PM-INR: Prior-Rich Multi-Modal Implicit Large-Scale Scene Neural Representation

Yiying Yang[1], **Fukun Yin**[2], **Wen Liu**[3],
Jiayuan Fan[1*], **Xin Chen**[3], **Gang Yu**[3], **Tao Chen**[2]
[1] Academy for Engineering and Technology, Fudan University
[2] School of Information Science and Technology, Fudan University
[3] Tencent PCG
{yiyingyang23, fkyin21}@m.fudan.edu.cn, liuwen@shanghaitech.edu.cn,
jyfan@fudan.edu.cn, chenxin2@shanghaitech.edu.cn, iskicy@gmail.com, eetchen@fudan.edu.cn

Abstract

Recent advancements in implicit neural representations have contributed to high-fidelity surface reconstruction and photo-realistic novel view synthesis. However, with the expansion of the scene scale, such as block or city level, existing methods will encounter challenges because traditional sampling cannot cope with the cubically growing sampling space. To alleviate the dependence on filling the sampling space, we explore using multi-modal priors to assist individual points to obtain more global semantic information and propose a prior-rich multi-modal implicit neural representation network, **PM-INR**, for the outdoor unbounded large-scale scene. The core of our method is multi-modal prior extraction and cross-modal prior fusion modules. The former encodes codebooks from different modality inputs and extracts valuable priors, while the latter fuses priors to maintain view consistency and preserve unique features among multi-modal priors. Finally, feature-rich cross-modal priors are injected into the sampling regions to allow each region to perceive global information without filling the sampling space. Extensive experiments have demonstrated the effectiveness and robustness of our method for outdoor unbounded large-scale scene novel view synthesis, which outperforms state-of-the-art methods in terms of PSNR, SSIM, and LPIPS.

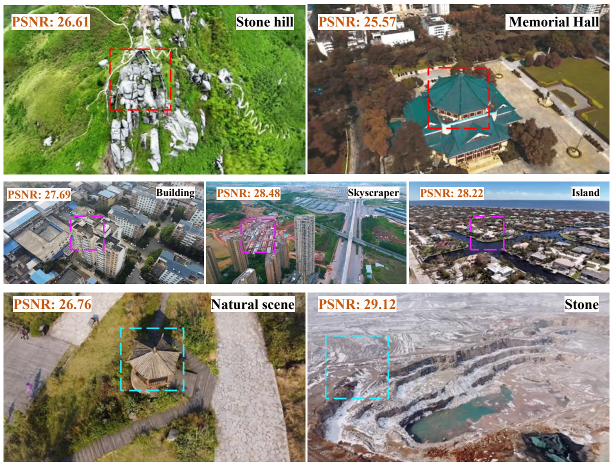

Figure 1: PM-INR is capable of handling outdoor unbounded large-scale scenes, and we have demonstrated this capability through experiments with scenes from the OMMO dataset (Lu et al. 2023). Impressively, PM-INR shows superior performance in various large scenes such as stone hills, memorials, buildings, etc.

1 Introduction

Implicit neural representations have shown promising performance in surface reconstruction and novel view synthesis for single objects or object-centric small-scale scenes under sparse or limited posed camera images (Barron et al. 2021; Martin-Brualla et al. 2021; Park et al. 2021) and have been widely applied in the field of virtual reality and augmented reality. However, the difficulty exacerbates at the cubic level when the sampling space increases from a small-scale scenario or object to an outdoor unbounded large-scale scene. The core of the problem is that the existing implicit neural representation networks only model the scene by sampling points according to the ray direction from the entire scene space. Hence, Neural Radiance Fields (NeRF) methods designed for small-scale scenes (Mildenhall et al. 2021; Verbin et al. 2022) are challenging to fill the sampling space for large-scale scenes and will synthesize rough geometry and blurry images.

Fortunately, Some methods have also noticed this problem and sample small regions rather than individual sample points to alleviate exploding sampling spaces to some extent (Barron et al. 2021, 2022; Ding et al. 2023). Moreover, compared to sampling individual points, sampling a small region allows a region of space to be compactly featured, which can help improve NeRF's (Mildenhall et al. 2021) ability to represent fine details. While for the block or city-level scenes, view synthesis quality degrades as the camera is moved far from the center of the scene. Meanwhile, inspired by codebook-assisted vision tasks, which learn a representative codebook to denote valuable prototypes and have been applied to image segmentation (You et al. 2022; Rahebi 2022; Zhou et al. 2022; Yin and Zhou 2020; Ye et al. 2022a; Yin et al. 2023b; Wu et al. 2023), image synthesis (Esser, Rombach, and Ommer 2021; Zhang et al. 2021;

[*] Corresponding author.

