

Diffusion-guided lightweight remote-sensing semantic models for on-board AI

Xinyue Li¹, Hongwei Li¹, Jingwen Wang¹, Chunhui Wang¹, Chaojie Zhang¹

¹ College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China

Corresponding author: Chaojie Zhang, zhangcj@zju.edu.cn

Earth observation missions increasingly benefit from interpreting imagery close to the sensor, but high-capacity semantic models are difficult to operate under embedded compute, memory, and power constraints. This paper presents a diffusion-guided semantic model compression framework for resource-constrained on-board AI. During ground training, a frozen pre-trained diffusion U-Net supplies multi-scale structural priors. Feature, attention, and response-level knowledge is then transferred to compact student networks; only the selected student and task head are retained for deployment. The task head combines channel-spatial attention with class imbalance-aware supervision, and the distilled model is converted through ONNX and TensorRT for FP16/INT8 execution. On ISPRS Vaihingen, the ResNet-34 U-Net student obtains 71.33% mIoU and 83.21% mF1, versus 71.70% and 83.73% for the 566.09-M-parameter teacher. Its 22.1 M parameters and 40.3 GFLOPs correspond to reductions of 96.10% and 98.94%, while throughput on an RTX 3090 increases from 1.2 to 32 FPS. On Jetson AGX Orin under the 15 W profile, TensorRT INT8 occupies 22.2 MB, reaches 22 FPS at 12.5 W, retains 70.74% mIoU, and improves energy efficiency from 0.61 to 1.76 FPS/W. The full diffusion representation also records the smallest mean relative mIoU drop (20.62%) under haze, shadow, and mixed sensor noise. These results support diffusion-to-student knowledge transfer as a model-compression path for on-board Earth observation semantic analysis.

Keywords: Earth observation, embedded deployment, knowledge distillation, low-bit inference, model compression, on-board AI, semantic representation, space computing

1. INTRODUCTION

1.1 Background and motivation

Low Earth orbit constellations and high resolution Earth observation payloads are transforming satellites from passive sensing platforms into distributed sensing-and-computing nodes. Space computing and on-board AI allow imagery to be screened close to the sensor, so that mission-relevant regions and preliminary semantic products can be generated before the next ground-contact window [1, 2, 3, 4, 5]. This capability is particularly valuable for disaster response, maritime monitoring, environmental assessment, and other time-sensitive applications in which delayed delivery reduces the operational value of the observation.

The resulting challenge is not limited to the volume of captured imagery. Semantic interpretation must also be performed repeatedly under strict limits on computation, memory, power, and heat dissipation. A high capacity model designed for a ground GPU can provide accurate maps, but its parameter count, activation memory, and floating-point workload make sustained on-board execution difficult. Conversely, a small network trained only from labels may fit the hardware budget but lose the boundary and small object cues that are important in remote-sensing scenes.

In this paper, compression is therefore used in the sense of semantic-model and representation compression. Knowledge acquired by a high capacity teacher is transferred to a compact student, and the student produces task-oriented semantic outputs that can support mission-dependent product selection. The work does not introduce a reconstructive image bitstream, replace standards such as JPEG2000 or CCSDS 123 [6, 7], or claim rate distortion superiority over image codecs. This scope matches the space-AI problem considered here: reducing the cost of Earth observation inference through knowledge distillation, compact representations, and low-bit deployment.

1.2 Related work and research gap

Existing on-board AI pipelines face four recurring limitations. First, high capacity networks developed for ground GPUs commonly require substantial parameter storage, activation memory, and memory traffic. Second, directly trained lightweight networks can sacrifice boundary detail and small object recognition. Third, model compression studies often emphasize parameter count or floating-point operations without jointly reporting latency, memory, power, and energy efficiency on an embedded platform. Fourth, semantic accuracy is often measured only on clean imagery, even though Earth observation inputs may contain haze, shadow, and sensor noise. These limitations call for an evaluation that links semantic accuracy to deployment cost and environmental robustness.

Knowledge distillation, quantization, compact architectures, and hardware-aware search provide established routes for reducing the neural network cost [8, 9, 10, 11, 12]. Recent surveys and on-board Earth observation studies further emphasize that useful compression must be evaluated on the target execution chain, because parameter count alone does not determine latency, memory demand, or energy efficiency [2, 3, 4, 5]. However, existing work rarely combines multilevel transfer from a generative teacher, task-specific robustness analysis, and staged embedded deployment within one remote-sensing evaluation.

Learned image compression addresses a complementary problem by optimizing reconstruction quality and bit rate through entropy models, hierarchical priors, hybrid backbones, or diffusion priors [13, 14, 15, 16, 17, 18, 19, 20, 21, 22], while task-oriented coding considers machine-consumable features [23]. The present study does not produce a coded image bitstream or a rate distortion curve. Its evidence instead concerns semantic accuracy, model size, computation, embedded latency, memory, power, and robustness. This distinction prevents model compression from being conflated with conventional image coding.

Diffusion models provide rich multi-scale structural priors, but their large U-Net backbones and iterative sampling process are unsuitable for direct on-board use [24, 25]. We therefore employ the diffusion U-Net only as a frozen ground-training teacher. Feature, attention, and response-level supervision transfers its representation to a compact discriminative student, and the teacher is discarded after training. The resulting research question is whether this transfer can preserve remote-sensing semantic quality while reducing computational cost sufficiently for embedded pre-verification.

1.3 Contributions

The main contributions of this work are as follows:

- We formulate a resource-aware Earth observation processing chain that couples compact semantic inference with mission-dependent selection of semantic products.
- We introduce diffusion-guided multi-granularity distillation with explicit feature, attention, and response-level objectives, and state the loss weights used in the reported study.
- We quantify the accuracy–efficiency trade-off across three student architectures and report controlled ablations for diffusion features, attention, task supervision, and the individual distillation components.
- We evaluate FP32, ONNX, FP16, and INT8 execution on an RTX 3090 workstation and a Jetson AGX Orin embedded platform, reporting model size, memory, latency, throughput, power, energy efficiency, and retained mIoU; the Jetson results are explicitly treated as pre-verification rather than flight-qualified performance.
- We assess semantic robustness under controlled haze, shadow, and mixed sensor-noise degradation using the same task metric as the clean-image evaluation.

The remainder of this paper is organized as follows. Section 2 defines the semantic task, evaluation dimensions, and joint objective. Section 3 presents the diffusion-guided representation, distillation objectives, student architectures, and low bit conversion. Section 4 positions the embedded platform relative to flight hardware. Section 5 reports the experiments, and Section 6 discusses limitations and concludes the paper.

