UNIMAC: UNIFIED MULTI-MODAL AUXILIARY LEARNED IMAGE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned image compression methods typically operate solely on RGB data, limiting their ability to benefit from complementary information provided by other sensing modalities. Although recent approaches have explored auxiliary inputs such as infrared images, depth maps, point clouds, and stereo views, they generally require a dedicated architecture for each modality, making them difficult to scale and deploy across heterogeneous sensing scenarios. To address this limitation, we propose UniMAC (**Uni**fied **M**ulti-modal **A**uxiliary Learned Image **C**ompression), a unified framework that enables a single compression model to leverage diverse auxiliary modalities. We first employ modality-specific input heads to account for the distinct characteristics of heterogeneous sensor data and project them into a unified feature space. The resulting features are subsequently processed by a shared cross-sensor auxiliary feature extractor, which distills modality-agnostic representations relevant to image compression. These representations are then integrated into the primary RGB compression pipeline to improve its rate–distortion performance. Experiments across diverse cross-sensor settings demonstrate that UniMAC effectively exploits complementary auxiliary cues while avoiding the need to design or train a separate compression model for each modality.

1 INTRODUCTION

Driven by the rapid progress of the Internet-of-Things (IoT) and edge intelligence, sensors have become indispensable in numerous applications. In practice, multiple sensors are often deployed cooperatively to capture richer information than a single sensor could offer. For example, autonomous vehicles are commonly equipped with LiDARs, cameras, millimeter-wave radars, and ultrasonic sensors to collect multi-modal observations, and similar trends appear in smart manufacturing, process control, and robotics. However, the massive data streams produced by these sensors place substantial pressure on storage, bandwidth, and energy consumption.

Among these modalities, cameras are particularly important, as visual information is essential for tasks such as object detection and semantic segmentation. Owing to its high information density, image data accounts for a large share of storage and bandwidth consumption, making efficient image compression an urgent problem. Notably, although sensors in a multi-sensor system differ in modality, they observe the same underlying scene. This inherent correlation makes it feasible to exploit auxiliary information from other sensors to facilitate image compression.

Conventional codecs such as JPEG (Pennebaker & Mitchell, 1992), JPEG-2000 (Charrier et al., 1999), BPG (Bellard, 2015), and VVC (Bross et al., 2021) deliver strong performance, but their handcrafted designs restrict scalability and hinder the incorporation of cross-sensor information. Learned image compression frameworks (Ballé et al., 2018; Minnen et al., 2018; Cheng et al., 2020; He et al., 2021; 2022; Li et al., 2025) have shown remarkable capability, with several (Cheng et al., 2020; He et al., 2021; 2022; Li et al., 2025) reaching performance on par with VVC (Bross et al., 2021), and their end-to-end optimizability offers greater flexibility for fusing cross-sensor information. However, most existing learned approaches still compress data from a single source, leaving inter-sensor correlations largely unexploited.

To exploit multi-sensor information, several works (Deng et al., 2021; Wödlinger et al., 2022; Deng et al., 2023) extend learned compression to stereo by jointly compressing left and right views,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

while another line leverages auxiliary modalities: infrared images as side information (Lu et al., 2022), depth maps from smartphone point clouds (Gnutti et al., 2024), camera images to aid point cloud compression (Lin et al., 2023), and sparse point clouds to enhance image compression in autonomous driving (Jiang et al., 2024). These methods, however, are tailored to particular sensor combinations and generalize poorly, requiring the framework to be rebuilt when the setting changes. Jiang *et al.* (Jiang et al., 2026) propose a unified processing framework for diverse sensor modalities that improves generalization to some extent, but still requires separate weights per scenario without cross-scenario weight sharing.

In this paper, we propose UniMAC (**Uni**fied **M**ulti-modal **A**uxiliary Learned Image **C**ompression), a unified framework in which a single shared-weight compression model leverages diverse auxiliary modalities. A modality-specific input adaptation module first projects heterogeneous auxiliary inputs into a unified feature space while preserving their distinct characteristics, and a shared multi-scale auxiliary extraction module then distills modality-agnostic multi-scale representations. An auxiliary-conditioned compression scheme injects these representations in a residual manner into the transforms, entropy model, and reconstruction stage of a pretrained network, exploiting cross-sensor redundancy without disturbing the pretrained solution. A unified multi-modal training strategy with zero-initialized residual branches and modality-alternating batch sampling ensures stable optimization, complemented by a no-auxiliary recovery stage that restores RGB-only operation. Experiments on five multi-sensor scenarios, namely infrared images, depth maps, stereo views, segmentation maps, and sparse point clouds, show consistent rate–distortion gains over the baseline with a single shared model. Moreover, we further conduct pairwise experiments in which only two auxiliary modalities are trained at a time, revealing three qualitatively distinct interaction regimes: mutually conflicting, asymmetric where one modality improves while the other degrades, and mutually beneficial. This suggests that auxiliary modalities are not uniformly compatible, and that a unified framework should absorb such heterogeneity rather than assume it away. In summary, our contributions are as follows:

- We propose UniMAC, a unified multi-modal auxiliary learned image compression framework that enables a single model with shared weights to exploit diverse auxiliary modalities, eliminating the need for separate codecs per scenario.
- We design four key stages, namely modality-specific input adaptation, shared multi-scale auxiliary extraction, auxiliary-conditioned compression, and a unified multi-modal training strategy, to fully harness cross-sensor information while ensuring stable optimization.
- Extensive experiments across five multi-sensor scenarios demonstrate consistent rate-distortion improvements over the baseline, validating the effectiveness and general applicability of UniMAC.
- We further conduct a systematic pairwise analysis of inter-modality interactions under the five-modality setting, identifying mutually conflicting, asymmetric, and mutually beneficial regimes.

2 Related Work

2.1 Learned Image Compression

Learned lossy image compression methods (Ballé et al., 2017; 2018; Cheng et al., 2020; He et al., 2021; 2022; Li et al., 2025) aim to establish a rate-distortion optimization approach, with the objective of compressing images at the lowest possible bitrate $\mathcal{R}$ while maintaining an acceptable level of distortion $\mathcal{D}$. A typical learned lossy compression framework consists of three main components: transformation, quantization, and entropy estimation. Unlike classical image compression, learned image compression jointly optimizes the three modules in an end-to-end manner, rather than relying on hand-crafted designs, thereby offering greater flexibility. This flexibility enables the integration of cross-modal information to achieve performance gains.

