ReliefSAM: A Geometry-Augmented Multi-Prior Adapter for Bas-Relief Segmentation

Yihang Chen[1], Xiang Lyu[2], Rui Xu[2], Jiao Pan[3], Fadjar Ibnu Thufail[4], Brahmantara[5], Jiaqing Liu[1], Satoshi Tanaka[1], and Liang Li[1⋆]

[1] Ritsumeikan University, Japan
liliang@fc.ritsumei.ac.jp
[2] Dalian University of Technology, China
[3] University of Science and Technology Beijing, China
[4] National Research and Innovation Agency, Indonesia
[5] Indonesian Heritage Agency, Indonesia

Abstract. Bas-relief sculptures are vital historical records. However, many are inaccessible for modern 3D scanning due to damage or occlusion, leaving archival monocular photographs as the only available visual data. Semantic segmentation, such as isolating human figures in these images, is essential for digital documentation and subsequent archaeological analysis, yet remains highly challenging for foundation models like the Segment Anything Model (SAM). The intrinsic properties of bas-reliefs, including extreme material homogeneity, subtle depth variations, and indistinct soft edges, result in a severe lack of both chromatic and geometric contrast. Consequently, models pretrained on natural imagery struggle to perceive the shallow 2.5D structural cues embedded in relief surfaces, leading to imprecise boundary delineation. To address this limitation, we propose ReliefSAM, a prompt-free and parameter-efficient framework that explicitly incorporates image-aligned depth and edge priors derived from a single monocular RGB photograph. These geometric cues are encoded via a lightweight Multi-Prior Feature Encoder (MFE) and injected into a frozen SAM backbone through adapter-based interaction blocks. For high-resolution archival imagery, we further adopt overlapped sliding-window inference with Gaussian-weighted merging to ensure spatial consistency. Extensive experiments demonstrate that while adapting the frozen SAM image encoder with lightweight adapters establishes a highly competitive pure-RGB baseline that outperforms standard decoder-only fine-tuning, naively injecting a single geometric prior can induce modality interference. Crucially, ReliefSAM's joint integration of depth and soft-edge priors achieves a geometric consensus in which the two priors counterbalance and neutralize prior-specific biases. This synergistic 2.5D guidance breaks the RGB-only performance ceiling, effectively bridging the gap between natural RGB appearance and heritage-specific structures without requiring large-scale retraining of the foundation model.

⋆ Corresponding author.

Keywords: Bas-relief segmentation · SAM · Parameter-efficient fine-tuning · Geometric priors · Cultural heritage

1 Introduction

Bas-relief sculptures are an important form of cultural heritage: they preserve religious narratives, historical events, and social life in densely composed carved scenes. A representative example is the Borobudur (Barabudur) temple in Central Java, whose extensive narrative reliefs have been systematically documented in archaeological studies and art-historical accounts [14, 21]. In such reliefs, *human figures* are often the primary semantic carriers of the story. Accurate human figure segmentation is therefore a key step toward scalable digital documentation, retrieval and indexing, and downstream visual analysis for archaeological and art-historical research.

Human figure segmentation on relief photographs is fundamentally different from that on natural images. Relief surfaces are usually made of homogeneous material with similar color and texture, so foreground–background separation cannot reliably rely on RGB appearance alone. Instead, the boundary is mainly expressed by subtle geometric cues, which are frequently weakened by weathering, erosion, occlusion, and physical damage. In practice, many sites cannot be revisited or re-measured due to conservation restrictions, and even when modern scanning is possible, long-term degradation means the current geometry may deviate from the historical state. As a result, a large portion of relief analysis must rely on *single-view archival photographs* as the only available record, making it unrealistic to assume access to depth sensors, LiDAR, or multi-view reconstruction.

Recent foundation models for segmentation, especially the Segment Anything Model (SAM) [13], provide strong generalization on natural imagery but are not optimized for relief-specific geometry signals. For instance, Galanakis *et al*. explored zero-shot SAM for cultural heritage scan-to-structural analysis, where the model tends to produce brick-/block-level regions with relatively clear boundaries rather than fine-grained relief semantics such as persons [6]. Réby *et al*. combined SAM with GroundingDINO [16] to segment architectural components of Notre-Dame de Paris [25]; however, the targets are visually separable components (not shallow carved figures), and the pipeline still relies on SAM's original appearance-biased representation as well as text prompts and detector-generated boxes. Beyond foundation models, Ji *et al*. demonstrated the usefulness of geometry-aware guidance for Borobudur relief segmentation [10], but such approaches typically depend on additional survey products and task-specific training data, which may be limited in scale and generalization. Relatedly, adapter-based designs have been explored in other heritage segmentation tasks (*e.g*., mural damage) [29], suggesting the potential of lightweight model adaptation, yet relief human figure segmentation under a *single-image, prompt-free* constraint remains underexplored.

To bridge this gap, we propose ReliefSAM, a prompt-free framework for binary human figure segmentation on bas-relief photographs using a *single monocular RGB image* as the only input. The key idea is to recover relief-relevant geometry signals that are *implicitly* encoded in a photo by deriving image-aligned *depth* and *edge* priors from the same RGB input, and inject them into SAM to enhance boundary and shape sensitivity. Concretely, we encode the priors into compact spatial tokens with a lightweight Multi-Prior Feature Encoder (MFE), and fuse them with pretrained SAM image tokens via adapter-based interaction blocks. We freeze the SAM image and prompt encoders, and fine-tune the mask decoder together with our lightweight prior modules. To support full-resolution relief photographs, we adopt patch-based training and perform inference with overlapped sliding-window prediction followed by Gaussian-weighted merging for seamless high-resolution masks.