2. PROBLEM SETUP AND EVALUATION

2.1 Task definition and scope

Let $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ denote an NIR–R–G remote-sensing tile and $\mathbf{g}$ its pixel-level reference map. The semantic model produces a dense prediction over the five evaluated

foreground classes: impervious surfaces, buildings, low vegetation, trees, and cars. Background is excluded from the reported mIoU and mF1 under the adopted ISPRS evaluation protocol. The operational output considered in this paper is therefore a semantic map or a task result, rather than a reconstructed image bitstream.

The model compression objective separates ground training from embedded inference. A frozen diffusion teacher T supplies supervision during training, whereas the deployed path retains only a compact student S and its task head h_{task} . Accordingly, this study evaluates semantic quality, model and computational cost, embedded execution efficiency, and robustness. It does not measure an orbital communication link, bit rate, or reconstruction quality, so no numerical downlink-saving claim is made.

2.2 Deployment evaluation dimensions

The deployment-oriented assessment reports measured quantities rather than nominal hardware capability. Semantic quality is paired with static complexity, runtime cost, energy efficiency, and robustness so that a reduction in model size is not treated as sufficient evidence of deployability by itself. Table 1 summarizes the reported quantities and their roles in the assessment.

In this study, “lightweight” does not refer to a fixed parameter-count threshold. Instead, it describes a task and platform-dependent balance between semantic performance and resource efficiency. We assess it relative to the teacher model in terms of parameter count, FLOPs, model size, runtime memory, latency, and energy consumption under the specified deployment conditions.

2.3 Joint training objective

Let T denote the frozen diffusion teacher, S the student representation network, h_{task} the lightweight prediction head, and $\mathbf{g}$ the semantic label. The upstream representation study uses cross-entropy with focal supervision to address class imbalance. In the subsequent model-compression stage, supervised task learning is combined with feature, attention, and response-level transfer from T to S :

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{feat}} + \beta \mathcal{L}_{\text{attn}} + \gamma \mathcal{L}_{\text{resp}}. \quad (1)$$

Here, $\mathcal{L}_{\text{CE}}$ is the supervised segmentation loss, $\mathcal{L}_{\text{feat}}$ aligns intermediate features, $\mathcal{L}_{\text{attn}}$ aligns attention responses, and $\mathcal{L}_{\text{resp}}$ aligns output distributions. The reported validation study selects $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 0.1$. Section 3.3 gives the mathematical definitions, and Section 5.4 reports the corresponding weight-

selection and component ablations. The teacher T remains frozen, and only S and h_{task} are retained after training. ONNX/TensorRT conversion and FP16/INT8 optimization are post-training deployment operations and do not alter Equation (1).

3. DIFFUSION-GUIDED COMPRESSION

3.1 Frozen diffusion teacher

The teacher is the frozen U-Net feature extractor of a pre-trained diffusion model [24, 25]. Its parameters are unchanged throughout representation learning and knowledge distillation. The NIR, red, and green bands are stacked in this order to form the three-channel input; no blue-band information is used. A normalized 256×256 NIR-R-G tile $\mathbf{x}_0$ is perturbed with the DDPM forward-noising operation, and the resulting $\mathbf{x}_t$ and timestep embedding are passed through the U-Net. The final noise-prediction output is discarded; only intermediate decoder activations are used as semantic supervision.

The direct teacher representation experiments in sections 5.2 and 5.3 use a fixed selected timestep and a broader decoder feature pyramid. Activations at decoder block boundaries (layers 4, 8, 12, and 16) are bilinearly resized to 64×64 and concatenated into an 8448-channel per-pixel descriptor. To control memory, the downstream pixel-wise head processes this descriptor in chunks of 2048 pixels; the 256×256 input tile and its teacher feature extraction remain intact. The direct teacher representation pipeline is illustrated in Fig. 1.

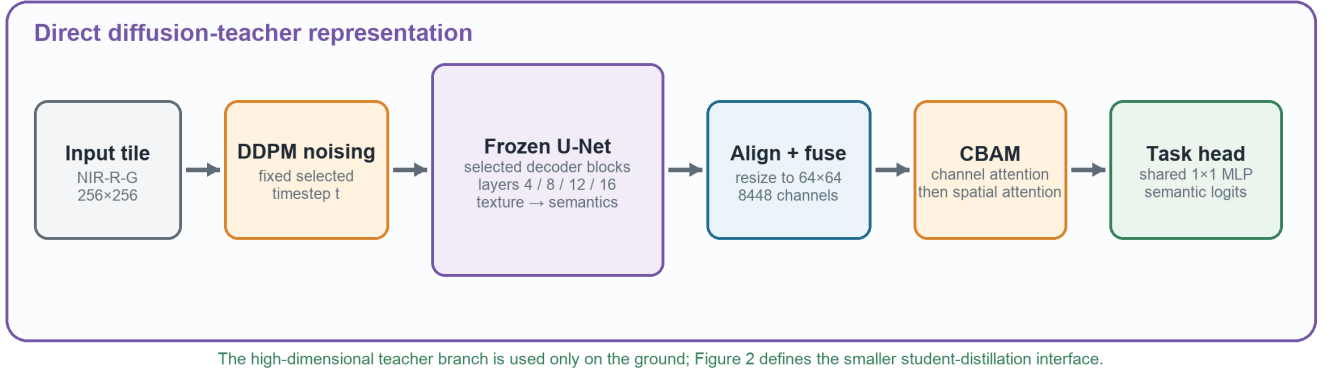
The student distillation interface is deliberately smaller. For each distillation batch, t is sampled uniformly from $\{50, 100, 150\}$, and the teacher exposes decoder features at 64×64 , 32×32 , and 16×16 . These three tensors provide boundary, local structure, and semantic-context targets, respectively. Neither the diffusion U-Net nor any timestep-conditioned teacher operation is retained in the deployed student.

3.2 Task-aware feature aggregation

For the direct representation branch, the selected decoder activations are bilinearly resized to a common 64×64 grid and concatenated along the channel dimension. CBAM [26] then performs sequential channel and spatial recalibration. Channel attention is obtained from shared-MLP transforms of global average and max-pooled descriptors; spatial attention is obtained by applying a 7×7 convolution and a sigmoid to the concatenated channel-wise average and maximum maps. The two masks are multiplied successively with the fused tensor, emphasizing task-relevant channels and regions while suppressing redundant activations.

Table 1 – Evaluation dimensions used for deployment-oriented assessment.

Evaluation dimension	Reported quantity	Role in the assessment
Semantic quality	mIoU and mF1	Task utility over five foreground classes
Static complexity	Parameters, model size, and FLOPs	Storage demand and theoretical workload
Runtime performance	FPS, latency, and peak GPU memory	Measured embedded execution cost
Energy efficiency	Power and FPS/W	Performance under the 15 W power profile
Robustness	mIoU under haze, shadow, and sensor noise	Stability under non-ideal imagery

**Figure 1** – Direct diffusion teacher representation used for ground-side training and evaluation. Selected decoder activations are resized, concatenated, recalibrated by CBAM, and mapped to semantic predictions by a lightweight task head.

