Enhancing Remote Sensing Vision-Language Models for Zero-Shot Scene Classification

Karim El Khoury*[1], Maxime Zanella*[1,2], Benoît Gérin*[1], Tiffanie Godelaine*[1]
Benoît Macq[1], Saïd Mahmoudi[2], Christophe De Vleeschouwer[1], Ismail Ben Ayed[3]

[1]UCLouvain, Belgium [2]UMons, Belgium [3] ETS Montreal, Canada

*Denotes Equal Contribution

#1200

▸ Paper CONTRIBUTION (TL;DR)

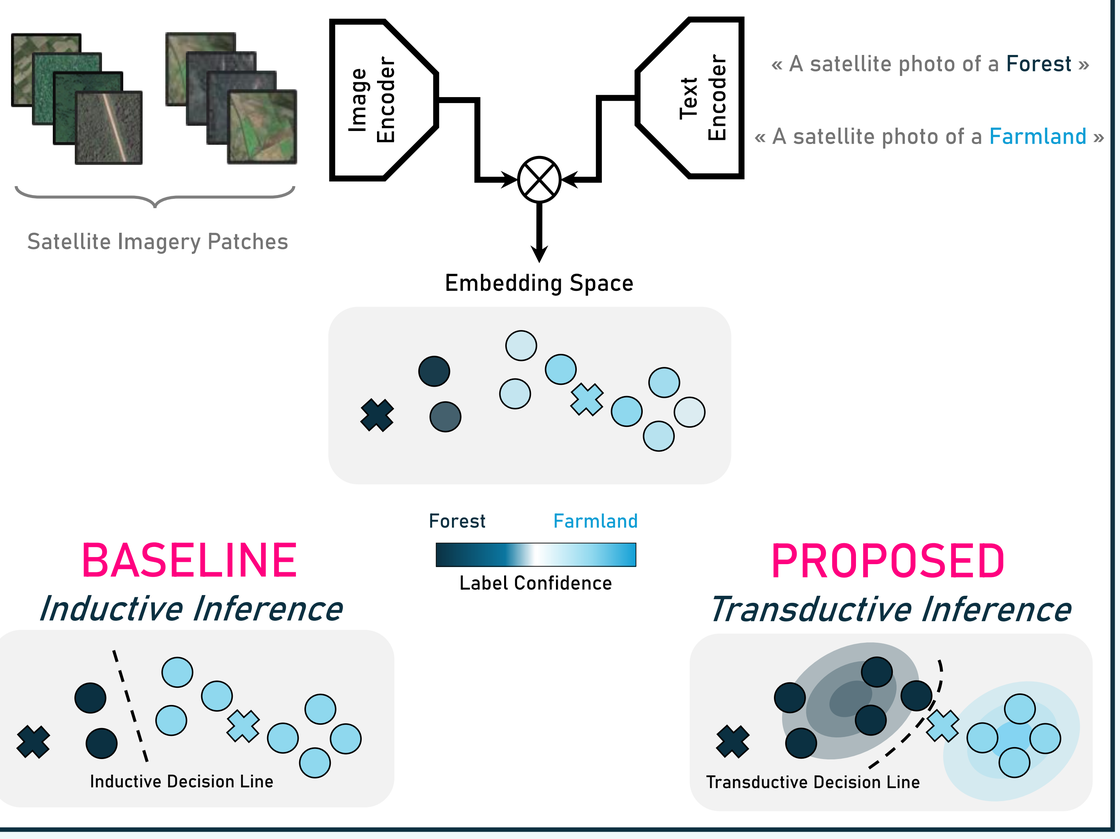
RS-TransCLIP: Unsupervised clustering boosting Remote Sensing Vision-Language Models **by 17%** in only **0.3 seconds**

▸ CONTEXT

◊ **RS-VLMs[1,2,3]** rely on **inductive inference** where each patch is **predicted individually**

◊ We employ **transductive inference[4]** to leverage the data structure within the feature space and **predict all patches simultaneously**

▸ TRANSDUCTIVE VS INDUCTIVE INFERENCE



▸ RS-TRANSCLIP OBJECTIVE FUNCTION

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underbrace{-\frac{1}{|\mathbb{Q}|}\sum_{i \in \mathbb{Q}} \boldsymbol{z}_i^{\mathrm{T}} \log(\boldsymbol{p}_i)}_{\text{(A)}} \underbrace{- \sum_{i \in \mathbb{Q}} \sum_{j \in \mathbb{Q}} \omega_{ij} \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{z}_j}_{\text{(B)}} \underbrace{+ \sum_{I \in \mathbb{Q}} \mathbf{KL}(\boldsymbol{z}_i || \hat{\boldsymbol{y}}_i}_{\text{(C)}}$$

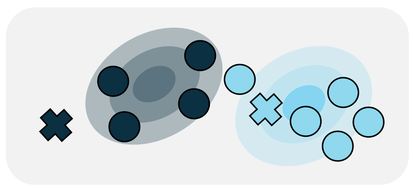
(A) Gaussian Mixture Model Clustering

(B) Laplacian Regularization

(C) Kullback-Leibler Regularization

(A) Gaussian Mixture Model Clustering

→ Model the likelihood of target data as a balanced mixture of Gaussian distributions