Extensive experiments and ablation analysis demonstrate that ReliefSAM outperforms both SOTA generic networks and established foundation model baselines in segmenting human figure regions.

Our contributions can be summarized as follows:

- We propose ReliefSAM, a parameter-efficient adaptation of SAM that injects *multi-prior geometry* derived from the same RGB image via a lightweight MFE and adapter-based interaction blocks.
- We present a practical high-resolution protocol for relief photographs, using patch-based learning and overlapped sliding-window inference with Gaussian-weighted merging to enable seamless full-resolution prediction.
- We build a high-resolution bas-relief dataset for human figure segmentation. Under the guidance of Indonesian heritage experts, we selected and annotated human figure regions for 156 panels of the "Hidden Relief" (Karmawibhangga) [5] from Borobudur, establishing a crucial benchmark for reburied and inaccessible cultural heritage.

2 Related Work

Cultural Heritage Segmentation with 3D Data. A large portion of cultural heritage (CH) segmentation and structural analysis has been built upon explicit 3D acquisition, *e.g.*, LiDAR, structured-light scanning, photogrammetry, and their resulting meshes or point clouds. For instance, recent studies have successfully applied deep neural networks to semantically parse historical monuments from dense 3D point clouds [20,24], while more recent work has further explored dedicated point-cloud semantic segmentation networks for ancient architectural heritage [30]. Such pipelines benefit from strong geometric cues, and are particularly effective when the target elements form separable components in 3D space. However, they implicitly assume on-site accessibility and dedicated equipment, which are often unavailable for restricted, degraded, or reburied heritage assets. In contrast, our problem setting focuses on archival *single-view* photographs, where no external 3D sensing is assumed.

RGB Semantic Segmentation and Foundation Models in CH. RGB semantic segmentation has evolved significantly from early fully convolutional networks (FCNs) [17] and the pioneering U-Net architecture [26]. To tackle the diverse and complex demands of various application domains, researchers have continually engineered a myriad of conventional network variants. Notable examples include nested architectures like U-Net++ [31], spatial pyramid pooling designs like DeepLabV3+ [2], and efficient attention-based representations like SegFormer [28]. More recently, the widespread success of Vision Transformers [4] in natural images has sparked growing interest across various fields, catalyzing a paradigm shift from task-specific networks toward massive foundation models. Following this trajectory, the Segment Anything Model (SAM) [13] offers strong pretrained representations with a promptable mask decoder and has been explored for CH analysis. For example, Galanakis *et al.* [6] investigate SAM in scan-to-structural analysis pipelines, where targets often exhibit relatively clear boundaries. Réby *et al.* [25] study CH semantic segmentation on Notre-Dame de Paris using foundation models (*e.g.*, SAM with GroundingDINO [16]), but the pipeline relies on textual prompts and detector-proposed boxes and does not explicitly enhance SAM with geometry-aware representations. These studies motivate CH-specific adaptation of foundation models toward more challenging heritage imagery and automated usage.

Parameter-Efficient Adaptation and Geometry-Aware Constraints. As pretrained large-scale models continue to grow in capacity, adapting them through full-network fine-tuning becomes increasingly costly, which has motivated the development of parameter-efficient fine-tuning (PEFT) methods. To address this bottleneck, parameter-efficient fine-tuning (PEFT) has emerged as a crucial paradigm, enabling the domain adaptation of massive foundation models by updating only a minuscule fraction of parameters. Among the most prominent PEFT techniques, Adapters [7] sequentially insert lightweight trainable bottleneck layers within frozen transformer blocks, while Low-Rank Adaptation (LoRA) [8] approximates the weight updates of dense layers using trainable low-rank decomposition matrices. These parameter-efficient strategies have matured extensively and become standard practice for knowledge injection across various computer vision domains [3, 11]. Following this success, in the cultural heritage (CH) sector, lightweight adaptation has also been explored for related restoration and segmentation tasks; *e.g.*, Zhang *et al.* [29] propose a convolutional adapter design for mural damage segmentation. Separately, geometry-aware constraints have been shown beneficial for relief understanding: Ji *et al.* [10] enhance relief segmentation with soft-edge cues, but their learning relies on explicit 3D measurements. Different from prior geometry-dependent pipelines, we operate under a strict monocular setting and derive image-aligned depth and soft-edge priors *from the same RGB photograph* (inspired by recent single-image relief reconstruction studies [23]), injecting them into a frozen SAM backbone via lightweight interaction adapters for geometry-aware, prompt-free segmentation.

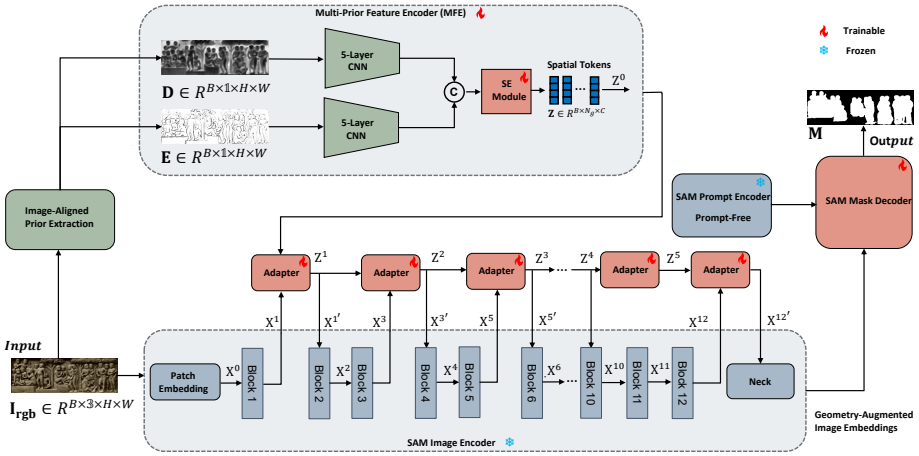

Fig. 1: Overview of our multi-prior framework. Given an RGB input, we extract image-aligned depth and edge priors, encode them into geometric tokens via the MFE, and inject them into a frozen SAM image encoder through our relief-specific geometric adapters for prompt-free segmentation.

