Spectrally Distilled Representations Aligned with Instruction-Augmented LLMs for Satellite Imagery

Supplementary Material

The Supplementary Material is organized as follows. Sec. 7 discusses limitations and future work; Sec. 8 reports additional quantitative results on downstream tasks; Sec. 9 provides further qualitative examples; Sec. 10 describes the baselines and downstream datasets; and Sec. 11 details the implementation to facilitate reproducibility.

7. Limitations and Future Work

This work is constrained by two limitations: reliance on optical (RGB) imagery and the cost of text encoding at inference. Our method studies image-text alignment in optical remote sensing and assumes RGB-only inputs at test time. Therefore, sensors such as SAR or thermal are not considered, which can limit robustness under cloud cover, haze, or night conditions. We chose this setting because RGB is the most commonly available modality. A promising extension is to incorporate additional sensors into the teacher-alignment framework so that the RGB pathway inherits cross-modal priors. On the efficiency side, using LLMs as the text encoder increases latency and memory during inference. However, in typical deployments this cost is largely amortized by pre-computing and caching text embeddings for the label vocabulary and prompt sets. In addition, the method is compatible with model compression techniques such as quantization, which can reduce runtime and memory footprint while preserving the accuracy.

8. Additional Quantitative Results

We report F1 scores for zero-shot classification and for the linear-probe setting.

In Tab. 5, we compare zero-shot classification F1 across three satellite benchmarks, namely EuroSAT, BigEarthNet, and ForestNet, grouped by training paradigm (generative vs. contrastive) and input modality (RGB vs. MS). Generative models underperform the best contrastive approaches on all datasets. Among contrastive baselines, remote sensing-aware methods such as GeoRSCLIP, SkyCLIP, and DOFA-CLIP reduce the gap to general-purpose vision-language models, and multi-spectral inputs help in some cases, for example DOFA-CLIP with multi-spectral input on ForestNet. Despite using only RGB inputs, SATtxt with RGB input attains the best results on all three datasets, reaching 73.40 on EuroSAT, 60.18 on BigEarthNet, and 19.89 on ForestNet, and surpassing both RGB and multi-spectral competitors. These results highlight the effectiveness of spectral-aware pre-training even without MS inputs.

Table 5. Zero-shot classification (F1↑) on three satellite benchmarks. "-" denotes an unsupported task or an inapplicable dataset.

Training Approach	Model	Input	Zero-shot Classification		
			EuroSAT	BigEarthNet	ForestNet
Generative	GeoChat	RGB	30.82	14.62	6.15
	EarthDial	MS	57.26	31.80	8.11
Contrastive	CLIP	RGB	40.35	54.85	8.30
	RemoteCLIP	RGB	27.34	52.28	8.50
	GeoRSCLIP	RGB	47.04	58.80	8.27
	SkyCLIP	RGB	48.50	52.88	9.73
	DINOv3txt	RGB	46.29	57.30	14.29
	FT-DINOv3txt	RGB	53.26	58.14	15.37
	DOFA-CLIP	RGB	63.73	58.96	12.33
	DOFA-CLIP	MS	42.18	56.58	15.35
	Llama3-MS-CLIP	MS	64.27	59.63	-
	SATtxt (ours)	RGB	**73.40**	**60.18**	**19.89**

Table 6. Linear-probe (F1 ↑) on EuroSAT, BigEarthNet (10% and 100%), and ForestNet. "-" denotes an inapplicable dataset.

Training Approach	Model	Input	EuroSAT	BigEarthNet		ForestNet
				10%	100%	
MIM	SpectralGPT	MS	95.96	60.70	69.93	-
	Terramind	MS	91.65	65.49	74.08	37.30
Contrastive	CLIP	RGB	91.83	45.32	64.12	24.02
	RemoteCLIP	RGB	95.02	45.40	64.24	18.35
	GeoRSCLIP	RGB	95.76	60.71	70.65	34.74
	SkyCLIP	RGB	93.13	58.21	44.15	16.09
	DINOv3txt	RGB	94.76	26.86	45.28	41.49
	FT-DINOv3txt	RGB	95.46	58.31	48.64	41.13
	DOFA-CLIP	RGB	96.88	59.35	65.59	12.17
	DOFA-CLIP	MS	94.50	66.21	69.41	13.35
	Llama3-MS-CLIP	MS	94.85	68.15	68.04	-
	SATtxt (ours)	RGB	**97.99**	**70.43**	**74.49**	**45.98**