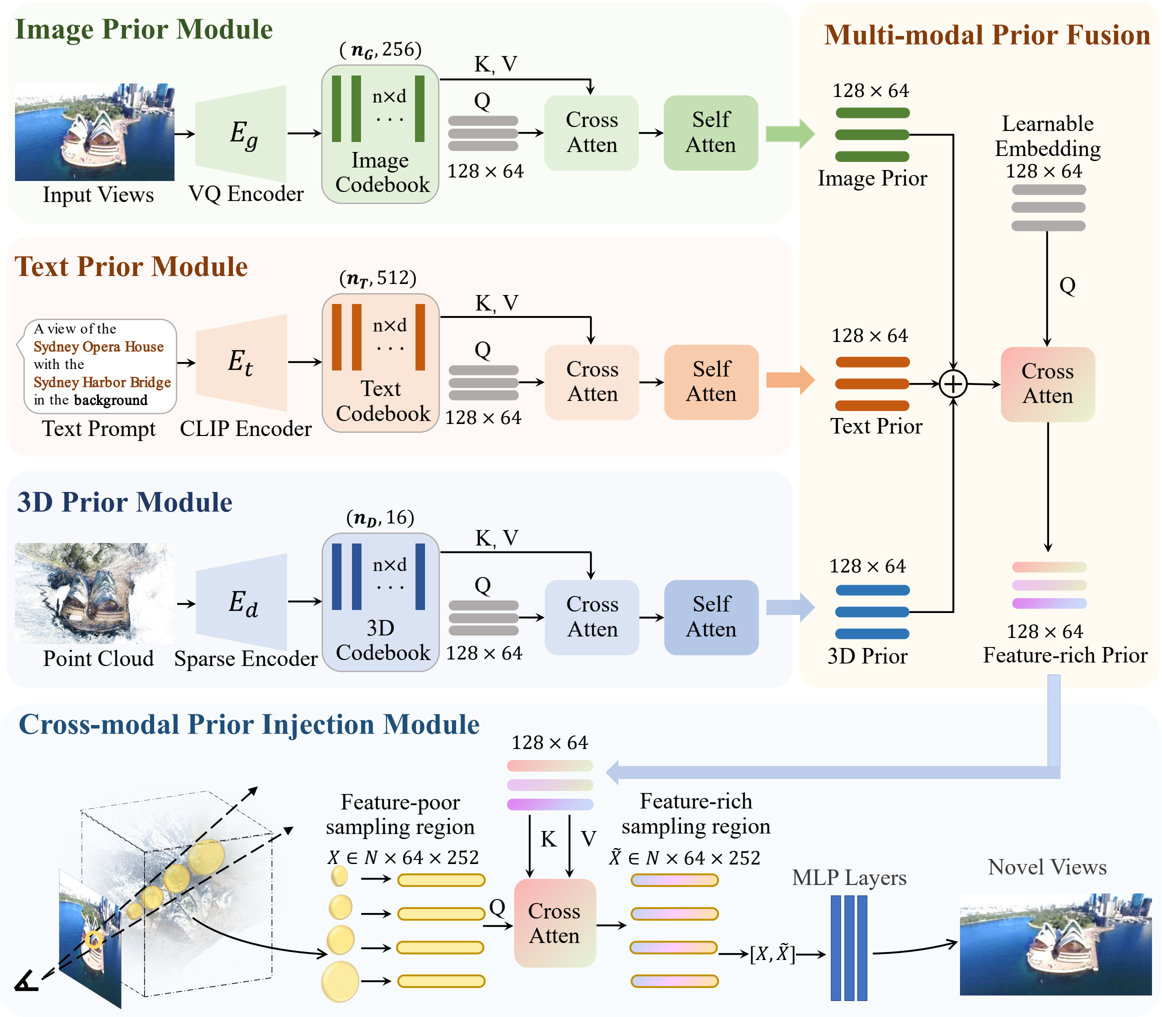

Figure 2: Multi-modal prior extraction and fusion module. Multi-modal priors are extracted from three parallel modules benefiting from valuable codebooks obtained from pre-trained models, as shown on the left. Feature-rich priors are fused to ensure cross-modal scene consistency and preserve each modality-specific feature, as shown on the right.

Esser et al. 2021; Yin et al. 2023a), and small-scale scene implicit representation (Yin et al. 2022; Shen, Ma, and Wang 2022; Yang et al. 2023; Ye et al. 2022b), rich global prior knowledge extracted from cross-modal codebooks coupled with local sampling regions seems able to cope with outdoor large-scale scene implicit representations. For local-sampled NeRF, prior knowledge can offer valuable global insights, which is lacking in ray-sampling based networks and is necessary for scene understanding and reconstruction. Therefore, equipping each region with rich priors is an unreached and promising approach for implicit large-scale scene neural representation.

In this paper, we propose **PM-INR:** A **P**rior-rich **m**ulti-modal **I**mplicit **N**eural **R**epresentation that aims to extract and fuse prior knowledge across multiple modalities to facilitate the implicit neural representation of large-scale scenes. To achieve this, we first extract various priors from codebooks obtained from different modal inputs, including, **Image Prior**, extracted from the image codebook encoded by Vector Quantised-Variational AutoEncoder (VQ-VAE) (Van Den Oord, Vinyals et al. 2017), contains valuable global semantic and appearance information for each scene and unique surface texture patterns for each training view; **Prompt Prior**, with the help of the pre-trained Contrastive Language-Image Pre-Training (CLIP) (Radford et al. 2021) model, the text prompts of each training view are converted into a more accessible format to form a text codebook, and then extract prototypes rich in scene layout and positional relationships, which are high-level properties that are not easy to see just from visualizing data; **Geometry (3D) Prior**, benefiting from Multi-view Stereo (MVS) (Seitz et al. 2006) methods and pre-trained MinkowskiEngine (Choy, Gwak, and Savarese 2019) convolutions, we encode the geometric codebook from reconstructed sparse point clouds and then filter out geometric priors with scene structure and topology properties. To reduce the distance between different modalities while maintaining cross-modal scene consistency, we propose a multi-modal prior fusion module, which fuses priors from different modalities into a feature-rich cross-modal prior. Some of the fused prior prototypes are shared by all modalities while others are unique, where the former can maintain scene consistency across modalities, and the latter provides additional information to enhance feature representation from different modalities. Finally, cross-modal priors are injected into sampling regions to perceive global semantic information and cope with the exploding sampling space.

