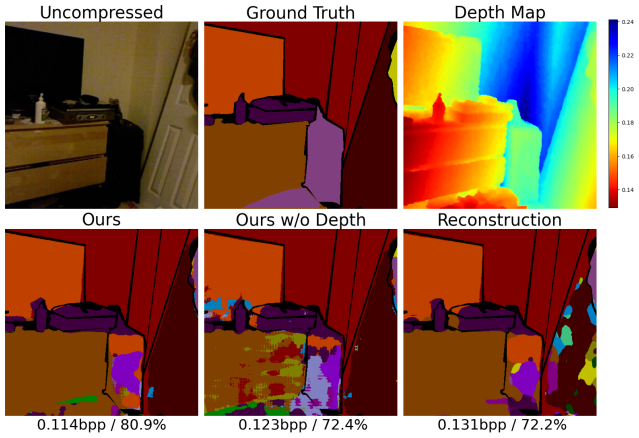
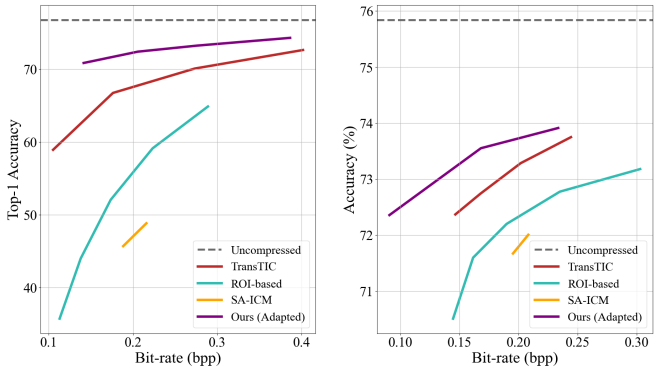


Fig. 7. Visualization of the recognition results on RGB-D semantic segmentation with corresponding depth map. The bits-per-pixel and pixel accuracy are denoted under each image. Our proposed method consistently shows superior prediction at a lower bitrate.

stages for ResNet50 [40] and Deeplabv3 [44] are skipped; likewise, the first two stages for Faster R-CNN [41], Mask R-CNN [42], and ShapeConv [15] are removed. To ensure compatibility, the output channel size of the final convolution layer of the proposed latent bridge is configured to match that of the input of each pruned recognition network.

1) RGB recognition networks: Figure 4 presents the rate-accuracy comparison for recognition networks using RGB-only inputs, covering four representative architectures across different datasets. Each point in the curves represents an averaged bits per pixel (bpp) and a task performance across all the images in the testing dataset. The overall quantitative performance are summarized in Table III using the Bjontegaard metrics. BD-Rate indicates the average bitrate differences under the same task performance, while BD-metric (e.g., BD-Accuracy, BD-mAP) quantifies the average improvement of task metric under equal bitrate. Based on the comparison, we draw the following observations. (1) The proposed framework, integrating both LiDAR-augmented encoding and latent bridge decoding, achieves significant performance improvements over the **Reconstruction** baseline, which feeds directly reconstructed images from the base codec into pre-trained



(a) RGB Classification (b) RGB-D Semantic Seg.

Fig. 8. Rate-accuracy comparison with other CfM methods.

downstream networks. This highlights the negative impact of standard compression artifacts on recognition accuracy and emphasizes the importance of task-specific framework design. (2) Incorporating auxiliary depth information via the LiDAR-augmented encoder further enhances the task performance at the same bitrate, demonstrating how depth information can effectively guide the encoding process to ensure semantic relevance. (3) By allowing the base codec to be re-trained, the **FT** approach achieves the best rate-accuracy performance. However, it comes at the cost of significantly higher computational and storage complexity, as will be discussed in Sec. IV-E. Notably, even in this setting, integrating depth information into the method continues to yield improvements, highlighting its benefits.

To further illustrate the impact of the proposed method, we provide a visualization analysis. Since the proposed approach does not require reconstruction, we compare the bit allocation maps of the latent representation in Fig. 5 for the image classification task. It is evident that *Reconstruction*, optimized for human vision, allocates more bits to complex regions (e.g., grass), even though they are in the background and less relevant to the recognition task. In contrast, the proposed method prioritizes foreground objects, resulting in superior rate-accuracy performance. Specifically, depth information provides strong cues about object locations and boundaries.

2) RGB-D recognition networks: To further evaluate the generality of our method, the analysis has been extended to specific RGB-D recognition tasks, where downstream machine vision networks are pre-trained to process both RGB and depth input data. The rate-accuracy curves are shown in Fig. 6 (see again Table III for BD-Rate and BD-metrics).