Tab. 6 presents linear-probe F1 on EuroSAT, BigEarthNet with 10% and full supervision, and ForestNet. The comparison includes masked-image-modeling pre-training, represented by SpectralGPT and Terramind, together with contrastive baselines. MIM approaches are competitive, particularly on EuroSAT and BigEarthNet, yet several contrastive models show substantial gains when increasing label fractions. SATtxt with RGB input achieves the best performance in every setting, namely 97.99 on EuroSAT, 70.43 and 74.49 on BigEarthNet with 10% and 100% labels, and 45.98 on ForestNet, outperforming the strongest multi-spectral and RGB baselines including DOFA-CLIP and Llama3-MS-CLIP. Gains in both low-label and full-label regimes indicate that SATtxt learns features that are label efficient and linearly separable, and the strong ForestNet result demonstrates robustness real-world forest categories. Overall, the linear-probe trends align with the zero-

shot findings, showing that spectral-aware contrastive pretraining yields consistently strong and robust representations across multiple optical satellite datasets.

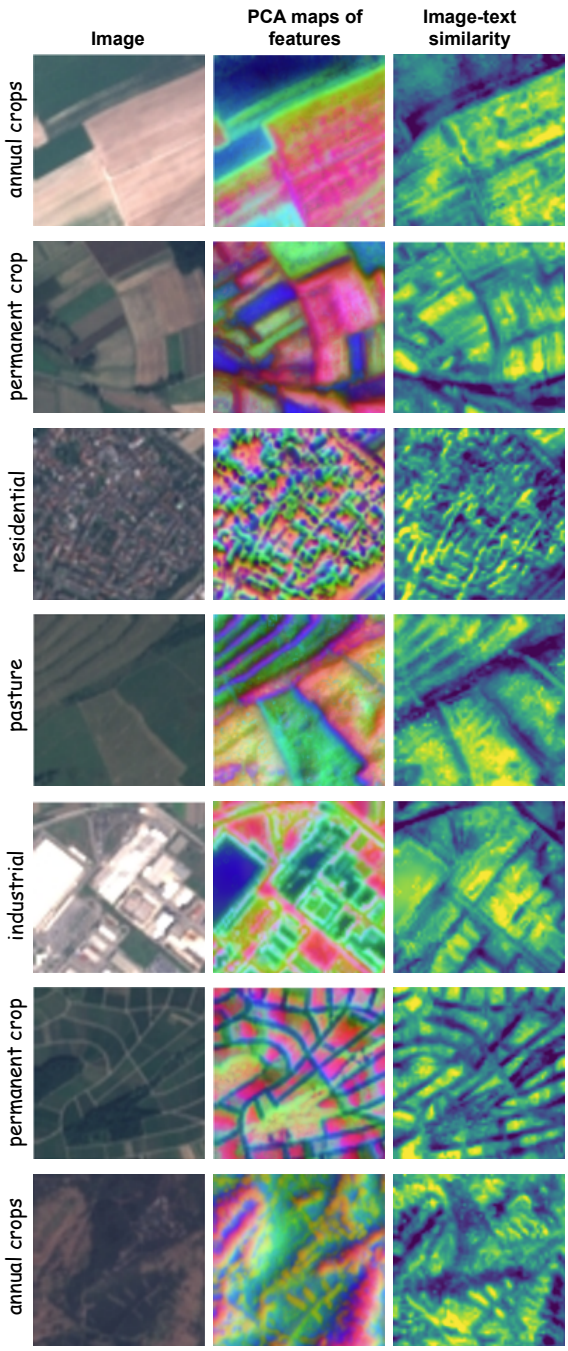

Figure 8. PCA-based feature map visualizations and patch-wise image-text similarity maps with labels.