3 Method

This section presents our multi-prior framework for bas-relief human figure segmentation. Starting from a single RGB image, we extract image-aligned depth and edge priors and inject them into a frozen SAM image encoder through a *relief-specific geometric adapter* (detailed in Sec. 3.3), enabling robust prompt-free mask prediction. We first summarize the end-to-end pipeline and notation in Sec. 3.1. We then describe prior extraction in Sec. 3.2, and detail the geometry-augmented SAM encoder and prompt-free decoder in Secs. 3.3 and 3.4. Finally, we introduce the training objective and the high-resolution inference protocol in Secs. 3.5 and 3.6.

3.1 Framework Overview

Problem Setup and Notation. Given a monocular RGB bas-relief photograph $\mathbf{I}_{rgb} \in \mathbb{R}^{B\times 3\times H\times W}$, our goal is to predict a binary human figure mask probability map $\hat{\mathbf{Y}} \in [0,1]^{B\times 1\times H\times W}$. To resolve the foreground–background camouflage caused by homogeneous material appearance, we extract two image-aligned geometric priors from the exact same RGB input: a depth map $\mathbf{D}$ and an edge map $\mathbf{E}$ (Sec. 3.2).

End-to-End Pipeline. As shown in Fig. 1, we encode the depth and edge maps with a lightweight Multi-Prior Feature Encoder (MFE) to produce a compact sequence of geometric tokens $\mathbf{Z} \in \mathbb{R}^{B\times N_g\times C}$. We inject $\mathbf{Z}$ into a frozen

SAM image encoder via our adapters at selected transformer stages, yielding geometry-enhanced image embeddings. Finally, SAM's mask decoder predicts a single foreground mask in a prompt-free manner.

3.2 Image-Aligned Depth and Edge Prior Extraction

Depth Prior Generation. To exploit the intrinsic 2.5D structure of basreliefs [27], we employ the monocular depth estimation framework proposed by Pan *et al.* [23]. Unlike general-purpose estimators that often suffer from depth value compression on low-contrast relief surfaces, this method incorporates an edge-matching module to constrain depth prediction specifically within curvature regions, effectively preserving fine-grained morphological details such as facial features.

Edge Prior Generation. Traditional edge detectors (*e.g.*, Canny [1]) typically fail to capture the subtle, low-curvature soft edges characteristic of weathered relief surfaces. To address this, we also employ the framework in [23] to derive image-aligned edge maps from the RGB input. This framework is specifically optimized to predict edge confidence maps that emulate the soft edges extracted in stochastic point-based rendering (SPBR) [12]. By capturing these delicate geometric gradients, the resulting edge prior serves as a reliable boundary constraint for human figure segmentation even when chromatic contrast is absent.

Image Alignment and Normalization. To ensure strict pixel-wise correspondence, any spatial resizing or cropping is applied consistently across all modalities. We represent the resized RGB input as a channel-first tensor $\mathbf{I}_{rgb} \in \mathbb{R}^{B\times 3\times H\times W}$. Following SAM's standard preprocessing, we first scale RGB values to $[0, 1]$ and then apply channel-wise ImageNet mean–std normalization: $\mathbf{I}_{rgb} \leftarrow (\mathbf{I}_{rgb}/255 - \boldsymbol{\mu})/\boldsymbol{\sigma}$, where $\boldsymbol{\mu} = (0.485, 0.456, 0.406)$ and $\boldsymbol{\sigma} = (0.229, 0.224, 0.225)$.

For the geometric priors, we directly represent them as single-channel float tensors $\mathbf{D}, \mathbf{E} \in \mathbb{R}^{B\times 1\times H\times W}$ and linearly normalize them to the $[0, 1]$ range via simple linear scaling $\mathbf{D} \leftarrow \mathbf{D}/255$ and $\mathbf{E} \leftarrow \mathbf{E}/255$. This ensures numerical consistency before feeding them into the network.

3.3 Geometry-Augmented SAM Image Encoder

Rather than redesigning the segmentation network, we preserve the original SAM architecture and extend only its frozen image encoder to accept geometric priors. Specifically, the priors are first tokenized by a Multi-Prior Feature Encoder and then injected into the backbone via lightweight, relief-specific geometric adapters.

