ACM multimedia
Dublin, Ireland 27-31.10.2025

Advancing Complex Wide-Area Scene Understanding with Hierarchical Coresets Selection

Jingyao Wang, Yiming Chen, Lingyu Si*, Changwen Zheng

Introduction

Scene understanding is one of the core tasks in computer vision, aiming to extract semantic information from images to identify objects, scene categories, and their interrelationships. Despite the progress in this field driven by the development of Vision-Language Models (VLMs), existing models still face challenges when adapting to unseen complex wide-area scenes, such as deep-sea scenes. To address this, this paper proposes a Hierarchical Coreset Selection (HCS) mechanism to enhance the adaptability of VLMs in complex wide-area scenes. Based on an importance function with theoretical guarantees, HCS comprehensively considers utility, representativeness, robustness, and synergy to gradually optimize the selected regions. Without additional fine-tuning, HCS can guide VLMs to efficiently understand unseen scenes of any scale using minimal and interpretable regions, while alleviating the problem of insufficient feature density. The proposed method has plug-and-play capability and is compatible with any VLM. Experimental results demonstrate that HCS exhibits superior performance and wide applicability across multiple tasks.

Contributions

1. We explore a more challenging problem of complex wide-area scene understanding, and reformulate it as a coreset feature selection problem, aiming for accurate and stable understanding with fewer interpretable regions.
2. We propose a hierarchical coreset selection mechanism (HCS) for precise wide-area scene understanding. It employs a theoretically validated importance function to assign weights and implements an efficient refinement strategy for coreset selection. HCS is plug-and-play and allows any VLM to achieve training-free understanding with only a few interpretable regions.
3. Extensive experiments on various datasets and tasks demonstrate the effectiveness of HCS on VLMs for scene understanding.

Motivation and Analysis

Existing Vision-Language Models (VLMs) face significant challenges in adapting to complex wide-area scenes. Compared to urban or indoor environments, such scenes—e.g., geographically intricate deep-sea regions—exhibit greater semantic diversity and sparser object distributions. They often contain numerous unknown or low-frequency objects (e.g., rare marine species), resulting in severe long-tail effects, and are dominated by homogeneous, high-frequency background features that obscure critical semantics. Although VLMs possess strong semantic extraction capabilities, their global attention mechanisms tend to prioritize frequent patterns under heterogeneous distributions and missing categories, leading to degraded performance in unseen wide-area scenarios. Furthermore, the high computational cost of training on large-scale, high-resolution data limits their scalability in practice. To address these issues, this paper re-examines scene understanding from a compression-and-selection perspective, aiming to develop an adaptive and interpretable region selection mechanism for efficient VLM-based analysis of wide-area scenes. Toward this goal, we propose a theory-driven Hierarchical Coreset Selection (HCS) framework to support scene understanding.