9. Additional Qualitative Results

We visualize the model's feature maps following the DINOv3 protocol. Using a patch size of 16, we first resize each image to 1024×1024 pixels, resulting in feature maps of spatial resolution 64×64 at the last layer. We then perform principal component analysis (PCA) on these feature maps and project them onto the first three principal components to obtain pseudo-RGB visualizations, which are bilinearly upsampled back to the original image resolution. In addition, we visualize patch-wise image-text similarity with labels. As shown in Fig. 8, the SATtxt produces well-structured PCA feature maps and object-focused image-text similarity maps.

10. Baseline & Dataset Details

Baseline. Tab. 7 summarizes the baseline models evaluated in this study. We include both general-purpose vision-language models and those specifically adapted for remote sensing imagery. All models are configured using their publicly available checkpoints that provide the best performance.

Table 7. Model configurations used in our experiments.

Model	Backbone
CLIP	ViT-L-14
RemoteCLIP	ViT-L-14
GeoRSCLIP	ViT-B-32
SkyCLIP	ViT-L-14
Llama3-MS-CLIP	ViT-B-16
DOFA-CLIP	ViT-L-14
DINOv3txt	ViT-L-16
SATtxt (ours)	ViT-L-16

Dataset. Tab. 8 summarizes the downstream datasets evaluated in this study. We consider four publicly available benchmarks spanning two optical satellite sensors: Sentinel-2 and Landsat-8. For each dataset, we report the number of training/validation/test samples, spectral bands, and classes. Note that BigEarthNet-10% and BigEarthNet-100% share the same test split. EuroSAT includes 13 spectral bands, whereas BigEarthNet includes 12 bands. For multi-spectral inputs on EuroSAT and BigEarthNet, following Clive *et al.* [27], we exclude bands with mismatched spatial resolutions or redundant information [7] and use the 10-band subset {B02 - B08, B8A, B11, B12} to ensure fair and consistent comparisons across multi-spectral models. To assess cross-sensor generalization, we additionally evaluate on ForestNet [15], collected with Landsat-8, which provides 5 spectral bands and 12 classes in a single-label setting.

Table 8. Downstream datasets used in this study.

Details	EuroSAT	BigEarthNet-10%	BigEarthNet-100%	ForestNet
# Training Samples	16,200	25,000	269,695	6,464
# Validation Samples	5,400	10,000	123,723	989
# Test Samples	5,400	125,866	125,866	993
# Spectral Bands	13	12	12	5
# Classes	10	19	19	12

11. Additional Implementation Details

Pre-training Implementation. In the first Stage (SRD), the multi-spectral (MS) teacher (*i.e.*, SpectralGPT) operates at the fixed 128×128 resolution imposed by its pre-trained checkpoint. Accordingly, we set the teacher's global resolution to 128×128 with two views. Following DINOv3, the student receives eight local views at 96×96. We use softmax temperatures of 0.1 for the student and 0.06 for the teacher, with a center momentum of 0.9. For data augmentation, we employ multi-crop augmentation that samples two shared random-resized global crops for RGB and MS plus eight RGB-only local crops, applying view-dependent color jitter, grayscale, blur and solarization to RGB while leaving MS geometrically aligned but photometrically clean, followed by per-modality normalization. Training proceeds for 5 epochs with a batch size of 128 using AdamW, an initial learning rate of 5×10^{-4}, and cosine decay. In the second stage (SGI-LLM), we employ the simple instruction set summarized in Tab. 9. Training runs for 10 epochs with a batch size of 1024 under AdamW optimizer, with an initial learning rate of 4×10^{-5} and cosine decay. Complete configuration details for Stage 1 (SRD) and Stage 2 (SGI-LLM) are provided in Fig. 9 and Fig. 10, respectively.