For the *Reconstruction* baseline, both the RGB image and the depth information are explicitly encoded and reconstructed before being passed into the downstream task network, leading to significantly increased bitrates due to the need to transmit both modalities. In contrast, the proposed method employs the LiDAR-augmented encoder, which implicitly incorporates the depth information into the compressed latent representations. As such, our method does not need to explicitly signal the depth map in a separate channel. The resulting latents are then transformed via the bridge to directly connect with the pruned (RGB-D) recognition network, effectively bypassing the input

TABLE IV
BD-RATE AND BD-METRIC COMPARISON (RECONSTRUCTION AS THE ANCHOR).

			No depth (Adapted)		Ours (Adapted)		No depth (FT)		Ours (FT)	
		Metric	BD-metric↑	BD-Rate↓	BD-metric↑	BD-Rate↓	BD-metric↑	BD-Rate↓	BD-metric↑	BD-Rate↓
MLLM	REC	Acc.	2.36	-26.77	3.32	-43.53	7.89	-77.07	8.11	-77.15
	VQA	Score	1.92	-46.43	2.33	-60.39	3.98	-82.61	3.88	-81.75

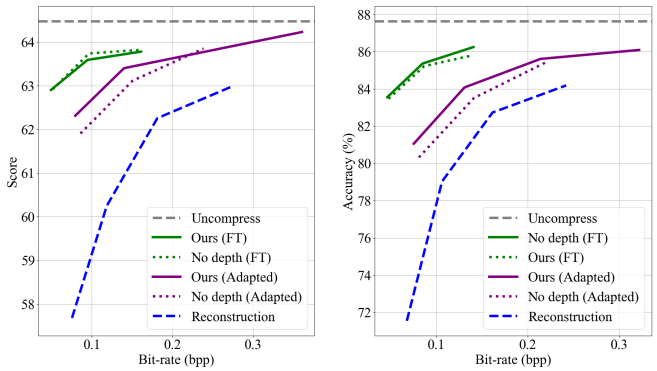
constraint of requiring separate RGB-D data. Notably, our system achieves high task performance, especially when operating at lower bitrates. In cases where depth information is absent during compression (**No depth**), the performance degrades drastically, underscoring the critical role of depth in RGB-D models. These findings reinforce the effectiveness of our approach in leveraging depth information for coding efficiency, while maintaining compatibility with multimodal downstream models.

Finally, the recognition results for RGB-D semantic segmentation are visualized in Fig. 7. It is first observed that the absence of depth information (**No depth (Adapted)**) leads to poor segmentation results, as the downstream network is pre-trained using both RGB and depth data. In contrast, with the aid of the depth map, which provides cues about the objects present in the scene, the proposed method achieves more precise segmentation, particularly at object boundaries. This leads to higher segmentation accuracy at lower bitrate.

C. Comparison with Existing CfM Methods

To further demonstrate the coding efficiency of our LiDAR-guided framework, we compare against three representative CfM baselines on RGB image classification and RGB-D semantic segmentation. In particular, TransTIC [5] serves as a transferring-based CfM baseline. For methods leveraging auxiliary information, we adopt an ROI-based approach [9] and SA-ICM [14]. Following [5], ROI maps for [9] are generated using Grad-CAM for classification and predicted segmentation maps for other tasks, and the RD curves are obtained by assigning different weights to ROI regions. For SA-ICM, SAM-generated segmentation maps are used to guide compression toward preserving object boundaries during training. For all these methods, the reconstructed images are directly passed into downstream networks for recognition.

Fig. 8 reports the rate–accuracy performance. Our method consistently achieves superior performance compared with existing approaches. The integration of LiDAR depth with lightweight adaptation modules provides more informative guidance than TransTIC, which does not use auxiliary information. ROI-based and SA-ICM methods, while leveraging auxiliary information, show limited rate-accuracy performance. The quality of generated ROI maps heavily affect the compression process. SA-ICM is trained to focus solely on object boundaries discarding the textures, which leads to degraded performance in classification and also poor transferability to unseen datasets (e.g., NYU-v2). Furthermore, none of these baselines support efficient compression for RGB-D tasks, as they require explicit transmission of the depth map, whereas our method integrates depth implicitly within the latents at no additional bitrate.



(a) VQA (b) REC

Fig. 9. Rate-accuracy comparison in MLLM-based computer vision tasks.