The enhanced tensor is mapped by a lightweight pixel-wise head implemented as shared 1×1 projections, equivalent to applying an MLP at every spatial location. Hidden projections use batch normalization and ReLU, and the final projection produces $C = 6$ logits for the ISPRS labels, including background/clutter. Bilinear upsampling restores the input-tile resolution. Training losses use all six logits, whereas mIoU and mF1 average only the five foreground classes. Let $\mathbf{p}_T = h_{\text{task}}(T(\mathbf{x}))$ denote the teacher-derived semantic logits. The representation objective is

$$\mathcal{L}_{\text{rep}} = \mathcal{L}_{\text{CE}}(\mathbf{p}_T, \mathbf{g}) + \eta \mathcal{L}_{\text{focal}}(\mathbf{p}_T, \mathbf{g}). \quad (2)$$

In Equation (2), $\mathcal{L}_{\text{CE}}$ provides stable pixel-wise supervision and $\mathcal{L}_{\text{focal}}$ uses focusing parameter 2 to reduce the influence of easy majority class pixels. The coefficient η is searched over $\{0.1, 0.2, \dots, 0.9\}$; $\eta = 0.3$ gives the highest validation mIoU in the representation study. The subsequent student compression stage uses the joint objective in Equation (1). Background is excluded from the reported five-class mIoU and mF1 metrics.

3.3 Multi-granularity distillation

Three complementary losses transfer the frozen teacher representation to the student. Let $\mathcal{T} = \{50, 100, 150\}$ denote the teacher timesteps and $\mathcal{S} = \{64, 32, 16\}$ the aligned spatial scales. For each scale ℓ , a trainable 1×1 projection A_ℓ maps the student feature F_S^ℓ to the teacher's channel dimension. The feature-alignment loss averages the squared error over a uniformly sampled timestep, the three scales, and N_ℓ tensor elements, preventing the largest feature map from dominating optimization:

$$\mathcal{L}_{\text{feat}} = \mathbb{E}_{t \sim \mathcal{U}(\mathcal{T})} \left[\frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \frac{1}{N_\ell} \times \left\| A_\ell(F_S^\ell) - F_T^{t,\ell} \right\|_2^2 \right]. \quad (3)$$

Second, attention transfer matches the channel and spatial responses produced by CBAM [26] at each aligned scale. W_c and M_s denote the channel-weight vector and spatial attention map, respectively; ψ_ℓ adapts a student channel-weight vector when its dimension differs from the teacher. The attention loss therefore transfers where and through which channels the teacher concentrates, rather than merely forcing raw feature equality:

$$\mathcal{L}_{\text{attn}} = \mathbb{E}_{t \sim \mathcal{U}(\mathcal{T})} \left[\frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \left(\left\| W_{c,T}^{t,\ell} - \psi_{\ell}(W_{c,S}^{\ell}) \right\|_2^2 + \left\| M_{s,T}^{t,\ell} - M_{s,S}^{\ell} \right\|_2^2 \right) \right]. \quad (4)$$

Third, response distillation [9] aligns teacher and student pixel-wise logits using temperature-scaled Kullback–Leibler divergence:

$$\mathcal{L}_{\text{resp}} = \mathbb{E}_{t \sim \mathcal{U}(\mathcal{T})} \left[\tau^2 D_{\text{KL}} \left(\sigma \left(\frac{\mathbf{z}_T^t}{\tau} \right) \parallel \sigma \left(\frac{\mathbf{z}_S}{\tau} \right) \right) \right]. \quad (5)$$

In Equation (5), σ denotes Softmax, $\mathbf{z}_T^t$ and $\mathbf{z}_S$ are the corresponding six-class logits, and $\tau > 1$ is the response-softening temperature; the factor τ^2 preserves the gradient scale. Equations (3)–(5) are expectations over the uniform timestep distribution, estimated in training with one sampled timestep per batch.

The coefficient notation is used consistently with Table 6: the coefficient of $\mathcal{L}_{\text{CE}}$ is fixed at 1.0, α weights feature alignment, β weights attention transfer, and γ weights response distillation. Validation selects $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 0.1$. Teacher features and logits are stop-gradient targets, and only the student, its 1×1 adapters, CBAM modules, and prediction head are updated.

Training uses two stages. The student is first warmed up with $\mathcal{L}_{\text{CE}}$ only and is then optimized with the complete objective in Equation (1). The implementation uses Adam with an initial learning rate of 1×10^{-4} , cosine decay to 1×10^{-6} , batch size 8, automatic mixed precision, at most 100 epochs, and validation early stopping with patience 10. During the distillation stage, one teacher timestep is sampled per batch and all three spatial scales are supervised jointly. The complete multi-granularity distillation framework is shown in Fig. 2.

3.4 Student architectures

All students expose features at 64×64 , 32×32 , and 16×16 for alignment with the teacher. After CBAM processing, the three outputs are resized to 64×64 , concatenated, projected to 256 channels, and mapped to C classes by a final convolution. This common interface keeps the distillation targets and task head unchanged across architectures.

Small-UNet is the lowest-resource option based on the U-Net topology [27]. It retains a four-stage encoder–decoder with skip connections; encoder widths are 32, 64, 128,

and 256 channels, followed by a 512-channel bottleneck. CBAM is attached to the three decoder outputs used for distillation. The resulting model has 7.8 M parameters.

ResNet-34 U-Net uses an ImageNet-pretrained ResNet-34 encoder [28] and a symmetric decoder. Features from encoder stages 2–4 enter the corresponding skip connections, and the decoder produces the three aligned resolutions. With 22.1 M parameters, it is the selected accuracy–efficiency operating point for embedded evaluation.

ResNet-50 DeepLabV3+ [29] represents a heterogeneous student. A channel-adapted stage-2 feature supplies the 64×64 level, the atrous spatial pyramid pooling output supplies the 16×16 level, and an auxiliary upsampling path produces the 32×32 level. CBAM follows each output. This configuration has 27.0 M parameters and tests whether the same distillation losses transfer to a non-symmetric topology.

3.5 Model conversion and quantization

After distillation, the selected ResNet-34 U-Net is exported from PyTorch to ONNX with a fixed 256×256 input and batch size 1. ONNX Runtime provides a framework-independent graph and applies static graph optimizations. TensorRT [30] then builds FP16 and INT8 engines: FP16 reduces storage and memory traffic, while INT8 uses calibration-based dynamic-range quantization on representative 256×256 calibration tiles.