2.2 Multi-sensor Compression

Multi-sensor compression integrates data from multiple sensors, of the same or different modalities, to improve compression performance. A representative same-modality application is stereo image

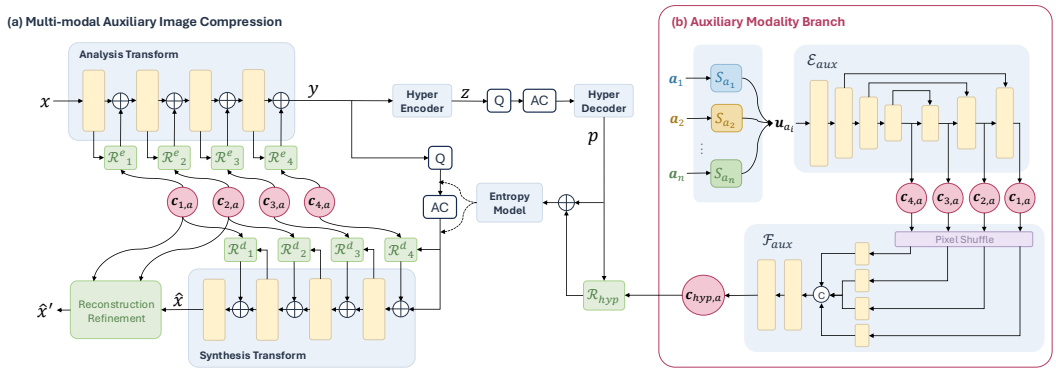


Figure 1: Overview of the UniMAC architecture. Features extracted from auxiliary modalities are integrated into the learned image compression network to enhance coding performance. Q and AC denote quantization and arithmetic coding, respectively. Yellow modules represent basic building blocks, including convolutions, residual blocks, and attention modules.

compression, which jointly compresses image pairs captured from different viewpoints; beyond extensively studied traditional methods (Bezzine et al., 2018; Kadaikar et al., 2018), learning-based approaches (Wödlinger et al., 2022; Deng et al., 2023; Wödlinger et al., 2024) are also advancing rapidly. Cross-modality compression covers even more diverse scenarios: handcrafted methods have explored natural images with medical images (Brahimi et al., 2017) and multi-view videos with depth maps (Varadarajan et al., 2012; Chen et al., 2020), while learning-based methods have been investigated for infrared images (Lu et al., 2022) and point clouds (Lin et al., 2023; Gnutti et al., 2024; Jiang et al., 2024). Jiang et al. (2026) employs a unified model architecture across diverse multi-modal assisted image compression scenarios, yet still requires scenario-specific weight training. None of these methods can handle multiple scenarios within a single framework using a shared set of weights.

3 Proposed Method

3.1 Overview

Let $\boldsymbol{x}$ and $\boldsymbol{a}$ denote two modalities of the same physical scene, where $\boldsymbol{x}$ is a natural image and $\boldsymbol{a}$ is an auxiliary modality independently transmitted and thus perfectly known at both the encoder and decoder. Our goal is to leverage $\boldsymbol{a}$ to facilitate the compression of $\boldsymbol{x}$. Since $\boldsymbol{a}$ may belong to diverse modalities, we aim for a unified model that adapts to various auxiliary inputs without modality-specific architectural redesigns:

$$\hat{\boldsymbol{x}} = \mathcal{C}(\boldsymbol{x}|\boldsymbol{a}), \quad \boldsymbol{a} \in \mathbb{A} \tag{1}$$

where $\hat{\boldsymbol{x}}$ is the reconstructed image, $\mathcal{C}(\cdot)$ is the compression network, and $\mathbb{A}$ denotes a collection of auxiliary modalities such as infrared images, depth maps, stereo images, semantic segmentation maps, and sparse point clouds.

As illustrated in Figure 1, we augment a pretrained image compression network with a unified auxiliary modality branch. Modality-specific adaptation first projects heterogeneous auxiliary inputs into a unified latent space, enabling subsequent processing by shared components. A shared feature extractor then yields multi-scale features tailored for end-to-end image compression, which are injected into the primary backbone to fuse with intrinsic image features and enhance representation learning. A dedicated training strategy further stabilizes the optimization of the proposed modules. This design allows a single model to accommodate heterogeneous sensing configurations without maintaining separate codecs per sensor.

3.2 Modality-specific Input Adaptation

Directly sharing an encoder across raw sensor measurements is ineffective because their channel layouts, value distributions, and sparsity patterns differ substantially. For modality $\boldsymbol{a}$, we therefore

introduce a stem $\mathcal{S}_a$ that maps its preprocessed input $\boldsymbol{a}$ to a common space:

$$\boldsymbol{u}_a = \mathcal{S}_a(\boldsymbol{a}) \tag{2}$$

Each stem consists of a modality-specific input projection followed by lightweight gated residual blocks. Its input projection is specialized as follows:

Infrared Images. We concatenate the normalized infrared intensity with its gradient magnitude, explicitly preserving both appearance and structural boundaries. Instance normalization reduces sensor- and environment-dependent intensity shifts before three convolutional layers map the two-channel input to the shared space.

Depth Maps. We remove invalid and extreme measurements using the valid range between the 1st and 99th percentiles. The resulting depth, inverse depth, validity mask, and gradient magnitude form a four-channel geometric representation. Unlike the infrared stem, the depth stem does not apply instance normalization, thereby preserving metric relations; dilated convolution enlarges its geometric receptive field.

Stereo Images. The auxiliary RGB view is normalized and processed using both standard and horizontally elongated convolutions. The latter reflects the predominantly horizontal displacement induced by rectified stereo cameras. To capture dataset-level crop and calibration bias, we further apply a learnable global translation to the stem feature. For stereo data, its initial horizontal displacement is selected by a global search on the training set and subsequently optimized end to end.

Semantic Segmentation Maps. For a discrete label map, a learnable embedding converts every class index into a C_0-dimensional vector. We derive a binary boundary map from changes between horizontally and vertically adjacent labels, project it with a convolution, and add it to the label embedding. The resulting representation retains both region semantics and object boundaries.

Sparse point clouds. Following Jiang et al. (2024), we project the point cloud onto the image plane and represent it using projected values together with validity and density cues. The modality-specific stem employs confidence-aware sparse-to-dense feature extraction, preserving reliable measurements while conservatively propagating geometric context to unobserved regions. Detailed preprocessing and architectural specifications are provided in the supplementary material.

To retain modality identity after entering the shared pathway, we use a modality embedding to predict feature-wise affine parameters $(\boldsymbol{\gamma}_a, \boldsymbol{\beta}_a)$ and compute:

$$\bar{\boldsymbol{u}}_a = (1 + \boldsymbol{\gamma}_a) \odot \boldsymbol{u}_a + \boldsymbol{\beta}_a \tag{3}$$

Both parameters are initialized to zero, making this operation an identity mapping at initialization. When enabled, a bounded learnable global translation is then applied to $\bar{\boldsymbol{u}}_a$ by bilinear grid sampling. It provides coarse sensor alignment before shared feature extraction, while the fusion modules described below handle residual local misalignment.

3.3 Shared Multi-scale Auxiliary Extraction

The adapted inputs is processed by a fully shared extraction module $\mathcal{E}_{\text{aux}}$, which is a U-Net-like architecture. Its downsample-path progressively reduces the spatial resolution, and its upsample-path combines upsampled representations with same-scale features from the downsample-path through skip connections. We use the three upsampled outputs and the bottleneck as the auxiliary feature pyramid:

$$(\boldsymbol{c}_{1,a}, \boldsymbol{c}_{2,a}, \boldsymbol{c}_{3,a}, \boldsymbol{c}_{4,a}) = \mathcal{E}_{\text{aux}}(\bar{\boldsymbol{u}}_a) \tag{4}$$

where $\boldsymbol{c}_{i,a}$ denotes extracted features with different resolutions. Using upsample-path features at the higher resolutions allows each output to combine local details with the large-receptive-field context accumulated at the bottleneck.