D. Performance on Multimodal Large Language Models

Given the growing adoption of MLLMs in real-world applications, we further explore the application of the proposed method to large-scale models, which serve as the downstream task networks. To address the challenge of training infeasibility due to the massive number of parameters in MLLMs, we adopt the strategy proposed in [31], where the task loss is omitted and the perceptual loss is computed solely through the visual encoder of the MLLM, thereby avoiding backpropagating the task loss through the large-scale multimodal language model.

The evaluation is performed with two different MLLMs [49], [50], both of which adopt the CLIP ViT-L/14 visual encoder. The pre-trained ELIC [3], a convolutional neural network-based image codec, is employed as the base codec. Following [31], the training utilizes ImageNet training split, and the evaluation is conducted on SEED Benchmark [51] and RefCOCO [52]. In particular, we evaluate Referring Expression Comprehension (REC) with the RefCOCO [52] validation set and Visual Question Answering (VQA) with SEED-Bench [51]. Similarly, the first two layers of the visual encoder are removed following [38].

The rate-accuracy performance comparison is presented in Fig. 9, with BD-rate and BD-metric summarized in Table IV. Consistent with the observations for traditional recognition networks, the proposed method not only greatly improves upon the base codec optimized for human perception, but it also validates the benefits of auxiliary depth information for **Ours (Adapted)**. In the VQA task, the **Ours (FT)** model without depth already achieves performance near saturation, which limits the potential gains from incorporating depth. Addition-

TABLE V
COMPUTATIONAL COMPLEXITY COMPARISON. TOTALS INCLUDE ENCODER AND DECODER (OR ALL MODULES FOR OUR METHOD).

Method / Module	kMAC/pixel	Model Size (M)
Base Codec (TIC [46])		
Encoder	145.41	4.77
Decoder	186.97	2.74
Total	**332.38**	**7.51**
TransTIC [5]		
Encoder	332.03	5.24
Decoder	202.60	3.89
Total	**534.63**	**9.13**
ROI-based [9]		
Encoder	800.36	21.91
Decoder	679.80	5.65
Total	**1480.16**	**27.56**
SA-ICM [14]		
Encoder	664.01	10.97
Decoder	1123.32	65.60
Total	**1787.33**	**76.57**
Ours		
Base Encoder	145.41	4.77
Depth Branch	108.79	1.18
L-SFMAs	22.21	0.20
Latent Bridge	29.11	2.22
Total	**305.52**	**8.37**

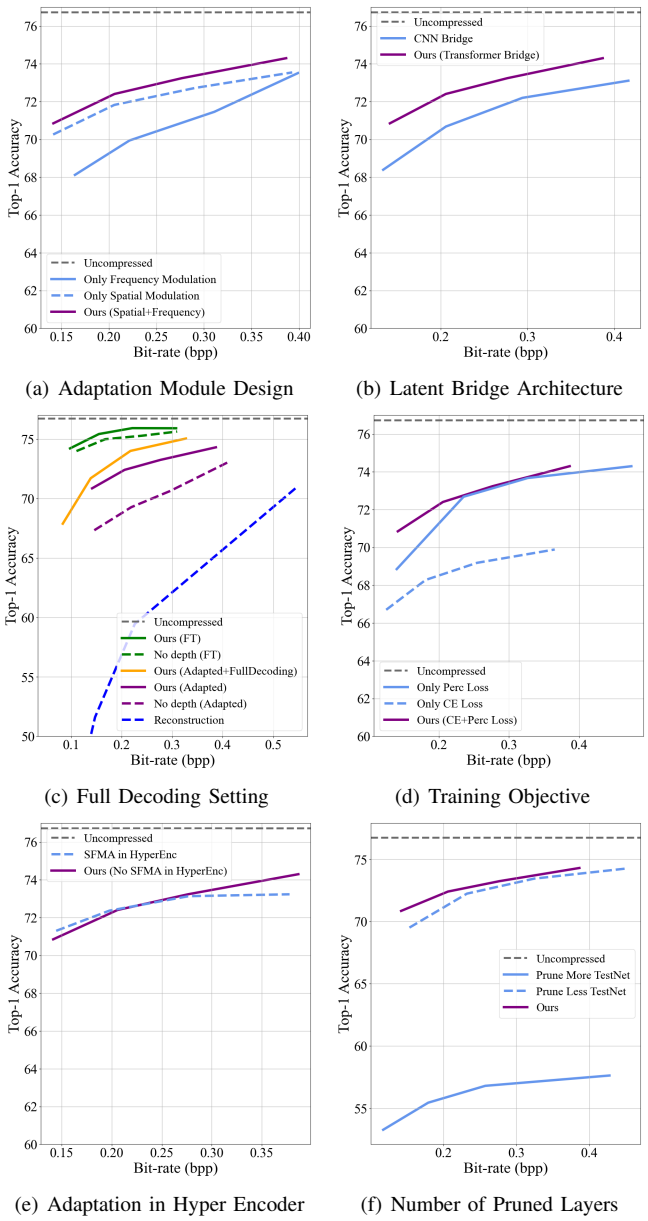
TABLE VI
DECODING COMPLEXITY COMPARISON BETWEEN FULLDECODING AND THE PROPOSED LATENT BRIDGE.