Extensive experiments show the effectiveness of multi-modal prior in large-scale implicit neural representation, which outperforms state-of-the-art method Mip-NeRF 360 (Barron et al. 2022) by more than 17% on each evaluation metric. We summarize the contributions as follows:

- We propose an effective implicit neural representation pipeline to cope with the cubically growing sampling space of outdoor unbounded large-scale scenes by extracting rich priors from multi-modal inputs and equipping sampling regions;
- A multi-modal prior fusion module is proposed to ensure scene cross-modal consistency while enriching regional feature representations;
- Extensive experiments demonstrate that our PM-INR outperforms state-of-the-art methods, including robustness to large-scale outdoor scene representation and the capability to synthesize more photo-realistic novel views. Our code and models will be available.

2 Related Work

Large-scale Neural Scene Representation. Large-scale scene representation is a crucial aspect of Implicit Neural Representation (INR) research, involving capturing and modeling complex scenes that encompass extensive spatial extents, such as urban environments, landscapes, or virtual worlds. NeRF++ (Zhang et al. 2020) handles the unbounded scenes by separately modeling foreground and background, and Mip-NeRF 360 (Barron et al. 2022) uses a non-linear scene parameterization to model large-scale unbounded scenes. Block-NeRF (Tancik et al. 2022) and Mega-NeRF (Turki, Ramanan, and Satyanarayanan 2022) decompose a scene into several partitions spatially and train model for each partition in parallel. BungeeNeRF (Xiangli et al. 2022) introduces an approach that progressively adds residual blocks to the network representation. Mega-NeRF (Turki, Ramanan, and Satyanarayanan 2022) and BungeeNeRF (Xiangli et al. 2022) address the challenges of modeling and rendering large-scale scenes, spanning from buildings to multiple city blocks and utilizing thousands of images captured from drones. However, when the camera is posed far away from the scene's center, such as in the urban environment, the visual synthesis quality will dramatically degrade. Consequently, existing methods face constraints in applying neural implicit reconstruction to expansive, outdoor, and unbounded scenes. To overcome these challenges, a robust and scalable solution is urgently in demand.

Prior Information in Implicit Neural Representation. Recently, the application of prior information in implicit neural representation has attracted significant attention as researchers aim to improve the performance in 3D scene reconstructions. Prior information helps the implicit neural representation to leverage domain knowledge, such as geometry, materials, lighting, and semantics. Pixel-NeRF (Yu et al. 2021) addresses the challenge of learning neural radiance fields from limited input images by incorporating prior information from a pre-trained 2D convolutional neural network(CNN). Mixture of volumetric primitives (MVP) (Lombardi et al. 2021) designs an unsupervised method for learning implicit shape representations using a MVP as prior information, enabling high-fidelity 3D reconstructions without explicit 3D supervision. DeRF (Rebain et al. 2021) designs a method to decompose a scene into multiple depth layers, each represented by its own neural radiance field. FastNeRF (Garbin et al. 2021) addresses the issue of sampling artifacts in the original NeRF model by incorporating prior information about the scene's density distribution. Point-NeRF (Xu et al. 2022) introduces surface point clouds as priors to guide point sampling and achieve scene generalization. Recently, CoCo-INR (Yin et al. 2022) learns a representative codebook from the large-scale 2D dataset ImageNet (Deng et al. 2009) with limited global features of the scene, which leads to inconsistencies between different views and rendering artifacts. These works have demonstrated the potential of incorporating prior information into implicit neural representation. However, the utility of prior knowledge in the large-scale implicit neural representation remains unexplored, especially multi-modal prior knowledge.

Inspired by the above work, our method attempts to explore the effect of multi-modal prior knowledge in the large-scale implicit neural representation for the first time, and proposes to enhance the understanding of scenes by extracting and fusing multiple modal priors.

3 Methodology

In this paper, we aim to develop a prior-rich multi-modal implicit neural representation to improve the capacity to represent the outdoor unbounded large-scale scenes. We first design a multi-modal codebook and prior extraction module to establish a codebook and extract prior knowledge from multiple modalities ($c.f.$ Sec.3.1 and Fig.2). Then we propose a multi-modal prior fusion module to incorporate heterogeneous prior knowledge and develop cross-modal prior features with rich scene-level semantics, contextual knowledge as well as the geometric properties of the scene ($c.f.$ Sec.3.2 and Fig.2). Next, we inject the feature-rich cross-modal prior into each sampling region, which will help better model the outdoor unbounded large-scale scenes ($c.f.$ Sec.3.3 and Fig.2). Finally, the implementation will be introduced ($c.f.$ Sec.3.4).