The entropy model requires a single feature with lower resolution. We first reduce the channel dimension of each $\boldsymbol{c}_{i,a}$ using a point-wise projection, align all scales using pixel unshuffle, and concatenate them. A depth-wise convolution followed by a point-wise projection produces:

$$\boldsymbol{c}_{hyp,a} = \mathcal{F}_{\text{aux}}(\boldsymbol{c}_{1,a}, \boldsymbol{c}_{2,a}, \boldsymbol{c}_{3,a}, \boldsymbol{c}_{4,a}) \tag{5}$$

3.4 AUXILIARY-CONDITIONED COMPRESSION

We inject the extracted auxiliary representation into the transforms and the entropy model, and finally refines the decoded image. All injections are residual, preserving the representation learned by the pretrained codec.

Analysis and Synthesis Transforms. At the three higher-resolution stages, we concatenate a main-path feature $\boldsymbol{f}_i$ with its same-scale auxiliary feature and predict a residual correction:

$$\boldsymbol{f}_i' = \boldsymbol{f}_i + \boldsymbol{g}_i \odot \mathcal{R}_i([\boldsymbol{f}_i; \boldsymbol{c}_{i,a}]), \quad i \in \{1, 2, 3\} \tag{6}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation, $\mathcal{R}_i$ is a lightweight convolutional fusion module, and $\boldsymbol{g}_i$ is a learnable channel-wise gate. At the lowest-resolution stage, where spatial correspondence is more ambiguous, we instead perform cross-attention with the main feature as the query and auxiliary features within a horizontal window as keys and values:

$$\boldsymbol{f}_4' = \boldsymbol{f}_4 + \boldsymbol{g}_4 \odot \mathrm{Attn}(\boldsymbol{f}_4, \boldsymbol{c}_{4,a}) \tag{7}$$

The horizontal search window exploits epipolar geometry for stereo inputs and acts as local cross-modal aggregation for already registered modalities. Symmetric fusion blocks are inserted into both the analysis and synthesis transforms.

Entropy-model Conditioning. Let $\boldsymbol{p} = h_s(\hat{\boldsymbol{z}})$ be the distribution parameters produced by the hyperprior decoder. We condition these parameters on the fused auxiliary feature using another residual fusion block:

$$\boldsymbol{p}' = \boldsymbol{p} + \boldsymbol{g}_{hyp} \odot \mathcal{R}_{hyp}([\boldsymbol{p}; \boldsymbol{c}_{hyp,a}]) \tag{8}$$

The corrected parameters $\boldsymbol{p}'$, together with the entropy model of the pretrained image codec, determine the distributions used to encode and decode $\hat{\boldsymbol{y}}$. The encoder and decoder independently recompute the same auxiliary features from $\boldsymbol{a}$; therefore, the bitstream still contains only the strings for $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$.

Auxiliary-guided Reconstruction Refinement. After the synthesis transform produces an initial reconstruction $\hat{\boldsymbol{x}}$, a two-scale U-Net predicts a residual image using $\boldsymbol{c}_{1,a}$ and $\boldsymbol{c}_{2,a}$:

$$\hat{\boldsymbol{x}}^{'} = \hat{\boldsymbol{x}} + \mathcal{N}_{\mathrm{ref}}(\hat{\boldsymbol{x}}, \boldsymbol{c}_{1,a}, \boldsymbol{c}_{2,a}) \tag{9}$$

Before each fusion, a dense offset field is predicted from the reconstruction and auxiliary features, and bilinear warping aligns the auxiliary feature to the RGB coordinate system. Multi-head local cross-attention then compensates for the remaining alignment error. This coarse-to-fine combination of global translation, dense warping, and local attention makes the framework applicable to both registered modalities and stereo pairs with non-negligible disparity.

3.5 TRAINING STRATEGY

Initialization and Optimization. We adopt HPCM (Li et al., 2025) as our baseline model and initialize it with the officially released pretrained weights. Newly introduced residual paths are constructed to preserve the pretrained solution: the last layer (or residual gate) of each fusion block and the output layer of the refinement head are initialized to zero, while the modality-wise affine embedding is initialized as the identity. Hence, before optimization, the auxiliary corrections vanish and the proposed method reproduces the pretrained codec. We organize pretrained backbone parameters and newly introduced auxiliary parameters into separate optimizer groups, allowing their learning rates to be controlled independently.

Rate-distortion Objective. We adopt the standard rate-distortion loss as our optimization objective, which can be formulated as follows:

$$\mathcal{L} = \underbrace{\mathbb{E}[-\log_2 p_{\hat{\boldsymbol{y}}}(\hat{\boldsymbol{y}})] + \mathbb{E}[-\log_2 p_{\hat{\boldsymbol{z}}}(\hat{\boldsymbol{z}})]}_{R(\hat{\boldsymbol{y}})+R(\hat{\boldsymbol{z}})} + \lambda \underbrace{\mathrm{MSE}(\boldsymbol{x}, \hat{\boldsymbol{x}}^{'})}_{D(\boldsymbol{x},\hat{\boldsymbol{x}}')} \tag{10}$$

Table 1: BD-Rate (%) relative to HPCM across five modalities. Negative values indicate bitrate savings. Δ denotes the BD-Rate difference with respect to UniMAC-Zeros (negative is better).

Method	Param (M)	PSNR BD-Rate (%)					Shared Weight	Codec w/o Aux
		Stereo.	Depth.	Infra.	Point.	Seg.		
HPCM	68.5	0.00	0.00	0.00	0.00	0.00	✓	✓
SALIC	87.4	-	-	-	-56.64	-	✗	✗
CALIC	93.9	-	-13.68	-15.92	-58.56	-	✗	✗
UniMAC-Zeros	76.6	-22.21	-8.23	-6.73	-50.53	-8.88	✓	✓
UniMAC-Shuffle	76.6	-21.34	-8.56	-6.64	-50.06	-8.97	✓	✓
UniMAC (w/o Aux)	76.6	-3.84	6.57	4.10	-36.81	-3.08	✓	✓
UniMAC	76.6	**-29.45**	**-10.33**	**-7.78**	**-51.97**	**-11.15**	✓	✓
Δ vs. Zeros		-7.24	-2.10	-1.05	-1.44	-2.27		

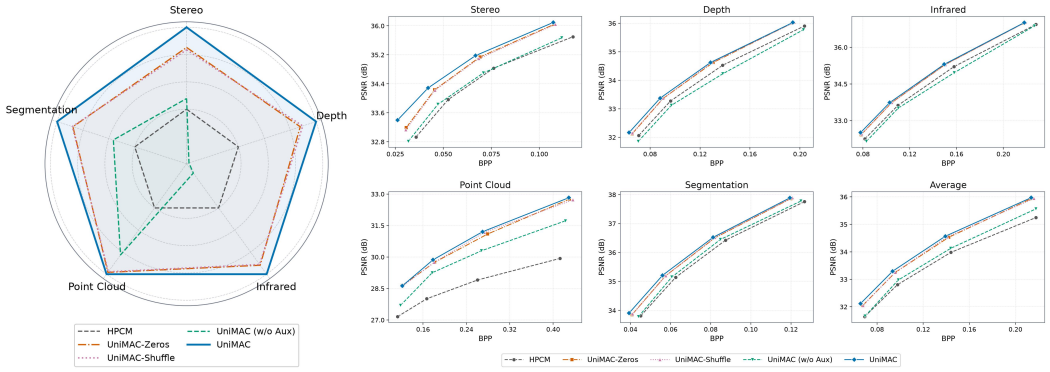


Figure 2: Rate-distortion performance across five modalities. **Left:** BD-Rate savings over HPCM (larger area is better). **Right:** RD curves on five modalities and their average. UniMAC consistently outperforms all variants.