This conversion is post-training optimization rather than quantization-aware training. Consequently, the semantic and distillation objectives are unchanged after model export, and accuracy is remeasured for each backend instead of being assumed invariant. Only the student graph and prediction head are converted; the diffusion teacher, feature adapters used only for supervision, and all distillation losses are absent from the deployed engine.

4. EMBEDDED PRE-VERIFICATION

4.1 Jetson AGX Orin pre-verification

A 64 GB Jetson AGX Orin developer kit is used as a representative embedded AI platform for ground-based algorithmic pre-verification. It is neither radiation-hardened nor flight-qualified. The deployment path retains only the ResNet-34 U-Net student and task head, exports the graph through ONNX, and builds TensorRT FP16/INT8 engines for a fixed 256×256 input with batch size 1. CUDA, ONNX Runtime, TensorRT, and programmable power profiles make the platform suitable for measuring application-level throughput, latency, memory demand,

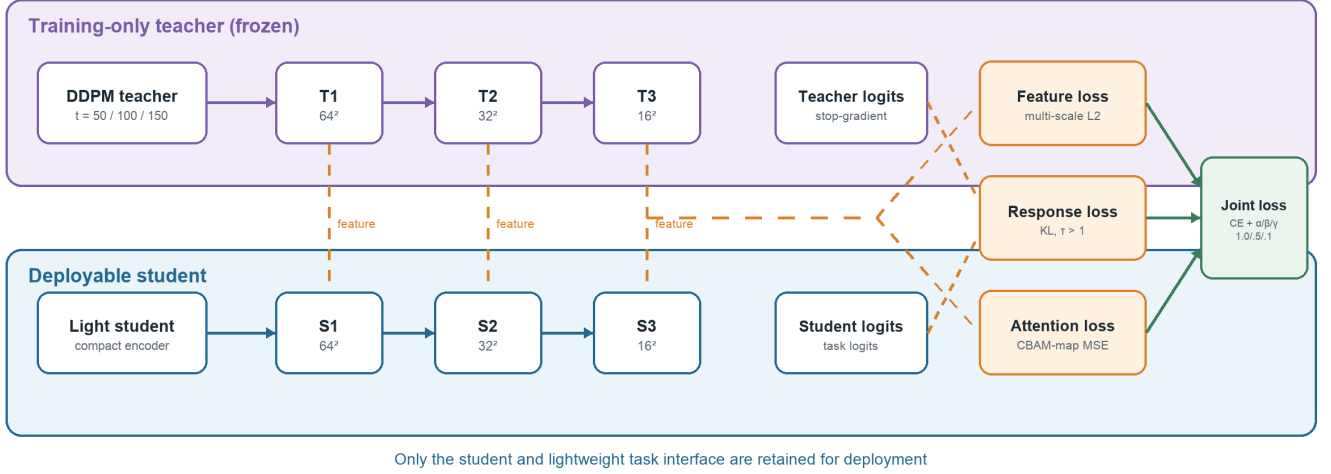


Figure 2 – Multi-granularity diffusion-to-student distillation. For each training batch, the frozen teacher is evaluated at one timestep sampled from {50, 100, 150}; its 64×64 , 32×32 , and 16×16 features supervise the corresponding student stages through feature, attention, and response losses. Only the student branch is retained for deployment.

and system power under controlled laboratory conditions [30].

The role of this platform is deliberately limited. These measurements verify graph portability, operator support, low-precision execution, and the extent to which model compression produces embedded execution gains. They do not demonstrate tolerance to radiation, launch vibration, thermal vacuum cycling, long duration operation, or integration with a spacecraft real-time control and fault-management stack. Accordingly, the Jetson results are pre-verification evidence rather than flight qualification.

4.2 Gap to flight-qualified hardware

Commercial Jetson devices and flight-qualified processors differ in component qualification, radiation and environmental tolerance, fault handling, software assurance, compute architecture, and sustained power and thermal constraints [2, 4, 12]. Peak TOPS and CPU DMIPS describe different operations and cannot be used to infer equivalent application throughput. Even for nominally similar arithmetic capability, an on-board implementation must establish deterministic latency, bounded memory behavior, recovery from faults, and performance over the mission temperature and power envelope.

Migration to flight hardware would therefore require porting the student graph to the selected qualified processor or FPGA, verifying every deployed operator and low-precision kernel, and repeating latency, memory, accuracy, and power measurements on that target. Radiation testing, fault injection, memory protection and watchdog verification, thermal vacuum testing, and long duration stress testing remain separate qualification ac-

tivities. None of these activities is claimed by the present Jetson evaluation. Table 2 summarizes the boundary between the reported pre-verification results and the evidence required for flight deployment.

4.3 Tiled processing and task output

The measured deployment workload consists of preprocessed 256×256 tiles with batch size 1, matching the student-model experiments. The reported FPS and latency therefore describe one tile after data preparation; they do not include full scene tiling, overlap handling, input/output transfer, mosaicking, or mission-level scheduling. End-to-end latency for an original large area scene should not be inferred by treating the per-tile measurement as a complete payload-processing pipeline.

The deployed model produces a six-class semantic map, with five foreground classes used for the reported mIoU and mF1. Such a map could support subsequent on-board screening, alert generation, region selection, or selective downlink. However, the present experiments do not implement an orbital communication link, image or feature bitstream, packetization stage, or rate-task evaluation. The paper therefore makes no numerical claim about downlink savings or communication latency.

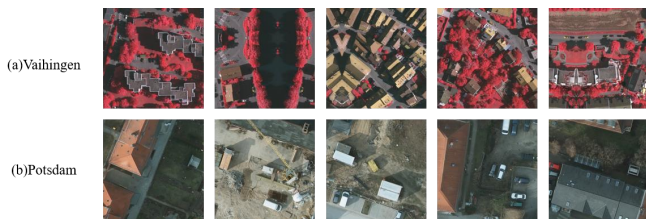
5. EXPERIMENTAL RESULTS

5.1 Datasets and protocols

Experiments use ISPRS Vaihingen and Potsdam. Vaihingen contains 33 NIR-R-G tiles with 9 cm ground sampling distance; Potsdam contains 38 tiles with 5 cm ground

Table 2 – Embedded validation versus flight deployment.

Dimension	Jetson pre-verification	Flight deployment requirement
Platform status	64 GB commercial developer kit	Mission-selected flight-qualified compute module
Radiation and environment	No TID/SEE, vibration, or thermal-vacuum qualification established here	Mission-specific TID/SEE tolerance and environmental qualification
Software and timing	CUDA/ONNX/TensorRT laboratory toolchain	Verified operators, deterministic scheduling, and bounded worst-case latency
Power and thermal	Programmable 15–50 W laboratory profiles	Spacecraft power budget, duty cycle, and qualified thermal envelope
Fault tolerance	Laboratory operation without flight fault testing	Protected memory, redundancy, watchdog, recovery, and fault-injection evidence
Interpretation	Application-level algorithmic efficiency evidence	No direct in-orbit performance or reliability guarantee

**Figure 3** – Representative NIR-R-G samples from Vaihingen (top) and Potsdam (bottom), illustrating differences in scene scale, building regularity, vegetation density, and object size.

sampling distance, of which the NIR-R-G bands are used for a common input configuration. The evaluated foreground classes are impervious surfaces, buildings, low vegetation, trees, and cars; background is excluded from mIoU and mF1. Representative NIR-R-G samples from the two datasets are shown in Fig. 3.