3.1 Multi-modal Codebook and Prior Extraction

Image Codebook. To make full use of the rich texture features provided by images, which are difficult to obtain in point-wise sampled implicit neural representations but necessary for synthesizing photo-realistic images, especially for large-scale scenes, we derive the image codebook from training views by Vector-Quantized Variational AutoEncoder (VQVAE) (Van Den Oord, Vinyals et al. 2017). We denote our image codebook as $B_g = \{p_0, p_1, ..., p_{N_g}\} \in \mathbb{R}^{Ng \times c_g}$, where N_g is the number of prototype vectors, c_g is the dimension of each vector, and p_i is each embedding vector. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, VQVAE leverages an encoder E_g to obtain a set of continuous feature $\hat{z} = E_g(I) \in \mathbb{R}^{h \times w \times c_g}$, where h and w are the height and width of the feature map. Then a quantization Q_g is performed onto its closest codebook entry p_i for the continuous feature map $\hat{z}$ to obtain the discrete representation z_q :

$$z_q = Q_g(\hat{z}) := \underset{p_k \in \mathcal{E}}{\operatorname{argmin}} \|\hat{z}_{ij} - p_k\|_2 , \tag{1}$$

where $\hat{z}_{ij} \in \mathbb{R}^{c_g}$. Next, the reconstructed image $\hat{I}$ is given by the decoder D_g:

$$\hat{I} = D_g(z_q) = D_g(Q_g(E_g(I))) \tag{2}$$

Since our VQVAE can be optimized by reducing the loss between the original image I and the reconstructed image $\hat{I}$:

$$\mathbb{L} = \|I - \hat{I}\|^2 + \|sg(z_q) - \hat{z}\|_2^2 + \|sg(\hat{z}) - z_q\|_2^2 \tag{3}$$

where $sg(.)$ denotes the stop-gradient operation.

Text Codebook. Text prompts contain rich human-annotated global descriptions consistent with human perceptual and visual systems, and connecting prompts and visual domains enables implicit neural representation models to capture a more comprehensive understanding of scene context. Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) is a neural network that efficiently learns visual concepts from natural language supervision. CLIP is designed to leverage large datasets of images and text pairs to train a model in a self-supervised manner, learning to break the gap between visual and textual modalities. We leverage the pre-trained CLIP encoder model E_t to produce text embeddings as our text codebook B_t given the input text prompts L.

$$B_t = E_t(L) \in \mathbb{R}^{N_c \times c_t} \tag{4}$$

where N_c is the number of text prompts and c_t is is the dimension of each text embedding.

3D Codebook. Point clouds are data structures that encapsulate multiple geometric information, including the exact coordinates of points within a three-dimensional space, which help inform and enhance the process of implicit neural representation. The Minkowski Engine (Choy, Gwak, and Savarese 2019) is an auto-differentiation library for sparse tensors, which is proposed to provide an efficient and flexible framework to represent and process point clouds, which enables the model to obtain the geometric information of the scene. We leverage the Minkowski sparse tensor to build our 3D codebook by aggregating the geometric patterns and relationships present in the point clouds. Given the original point cloud $P = \{d_0, d_1, ..., d_n\}$ where d_i represents the i-th point. We utilize the Minkowski Engine to transform this cloud into a sparse tensor representation $S = (s_1, f_1), (s_1, f_1), ..., (s_n, f_n)$, composed of tuples (s_i, f_i), where s_i and f_i denote spatial position and feature vectors, respectively. To capture and quantize unique geometric patterns within this data, we construct a 3D codebook $B_d = \{m_1, m_2, ..., m_{N_D}\} \in \mathbb{R}^{N_D \times c_d}$, constructed by encoding S, where m_i represents each item in the codebook B_d. Each feature vector f_i is then mapped to its closest entry in the codebook, ensuring a compact and efficient representation of the original geometric information.

Prior Extraction. Since the multi-modal codebook might contain a large number of redundant or unrelated prototypes for implicit neural representations, we designed a prior extraction module to query valuable prototypes for scene representation and novel view synthesis from each modality codebook. Given a pre-trained codebook B (could be B_g, B_t or B_d) and learnable query embedding vectors $q = \{q_1, q_2, ..., q_M\}$, each embedding vector q_i queries the valuable prior Z^0 information from the given codebook via a cross-attention mechanism:

$$\mathrm{Q} \leftarrow f_{\mathrm{Q}}(q), \qquad \mathrm{K} \leftarrow f_{\mathrm{K}}(B), \qquad \mathrm{V} \leftarrow f_{\mathrm{V}}(B)$$
$$\mathcal{Z}^0 \leftarrow \text{Cross-Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \text{Softmax}(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{d_k}}) \tag{5}$$

where f_Q, f_K, and f_V are the query, key, and value linear projections, respectively. Then we apply a self-attention module on the initial prior $\mathcal{Z}^0$ to improve prior feature representations further and obtain the final prior $\mathcal{Z}$:

$$\mathcal{Z} \leftarrow \text{Self-Attention}(f_q(\mathcal{Z}^0), f_k(\mathcal{Z}^0), f_v(\mathcal{Z}^0)) \tag{6}$$

where f_q, f_k, and f_v are the query, key, and value linear projections, respectively. We apply the above method to the multi-modal codebooks, image codebook B_G, text codebook B_t, and 3D codebook B_d, and obtain the corresponding priors, image prior $\mathcal{Z}_G$, text prior $\mathcal{Z}_T$, and 3D prior $\mathcal{Z}_D$, which contain respective modality-specific scene-relevant prior information.

3.2 Multi-modal Prior Fusion

The extracted priors of each modality contain rich features, some of which are shared, describing the same object from different modalities, and others are unique, providing additional supplements from their respective modalities. However, this will also bring certain hidden dangers, such as scene inconsistency that may be caused by different modal prior features when describing the same object.