Linear-probe Implementation. Unlike zero-shot evaluation, linear-probe results can vary across models because of differences in training hyperparameters and preprocessing. To ensure comparability, we adopt the PANGAEA configuration with minimal preprocessing: resizing, selecting the RGB channels for models that accept RGB inputs, and standard mean-std normalization. All linear probes are trained for 30 epochs with a batch size of 128 using AdamW (learning rate 1×10^{-4}, $\beta = (0.9, 0.999)$, weight decay 0.05) and a multi-step learning-rate scheduler.

```yaml
net:
  _target_: models.SRD_module.SRD
  student_encoder_backbone: facebook/dinov3-vitl16-pretrain-sat493m.pth
  teacher_encoder_backbone: danfenghong/IEEE_TPAMI_SpectralGPT/SpectralGPT+.pth
  student_projector: transformer
  student_projector_nlayers: 2
  embed_dim: 1024
  lock_student: true
  lock_teacher: true

criterion:
  _target_: losses.SRDLoss
  feature_dim: 1024
  prototypes_dim: 8192
  T_teacher: 0.06      # teacher temperature
  T_student: 0.10      # student temperature
  center_momentum: 0.9 # EMA momentum for logit center
  eps: 1e-6

data:
  _target_: data.ssl4eos12.Ssl4eos12
  batch_size: 128
  modalities: [S2RGB, S2L2A]
  S2RGB_means: [0.48145466, 0.4578275, 0.40821073]
  S2RGB_stds: [0.26862954, 0.26130258, 0.27577711]
  S2L2A_means: [1924.863, 2184.553, 2340.936, 2671.402,
                3240.082, 3468.412, 3563.244, 3627.704, 3416.714, 2849.625]
  S2L2A_stds: [1201.092, 1219.943, 1397.225, 1400.035,
               1373.136, 1429.17, 1485.025, 1447.836, 1471.002, 1365.307]
  global_size: 128
  n_globals: 2 # number of global views
  local_size: 96
  n_locals: 8 # number of local views

optimizer:
  _target_: torch.optim.AdamW
  lr: 5e-4
  weight_decay: 0.01
  betas: [0.9, 0.999]
  eps: 1e-8

scheduler:
  _target_: schedulers.WarmupCosineWithGroupMultipliers
  warmup_epochs: 1
  max_epochs: 5
  warmup_start_lr: 1e-6
  eta_min: 1e-6
```

Figure 9. Stage 1 (SRD) configuration details.

- "Represent this satellite caption to align with its image"
- "Represent this overhead description for image-text retrieval"
- "Remote sensing caption to match its satellite image"
- "Overhead scene description for image-text alignment"
- "Produce a caption representation suitable for visual search over satellite images"

Table 9. The list of instructions for vision-language alignment.

```yaml
net:
  _target_: models.SATtxt_module.SATtxt_Module
  embed_dim: 2048          # vision embedding dimension
  text_hidden: 4096       # text embedding dimension
  image_backbone: facebook/dinov3-vitl16-pretrain-sat493m.pth
  vision_projector: SRD_pretrained_projector.pt
  text_backbone: McGill-NLP/LLM2Vec-Meta-Llama-3-8B-Instruct-mntp
  text_proj_type: linear
  lock_vision: true       # freeze image backbone
  lock_text: true         # freeze text backbone

criterion:
  _target_: losses.InfoNCELoss
  initial_temperature: 0.07
  learnable_temperature: true
  symmetric_loss: true

data:
  _target_: data.ssl4eos12.Ssl4eos12
  modalities: [S2RGB, precomputed_text_embeddings]
  batch_size: 1024
  channels: 3
  image_size: 224
  S2RGB_means: [0.48145466, 0.4578275, 0.40821073]
  S2RGB_stds: [0.26862954, 0.26130258, 0.27577711]

optimizer:
  _target_: torch.optim.AdamW
  lr: 4e-5
  weight_decay: 0.01
  betas: [0.9, 0.999]
  eps: 1e-8

scheduler:
  _target_: schedulers.LinearWarmupCosineAnnealingLR
  warmup_epochs: 2
  max_epochs: 10
  warmup_start_lr: 1e-5
  eta_min: 1e-6
```

Figure 10. Stage 2 (SGI-LLM) configuration details.