Multi-Prior Feature Encoder (MFE). Given the image-aligned priors from Sec. 3.2, we design a lightweight Multi-Prior Feature Encoder to convert the depth prior $\mathbf{D}$ and edge prior $\mathbf{E}$ into a sequence of geometric tokens. We represent

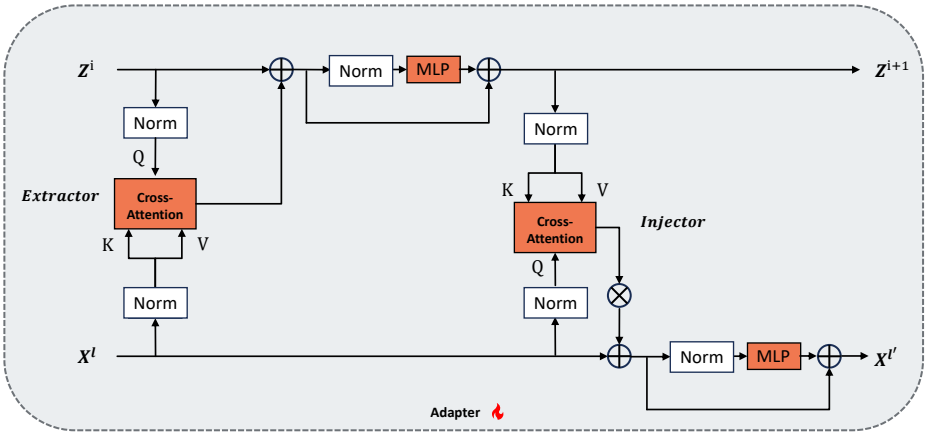

Fig. 2: Architecture of the proposed relief-specific geometric adapter. It employs an extract-then-inject bidirectional cross-attention mechanism explicitly designed to fuse geometry priors.

$\mathbf{D}, \mathbf{E} \in \mathbb{R}^{B\times 1\times H\times W}$ and set the token embedding dimension C to match that of SAM image tokens, enabling direct fusion inside the image encoder.

Dual-Branch CNN Encoding. The MFE employs two parallel convolutional branches to process the depth and edge inputs. Each branch consists of five 3×3 Conv–BN–ReLU stages, producing feature maps $\mathbf{F}_d \in \mathbb{R}^{B\times C_d\times H_g\times W_g}$ and $\mathbf{F}_e \in \mathbb{R}^{B\times C_e\times H_g\times W_g}$, where $C_d = \lfloor C/2 \rfloor$, $C_e = C - C_d$, and (H_g, W_g) is the downsampled resolution. If $\mathbf{F}_d$ and $\mathbf{F}_e$ have mismatched spatial sizes due to rounding, we resize one to match the other using bilinear interpolation.

Fusion and Channel Recalibration. We fuse the two modalities by channel concatenation $\mathbf{F} = [\mathbf{F}_d; \mathbf{F}_e] \in \mathbb{R}^{B\times C\times H_g\times W_g}$, followed by a Squeeze-and-Excitation (SE) module [9] for channel recalibration. Specifically, global average pooling yields a channel descriptor $\mathbf{a} \in \mathbb{R}^{B\times C}$, and channel weights are computed by a two-layer MLP with reduction ratio r=4:

$$\mathbf{s} = \sigma\big(\mathbf{W}_2\,\delta(\mathbf{W}_1\mathbf{a})\big) \in (0,1)^{B\times C}, \tag{1}$$

where $\delta(\cdot)$ and $\sigma(\cdot)$ denote ReLU and Sigmoid, respectively. We then broadcast $\mathbf{s}$ over spatial dimensions and reweight the fused features as $\tilde{\mathbf{F}} = \mathbf{F} \odot \mathbf{s}$. During single-prior ablations, we bypass the SE module such that the fusion reduces to an identity mapping for the available branch.

Tokenization. Finally, we flatten the spatial dimensions and form a token sequence. Concretely, we reshape $\tilde{\mathbf{F}}$ to $B \times C \times (H_g W_g)$ and then transpose it to obtain $\mathbf{Z} \in \mathbb{R}^{B\times N_g\times C}$ with $N_g = H_g W_g$. These geometric tokens are then fed into the SAM image encoder and fused via the adapter modules described next.

Relief-Specific Geometric Adapter. Let $\mathbf{Z}^0 \in \mathbb{R}^{B\times N_g\times C}$ denote the initial sequence of geometric tokens produced by the MFE. As this geometric stream

passes through the network, it is sequentially updated by a series of our relief-specific geometric adapters (Fig. 2). Let $\mathbf{X}^l \in \mathbb{R}^{B\times N_i\times C}$ denote the image token sequence output from the l-th transformer block, where N_i is the number of image tokens.

Extract-Then-Inject Geometric Prior Interaction. Naive feature fusion (*e.g.*, element-wise addition) can contaminate SAM's robust visual features with spurious structures and noise from the geometric priors (e.g., weathering cracks). To inject geometry without degrading the visual representation, we perform bidirectional feature fusion at selected stages $\mathcal{S}$ using two pre-normalized cross-attention blocks. Following an *extract-then-inject* design, we first extract geometry-relevant cues via cross-attention, and then inject only the filtered geometric information back into the SAM feature stream.

First, an extractor updates the geometric tokens by attending to the image tokens:

$$\mathbf{Z} \leftarrow \mathbf{Z} + \mathrm{CA}\big(\mathrm{LN}(\mathbf{Z}),\, \mathrm{LN}(\mathbf{X}^{(l)}),\, \mathrm{LN}(\mathbf{X}^{(l)})\big), \tag{2}$$

so that $\mathbf{Z}$ can absorb instance-aware cues from the current image representation.

Then, an injector feeds the updated geometric information back into the image tokens:

$$\mathbf{X}^{(l)} \leftarrow \mathbf{X}^{(l)} + \mathrm{CA}\big(\mathrm{LN}(\mathbf{X}^{(l)}),\, \mathrm{LN}(\mathbf{Z}),\, \mathrm{LN}(\mathbf{Z})\big). \tag{3}$$

Injection Schedule. For ViT-B [4] with 12 transformer blocks, we insert interaction blocks at $\mathcal{S} = \{2, 4, 6, 8, 10\}$ and apply an additional final injection after the last transformer block.