Therefore, finding common ground while reserving differences and combining multi-modal priors to create a single, unified representation is crucial, which can ensure scene consistency across modals and help leverage the complementary and supplementary information present in each modality to improve the overall performance of scene reconstruction. Considering the feature distance between image prior $\mathcal{Z}_G$, text prior $\mathcal{Z}_T$, and 3D prior $\mathcal{Z}_D$, we first employ linear layers to obtain embedding priors $\hat{\mathcal{Z}_G}$, $\hat{\mathcal{Z}_T}$, and $\hat{\mathcal{Z}_D}$ respectively, and concatenate the elements along the zeroth dimension to construct a cohesive and unified prior $\hat{\mathcal{U}}$:

$$\hat{\mathcal{U}} \leftarrow \text{concat}(\hat{\mathcal{Z}_G}, \hat{\mathcal{Z}_T}, \hat{\mathcal{Z}_D})) \tag{7}$$

We then define a learnable query embedding $\mathcal{U}_q = \{q_1, q_2, ..., q_M\}$ to query scene-consistent cross-modal priors $\mathcal{U}$ by effectively capturing multimodal relations and dependencies:

$$\mathcal{U} \leftarrow \text{Cross-Attention}(f_{\mathrm{Q}}(\mathcal{U}_q), f_{\mathrm{K}}(\hat{\mathcal{U}}), f_{\mathrm{V}}(\hat{\mathcal{U}})) \quad (8)$$

where f_{Q}, f_{K}, and f_{V} are the query, key, and value linear projections, respectively. Through the aforementioned multi-modal prior fusion module, we will arrive at a series of representative cross-modal scene priors $\mathcal{U}$, which preserve consistent features from multi-modal priors to synthesize view-consistent images while retaining the unique features of each modality to enrich image details and ensure realism.

3.3 Cross-modal Feature Injection and Implicit Neural Representations

In this subsection, we inject feature-rich cross-modal priors into sampled regions for outdoor unbounded large-scale scene implicit neural representations. We follow the sampling strategy of Mip-NeRF 360 (Barron et al. 2022), which uses a non-linear scene parameterization to model large-scale unbounded scenes and samples a region with more local features in the sampling space instead of sampling a single individual point. Although sampling regions can obtain more local features, global features are still lacking, especially for outdoor large-scale scenes. Our unified prior can overcome this challenge with rich global features from different modalities.

As shown in Fig.2, we inject the feature-rich multi-modal prior $\mathcal{U}$ derived in Sec.3.2 into the sampling domain via a cross-attention module. Gradually propagating cross-modal representative prototypes to each sampled region results in rich global scene features and representative representations for each sampling region. With the rich global and representative features, our method can better understand and represent outdoor unbounded large-scale scenes and synthesize high-fidelity and more detailed novel views.

Our backbone follows Mip-NeRF 360. We use the same sampling method and loss function for a fair comparison but inject our cross-modal prior information into each sampling region and apply it to outdoor large-scale scene datasets.

3.4 Implementation Details

We train the VQ-VAE network for 20k iterations with a batch size of 16 accumulated over 21 batches, which needs about 1 day on two A100 GPUs. The dimensions of image codebook B_g, text codebook B_t, and 3D codebook B_d are 256, 512, and 16, respectively. In our multi-modal codebook and prior extraction module, the number of learnable query embeddings is 128, and each embedding has a dimension $q_i \in \mathbb{R}^{64}$. Hence, the size of all priors is 128×64.

We apply one cross-attention mechanism in the prior extraction module, followed by one self-attention block. In the multi-modal prior feature fusion module, we apply linear layers to three modalities prior to initially reduce the feature distribution gap and then concatenate the processed prior embedding into a unified multi-modal prior embedding. We apply one cross-attention to the multi-modal prior embedding to derive the feature-rich cross-modal prior embedding. In the multi-modal feature injection module, we perform one cross-attention operation in the Mip-NeRF 360's module of predicting density. All attention modules are transformer-based with a multi-head attention mechanism, Layer Normalization, Feed-Forward Network, and GELU activation.

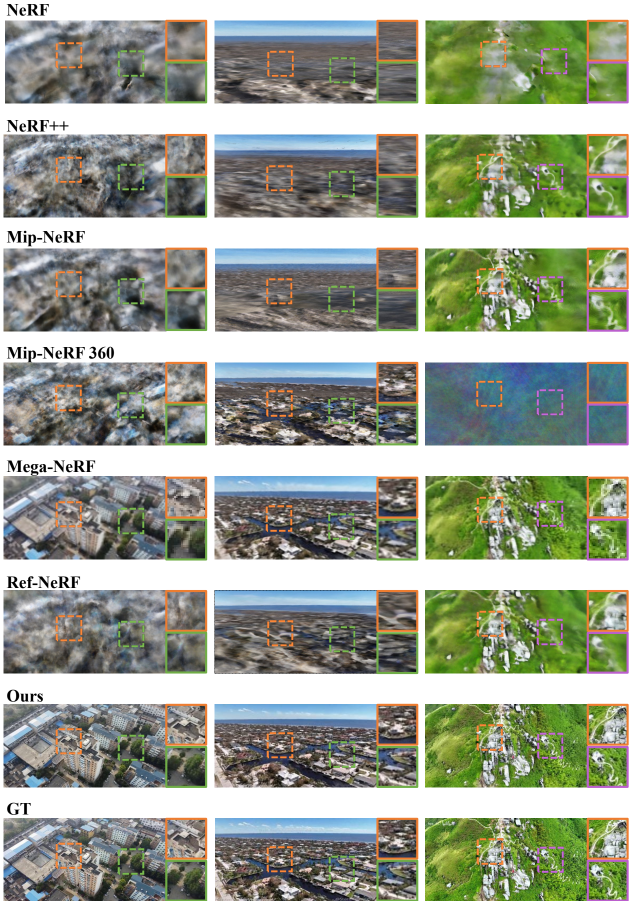

Figure 3: Qualitative results sampled from the OMMO dataset. For each scene, we present a visualization of a synthetic novel view and zoom in on two regions.