Method	kMAC/pixel
FullDecoding	186.97
Latent Bridge (Ours)	29.11

ally, because VQA relies more on linguistic reasoning than on spatial depth cues, the impact of depth information is naturally reduced. These results demonstrate that the proposed approach generalizes effectively across a wide range of downstream applications and networks.

E. Computational Complexity Comparison

Table V summarizes the computational complexity of the proposed method compared to baseline methods for traditional RGB/RGB-D recognition tasks, where our method adopts TIC [46] as the base codec. Two key metrics are reported: the kilo-multiply-accumulate-operations per pixel (kMACs/pixel) and model size (measured in millions of parameters). In comparison to the base codec, the proposed method significantly reduces the computational cost on the decoder side, requiring only around 15% of kMACs needed for full image reconstruction. Additionally, **Ours (Adapted)** is built on top of a fixed pre-trained base codec; only the adaptation and bridge modules are trained, accounting for less than half of the model size of the original codec. In contrast, the **FT** settings requires full training of the entire proposed system for each individual task, resulting in an n-fold increase in required model size for n separate tasks.



(a) Adaptation Module Design (b) Latent Bridge Architecture

(c) Full Decoding Setting (d) Training Objective

(e) Adaptation in Hyper Encoder (f) Number of Pruned Layers

Fig. 10. Ablation expeirments on our proposed system.

Compared to existing CfM baselines, our framework also exhibits noticeably lower complexity. A key reason is the adoption of compressed-domain recognition via the latent bridge, which avoids full image reconstruction. By contrast, [9], [14] rely on reconstructing task-oriented images using larger base codecs, resulting in heavier computation. Although our LiDAR-guided adapters introduce a modest computation overhead to the base encoder, this cost is insignificant relative to performing full recognition on the encoder side. This, along with the comparison with prior works, confirms that our proposed framework achieves favorable trade-offs between computational efficiency and recognition accuracy.

V. ABLATION AND GENERALIZATION ANALYSIS

In this section, we provide additional analyses to further validate the proposed LiDAR-guided coding-for-machine

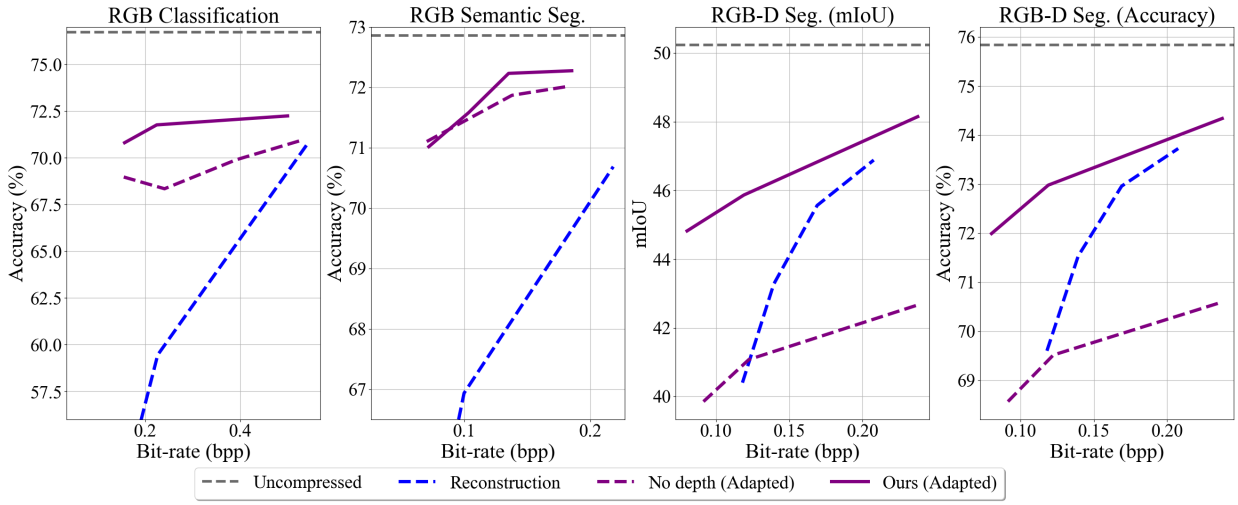


Fig. 11. Performance comparison using ELIC as the base learned image compression model.