Unified Multi-modal Training and No-auxiliary Recovery. Each mini-batch contains samples from one modality, while the batch sampler alternates among modalities during training. The shared modules thus receive gradients from every sensor type without mixing incompatible tensor layouts within a batch. After multi-modal training, we additionally construct a no-auxiliary branch by feeding a three-channel zero tensor through a dedicated stem. All previously learned parameters are frozen; only this new stem and its corresponding rows in the modality embedding and global-shift table are optimized. This procedure recovers RGB-only operation without modifying the behavior of any learned auxiliary branch.

4 Experimental Results

4.1 Experiment Settings

We train and evaluate our method on five distinct multi-sensor scenarios, with datasets drawn from M3FD (infrared images) (Liu et al., 2022), SUNRGBD (depth images) (Song et al., 2015), InStereo2K (stereo images) (Bao et al., 2020), CityScapes (semantic segmentation maps) (Cordts et al., 2016), and KITTI (sparse point clouds) (Geiger et al., 2013). To balance the contribution of each scenario in the training set, we take the smallest dataset as the reference and randomly sample an equal number of instances from the remaining datasets at each training iteration, thereby ensuring that all scenarios maintain approximately equal proportions within every training epoch. We train our model for a total of 100 epochs. The initial learning rate is set to 10^{-4} for the pretrained compression backbone, 10^{-2} for the modality-specific stem modules, and 10^{-3} for the shared fusion components. All other training configurations remain consistent with those reported in the original

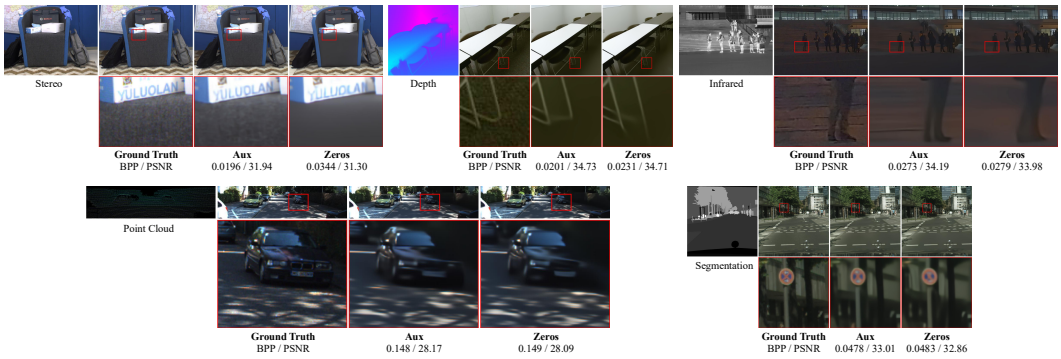


Figure 3: Comparison of reconstructed images, where "Aux" denotes the reconstruction of UniMAC with cross-modal auxiliary information, and "Zeros" denotes the reconstruction of UniMAC-Zeros with all-zero input. The metric is [BPP↓ / PSNR↑].

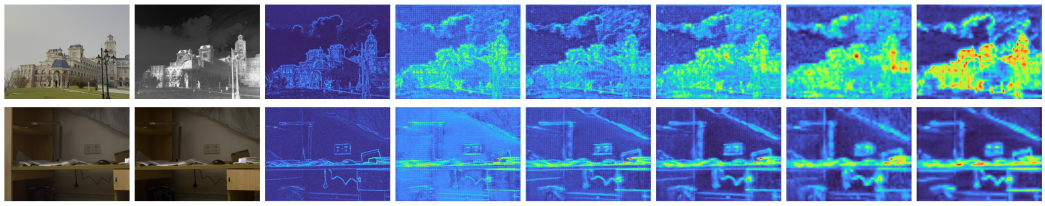


Figure 4: Feature visualization. **Top:** Infrared image. **Bottom:** stereo image. From left to right: (a) original image $\boldsymbol{x}$, (b) auxiliary input $\boldsymbol{a}$, (c) output after modality-specific input adaptation $\bar{\boldsymbol{u}}_a$, and (d–h) extracted multi-scale features $\boldsymbol{c}_{1,a}, \boldsymbol{c}_{2,a}, \boldsymbol{c}_{3,a}, \boldsymbol{c}_{4,a}, \boldsymbol{c}_{hyp,a}$.

HPCM paper. For SALIC (Jiang et al., 2024) and CALIC (Jiang et al., 2026), we implement the models following their original setups and train them for the same total number of epochs as used in our training.

4.2 Rate-distortion Performance

Performance Gains from Cross-modal Auxiliary Information. We use the officially pretrained HPCM as our anchor to evaluate three primary experimental settings: (1) training with correctly paired auxiliary sensor inputs (UniMAC), (2) training with mismatched auxiliary inputs (UniMAC-Shuffle), and (3) training with all-zero inputs (UniMAC-Zero). All three variants maintain identical model architectures, computational complexity, and training configurations. This setup allows us to directly quantify the performance gains attributable to the cross-modal auxiliary information.

As shown in Figure 2 and Table 1, UniMAC-Shuffle and UniMAC-Zero exhibit comparable performance, confirming that mismatched auxiliary information yields no measurable gain. In contrast, UniMAC achieves consistent improvements over UniMAC-Zero across all five scenarios, validating the effectiveness of correctly paired auxiliary information within our framework. The performance gains vary from 1.05% to 7.24% depending on the scenario. Generally, stronger cross-modal correlation (e.g., stereo images) yields larger gains. Notably, despite their high perceptual correlation with RGB, infrared images yield lower gains than depth or semantic maps, suggesting that perceptual correlation does not guarantee network-exploitable features. However, the proportional gains in Figure 2 (left) align with expectations: sparse point clouds contribute the least, while stereo images provide the largest share. Since UniMAC employs a single set of weights to accommodate all scenarios, its performance in specific scenarios is lower than that of SALIC and CALIC. Given its superior generalizability and smaller parameter count, this performance gap is acceptable.