The general representation study uses a 7:2:1 division for the training, validation, and test sets. Inputs are cropped to 256×256 pixels, with random rotation and horizontal flipping during training. The representation models are implemented in PyTorch on an RTX 3090 and optimized with Adam from an initial learning rate of 1×10^{-3} using cosine annealing. Comparison networks use image batch size 1, whereas the proposed pixel-wise task head processes the fused descriptor in 2048-pixel chunks.

The deployment-oriented distillation study uses a fixed Vaihingen partition comprising 16 training tiles, 5 validation tiles, and 12 test tiles; its results are reported separately. Distillation models are trained for up to 100 epochs with Adam, initial learning rate 1×10^{-4} , cosine decay to 1×10^{-6} , batch size 8, automatic mixed precision, and validation early stopping with patience 10.

5.2 Representation component ablation

Table 3 reports the contribution of the diffusion features, CBAM attention, and class imbalance-aware task loss.

Table 3 – Ablation of diffusion features, attention, and task loss on Vaihingen.

Configuration	DDPM	CBAM	Loss	mIoU (%)	mF1 (%)
CNN baseline	No	No	CE	68.03	80.78
Diffusion representation	Yes	No	CE	70.15	82.41
+ attention	Yes	Yes	CE	71.02	83.25
Full representation	Yes	Yes	CE + focal	71.70	83.73

Replacing the CNN representation with frozen DDPM features raises mIoU by 2.12 points. CBAM contributes a further 0.87 points, and the class imbalance-aware task loss adds 0.68 points. The complete configuration improves mIoU and mF1 by 3.67 and 2.95 points, respectively, over the CNN baseline.

5.3 Task accuracy on two benchmarks

Tables 4 and 5 compare the class-wise and overall semantic accuracy of the evaluated methods on Vaihingen and Potsdam, respectively.

Across the two benchmarks, the diffusion-guided representation achieves 71.70% mIoU and 83.73% mF1 on Vaihingen and 76.62% mIoU and 87.78% mF1 on Potsdam. The class-wise results show that the gains are not confined to one dominant land cover category, supporting the use of the frozen generative prior as a general multi-scale representation rather than a dataset-specific output refinement.

5.4 Distillation weights and ablation

Table 7 uses ResNet-34 U-Net student and validation protocol throughout. Feature alignment raises mIoU from 68.24% to 69.30% (+1.06 points). Adding response supervision increases the gain to 2.16 points, and attention transfer contributes the final 0.93-point increment, yielding 71.33% mIoU and 83.21% mF1. In Table 6, increasing the response weight beyond 0.1 reduces validation accuracy, indicating that overly strong output-distribution supervision can interfere with feature learning.

Table 4 – Vaihingen semantic accuracy.

Method	Impervious	Building	Low veg.	Tree	Car	mIoU	mF1
U-Net	79.38	84.73	58.79	72.91	44.35	68.03	80.78
TransUNet	78.35	85.34	57.35	74.21	47.68	68.58	81.25
Swin Transformer	78.71	85.16	63.29	74.37	47.82	69.87	81.58
ST-UNet	77.10	84.89	58.79	72.80	61.10	70.93	82.35
Diffusion representation	78.31	85.03	63.84	74.28	56.95	71.70	83.73

Table 5 – Potsdam semantic accuracy.

Method	Impervious	Building	Low veg.	Tree	Car	mIoU	mF1
U-Net	77.48	85.24	71.56	69.57	73.42	75.45	86.05
TransUNet	75.63	83.43	68.98	65.06	74.37	73.49	84.61
Swin Transformer	73.86	83.58	63.69	56.90	71.23	70.36	83.19
ST-UNet	77.51	85.63	68.96	66.74	72.78	74.32	85.19
Diffusion representation	79.31	87.03	72.34	71.48	72.95	76.62	87.78

Table 6 – Distillation-weight selection on the Vaihingen validation set.

α (feat.)	β (attn.)	γ (resp.)	mIoU	mF1
1.0	0.0	0.0	69.82	81.45
1.0	0.5	0.0	70.67	82.31
1.0	0.5	0.1	71.33	83.21
1.0	0.5	0.3	71.10	82.95
1.0	0.5	0.5	70.78	82.40
0.5	0.5	0.1	70.95	82.77
1.0	1.0	0.1	71.02	82.85

Table 7 – Controlled multi-granularity distillation ablation on the Vaihingen validation set.

Student configuration	mIoU (%)	mF1 (%)	Δ mIoU
CE-only baseline	68.24	81.12	—
+ $\mathcal{L}_{\text{feat}}$	69.30	81.80	+1.06
+ $\mathcal{L}_{\text{feat}}$ + $\mathcal{L}_{\text{resp}}$	70.40	82.50	+2.16
+ $\mathcal{L}_{\text{feat}}$ + $\mathcal{L}_{\text{resp}}$ + $\mathcal{L}_{\text{attn}}$	71.33	83.21	+3.09

5.5 Student accuracy-efficiency trade-off

As shown in Table 8, the ResNet-34 U-Net offers the strongest accuracy–efficiency balance. Relative to the 566.09-M-parameter, 3812.64-GFLOP teacher, it reduces parameters by 96.10% and FLOPs by 98.94%, retains 99.48% of teacher mIoU, and increases RTX 3090 throughput from 1.2 to 32 FPS (26.67 $\times$). Small-UNet provides a lower-resource operating point when throughput and model size are prioritized over peak accuracy. Fig. 4 qualitatively compares the diffusion teacher, the distilled student, and the representative segmentation baselines.

5.6 Embedded conversion and quantization

On Jetson, INT8 reduces model storage by 74.89% and GPU memory by 37.17% relative to the FP32 PyTorch student. Under the Table 9 deployment setting, throughput rises from 7 to 22 FPS (3.14 $\times$). The corresponding mIoU changes from 71.31% to 70.74%, a decrease of 0.57 percentage points and a retention ratio of 99.20%. Table 10 reports latency under the 15 W profile.