framework. First, in Sec. V-A we conduct a series of ablation studies to isolate the contribution of each architectural component and design choice. Then, in Sec. V-B we evaluate the generalizability of the proposed method by replacing the base codec with an alternative learned image compression model. Finally, in Sec. V-C we analyze the robustness of the framework with respect to depth quality, investigating how degradations in the LiDAR signal affect downstream task performance. These experiments provide deeper insight into the behavior, efficiency, and reliability of the proposed approach.

A. Ablation Studies

To validate each proposed component in our proposed LiDAR-augmented coding for machines framework, we conduct a series of ablation experiments. Each experiment focus on one specific design choice and compares the final adopted approach with alternative variants in terms of rate–accuracy performance on the RGB image classification task under **Adapted** configuration.

1) Adaptation Module Design: Inspired by [6], the proposed encoder modulation introduces LiDAR-augmented adaptation in both the spatial and frequency domains. This ablation study evaluates the effect of restricting adaptation to a single domain (spatial-only or frequency-only) in comparison to the proposed dual-domain approach. As shown in Fig. 10(a), adopting frequency-only modulation leads to a significant accuracy drop, likely because frequency-level adaptation alone are insufficient to ensure fine semantic information. Spatial-only modulation performs considerably better, but still lags behind the proposed dual-domain design. Overall, dual-domain modulation achieves the best RD performance, validating our design choice.

2) Latent Bridge Architecture: To bypass the costly full image reconstruction, the proposed latent bridge directly transforms compressed latents for downstream recognition. We adopt a lightweight Transformer-based bridge inspired by [31] to maintain a lightweight system. In Fig. 10(b), we compare this design with a CNN-based counterpart of similar complexity (comparable kMACs/pixel and model size). The Transformer-based bridge consistently outperforms the CNN-based one, demonstrating its superior capability in modeling long-range dependencies with lower complexity, an advantage also observed in prior works [53].

3) Re-trained Full Decoder: To further evaluate the necessity of the proposed latent bridge, we compare it against a *FullDecoding* setting. In this configuration, the latent bridge is removed and the complete image decoder is re-trained jointly within the framework. Recognition is then performed on reconstructed pixel-domain images. Fig. 10(c) reports the RD performance on the RGB classification task. FullDecoding achieves only marginally higher RD performance compared to the proposed latent bridge. However, this slight gain comes at a substantially higher computational cost. As shown in Table VI, the decoder in the FullDecoding setting requires 186.97 kMAC/pixel, whereas the proposed latent bridge introduces only 29.11 kMAC/pixel, resulting in more than a six-fold reduction in decoding complexity. These results demonstrate that operating directly in the compressed domain via the latent bridge achieves a significantly more favorable efficiency–accuracy trade-off for coding-for-machine applications.

4) Training Objective: Our training objective combines a task loss and a perceptual loss to jointly guide the model. This ablation experiment investigates the contribution of each term. Fig. 10(d) shows that using both terms achieves the highest performance overall. Training with only the task loss leads to noticeable degradation, suggesting that task supervision alone does not provide sufficient guidance for model learning. In contrast, the addition of perceptual loss ensures better latent alignment and yields significant improvements, justifying the adoption of the joint objective.

5) Adaptation in Hyper Encoder: In our framework, LiDAR-augmented adaptation is applied only in the main encoder g_a, while the hyperprior encoder h_a remains unchanged. Fig. 10(e) compares it with the setting where L-SFMA are also injected into hyperprior encoder. Although it achieves comparable accuracy, incorporating adaptation in