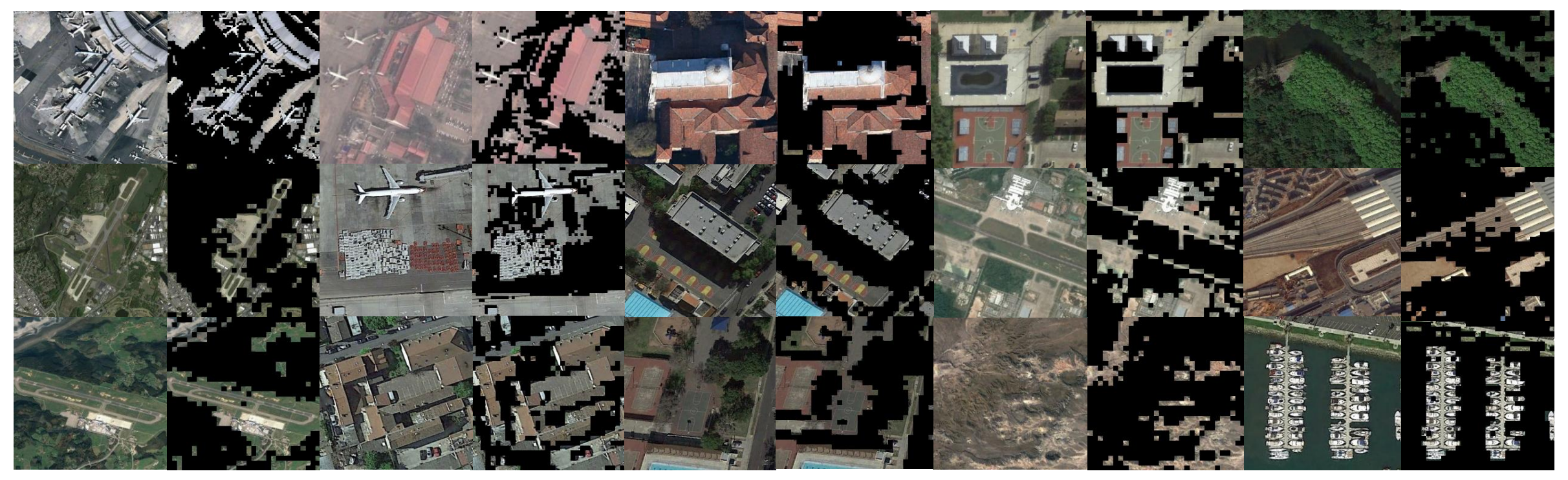

Figure 1: Visualization of samples and region selection.

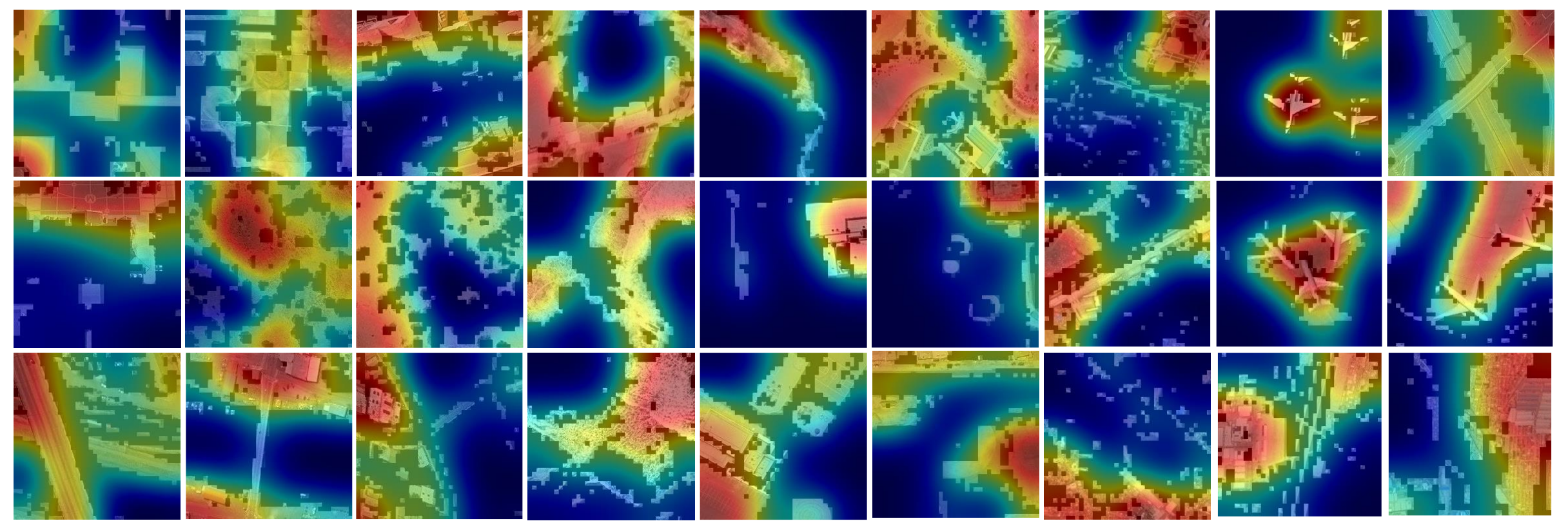

Figure 2: Visualization of interpretable regions.