3.4 Prompt-Free Mask Decoding

Given the geometry-augmented image embeddings from Sec. 3.3, we utilize SAM's original prompt encoder and mask decoder in a strictly prompt-free manner. To achieve fully automated processing without human intervention, we provide no manual prompts (*e.g.*, points or boxes) during inference. Instead, the model relies solely on SAM's default learned *no-mask* embeddings to initiate the decoding process. Furthermore, as our target is binary human figure segmentation, we explicitly set `multimask_output=False`. The decoder thus processes the geometry-augmented image tokens to directly predict a single low-resolution mask logit map, which is subsequently passed to the training objective.

3.5 Training Objective and Optimization

We train the network for binary human figure segmentation on relief sculptures. Given an input image, the model first predicts a low-resolution mask logit map $\mathbf{M}^{\mathrm{lr}} \in \mathbb{R}^{B\times 1\times H_s\times W_s}$. Since SAM predicts masks at a lower resolution, we bilinearly upsample the logits to the ground-truth resolution,

$$\mathbf{M} = \mathcal{U}(\mathbf{M}^{\mathrm{lr}}) \in \mathbb{R}^{B\times 1\times H\times W}, \tag{4}$$

and convert the logits to probabilities by the sigmoid function,

$$\hat{\mathbf{Y}} = \sigma(\mathbf{M}), \tag{5}$$

where $\hat{\mathbf{Y}} \in [0,1]^{B\times 1\times H\times W}$ and the ground-truth mask is $\mathbf{Y} \in \{0,1\}^{B\times 1\times H\times W}$.

Loss Function. We adopt a Dice–Focal objective. The soft Dice [22] loss is defined as

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\langle \hat{\mathbf{Y}}, \mathbf{Y}\rangle + \epsilon}{\|\hat{\mathbf{Y}}\|_1 + \|\mathbf{Y}\|_1 + \epsilon}, \tag{6}$$

where $\langle \hat{\mathbf{Y}}, \mathbf{Y}\rangle$ denotes the sum of element-wise products over all pixels, and ϵ is a small constant for numerical stability.

For focal loss [15], we use the standard p_t form applied per pixel. Let $\hat{y}_i$ and y_i denote the elements of $\hat{\mathbf{Y}}$ and $\mathbf{Y}$ at pixel $i \in \Omega$. We define $p_{t,i} = \hat{y}_i y_i + (1 - \hat{y}_i)(1 - y_i)$ and $\alpha_{t,i} = \alpha(1 - y_i) + (1 - \alpha)y_i$, where α weights the background class ($y_i = 0$) and $1 - \alpha$ weights the foreground ($y_i = 1$). The focal loss is

$$\mathcal{L}_{\text{focal}} = -\frac{1}{|\Omega|}\sum_{i\in\Omega} \alpha_{t,i}(1 - p_{t,i})^{\gamma} \log(p_{t,i} + \epsilon). \tag{7}$$

The final training objective is

$$\mathcal{L} = \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{focal}}\mathcal{L}_{\text{focal}}, \tag{8}$$

where λ_{dice} and λ_{focal} balance the contributions of the Dice and focal terms. The exact values for balancing weights and focal loss hyperparameters (α, γ) are detailed in the implementation section.

Trainable Parameters and Freezing Policy. We initialize SAM from pretrained checkpoints and keep the prompt encoder frozen. To preserve the pretrained representation, we freeze the original SAM image encoder weights and only train the newly introduced geometry modules (the MFE and the adapter interaction blocks) together with the SAM mask decoder. This yields a lightweight fine-tuning regime that modifies SAM minimally while enabling geometry-aware segmentation.

Optimization. We optimize the trainable parameters using AdamW [19] and adopt a cosine annealing learning-rate [18] schedule. In ablations, we toggle depth or edge modalities without changing the model interface by enabling or disabling the corresponding input streams in the data pipeline.

3.6 High-Resolution Inference and Evaluation

Relief images often exceed the fixed input size of our SAM-based model. Therefore, both validation and testing adopt an overlapped sliding-window inference scheme, which enables prediction on *arbitrarily high-resolution* images.

Given an image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ and its precomputed geometric priors ($\mathbf{D}$ and $\mathbf{E}$), we tile all modalities synchronously into overlapping $P \times P$ windows with stride S. Boundary tiles are zero-padded to $P \times P$ before inference, and the

predicted patch probability maps are cropped back to their valid regions to avoid padding artifacts.

To eliminate visible stitching seams, overlapping predictions are merged using a spatial Gaussian weighting strategy. Specifically, we accumulate a Gaussian-weighted probability sum $\mathbf{S}$ and a spatial weight map $\mathbf{W}$ over the full resolution and compute

$$\hat{\mathbf{Y}} = \frac{\mathbf{S}}{\mathbf{W} + \epsilon}, \tag{9}$$

followed by thresholding at 0.5 to obtain the final binary mask.

4 Experiments

In this section, we evaluate ReliefSAM on high-resolution bas-relief human figure segmentation under the single-image setting described in Sec. 3. Unless otherwise specified, all methods are trained on cropped patches and evaluated on full-resolution images, with model selection based on full-image validation IoU. To disentangle the effects of architectural adaptation and geometric priors, we compare ReliefSAM against two SAM-based baselines: a standard decoder-only fine-tuning baseline (SAM-DecoderFT) and a strong pure-RGB adapter baseline (SAM-Adapter). We then analyze the individual and joint effects of depth and edge priors under this strong RGB baseline. The experimental setup and compared methods are given in Secs. 4.1 and 4.2, followed by the main results in Sec. 4.3 and ablation studies in Sec. 4.4.