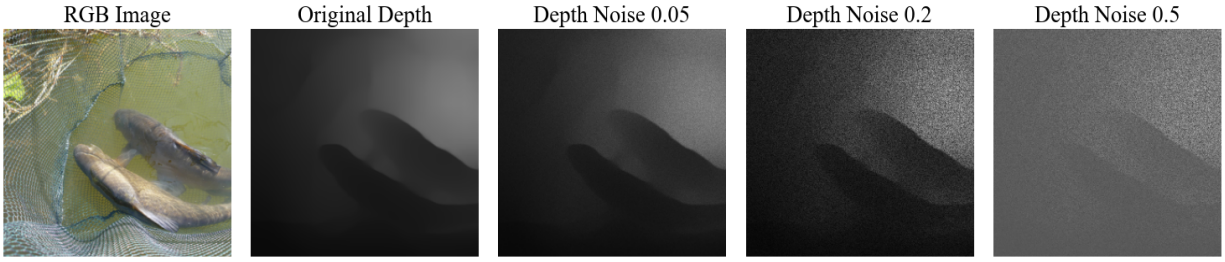


Fig. 12. Example of depth degradation under different Gaussian noise levels. From left to right: RGB image, original depth, and noisy depth maps with $\sigma = 0.05$, 0.2, and 0.5.

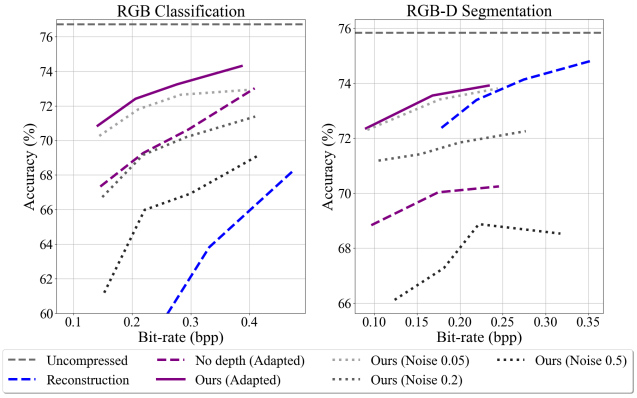


Fig. 13. Rate–distortion performance under different depth noise levels for RGB classification (left) and RGB-D segmentation (right). Increasing depth noise progressively degrades downstream accuracy, confirming the importance of reliable depth guidance.

the hyper encoder introduces additional computational cost clearly. Therefore, we opt for the simpler and more efficient design of restricting adaptation to the main encoder.

6) Number of Pruned Layers: Lastly, we investigate the effect of pruning different numbers of early layers in the downstream recognition networks. For classification, our default design removes the first three processing stages, keeping the most computationally intensive stage (stage 4) intact. Two alternatives are considered: pruning fewer layers (removing only the first two stages) and pruning more layers (retaining only the final stage). As shown in Fig. 10(f), pruning more layers results in a drastic performance drop, as the remaining computation is insufficient for adequate recognition. Pruning fewer layers still yields slightly worse accuracy compared to our setting, and even comes with increased complexity. This behavior can be explained by the fact that the recognition model is originally designed to process raw image pixels, whereas in our framework it receives latent feature representations produced by the LiDAR-guided encoder. If too few stages are removed, these feature-level representations are fed into early layers that are tailored for low-level pixel processing, leading to a representational mismatch that can hinder performance. Thus, pruning three stages provides the best balance between efficiency and accuracy, validating our design choice.

B. Generalization Across Different Codecs

To further assess the generalizability of the proposed framework across different learned image compression models, we extend our experiments by replacing the base codec TIC with ELIC [3] for the traditional RGB and RGB-D recognition tasks. ELIC is a CNN-based learned image compression model, architecturally distinct from TIC. By adopting ELIC as the base codec, we evaluate whether the proposed LiDAR-guided coding-for-machine framework is tied to a specific encoder design or can generalize across different compression backbones.

Fig. 11 reports the performance obtained for RGB and RGB-D recognition tasks. We compare the same four settings used in the main paper: (i) uncompressed images, (ii) reconstruction-based inference (decoded images fed to the recognition model), (iii) latent-domain adaptation without depth guidance, and (iv) the proposed LiDAR-guided approach.

Across all tasks—RGB classification, RGB semantic segmentation, and RGB-D segmentation (mIoU and accuracy)—a consistent trend can be observed. The reconstruction-based pipeline exhibits a clear performance degradation, reflecting the information loss introduced by compression when the codec is optimized primarily for reconstruction quality. Introducing latent-domain adaptation without depth guidance improves performance for RGB tasks, indicating the benefit of operating directly in the compressed domain. However, for RGB-D tasks, removing depth guidance leads to a significant degradation, as depth information is not exploited in this setting. In contrast, the proposed LiDAR-guided framework consistently achieves the best performance among the baselines across all evaluated bit-rates. The gap with respect to the reconstruction baseline is particularly evident at lower bit-rates. While the uncompressed setting still provides an upper performance bound, the proposed method substantially narrows this gap. Overall, the results obtained with ELIC follow trends similar to those observed with TIC, supporting the generality of the proposed framework across different learned compression backbones.