For a fair comparison, we adopt the optimizing strategies of Mip-NeRF 360, 250k iterations of optimization with a batch size of 2^{11}, using Adam (Kingma and Ba 2014) optimizer with a learning rate that is annealed log-linearly from 2×10^{-3} to 2×10^{-5} with a warm-up phase of 512 iterations, and gradient clipping to a norm of 10^{-3}. Our method is built with Pytorch framework. Each scene is trained on four Nvidia A100 GPU devices for around one day.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate our PM-INR on two datasets, namely the OMMO (Lu et al. 2023) and BlendedMVS (Yao et al. 2020) dataset. The OMMO dataset contains a total of 33 unbounded large-scale scenes with prompt annotations, tags, and 14k calibrated images. The BlendedMVS contains 17k pose images, covering 113 scenes, which are divided into the general scene part and the large scene part according to the scene scale. We conduct experiments on all scenes of the OMMO dataset and the five outdoor unbounded large-scale scenes of the BlendedMVS dataset due to the time cost constraints caused by per-scene optimization.

Baselines. We choose the recent state-of-the-art implicit large-scale scene neural representation methods, including NeRF (Mildenhall et al. 2021), NeRF++ (Zhang et al. 2020), Mip-NeRF (Barron et al. 2021), Mip-NeRF 360 (Barron et al. 2022), Mega-NeRF (Turki, Ramanan, and Satyanarayanan 2022), Ref-NeRF (Verbin et al. 2022) as the baselines. NeRF (Mildenhall et al. 2021) designs the first continuous MLP-based neural network to represent the scene, NeRF++ (Zhang et al. 2020) separately models the foreground and background neural representations to handle the unbounded scenes, Mip-NeRF (Barron et al. 2021) extends NeRF to represent the scene at a continuously-valued scale and improves NeRF's ability to represent fine details, Mip-NeRF 360 (Barron et al. 2022) uses a non-linear scene parameterization to model large-scale unbounded scenes, Mega-NeRF (Turki, Ramanan, and Satyanarayanan 2022) decomposes a scene into several spatially to train the model in parallel, Ref-NeRF (Verbin et al. 2022) improves the quality of appearance and normal in synthesized views of the scene by reparameterizing NeRF's directional MLP.

Evaluation Metrics To evaluate the performance of each method in large-scale implicit neural representation, we use three standard metrics: Peak Signal Noise Ratio (PSNR), Structural Similarity (SSIM) (Wang et al. 2004) and the VGG implementation of Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) on novel view synthesis. Higher PSNR and SSIM mean better performance, while a lower LPIPS means better.

4.2 Performance Comparison

OMMO dataset. Quantitative results on the OMMO dataset are reported in Tab. 1, which demonstrates that our method outperforms others on the average and most scenes in terms of PSNR, SSIM, and LPIPS. Among them, Mip-NeRF 360 and Mega NeRF are both aimed at unbounded scenes, and our average gain of the three evaluation metrics is 17% and 24% higher than the two of them, respectively, implying that our method is more effective for large-scale scenes. At the same time, our LPIPS, a metric that correlates more strongly with human-perceived distance, is over 43% higher than all baselines, demonstrating that our method can generate more photo-realistic novel views.

The qualitative results on the OMMO dataset are drawn in Fig.1. Our method can reconstruct finer texture in unbounded large-scale scenes, and some representative details are selected and zoomed in in Fig.1. It is worth noting that our method expresses better robustness for outdoor unbounded large-scale scene representation.

BlendedMVS dataset. To further demonstrate the performance of our method for large scale scenes, we conduct the experiments and make comparisons with Mip-NeRF 360 (Barron et al. 2022), which is aimed at unbounded scenes and overperforms other methods on the OMMO dataset. Quantitative results between Mip-NeRF 360 (Barron et al. 2022) and our method on the BlendedMVS dataset are reported in Tab. 2. We conduct experiments on five outdoor unbounded large scale scenes of the BlendedMVS dataset. Tab. 2 demonstrates that our method also outperforms others on the average and most scenes in terms of PSNR, SSIM, and LPIPS. We compare our method on the BlendedMVS dataset with the baseline Mip-NeRF 360, and our average gain of the three evaluation metrics is **30 percent** higher than Mip-NeRF 360 on the outdoor unbounded large scale scenes of the BlendedMVS dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
NeRF	18.72	0.48	0.600
NeRF++	21.45	0.58	0.538
Mip-NeRF	18.39	0.50	0.623
Mip-NeRF 360	23.10	0.67	0.419
Mega-NeRF	21.63	0.62	0.508
Ref-NeRF 360	21.28	0.55	0.574
PM-INR (Ours)	**27.10**	**0.81**	**0.239**

Table 1: Quantitative comparison results of our model PM-INR with baselines on the OMMO dataset. ↑ means the higher, the better, ↓ means the lower, the better.

Method	PSNR↑	SSIM↑	LPIPS↓
Mip-NeRF 360	23.10	0.67	0.419
PM-INR (Ours)	**27.10**	**0.81**	**0.239**

Table 2: Quantitative comparison results of our model PM-INR with baseline Mip-NeRF 360 on the five outdoor unbounded large-scale scenes of the BlendedMVS dataset. ↑ means the higher, the better, ↓ means the lower, the better.