5.7 Low-power energy efficiency

Within the same 15 W profile, INT8 execution increases throughput by 3.67 $\times$ and energy efficiency by 2.89 $\times$ relative to FP32 PyTorch; latency decreases by 72.71%. In a separate 30 W measurement, INT8 reaches 32 FPS at 21.8 W, corresponding to 1.47 FPS/W. For this tested workload, the 15 W result of 1.76 FPS/W therefore provides the higher measured energy efficiency, whereas the 30 W profile provides higher throughput.

5.8 Robustness under image degradation

As shown in Table 11, the diffusion-guided representation has the smallest average relative decline. Its mIoU remains 53.64% under haze, 66.94% under shadow, and 50.15% under mixed Poisson–Gaussian sensor noise. This consistency supports the use of generative priors as training time structural regularizers for non-ideal Earth observation imagery. Fig. 5 further compares the prediction stability of the evaluated models under representative image degradations.

Table 8 – Student architecture comparison after distillation (RTX 3090, 256 × 256 input, batch size 1).

Backbone	Baseline mIoU	Distilled mIoU	mF1	Params (M)	FLOPs (G)	FPS
ResNet-34 U-Net	68.24	71.33	83.21	22.1	40.3	32
ResNet-50 DeepLabV3+	67.33	70.83	82.23	27.0	55.0	26
Small-UNet	66.92	68.37	81.03	7.8	12.5	58

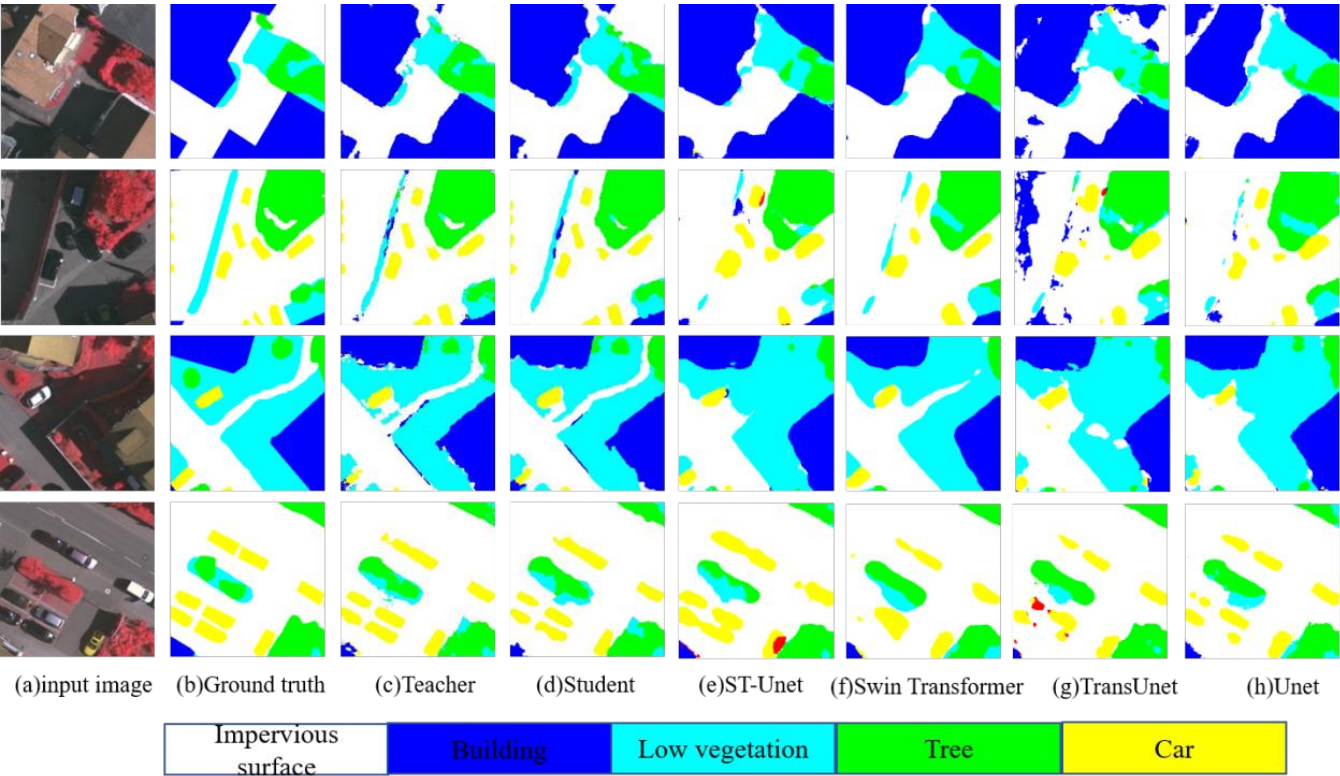


Figure 4 – Qualitative comparison between the teacher, the distilled ResNet-34 U-Net student, and representative baselines. The student retains most teacher boundary and region-level structure at a small fraction of the teacher complexity.

Table 9 – Deployment stages on RTX 3090 and Jetson AGX Orin (256 × 256, batch size 1).

Platform/backend	Precision	FPS	Model (MB)	mIoU (%)	GPU mem. (MB)
RTX 3090 teacher	FP32	1.2	2264	71.70	8500
RTX 3090 PyTorch student	FP32	32	88.4	71.33	1120
RTX 3090 ONNX	FP32	35	88.4	71.31	960
RTX 3090 TensorRT	FP16	55	44.2	71.20	821
RTX 3090 TensorRT	INT8	80	22.1	70.80	672
Jetson PyTorch	FP32	7	88.4	71.31	1380
Jetson ONNX	FP32	9	88.3	71.25	1220
Jetson TensorRT	FP16	14	44.1	71.13	1034
Jetson TensorRT	INT8	22	22.2	70.74	867

Table 10 – Jetson AGX Orin under the 15 W power profile.

Backend	Precision	FPS	Power (W)	FPS/W	Latency (ms)
PyTorch	FP32	6	9.8	0.61	166.7
ONNX Runtime	FP32	8	10.2	0.78	125.0
TensorRT	FP16	12	11.0	1.09	83.3
TensorRT	INT8	22	12.5	1.76	45.5