4.1 Experimental Setup

Dataset. We evaluate ReliefSAM on a novel high-resolution bas-relief benchmark specifically constructed for segmenting human figure regions. The primary data source is the "Hidden Relief" (Karmawibhangga) from the Borobudur temple, a UNESCO World Heritage site. Due to structural reinforcement in the late 19th century, most of the 160 narrative panels were permanently reburied and remain inaccessible for modern 3D scanning or on-site LiDAR acquisition. Consequently, researchers must rely solely on archival monocular photographs captured in 1890 as the only visual evidence. The raw archival photographs were provided by an official heritage conservation authority under a research agreement. These 156 monocular images possess a high native resolution of 3200×1024 pixels. Under the guidance of heritage conservation experts, we selected and annotated human figure regions for 156 panels to establish precise pixel-wise ground-truth masks. This establishes a crucial benchmark for reburied cultural heritage, characterized by extreme material homogeneity and degraded details from over a century of concealment. We partition the dataset into 116, 20, and 20 images for training, validation, and testing, respectively. To rigorously avoid information leakage during patch-based learning, this split is performed strictly at the image level, ensuring that all patches cropped from the same full-resolution photograph are assigned exclusively to the same set.

Implementation Details. The backbone is SAM ViT-B. SAM-DecoderFT freezes the image and prompt encoders and fine-tunes only the mask decoder, whereas ReliefSAM additionally trains the MFE and relief-specific geometric adapters inserted at blocks $\{2, 4, 6, 8, 10\}$ with one final post-encoder injection. Input normalization and prior scaling follow Sec. 3. Training uses 1024×1024 patches with stride 128 and lightweight aligned augmentation, while validation and testing follow the full-resolution sliding-window protocol of Sec. 3.6 with patch size 1024 and stride 64. We optimize with AdamW for 50 epochs (batch size 2, learning rate 1×10^{-4}, weight decay 1×10^{-4}) under a cosine schedule, using an equally weighted Dice–Focal loss ($\alpha = 0.25$, $\gamma = 2.0$). Experiments are run on a single NVIDIA GeForce RTX 5090 GPU (32 GB VRAM).

4.2 Compared Methods

We compare ReliefSAM against both conventional RGB segmentation baselines and SAM-based adaptation baselines under the same training and full-resolution evaluation protocol. As conventional baselines, we adopt DeepLabV3+ [2] and SegFormer [28]. For SAM-based comparisons, we include SAM-DecoderFT, a decoder-only fine-tuning baseline in which the SAM image encoder and prompt encoder are frozen and only the mask decoder is optimized, and SAM-Adapter, a strong pure-RGB baseline using the same relief-specific adapter design as our method but without depth or edge priors. Our full model, **ReliefSAM**, shares the same SAM backbone, prompt-free decoding protocol, and adapter-based adaptation framework as SAM-Adapter, while additionally introducing image-derived depth and edge priors through the proposed MFE.

4.3 Main Results

Quantitative Results. We summarize the quantitative comparison on the high-resolution bas-relief dataset in Tab. 1. Under our single-image setting, ReliefSAM takes only an RGB photograph as input, while the depth and edge cues are derived from the same image as geometric priors. All results are reported on full-resolution test images under the same sliding-window evaluation protocol.

ReliefSAM achieves the best overall performance, reaching 91.78% IoU and 95.65% Dice. Compared with the decoder-only SAM fine-tuning baseline (SAM-DecoderFT), the gain is substantial (+7.96 IoU and +4.70 Dice). More importantly, ReliefSAM also surpasses the strong pure-RGB baseline (SAM-Adapter) by +0.53 IoU and +0.28 Dice, showing that the improvement cannot be attributed merely to architectural adaptation. Instead, it comes from the effective fusion of complementary geometric priors derived from the same RGB image.

Qualitative Results. To complement the quantitative comparison, Fig. 3 presents one full-resolution comparison together with four enlarged local patches. The visual results show that conventional segmentation baselines and decoder-only SAM fine-tuning often produce incomplete masks or leak into surrounding relief

Table 1: Main quantitative comparison on the high-resolution bas-relief human figure segmentation dataset. All methods use full-resolution sliding-window inference and model selection by full-image validation IoU. The best results are highlighted in **bold**.

Method	Backbone	IoU (%) ↑	Dice (%) ↑
SegFormer [28]	MiT-B3	87.86	93.45
DeepLabV3+ [2]	ResNet-50	89.32	94.27
SAM-DecoderFT	ViT-B	83.82	90.95
SAM-Adapter (Ours, RGB-only)	ViT-B	91.25	95.37
ReliefSAM (Ours)	ViT-B	**91.78**	**95.65**

Table 2: Ablation of geometric priors under a fixed SAM-Adapter backbone (overall performance). All results are reported on the full-resolution test set. The best result is highlighted in **bold**.

Variant	IoU (%) ↑	Dice (%) ↑
SAM-Adapter (RGB-only)	91.25	95.37
+ Depth prior	89.97	94.57
+ Edge prior	90.52	94.95
ReliefSAM (Ours)	**91.78**	**95.65**

textures in ambiguous regions. SAM-Adapter already provides a strong pure-RGB baseline, while ReliefSAM further improves foreground completeness and boundary delineation. The enlarged patches particularly highlight the benefit of the proposed geometric priors in regions with weak relief boundaries and texture interference.