4.3 Ablation Studies and Analysis

Effectiveness of each modal prior. To verify the effectiveness of each modal prior, we conduct controlled experiments on different modalities and their pairwise combinations, including the image prior (G), text prior (T), 3D prior (D), image prior plus text prior (G+T), image prior plus 3D prior (G+D), and text prior plus 3D prior (T+D).

The comparison result in Fig.5 shows that every modal prior helps contribute to scene reconstruction while removing any single modal prior degrades the performance across all metrics. It's also observed that without any modal prior, the performance degrades considerably further compared to the model equipped with any modal prior.

Effectiveness of multi-modal prior fusion module. To demonstrate the effectiveness of the multi-modal prior fusion module, we conduct ablation studies on different fusion strategies: removing the module of cross-attention mechanism and directly injecting the initial multi-modal prior embedding, denoted as w/o cross; injecting the multiple modalities into the sampling region serially, denoted as serial; directly adding the multiple modalities to develop the cross-modal prior, denoted as plus; concatenating multi-model

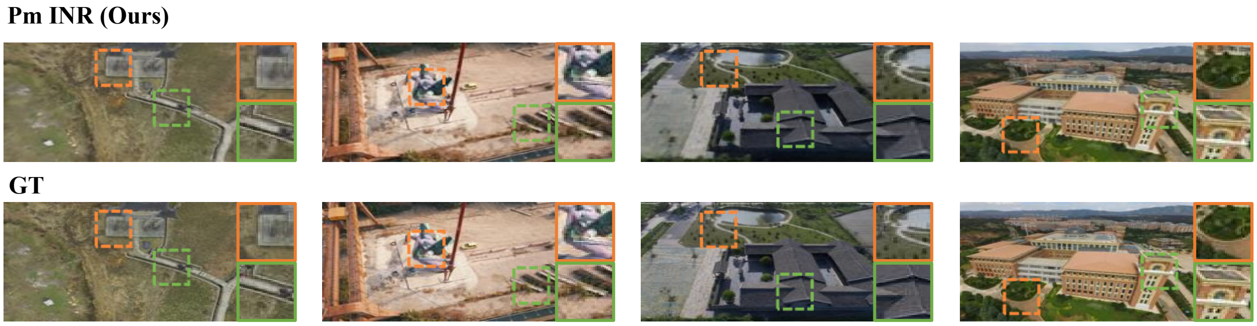

Figure 4: Qualitative results sampled from the large-scale part of the BlendedMVS dataset. For each scene, we present a visualization of a synthetic novel view and zoom in on two regions.

prior in the one dimension rather than the the zeroth dimension, denoted as hstack. The comparison results are shown in Tab. 3, which implies that our fusion strategy is the most advanced among them.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o cross	31.12	0.88	0.221
serial	29.36	0.89	0.212
plus	28.34	0.87	0.237
hstack	28.64	0.87	0.230
PM-INR (Ours)	**31.338**	**0.917**	**0.157**

Table 3: Experiment results about the effectiveness of multi-modal prior fusion module."w/o" cross represents removing the cross-attention mechanism of our method, "serial" represents serially injecting the multiple modalities into our network, "plus" represents directly plusing the multiple modalities prior, and "hstack" represents concatenating the multiple modalities prior in the dimension. ↑ means the higher, the better, ↓ means the lower, the better.

5 Conclusions and Limitations

In this paper, we propose PM-INR, a priori-rich multi-modal implicit neural representation network for outdoor unbounded large-scale scenes. Benefiting from our advanced multi-modal prior extraction and fusion modules, representative feature-rich priors are propagated to each sampling region. Therefore, without relying entirely on exploring sampling regions through individual sampling points, our PM-INR network can obtain global-level cross-modal semantics, which is lacking in current methods to cope with the exploding sampling space. Expensive experiments have demonstrated that our method surpasses the state-of-the-art method Mip-NeRF 360 by over 17% in various evaluation metrics. Meanwhile, abundant ablation experiments prove each multi-modal prior knowledge, and our fusion method can help the network generate more robust scene representations and synthesize more photo-realistic novel views.

With the help of multi-modal priors, our method can synthesize realistic novel views for outdoor unbounded large-scale scenes. However, our method still does not have any scene editing capabilities. We will explore the capability of scene editing via editing priors, which is considered a fascinating, valuable and promising endeavor.

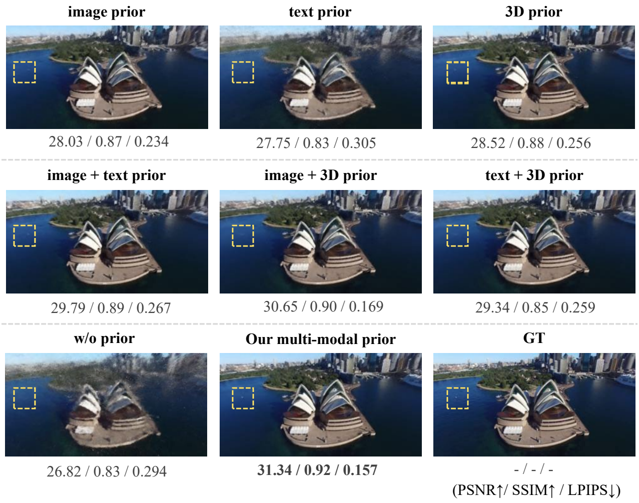

Figure 5: Qualitative visualization results for the effectiveness of each modal prior (zoom-in for the best of views) on the OMMO dataset. Obviously, using no prior or just a single modality prior will produce blurry images, while extracting cross-modal priors from both modalities can produce relatively realistic images.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62101137, 62071127, and U1909207), Shanghai Natural Science Foundation (No. 23ZR1402900), and Zhejiang Lab Project (No. 2021KH0AB05).