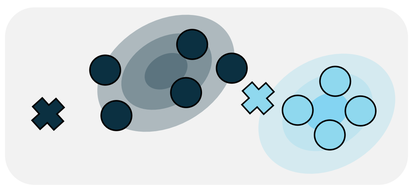


(B) Laplacian Regularization

→ Favor pair of samples with high affinity to have similar assignment variables

(C) Kullback-Leibler Regularization

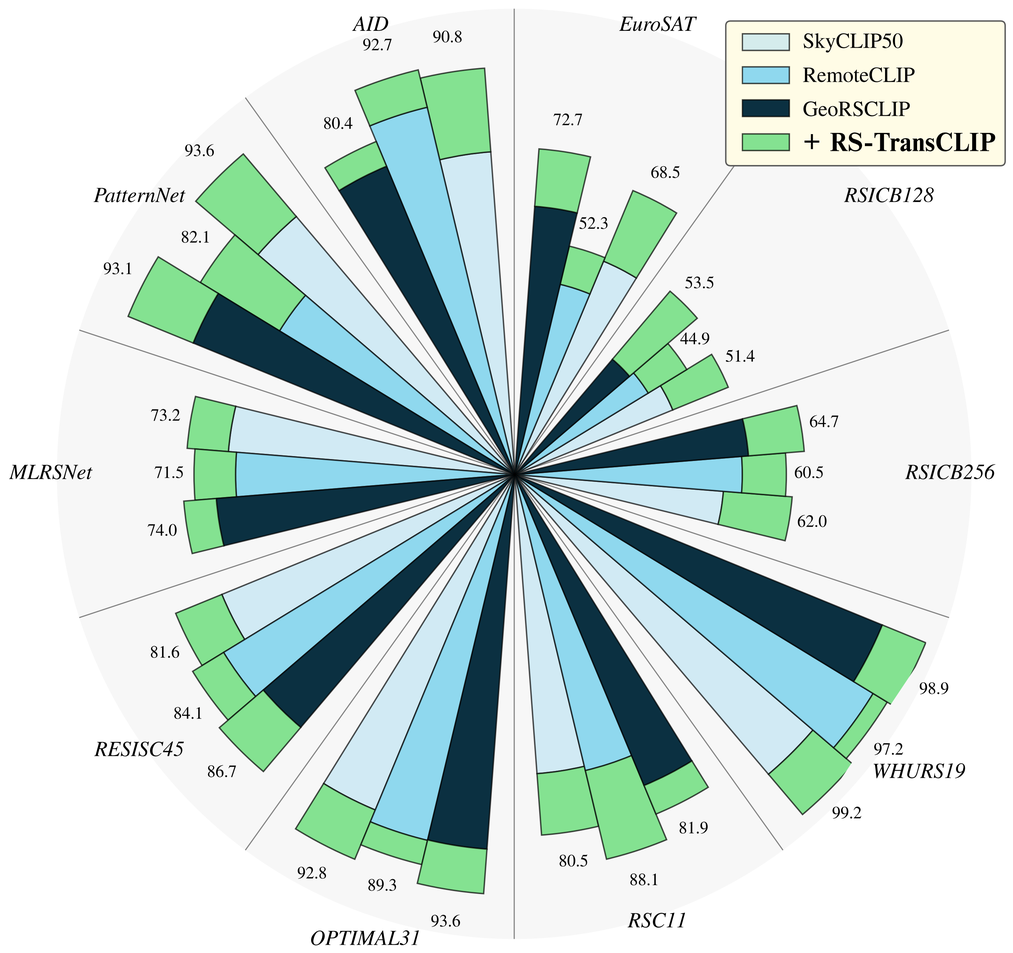
→ Prevent assignment variables to deviate significantly from initial pseudo-labels



▸ RESULTS

➔ RS-TransCLIP boosts zero-shot accuracy

even when inductive performance is high



➔ While maintaining fast inference

thanks to negligeable computational cost

RS dataset	Total patches	Features encoding time	+ RS-TransCLIP time
WHURS19	$\sim 10^3$	$\sim$ 8 seconds	$\sim$ 0.3 seconds
AID	$\sim 10^4$	$\sim$ 40 seconds	$\sim$ 2 seconds
MLRSNet	$\sim 10^5$	$\sim$ 6 minutes	$\sim$ 25 seconds