4.4 Ablation Studies

Effectiveness of Geometric Priors under a Strong RGB Baseline. To disentangle geometric prior gains from architectural adaptation gains, we conduct ablations under a fixed SAM-Adapter backbone. Specifically, all variants in this subsection share the same frozen SAM image encoder, prompt-free decoding protocol, and relief-specific geometric adapters; only the availability of depth and edge priors is changed.

As shown in Tab. 2, the RGB-only SAM-Adapter baseline already achieves strong performance (91.25% IoU / 95.37% Dice), indicating that adapter-based architectural adaptation substantially improves over decoder-only fine-tuning. However, introducing a *single* geometric prior does not monotonically improve upon this strong pure-RGB baseline.

To understand this behavior, Tab. 3 reveals that the two priors exhibit opposite tendencies. The Depth-only variant attains the highest Precision (95.37%) but the lowest Recall (93.89%), suggesting that depth acts as a conservative structural cue: it suppresses false positives, yet tends to miss fine or weak bound-

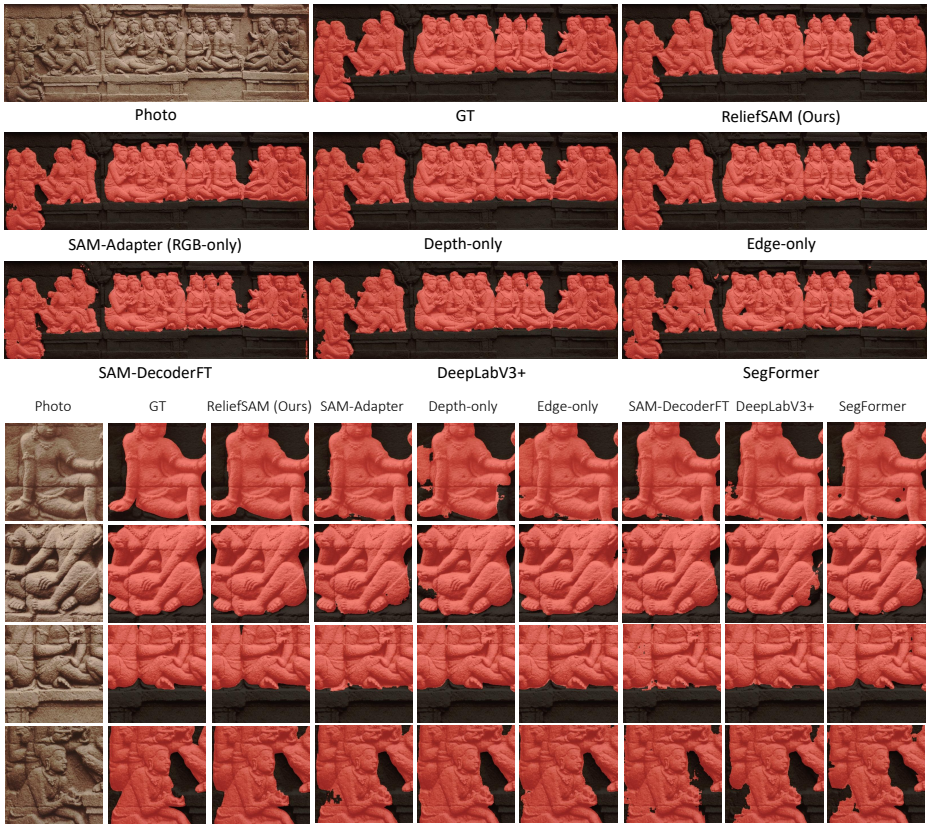

Fig. 3: Qualitative comparison on representative bas-relief images. Top: Full-resolution comparison on a representative relief panel. Bottom: Enlarged local patches highlighting difficult regions. From left to right: input RGB, ground truth, ReliefSAM (Ours), SAM-Adapter, Depth-only, Edge-only, SAM-DecoderFT, DeepLabV3+, and SegFormer. ReliefSAM produces more foreground regions and cleaner boundaries, while reducing leakage into surrounding relief textures in visually ambiguous areas.

aries due to the over-smoothed geometry of shallow relief surfaces. In contrast, the Edge-only variant attains the highest Recall (97.43%) but noticeably lower Precision (92.65%), indicating that edge cues aggressively recover boundaries while also introducing false positives caused by weathering cracks and other high-frequency artifacts.

Crucially, ReliefSAM combines both priors and achieves the best overall IoU and Dice (Tab. 2). This shows that the gain cannot be attributed merely to a stronger adapter architecture or additional parameters. Instead, depth and edge priors provide complementary geometric evidence: depth improves structural conservativeness, edge improves boundary sensitivity, and their joint fusion balances the Precision–Recall trade-off.

Table 3: Precision–Recall analysis of isolated priors under the same SAM-Adapter backbone. Depth-only is more conservative (higher Precision), whereas Edge-only is more aggressive (higher Recall). All results are reported on the full-resolution test set. The best result in each column is highlighted in **bold**.

Variant	Precision (%) ↑	Recall (%) ↑
SAM-Adapter (RGB-only)	94.50	96.31
+ Depth prior	**95.37**	93.89
+ Edge prior	92.65	**97.43**
ReliefSAM (Ours)	94.26	97.14

5 Conclusion

We presented ReliefSAM, a prompt-free and parameter-efficient framework for human figure segmentation in bas-relief sculptures under a realistic monocular setting. Instead of relying on external 3D scanning, ReliefSAM derives image-aligned depth and edge priors directly from a single RGB photograph and injects them into a frozen SAM backbone through lightweight relief-specific geometric adapters.