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3075–3084.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ding, Y.; Yin, F.; Fan, J.; Li, H.; Chen, X.; Liu, W.; Lu, C.; YU, G.; and Chen, T. 2023. PDF: Point Diffusion Implicit Function for Large-scale Scene Neural Representation. arXiv:2311.01773.

Esser, P.; Rombach, R.; Blattmann, A.; and Ommer, B. 2021. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34: 3518–3532.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Garbin, S. J.; Kowalski, M.; Johnson, M.; Shotton, J.; and Valentin, J. 2021. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14346–14355.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lombardi, S.; Simon, T.; Schwartz, G.; Zollhoefer, M.; Sheikh, Y.; and Saragih, J. 2021. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13.

Lu, C.; Yin, F.; Chen, X.; Chen, T.; Yu, G.; and Fan, J. 2023. A Large-Scale Outdoor Multi-modal Dataset and Benchmark for Novel View Synthesis and Implicit Scene Reconstruction. *arXiv preprint arXiv:2301.06782*.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7210–7219.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5865–5874.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rahebi, J. 2022. Vector quantization using whale optimization algorithm for digital image compression. *Multimedia Tools and Applications*, 81(14): 20077–20103.

Rebain, D.; Jiang, W.; Yazdani, S.; Li, K.; Yi, K. M.; and Tagliasacchi, A. 2021. Derf: Decomposed radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14153–14161.

Seitz, S. M.; Curless, B.; Diebel, J.; Scharstein, D.; and Szeliski, R. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, 519–528. IEEE.

Shen, Y.; Ma, W.-C.; and Wang, S. 2022. SGAM: Building a Virtual 3D World through Simultaneous Generation and Mapping. *Advances in Neural Information Processing Systems*, 35: 22090–22102.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.

Turki, H.; Ramanan, D.; and Satyanarayanan, M. 2022. Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs. arXiv:2112.10703.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Verbin, D.; Hedman, P.; Mildenhall, B.; Zickler, T.; Barron, J. T.; and Srinivasan, P. P. 2022. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5481–5490. IEEE.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, K.; Fan, J.; Ye, P.; and Zhu, M. 2023. Hyperspectral Image Classification Using Spectral–Spatial Token Enhanced Transformer With Hash-Based Positional Embedding. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Xiangli, Y.; Xu, L.; Pan, X.; Zhao, N.; Rao, A.; Theobalt, C.; Dai, B.; and Lin, D. 2022. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, 106–122. Springer.

Xu, Q.; Xu, Z.; Philip, J.; Bi, S.; Shu, Z.; Sunkavalli, K.; and Neumann, U. 2022. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5438–5448.

Yang, Y.; Liu, W.; Yin, F.; Chen, X.; Yu, G.; Fan, J.; and Chen, T. 2023. VQ-NeRF: Vector Quantization Enhances Implicit Neural Representations. *arXiv preprint arXiv:2310.14487*.

Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; and Quan, L. 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1790–1799.

Ye, P.; Li, B.; Chen, T.; Fan, J.; Mei, Z.; Lin, C.; Zuo, C.; Chi, Q.; and Ouyang, W. 2022a. Efficient joint-dimensional search with solution space regularization for real-time semantic segmentation. *International Journal of Computer Vision*, 130(11): 2674–2694.

Ye, P.; Li, B.; Li, Y.; Chen, T.; Fan, J.; and Ouyang, W. 2022b. b-darts: Beta-decay regularization for differentiable architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10874–10883.

Yin, F.; Chen, X.; Zhang, C.; Jiang, B.; Zhao, Z.; Fan, J.; Yu, G.; Li, T.; and Chen, T. 2023a. ShapeGPT: 3D Shape Generation with A Unified Multi-modal Language Model. *arXiv preprint arXiv:2311.17618*.

Yin, F.; Huang, Z.; Chen, T.; Luo, G.; Yu, G.; and Fu, B. 2023b. Dcnet: Large-scale point cloud semantic segmentation with discriminative and efficient feature aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yin, F.; Liu, W.; Huang, Z.; Cheng, P.; Chen, T.; and YU, G. 2022. Coordinates Are NOT Lonely–Codebook Prior Helps Implicit Neural 3D Representations. *arXiv preprint arXiv:2210.11170*.

Yin, F.; and Zhou, S. 2020. Accurate estimation of body height from a single depth image via a four-stage developing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8267–8276.

You, C.; Zhao, R.; Liu, F.; Dong, S.; Chinchali, S.; Topcu, U.; Staib, L.; and Duncan, J. 2022. Class-aware adversarial transformers for medical image segmentation. *Advances in Neural Information Processing Systems*, 35: 29582–29596.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.

Zhang, K.; Riegler, G.; Snavely, N.; and Koltun, V. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv:1801.03924.

Zhang, Z.; Ma, J.; Zhou, C.; Men, R.; Li, Z.; Ding, M.; Tang, J.; Zhou, J.; and Yang, H. 2021. UFC-BERT: Unifying multi-modal controls for conditional image synthesis. *Advances in Neural Information Processing Systems*, 34: 27196–27208.

Zhou, S.; Chan, K.; Li, C.; and Loy, C. C. 2022. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35: 30599–30611.