C. Robustness to Depth Quality

Since depth information plays a central role in the proposed LiDAR-guided modulation mechanism, we analyze how the quality of the depth map affects RD performance. To this

end, we inject additive Gaussian noise with different standard deviations ($\sigma = 0.05, 0.2, 0.5$) into the depth maps at inference time, while keeping the compression and recognition models unchanged.

Fig. 12 illustrates an example RGB image and the corresponding depth maps under different noise levels. As the noise variance increases, the structural cues in the depth map progressively deteriorate, eventually becoming unreliable at high noise levels.

The quantitative impact on downstream performance is reported in Fig. 13 for RGB classification and RGB-D segmentation. At low noise levels ($\sigma = 0.05$), the proposed system remains relatively robust, with only minor degradation compared to the clean-depth setting. As the noise increases to $\sigma = 0.2$, a noticeable drop in accuracy is observed, particularly at lower bit-rates. When the noise becomes severe ($\sigma = 0.5$), the performance degrades significantly, approaching or even falling below the depth-agnostic adapted baseline in some cases.

These results confirm that the quality of the depth map directly influences bit allocation and downstream task performance. While the proposed framework demonstrates robustness to mild depth perturbations, accurate depth estimation remains important for maximizing coding efficiency and task accuracy.

VI. Conclusion

In this paper, we proposed a novel learned image coding framework for machines that leverages readily-available, low-cost auxiliary LiDAR depth information. By introducing depth-augmented encoding and latent decoding via bridge, the method efficiently integrates depth cues to adapt the encoder, while reducing computational complexity bypassing full image reconstruction. Through extensive evaluation across a wide range of downstream machine vision tasks, including traditional recognition networks and MLLMs, we demonstrated the effectiveness and generalization capabilities of the proposed approach. Specifically, we showed that incorporating depth information consistently enhances image compression performance for machine vision, independently of the specific task. As the first work on exploring the use of LiDAR depth information on image coding for machines, this method opens up promising directions for learned compression in real-world contexts.

References

[1] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

[2] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[3] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. 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*, 2022.

[4] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[5] Y.-H. Chen, Y.-C. Weng, C.-H. Kao, C. Chien, W.-C. Chiu, and W.-H. Peng, "Transtic: Transferring transformer-based image compression from human perception to machine perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[6] H. Li, S. Li, S. Ding, W. Dai, M. Cao, C. Li, J. Zou, and H. Xiong, "Image compression for machine and human vision with spatial-frequency adaptation," in *European Conference on Computer Vision*, 2024.

[7] J. Liu, H. Sun, and J. Katto, "Learning in compressed domain for faster machine vision tasks," in *2021 International Conference on Visual Communications and Image Processing*, 2021.

[8] Y. Mei, F. Li, L. Li, and Z. Li, "Learn a compression for objection detection-vae with a bridge," in *2021 International Conference on Visual Communications and Image Processing*, 2021.

[9] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[10] B. Li, J. Liang, H. Fu, and J. Han, "Roi-based deep image compression with swin transformers," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[11] K. Iino, S. Akamatsu, H. Watanabe, S. Enomoto, A. Sakamoto, and T. Eda, "Improving image coding for machines through optimizing encoder via auxiliary loss," in *2024 IEEE International Conference on Image Processing*, 2024.

[12] J. Liu, Y. Wei, J. Lin, S. Zhao, H. Sun, Z. Chen, W. Zeng, and X. Jin, "Semantics disentanglement and composition for versatile codec toward both human-eye perception and machine vision task," *arXiv preprint arXiv:2412.18158*, 2024.

[13] ——, "Tell codec what worth compressing: Semantically disentangled image coding for machine with lmms," in *2024 IEEE International Conference on Visual Communications and Image Processing*, 2024.

[14] T. Shindo, K. Yamada, T. Watanabe, and H. Watanabe, "Image coding for machines with edge information learning using segment anything," in *2024 IEEE International Conference on Image Processing*, 2024.

[15] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[16] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," in *2021 IEEE international conference on robotics and automation*, 2021.

[17] B. Yin, X. Zhang, Z.-Y. Li, L. Liu, M.-M. Cheng, and Q. Hou, "Dformer: Rethinking rgbd representation learning for semantic segmentation," in *ICLR*, 2024.