Experiments on a new high-resolution benchmark, comprising 156 "Hidden Relief" panels of the Borobudur temple, demonstrate that a strong RGB-only adapter baseline already provides substantial gains over decoder-only fine-tuning, while isolated geometric priors exhibit opposite biases. In contrast, the joint fusion of depth and edge delivers the best overall performance, indicating that complementary geometric evidence is more effective than either single-prior guidance or architectural adaptation alone.

These results suggest that geometry-aware adaptation of foundation models is a promising direction for cultural heritage imagery, especially when only archival monocular photographs are available. In future work, we will explore more robust prior extraction, broader heritage categories, and extensions toward multi-class semantic segmentation.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62406024) and the Fundamental Research Funds for the Central Universities (Grant No. FRF-TP-26-060), and the Research Institute Support Fund of the Art Research Center, Ritsumeikan University (Grant No. 26CACK60014).

References

1. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **8**(6), 679–698 (1986). https://doi.org/10.1109/TPAMI.1986.4767851

2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 801–818 (2018)
3. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: AdaptFormer: Adapting vision transformers for scalable visual recognition. In: Advances in Neural Information Processing Systems. vol. 35, pp. 16664–16678 (2022)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
5. Fontein, J.: The Law of Cause and Effect in Ancient Java. Koninklijke Nederlandse Akademie van Wetenschappen (1989)
6. Galanakis, D., Lucho, S., Maravelakis, E., Bolanakis, N., Konstantaras, A., Vidakis, N., Petousis, M., Treuillet, S., Desquesnes, X., Brunetaud, X.: Segment Anything Model for scan-to-structural analysis in cultural heritage. In: 2024 5th International Conference in Electronic Engineering, Information Technology & Education (EEITE). pp. 1–7. IEEE (2024). https://doi.org/10.1109/EEITE61750.2024.10654401
7. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning (ICML). pp. 2790–2799 (2019)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
10. Ji, S., Pan, J., Li, L., Hasegawa, K., Yamaguchi, H., Thufail, F.I., Brahmantara, Sarjiati, U., Tanaka, S.: Semantic segmentation for digital archives of Borobudur reliefs based on soft-edge enhanced deep learning. Remote Sensing **15**(4), 956 (2023)
11. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
12. Kawakami, K., Hasegawa, K., Li, L., Nagata, H., Adachi, M., Yamaguchi, H., Thufail, F.I., Riyanto, S., Brahmantara, Tanaka, S.: Opacity-based edge highlighting for transparent visualization of 3D scanned point clouds. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **V-2-2020**, 373–380 (2020). https://doi.org/10.5194/isprs-annals-V-2-2020-373-2020
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
14. Krom, N.J.: Barabudur: Archaeological Description. Martinus Nijhoff, The Hague (1927)
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
16. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded

pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023). https://doi.org/10.48550/arXiv.2303.05499
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
18. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
20. Matrone, F., Lingua, A., Pierdicca, R., Malinverni, E.S., Paolanti, M., Grilli, E.: Comparing machine and deep learning methods for large 3D heritage semantic segmentation. ISPRS International Journal of Geo-Information **9**(9), 535 (2020). https://doi.org/10.3390/ijgi9090535
21. Miksic, J.N.: Borobudur: Golden Tales of the Buddhas. Periplus Editions (1990)
22. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. pp. 565–571. IEEE (2016)
23. Pan, J., Gu, X., Li, L., Li, W., Wang, Y., Bo, N., Yamaguchi, H., Hasegawa, K., Thufail, F.I., Brahmantara, Yao, C., Ban, X., Tanaka, S.: Enhanced 3D monocular reconstruction of relief-type cultural heritage via an edge-matching module. ACM Journal on Computing and Cultural Heritage (2025). https://doi.org/10.1145/3771097
24. Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E.S., Frontoni, E., Lingua, A.M.: Point cloud semantic segmentation using a deep learning framework for cultural heritage. Remote Sensing **12**(6), 1005 (2020). https://doi.org/10.3390/rs12061005
25. Réby, K., Guilhelm, A., De Luca, L.: Semantic segmentation using foundation models for cultural heritage: an experimental study on Notre-Dame de Paris. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 1689–1697 (2023)
26. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
27. Weyrich, T., Deng, J., Barnes, C., Rusinkiewicz, S., Finkelstein, A.: Digital bas-relief from 3D scenes. ACM Transactions on Graphics **26**(3), 32 (2007). https://doi.org/10.1145/1276377.1276417
28. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems. vol. 34, pp. 12077–12090 (2021)
29. Zhang, J., Wang, Z., Chen, W., Xu, Z., Xu, T., Liu, J., Li, X.: EdgeSAM-CASD: Lightweight mural damage segmentation via convolutional adapter. In: Huang, D.S., Li, B., Chen, H., Zhang, C. (eds.) Advanced Intelligent Computing Technology and Applications. ICIC 2025, Lecture Notes in Computer Science, vol. 15846, pp. 49–59. Springer, Singapore (2025). https://doi.org/10.1007/978-981-96-9794-6_5
30. Zhou, C., Dong, Y., Hou, M., Ji, Y., Wen, C.: MP-DGCNN for the semantic segmentation of chinese ancient building point clouds. Heritage Science **12**, 249 (2024)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 3–11. Springer (2018)