	Model	AID	EuroSAT	MLRSNet	OPTIMAL31	PatternNet	RESISC45	RSC11	RSICB128	RSICB256	WHURS19	Average
ResNet-50	CLIP	55.4	28.3	45.0	64.5	46.4	52.8	56.7	23.4	30.4	71.3	47.4
	+ RS-TransCLIP	69.6	48.1	54.2	79.6	69.0	69.6	77.8	34.3	46.8	95.9	64.5
	Δ	+14.2	+19.8	+9.3	+15.2	+22.6	+16.7	+21.0	+10.8	+16.4	+24.6	+17.1
	RemoteCLIP	89.1	26.7	43.0	64.0	43.6	51.6	67.0	15.0	36.4	95.4	53.2
	+ RS-TransCLIP	93.3	34.4	58.0	85.0	53.6	72.9	87.2	19.1	48.2	98.4	65.0
	Δ	+4.2	+7.8	+15.0	+21.0	+10.0	+21.2	+20.2	+4.1	+11.8	+3.0	+11.8
ViT-B/32	CLIP	66.4	45.3	51.2	73.0	59.6	60.7	55.5	27.7	40.3	81.1	56.1
	+ RS-TransCLIP	80.7	49.0	64.2	82.9	76.6	74.1	67.0	33.2	46.4	90.3	66.5
	Δ	+14.3	+3.6	+13.0	+9.9	+16.9	+13.4	+11.5	+5.6	+6.0	+9.3	+10.4
	GeoRSCLIP	70.3	53.4	65.0	79.6	75.8	68.8	68.3	29.0	46.5	88.8	64.5
	+ RS-TransCLIP	78.2	69.0	71.9	87.3	94.5	79.5	78.6	42.8	61.8	98.7	76.2
	Δ	+7.9	+15.5	+6.9	+7.7	+18.6	+10.7	+10.3	+13.8	+15.3	+10.0	+11.7
	RemoteCLIP	91.7	35.5	56.3	77.6	55.9	68.1	61.8	26.0	41.5	95.2	61.0
	+ RS-TransCLIP	95.6	51.0	65.8	87.8	70.7	79.4	79.7	31.1	49.2	97.9	70.8
	Δ	+3.9	+15.5	+9.5	+10.3	+14.8	+11.2	+17.9	+5.1	+7.7	+2.7	+9.9
	SkyCLIP50	70.3	52.6	63.2	79.5	73.8	66.7	61.2	39.0	47.1	91.0	64.5
	+ RS-TransCLIP	78.7	64.5	73.2	85.2	87.6	77.3	77.1	49.4	59.1	97.8	75.0
	Δ	+8.3	+11.9	+10.1	+5.8	+13.8	+10.6	+15.9	+10.4	+11.9	+6.8	+10.5
ViT-L/14	CLIP	69.7	60.1	64.1	80.6	74.7	71.3	67.3	37.9	47.2	85.5	65.8
	+ RS-TransCLIP	84.2	71.9	74.5	92.4	91.8	82.2	80.5	43.9	50.5	99.1	77.1
	Δ	+14.4	+11.9	+10.4	+11.7	+17.1	+10.9	+13.2	+5.9	+3.3	+13.6	+11.3
	GeoRSCLIP	74.4	59.9	66.7	83.7	77.4	73.8	75.0	33.7	52.2	88.5	68.5
	+ RS-TransCLIP	80.4	72.7	74.0	93.6	93.1	86.7	81.9	53.5	64.7	98.9	79.9
	Δ	+6.0	+12.8	+7.3	+9.9	+15.7	+12.9	+6.9	+19.9	+12.4	+10.4	+11.4
	RemoteCLIP	84.1	43.6	62.2	83.8	61.4	76.0	67.8	34.8	50.7	93.5	65.8
	+ RS-TransCLIP	92.7	52.3	71.5	89.3	82.1	84.1	88.1	44.9	60.5	97.2	76.3
	Δ	+8.6	+8.7	+9.3	+5.5	+20.7	+8.1	+20.3	+10.1	+9.8	+3.7	+10.5
	SkyCLIP50	72.1	51.5	64.0	80.9	75.3	70.5	66.8	38.0	46.6	87.5	65.3
	+ RS-TransCLIP	90.8	68.5	73.2	92.8	93.6	81.6	80.5	51.4	62.0	99.2	79.4
	Δ	+18.7	+17.0	+9.2	+11.9	+18.3	+11.1	+13.7	+13.4	+15.4	+11.7	+14.0
ViT-H/14	GeoRSCLIP	76.3	68.3	67.4	84.8	82.7	73.8	77.4	43.1	56.5	90.4	72.1
	+ RS-TransCLIP	83.8	91.2	78.1	94.5	96.2	88.0	83.3	54.8	72.8	99.7	84.2
	Δ	+7.5	+22.9	+10.7	+9.7	+13.5	+14.2	+5.9	+11.7	+16.3	+9.3	+12.1

▸ REFERENCES

[1] Liu et al., RemoteCLIP: A vision language foundation model for remote sensing, TGRS 2024

[2] Wang, et al., SkyScript: A large and semantically diverse vision-language dataset for remote sensing, AAAI 2024

[3] Zhang et al., GeoRSCLIP: A large scale vision-language model for remote sensing, TGRS 2024

[4] Zanella, Gérin and Ben Ayed, Boosting Vision-Language Models with Transduction, NeurIPS 2024

UCLouvain UMONS ÉTS TRAIL TRUSTED AI LABS

2025 ICASSP HYDERABAD, INDIA Celebrating Signal Processing