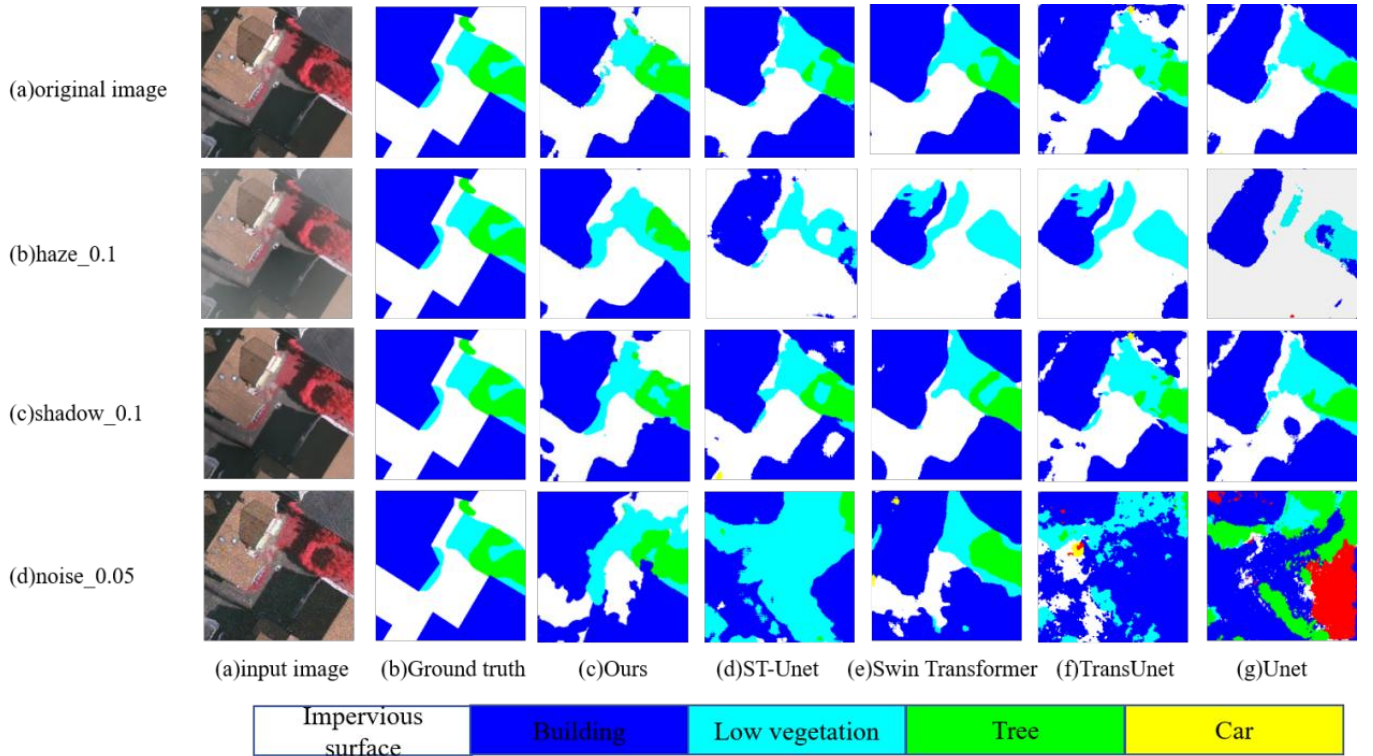
6. DISCUSSION AND CONCLUSION

6.1 Main findings

The experiments show that a high capacity diffusion representation can be compressed into a practical student without reproducing the teacher’s computational burden during embedded inference. Multi-granularity distillation provides a controlled 3.09-point mIoU gain over the CE-only student. The ResNet-34 U-Net uses

Table 11 – mIoU under haze, shadow, and mixed sensor noise.

Method	Clean	Haze	Shadow	Noise	Mean relative drop (%)
U-Net	68.03	26.94	58.75	21.48	49.08
TransUNet	68.58	48.53	56.65	21.32	38.51
Swin Transformer	69.87	22.66	58.31	35.00	44.33
ST-UNet	70.93	32.87	59.82	31.04	41.85
Diffusion representation	71.70	53.64	66.94	50.15	20.62

**Figure 5** – Qualitative robustness comparison under representative disturbances. The diffusion-guided representation preserves larger coherent regions and produces fewer isolated label artifacts than the selected CNN and transformer baselines.

3.90% of the teacher parameters and 1.06% of its FLOPs while preserving 99.48% of teacher mIoU. Post-training INT8 compression further reduces the student to 22.2 MB and reaches 1.76 FPS/W in the Jetson AGX Orin 15 W profile.

The contribution is therefore distinct from conventional reconstructive image coding. The framework compresses semantic knowledge and execution cost for resource-constrained Earth observation inference. Traditional image codecs remain relevant when visually faithful reconstruction is required; the proposed task-oriented path is intended for missions in which masks, alerts, or selected regions carry the primary value. The present evidence establishes this distinction at the algorithmic and embedded pre-verification levels, not through an operational satellite downlink experiment.

6.2 Limitations

First, the experiments focus on optical NIR–R–G imagery and semantic segmentation; multispectral, hyperspectral, and synthetic-aperture-radar validation remains future work. Second, the deployment measurements describe batch 1 inference on preprocessed 256×256 tiles and exclude full scene ingestion, tiling, transfer, mosaicking, scheduling, and communication. No reconstructive bitstream or rate–task curve is evaluated. Finally, Jetson AGX Orin is a commercial pre-verification device rather than flight-qualified hardware; the reported throughput, memory, accuracy, and power measurements are not in-orbit guarantees.

6.3 Future work

Future work will investigate rate-constrained semantic features, adaptive task-result coding, multimodal

Earth observation inputs, and implementation on flight-qualified reconfigurable accelerators. Reliability evaluation should include radiation-aware fault injection, state redundancy, memory scrubbing, and long duration thermal testing. Collaborative inference across satellite-ground and inter-satellite links is also a promising extension when several nodes share partial representations or mission tasks.

6.4 Conclusion

This paper presents a diffusion-guided lightweight semantic model compression framework for on-board Earth observation AI. The method combines frozen diffusion priors, explicit multi-granularity distillation, compact student architectures, and low-bit embedded deployment. Controlled ablations, benchmark evaluations, robustness analysis, and Jetson measurements establish a laboratory pre-verification path from high-capacity ground training to resource-aware embedded inference. Flight-hardware porting, environmental qualification, and in-orbit validation remain necessary before operational space deployment.

NOMENCLATURE

Table 12 – Metric and technical abbreviations.

Abbreviation	Definition
AI	Artificial intelligence
CBAM	Convolutional block attention module
DDPM	De-noising diffusion probabilistic model
FLOPs	Floating-point operations for one inference workload
FPS	Frames or image tiles processed per second
FP16/FP32	16-bit/32-bit floating-point precision
INT8	8-bit integer precision
mF1	Mean class-wise F1 score
mIoU	Mean intersection-over-union over foreground classes
ONNX	Open Neural Network Exchange representation
PTQ	Post-training quantization
SEE	Single-event effect
TID	Total ionizing dose
TOPS	Trillion operations per second; peak hardware capability, not measured application throughput

REFERENCES

[1] National Research Council. *Earth Observations from Space: The First 50 Years of Scientific Achievements*. Washington, DC: National Academies Press, 2008.

[2] L. Diana and P. Dini. "Review on Hardware Devices and Software Techniques Enabling Neural Network Inference Onboard Satellites". In: *Remote Sensing* 16.21, 3957 (2024). doi: [10.3390/rs16213957](#).