Compression Performance without Auxiliary Information. To broaden the applicability of our method, we introduce a no-auxiliary recovery mode that enables pure image coding without auxiliary inputs while preserving cross-modal auxiliary performance, denoted as UniMAC(w/o Aux). As

Method	Stereo.	Depth.	Infra.	Point.	Seg.
HPCM	0.00	0.00	0.00	0.00	0.00
SALIC	-	-	-	-56.64	-
CALIC	-	-13.68	-15.92	-58.56	-
UniMAC-Zeros	-28.50	-11.89	-9.69	-53.95	-12.20
UniMAC	**-47.39**	**-14.27**	**-11.22**	**-54.67**	**-15.80**
Δ vs. Zeros	-18.89	-2.39	-1.53	-0.72	-3.60

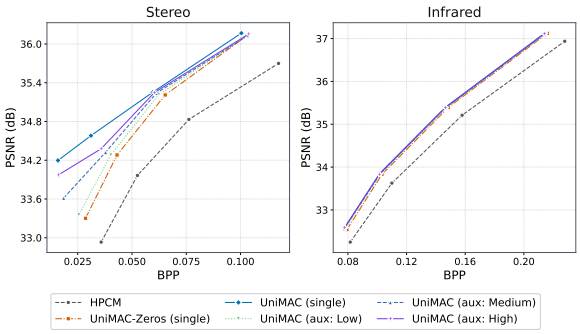


Figure 5: **Left (Table):** BD-Rate (%) under single-modal training; the last row reports the BD-Rate difference of UniMAC with respect to UniMAC-Zeros. **Right (Figure):** Rate-distortion curves of UniMAC with auxiliary information compressed at different qualities on Stereo and Infrared under single-modal training.

shown in Figure 2 and Table 1, even without cross-modal auxiliary inputs, UniMAC remains capable of image compression. Although it does not consistently outperform HPCM under individual scenarios, it achieves superior overall performance across the five scenarios, primarily benefiting from domain adaptation.

4.3 Reconstruction and Feature Visualization

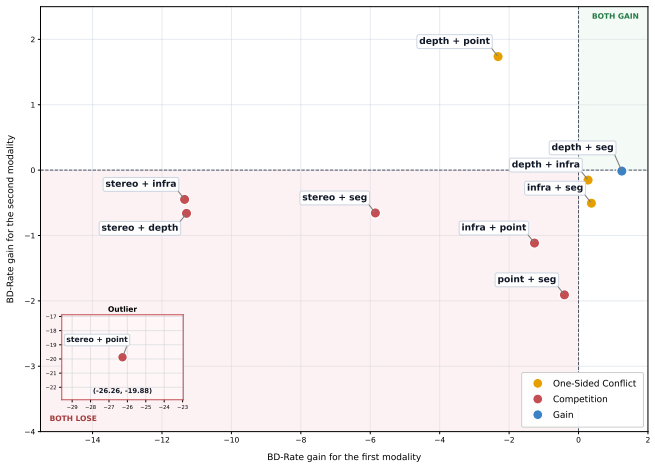
Reconstruction. Figure 3 presents a visual comparison of the reconstructed images produced by UniMAC and UniMAC-Zero. To visually demonstrate the gains from auxiliary information, we present reconstructed images at low bitrates. With auxiliary information, the model achieves higher PSNR at lower bitrates and better preserves structural details.

Feature Visualization. To further analyze the learned representations, we visualize the intermediate features of UniMAC. As shown in Figure 4, the information extracted from different auxiliary modalities is predominantly edge-structural. This observation is well justified, as cross-modal sharing is primarily structural, with edges serving as a commonality across all modalities.

4.4 Ablation Study

Optimized for a Single Scenario. UniMAC supports multi-modal auxiliary compression across five scenarios and pure image compression without auxiliary inputs, demonstrating strong generalization. To probe its performance upper bound, we further conduct single-scenario optimization. As shown in Figure 5 (left), scenario-specific training consistently improves the rate-distortion performance of both UniMAC and UniMAC-Zero over joint multi-scenario training, suggesting that distinct data distributions and modality-specific characteristics benefit from focused domain adaptation. After scenario-specific training, the performance gap between UniMAC and competing methods narrows substantially. Notably, except for sparse point clouds, UniMAC's gain over UniMAC-Zero is more pronounced under single-scenario training, indicating the improvement is not solely due to enhanced RGB-domain adaptation. The stereo setting shows the most significant improvement, with the relative gain rising from 7.24% to 18.89%. Conversely, sparse point clouds perform worse than joint training, suggesting that other modalities enhance the model's ability to handle this scenario.

Impact of Auxiliary Modality Quality on Performance. To investigate the impact of auxiliary information quality on compression performance, we vary the quality of auxiliary inputs fed into the cross-modal branch. Specifically, we focus on two information-rich modalities—infrared images and stereo images—to examine this effect. We compress these modalities using the original HPCM codec at different bitrates to obtain three quality levels: High, Medium, and Low, and then feed them into the cross-modal branch to evaluate the resulting performance. As shown in Figure 5 (right), for scenarios where RGB information plays a critical role, such as stereo images, the performance gain diminishes significantly as the quality of auxiliary information degrades. In contrast, for scenarios



Mod. 1	ΔG_1	Mod. 2	ΔG_2	Interaction
Stereo.	−11.29	Depth.	−0.66	Competition
Stereo.	−11.35	Infra.	−0.45	Competition
Stereo.	**−26.26**	Point.	**−19.88**	Competition
Stereo.	−5.85	Seg.	−0.66	Competition
Depth.	+0.28	Infra.	−0.15	One-sided
Depth.	−2.32	Point	+1.74	One-sided
Depth.	**+1.25**	Seg.	**−0.01**	Gain
Infra.	−1.27	Point.	−1.12	Competition
Infra.	+0.37	Seg.	−0.51	One-sided
Point.	−0.41	Seg.	−1.91	Competition

Figure 6: Cross-modality interactions under pairwise joint two-modality training relative to independent single-modality optimization. The left panel visualizes the per-modality BD-rate gain changes, while the right panel reports their numerical values. Positive and negative values indicate gains and losses, respectively. Blue and orange denote the best and worst overall interactions, respectively.

that share only structural similarity with the original image, such as infrared images, variations in auxiliary information quality have negligible impact on performance.

Optimized for Two Scenarios. To investigate potential conflicts across different scenarios, we select pairs of scenarios from the five available ones for joint optimization, resulting in a total of 10 experimental groups. As shown in Figure 6, joint training of two scenarios leads to performance degradation in both scenarios compared to single-scenario optimization in certain cases. The most severe degradation occurs in the combination of sparse point clouds and stereo images, where joint optimization incurs a performance drop of approximately 20% relative to single-modality optimization. This result is reasonable, as stereo images are dense and highly similar to the original image in the RGB domain, whereas point clouds are sparse and share only structural similarity with the original image. The substantial discrepancy between these two modalities accounts for the observed performance drop. The combination of depth maps and semantic segmentation maps yields the best overall performance: the depth scenario achieves a gain of 1.25%, while the semantic segmentation scenario remains nearly unchanged. For the remaining scenario pairs, we observe one-sided conflicts, where one scenario improves at the expense of the other. Notably, in the point cloud–depth map combination, the point cloud scenario gains 1.75%, whereas the depth scenario drops by 2.32%. This result is reasonable, as point clouds can be regarded as a sparse, low-quality form of depth maps.