[18] W. Zhang, Y. Jiang, K. Fu, and Q. Zhao, "Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection," in *2021 IEEE International Conference on Multimedia and Expo*, 2021.

[19] P. Dong and Q. Chen, *LiDAR remote sensing and applications*. CRC Press, 2017.

[20] "Apple unveils new ipad pro with breakthrough lidar scanner and brings trackpad support to ipados," https://www.apple.com/newsroom/2020/03/apple-unveils-new-ipad-pro-with-lidar-scanner-and-trackpad-support-in-ipados/, [Online; accessed 07-November-2023].

[21] A. Gnutti, S. Della Fiore, M. Savardi, Y.-H. Chen, R. Leonardi, and W.-H. Peng, "Lidar depth map guided image compression model," in *2024 IEEE International Conference on Image Processing*, 2024.

[22] H. Lee, M. Kim, J.-H. Kim, S. Kim, D. Oh, and J. Lee, "Neural image compression with text-guided encoding for both pixel-level and perceptual fidelity," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[23] S. Iwai, T. Miyazaki, and S. Omachi, "Semantically-guided image compression for enhanced perceptual quality at extremely low bitrates," *IEEE Access*, vol. 12, pp. 100 057–100 072, 2024.

[24] X. Zhang and X. Wu, "Attention-guided image compression by deep reconstruction of compressive sensed saliency skeleton," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 13 354–13 364.

[25] R. Feng, X. Jin, Z. Guo, R. Feng, Y. Gao, T. He, Z. Zhang, S. Sun, and Z. Chen, "Image coding for machines with omnipotent feature learning," in *European Conference on Computer Vision*, 2022.

[26] Y. Tian, G. Lu, and G. Zhai, "Free-vsc: Free semantics from visual foundation models for unsupervised video semantic compression," in *European Conference on Computer Vision*, 2024.

[27] R. Torfason, F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Towards image understanding from deep compression without decoding," *preprint arXiv:1803.06131*, 2018.

[28] L. D. Chamain, S. Qi, and Z. Ding, "End-to-end image classification and compression with variational autoencoders," *IEEE Internet of Things J.*, vol. 9, no. 21, 2022.
[29] J. Liu, H. Sun, and J. Katto, "Semantic segmentation in learned compressed domain," in *Proc. IEEE Picture Coding Symp.*, 2022.
[30] A. Alkhateeb, A. Gnutti, F. Guerrini, R. Leonardi, J. Ascenso, and F. Pereira, "Jpeg ai compressed domain face detection," in *2024 IEEE 26th International Workshop on Multimedia Signal Processing*, 2024.
[31] C.-H. Kao, C. Chien, Y.-J. Tseng, Y.-H. Chen, A. Gnutti, S.-Y. Lo, W.-H. Peng, and R. Leonardi, "Bridging compressed image latents and multimodal large language models," in *The Thirteenth International Conference on Learning Representations*, 2025.
[32] O. Stankiewicz, T. Grajek, S. Maćkowiak, J. Stankowski, S. Różek, M. Lorkiewicz, M. Wawrzyniak, and M. Domański, "Region-of-interest-based video coding for machines," in *2024 IEEE International Conference on Multimedia and Expo Workshops*, 2024.
[33] A. Alkhateeb, A. Gnutti, F. Guerrini, R. Leonardi, J. Ascenso, and F. Pereira, "Jpeg ai compressed domain face detection: a multi-scale bridging perspective," *IEEE Transactions on Multimedia*, pp. 1–10, 2025.
[34] E. Alshina, J. Ascenso, and T. Ebrahimi, "JPEG AI: The first international standard for image coding based on an end-to-end learning-based approach," *IEEE Multimedia*, vol. 31, no. 4, 2024.
[35] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.
[36] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.
[37] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.
[38] C.-H. Kao, Y.-C. Weng, Y.-H. Chen, W.-C. Chiu, and W.-H. Peng, "Transformer-based variable-rate image compression with region-of-interest control," in *2023 IEEE International Conference on Image Processing*, 2023.
[39] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017.
[43] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz, and E. Shulman, "ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, 2017.
[45] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *12th European Conference on Computer Vision*, 2012.
[46] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, "Transformer-based image compression," in *IEEE Data Compression Conference*, 2022.
[47] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics
[48] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
[49] J. Cha, W. Kang, J. Mun, and B. Roh, "Honeybee: Locality-enhanced projector for multimodal llm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
[50] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.
[51] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.
[52] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European Conference on Computer Vision*, 2016.
[53] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International conference on learning representations*, 2022.