[3] G. Guerrisi, F. Del Frate, and G. Schiavon. "AI-Based On-Board Image Compression for Phi-Sat-2". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), pp. 8063–8075.

[4] A. Duggan, B. Andrade, and H. Afli. "Advancing Earth Observation: A Survey on AI-Powered Image Processing in Satellites". In: *European Journal of Remote Sensing* (2025). doi: [10.1080/22797254.2025.2567921](#).

[5] N. Ruan et al. "Edge AI for Earth Observation". In: *IEEE Internet Computing* 29.3 (2025), pp. 31–40.

[6] D. Taubman and M. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Springer, 2002.

[7] M. Hernandez-Cabronero et al. "The CCSDS 123.0-B-2 Compression Standard". In: *IEEE Geoscience and Remote Sensing Magazine* 9.4 (2021), pp. 102–119.

[8] B. Jacob et al. "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 2704–2713.

[9] G. Hinton, O. Vinyals, and J. Dean. "Distilling the Knowledge in a Neural Network". In: *arXiv preprint* (2015). arXiv: [1503.02531](#).

[10] Y. Cheng, D. Wang, P. Zhou, and T. Zhang. "Model Compression and Acceleration for Deep Neural Networks". In: *IEEE Signal Processing Magazine* 35.1 (2018), pp. 126–136.

[11] Y. Guo. "A Survey on Methods and Theories of Quantized Neural Networks". In: *arXiv preprint* (2018). arXiv: [1808.04752](#).

[12] R. Del Prete et al. "Optimizing Deep Learning Models for On-Orbit Deployment through Neural Architecture Search". In: *Scientific Reports* 15 (2025). doi: [10.1038/s41598-025-21467-8](#).

[13] J. Ballé, V. Laparra, and E. P. Simoncelli. "End-to-End Optimized Image Compression". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2017.

[14] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston. "Variational Image Compression with a Scale Hyperprior". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2018.

[15] D. Minnen, J. Ballé, and G. Toderici. "Joint Autoregressive and Hierarchical Priors for Learned Image Compression". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[16] Z. Zhang et al. "COSMIC: Compress Satellite Images Efficiently via Diffusion Compensation". In: *Advances in Neural Information Processing Systems*. Vol. 37. 2024.

[17] D. He et al. "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 (CVPR)*. 2022, pp. 5718–5727.

[18] J. Liu, H. Sun, and J. Katto. "Learned Image Compression with Mixed Transformer-CNN Architectures". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 14388–14397.

[19] Q. Fang, J. Wang, J. Liu, and L. Zhang. "Lightweight Remote Sensing Multispectral Image Compression with Decoupled Interactive Knowledge Distillation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 19 (2026). doi: [10.1109/JSTARS.2026.3674657](#).

[20] C.-H. Huang et al. "Unveiling the Future of Human and Machine Coding: A Survey of End-to-End Learned Image Compression". In: *Entropy* 26.5, 357 (2024).

[21] Y. Xia, Y. Zhou, J. Wang, and B. Chen. "Towards Efficient Low-Rate Image Compression with Frequency-Aware Diffusion Prior Refinement". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 40. 13. 2026, pp. 10942–10950.

- [22] A. Ke et al. "Ultra Lowrate Image Compression with Semantic Residual Coding and Compression-Aware Diffusion". In: *Proceedings of the International Conference on Machine Learning (ICML)*. Vol. 267. Proceedings of Machine Learning Research. 2025.
- [23] R. Feng et al. "Image Coding for Machines with Omnipotent Feature Learning". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2022, pp. 510–528.
- [24] J. Ho, A. Jain, and P. Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 6840–6851.
- [25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10684–10695.
- [26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. "CBAM: Convolutional Block Attention Module". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19.
- [27] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015, pp. 234–241.
- [28] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.
- [29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 801–818.
- [30] NVIDIA. *TensorRT Optimization Guide: Accelerating Deep Learning Inference*. 2024.

AUTHORS



XINYUE LI is the first author. She received her B.E. degree in Electronic Information Engineering from North China Electric Power University in 2024. She is currently pursuing a M.S. degree with the School of Aeronautics and Astronautics, Zhejiang University, Hangzhou, China. Her research interests include deep learning, large models, and remote sensing, with a particular focus on remote sensing change detection.

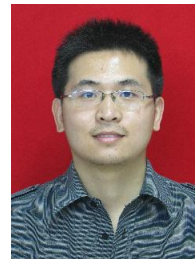


HONGWEI LI received his PhD degree in surveying and mapping science and technology from Tongji University in 2024. Following his doctoral studies, he joined the College of Information Science and Electronic Engineering at Zhejiang University, where he is currently conducting a two-year postdoctoral research program. His research interests span photogrammetry and remote sensing, computer vision, large-scale satellite-borne LiDAR point cloud processing, deep learning and intelligent spatial information

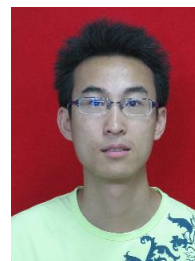
analysis, edge-intelligent algorithms, multimodal data processing, and UAV-based 3D reconstruction. Over the past five years, he has published more than twenty papers in international journals and conferences and holds five granted invention patents.



JINGWEN WANG received her B.E. degree in Communication Engineering from Harbin Institute of Technology in 2023, and master's degree in Electronic Information from Zhejiang University, Hangzhou, China, in 2026. Her research interests include remote sensing vision and diffusion models, with a particular focus on remote sensing image generation and visual interpretation.



CHUNHUI WANG received his bachelor's degree in communication engineering from Zhejiang University in 2004, and his PhD degree in electronic science and technology from Zhejiang University in 2009. After completing a doctorate, he carried out a two-year postdoctoral fellowship at Zhejiang University from 2009 to 2011. In 2011, he joined Zhejiang University as a faculty member, and was promoted to associate research fellow in 2015. Over the past years, he has been engaged in research on micro-satellite TT&C and power technology, as well as satellite formation measurement and communication technology. He has published several SCI/EI-indexed papers and holds 4 granted invention patents.



CHAOJIE ZHANG received his bachelor's degree in communication engineering from Zhejiang University in 2004, and his PhD degree in electronic science and technology from Zhejiang University in 2009. He held a postdoctoral fellowship at Zhejiang University from 2009 to 2012. In 2012, he joined Zhejiang University as a faculty member, and was promoted to associate research fellow in 2016. He was a visiting scholar at Nanyang Technological University from 2012 to 2013. His research interests include software-defined radio technology, satellite communication technology, and miniaturized TT&C transponder technology. He has published over 20 SCI/EI-indexed papers and holds 8 granted invention patents.