5 Conclusion

UniMAC demonstrates that a single learned image codec can effectively exploit diverse auxiliary modalities with one shared set of weights. Without any per-scenario retraining or architectural modification, it consistently improves rate-distortion performance across five heterogeneous scenarios. Modality-specific input adaptation projects heterogeneous sensor data into a unified feature space, while a shared multi-scale extractor distills modality-agnostic representations relevant to compression. Auxiliary-conditioned residual injection into the transforms, the entropy model, and the reconstruction stage exploits cross-sensor redundancy without disturbing the pretrained solution, and a unified training strategy with zero-initialized branches and modality-alternating sampling further enables stable optimization. However, our framework still relies on a manually designed input stem for each new modality, and the auxiliary data is assumed to be perfectly available at the decoder without accounting for its own transmission cost. Future work will explore automatic adaptation to unseen modalities and the joint optimization of auxiliary information transmission.

AI USE STATEMENT

We used AI to assist with language editing, drafting and revising portions of the Introduction, Related Work, and Conclusion, literature discovery, reference formatting, and LaTeX preparation. We also used it to obtain feedback on the research framing, the interpretation of existing results, and the design of ablation studies and controlled comparisons. These uses did not include generating experimental measurements or synthetic datasets, or proving mathematical claims. The authors are responsible for verifying all AI-assisted material and for the final content of the paper.

REFERENCES

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations*, 2017.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *6th International Conference on Learning Representations*, 2018.

Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63(11): 212101, 2020.

Fabrice Bellard. Bpg image format. In *https://bellard.org/bpg*, 2015.

I Bezzine, Mounir Kaaniche, Saadi Boudjit, and Azeddine Beghdadi. Sparse optimization of non separable vector lifting scheme for stereo image coding. *Journal of Visual Communication and Image Representation*, 57:283–293, 2018.

Tahar Brahimi, Larbi Boubchir, Régis Fournier, and Amine Naït-Ali. An improved multimodal signal-image compression scheme with application to natural images and biomedical data. *Multimedia Tools and Applications*, 76(15):16783–16805, 2017.

Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. doi: 10.1109/TCSVT.2021.3101953.

Maryline Charrier, DS Cruz, and Mathias Larsson. Jpeg2000, the next millennium compression standard for still images. In *Proceedings IEEE International Conference on Multimedia Computing and Systems*, volume 1, pp. 131–132. IEEE, 1999.

Siqi Chen, Qiong Liu, and You Yang. Adaptive multi-modality residual network for compression distorted multi-view depth video enhancement. *IEEE Access*, 8:97072–97081, 2020.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939–7948, 2020.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

Xin Deng, Yufan Deng, Ren Yang, Wenzhe Yang, Radu Timofte, and Mai Xu. Masic: Deep mask stereo image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 33 (10):6026–6040, 2023. doi: 10.1109/TCSVT.2023.3253702.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Alessandro Gnutti, Stefano Della Fiore, Mattia Savardi, Yi-Hsin Chen, Riccardo Leonardi, and Wen-Hsiao Peng. Lidar depth map guided image compression model. *arXiv preprint arXiv:2401.06517*, 2024.

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

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

Yiheng Jiang, Haotian Zhang, Li Li, Dong Liu, and Zhu Li. Sparse point clouds assisted learned image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5): 4000–4010, 2024.

Yiheng Jiang, Yuqi Li, Yifan Bian, Haotian Zhang, Dong Liu, and Li Li. A unified framework for cross-sensor assisted learned image compression. *IEEE Transactions on Multimedia*, pp. 1–11, 2026. doi: 10.1109/TMM.2026.3712229.

Aysha Kadaikar, Gabriel Dauphin, and Anissa Mokraoui. Joint disparity and variable size-block optimization algorithm for stereoscopic image compression. *Signal Processing: Image Communication*, 61:1–8, 2018.

Yuqi Li, Haotian Zhang, Li Li, and Dong Liu. Learned image compression with hierarchical progressive context modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18834–18843. IEEE, 2025.

Yuhuan Lin, Tongda Xu, Ziyu Zhu, Yanghao Li, Zhe Wang, and Yan 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. IEEE, 2023.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, 2022.

Guo Lu, Tianxiong Zhong, Jing Geng, Qiang Hu, and Dong Xu. Learning based multi-modality image and video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6083–6092, 2022.

David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing*, pp. 10794–10803, 2018.

William B Pennebaker and Joan L Mitchell. *JPEG: Still image data compression standard*. Springer Science & Business Media, 1992.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Karthik Mahesh Varadarajan, Kai Zhou, and Markus Vincze. Rgb and depth intra-frame cross-compression for low bandwidth 3d video. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 955–958. IEEE, 2012.

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

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

A Training and Evaluation Details

Datasets. We use data from five scenarios as our training set, and the size of each training set varies. If the dataset has already been split into training and testing sets, we directly adopt the official protocol; otherwise, we randomly sample a portion of the original data as the test set. The detailed data scales are as follows:

- M3FD (infrared images) (Liu et al., 2022): We divide the raw dataset into three parts. 2520 images for training, 840 images for validation, and 840 images for testing.
- SUNRGBD (depth images) (Song et al., 2015): We divide the raw datasets into three parts. 6201 images for training, 2067 images for validation, and 2067 images for testing.
- InStereo2K (stereo images) (Bao et al., 2020): We follow the official protocol. 2000 images for training, and 50 images for testing.
- CityScapes (semantic segmentation maps) (Cordts et al., 2016): We follow the official protocol. 2975 images for training, and 500 images for testing.
- KITTI (sparse point clouds) (Geiger et al., 2013): We divide the raw dataset into three parts. 12674 images for training, 3114 images for validation, and 3114 images for testing.

Supplementary Training Strategies. We train our model jointly on five RGB-auxiliary scenarios, all samples are cropped into patches of size 256×512 during training. To avoid domination by large datasets, we adopt modality-alternating sampling. For each epoch, at most 2,000 samples are drawn from each modality, and modalities are sampled with approximately equal frequency. For no-auxiliary recovery, we perform a separate recovery-stage fine-tuning starting from the jointly trained five-modality model. We introduce a dedicated "zero" stem for absent auxiliary inputs. During this stage, only the new zero-stem parameters are optimized; previously trained modality entries are protected by gradient masks and are verified to remain unchanged.

Zero-initialized Residual Adaptation. To incorporate auxiliary information without disrupting the pretrained HPCM codec, all newly introduced modules are formulated as zero-initialized residual branches. Let $\boldsymbol{x}_l$ be a feature in the pretrained codec and $\boldsymbol{F}_l$ be the auxiliary feature at the matched scale. Cross-modal conditioning is performed as

$$\widetilde{\boldsymbol{x}}_l = \boldsymbol{x}_l + \boldsymbol{g}_l \odot \Delta_l(\boldsymbol{x}_l, \boldsymbol{F}_l), \tag{11}$$

where $\Delta_l(\cdot)$ is a learnable fusion function and $\boldsymbol{g}_l$ is a learnable channel-wise gate. We initialize the residual path to output zero. For the concatenation fusion used in our main experiments, this is achieved by zero-initializing the weight and bias of the final 1×1 convolution:

$$\Delta_l(\boldsymbol{x}_l, \boldsymbol{F}_l) = \mathbf{0} \quad \text{at initialization.} \tag{12}$$

Therefore, $\widetilde{\boldsymbol{x}}_l = \boldsymbol{x}_l$ initially, regardless of the auxiliary input. The same mechanism is applied to the auxiliary injections in the analysis transform, synthesis transform, and hyperprior decoder.

Consequently, before fine-tuning, UniMAC produces the same latent representations, entropy-model parameters, and reconstructed image as the pretrained HPCM:

$$\hat{\boldsymbol{y}}_{\mathrm{UniMAC}} = \hat{\boldsymbol{y}}_{\mathrm{HPCM}}, \quad \hat{\boldsymbol{z}}_{\mathrm{UniMAC}} = \hat{\boldsymbol{z}}_{\mathrm{HPCM}}, \quad \hat{\boldsymbol{x}}_{\mathrm{UniMAC}} = \hat{\boldsymbol{x}}_{\mathrm{HPCM}}. \tag{13}$$

This initialization prevents randomly initialized auxiliary modules from degrading the pretrained codec at the start of joint optimization.

B Modality-specific Stem Modules

The details of our modality-specific stems for five modalities are as follows:

Infrared Images. For the infrared input $\boldsymbol{a}_{\mathrm{ir}} \in \mathbb{R}^{2\times H\times W}$, we first normalize each channel and then apply three 3×3 convolutions:

$$\mathcal{S}_{\mathrm{ir}}(\boldsymbol{a}_{\mathrm{ir}}) = \phi(\mathrm{Conv}_{3\times3}\left(\phi(\mathrm{Conv}_{3\times3}\left(\phi(\mathrm{Conv}_{3\times3}\left(\mathrm{IN}(\boldsymbol{a}_{\mathrm{ir}})\right)\right)\right)\right)\right), \tag{14}$$

where $\mathrm{IN}(\cdot)$ denotes instance normalization and $\phi(\cdot)$ denotes LeakyReLU.

Depth Maps. The depth input contains depth-related geometric channels, including depth, inverse depth, validity mask, and edge cues: $\boldsymbol{a}_{\text{depth}} \in \mathbb{R}^{4\times H\times W}$. To preserve its absolute geometric scale, no instance normalization is used:

$$\mathcal{S}_{\text{depth}}(\boldsymbol{a}_{\text{depth}}) = \phi(\text{Conv}_{3\times3}\left(\phi(\text{Conv}_{3\times3,d=2}\left(\phi(\text{Conv}_{3\times3}(\boldsymbol{a}_{\text{depth}}))\right)\right)\right), \tag{15}$$

where $\text{Conv}_{3\times3,d=2}$ is a dilated convolution with dilation rate 2.

Stereo Images. For the auxiliary stereo image $\boldsymbol{a}_{\text{stereo}} \in \mathbb{R}^{3\times H\times W}$, we introduce a horizontal convolution to encode the epipolar prior:

$$\mathcal{S}_{\text{stereo}}(\boldsymbol{a}_{\text{stereo}}) = \phi(\text{Conv}_{3\times3}\left(\phi(\text{Conv}_{1\times7}\left(\phi(\text{Conv}_{3\times3}\left(\text{IN}(\boldsymbol{a}_{\text{stereo}})\right)\right)\right)\right)\right). \tag{16}$$

The 1×7 convolution aggregates features along the horizontal direction, which is consistent with the epipolar geometry of rectified stereo pairs.

Semantic Segmentation Maps. Let $\boldsymbol{a}_{\text{sem}} \in \{0,\dots,Q-1\}^{H\times W}$ be a semantic label map with Q classes. We convert discrete labels into continuous features using a learnable embedding table $\boldsymbol{E} \in \mathbb{R}^{Q\times C}$, and add a boundary-aware feature:

$$\boldsymbol{e}_{\text{sem}}(i,j) = \boldsymbol{E}[\boldsymbol{a}_{\text{sem}}(i,j)], \tag{17}$$

$$\boldsymbol{b}_{\text{sem}}(i,j) = \mathbb{I}[\boldsymbol{a}_{\text{sem}}(i,j) \neq \boldsymbol{a}_{\text{sem}}(i,j-1)] \vee \mathbb{I}[\boldsymbol{a}_{\text{sem}}(i,j) \neq \boldsymbol{a}_{\text{sem}}(i-1,j)], \tag{18}$$

where invalid neighbors at image boundaries are ignored. The semantic projection is then formulated as

$$\mathcal{S}_{\text{sem}}(\boldsymbol{a}_{\text{sem}}) = \phi(\text{Conv}_{3\times3}\left(\phi(\text{Conv}_{3\times3}\left(\boldsymbol{e}_{\text{sem}} + \phi(\text{Conv}_{3\times3}(\boldsymbol{b}_{\text{sem}}))\right)\right)\right), \tag{19}$$

Sparse Point Clouds. We represent the projected sparse point cloud as $\boldsymbol{a}_{\text{lidar}} = [\boldsymbol{d}, \boldsymbol{m}, \boldsymbol{\rho}] \in \mathbb{R}^{3\times H\times W}$, where $\boldsymbol{d}$ is the sparse depth/value map, $\boldsymbol{m}$ is the validity mask, and $\boldsymbol{\rho}$ is the local sampling density. We first extract sparse geometric features and a completion prior:

$$\boldsymbol{f}_{\text{sp}} = \mathcal{F}_{\text{sp}}(\boldsymbol{a}_{\text{lidar}}), \qquad \boldsymbol{f}_{\text{comp}} = \mathcal{F}_{\text{comp}}(\boldsymbol{f}_{\text{sp}}), \tag{20}$$

where $\mathcal{F}_{\text{sp}}$ contains two 3×3 convolutional layers with GroupNorm and LeakyReLU, while $\mathcal{F}_{\text{comp}}$ contains dilated 3×3 convolutions with dilation rates 2 and 4. A confidence map is predicted from the validity mask and local density:

$$\boldsymbol{c}_{\text{lidar}} = \sigma(\mathcal{F}_{\text{conf}}\left([\boldsymbol{m}, \boldsymbol{\rho}]\right)), \tag{21}$$

and the final projection is

$$\mathcal{S}_{\text{lidar}}(\boldsymbol{a}_{\text{lidar}}) = \boldsymbol{c}_{\text{lidar}} \odot \boldsymbol{f}_{\text{sp}} + (1 - \boldsymbol{c}_{\text{lidar}}) \odot 0.25\, \boldsymbol{f}_{\text{comp}}. \tag{22}$$

C ADDITIONAL RECONSTRUCTED IMAGES

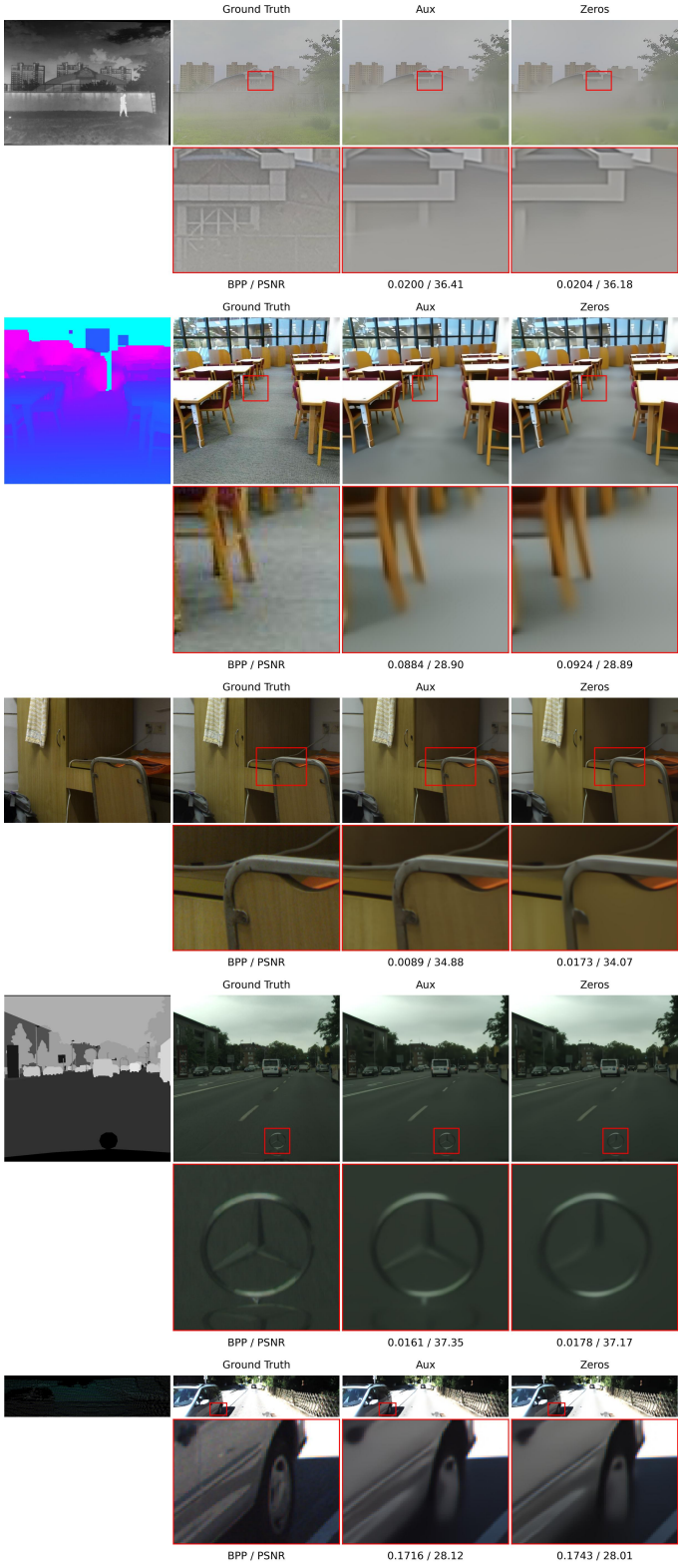


Figure 7: More reconstructed images.

D ADDITIONAL FEATURE VISULIZATION

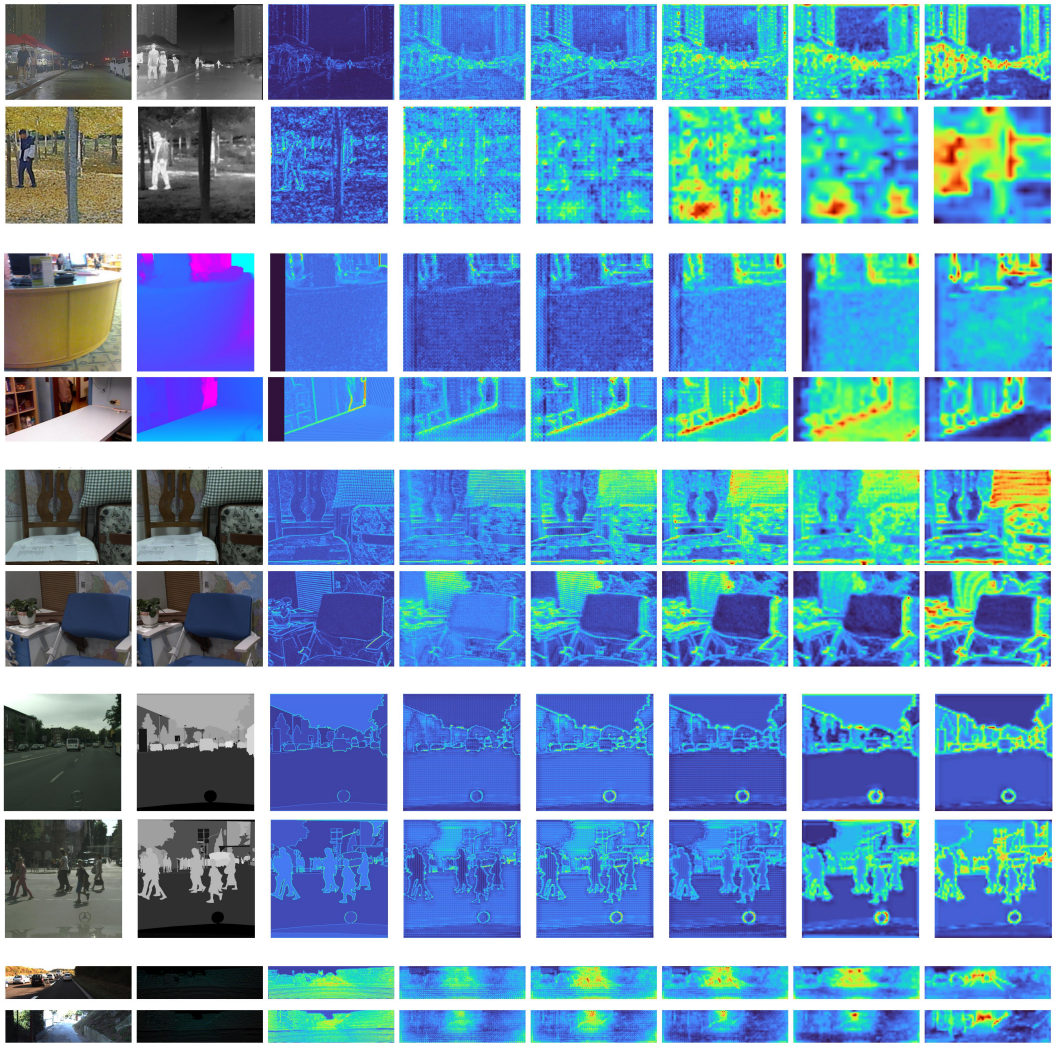


Figure 8: More feature visualizations.