Methodology

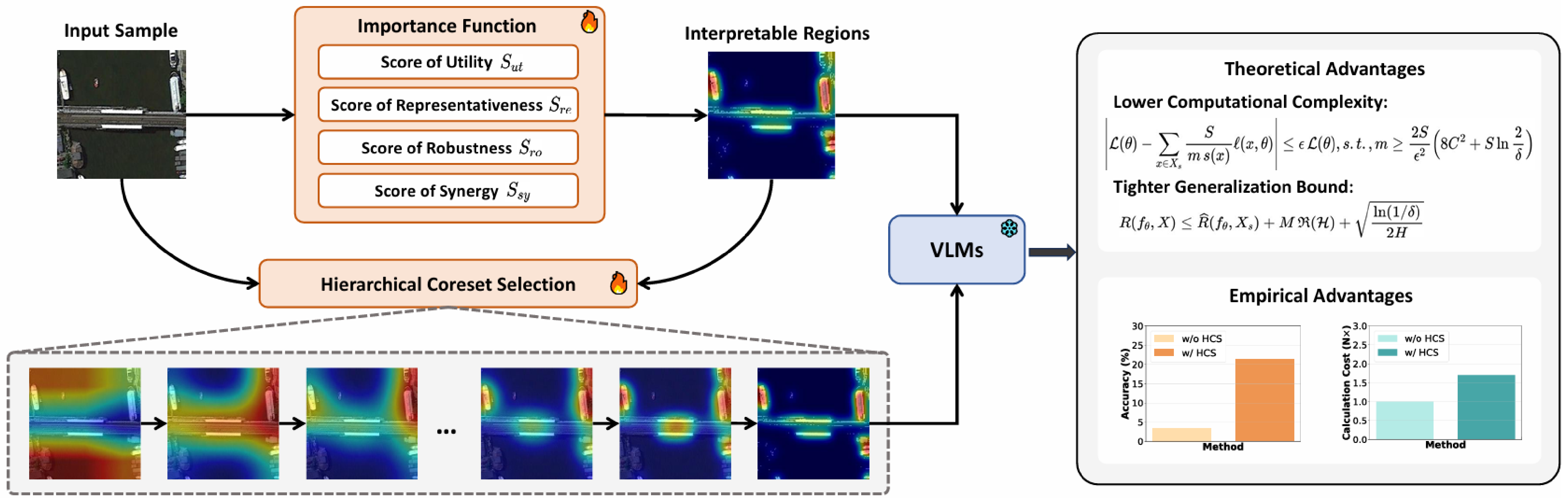
Input Sample
Importance Function
Score of Utility S_{ut}
Score of Representativeness S_{re}
Score of Robustness S_{ro}
Score of Synergy S_{sy}
Interpretable Regions
Hierarchical Coreset Selection
...
VLMs
Theoretical Advantages
Lower Computational Complexity:
$\left|\mathcal{L}(\theta) - \sum_{x\in X_s} \frac{S}{m\,s(x)}\ell(x,\theta)\right| \leq \epsilon\mathcal{L}(\theta), s.t., m \geq \frac{2S}{\epsilon^2}\left(8C^2 + S\ln\frac{2}{\delta}\right)$
Tighter Generalization Bound:
$R(f_\theta, X) \leq \widehat{R}(f_\theta, X_s) + M\,\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta)}{2H}}$
Empirical Advantages
Accuracy (%)
30
25
20
15
10
5
0
w/o HCS
w/ HCS
Method
Calculation Cost (N×)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
w/o HCS
w/ HCS
Method

Figure 3: The framework of the proposed HCS. Its integration allows VLMs, to be tested without fine-tuning (frozen), instead training this lightweight network (HCS) for coreset selection to enhance model performance.

Experimental Resutls

Experimental results on various benchmark datasets, e.g., NWPU-RESISC45, AID, RSI-CB, TikTok dances, TrashCan, and GTEA, demonstrate the effectiveness of the proposed HCS.

Table 1: Performance comparison (Accuracy %) of scene image classification on NWPU-RESISC45, AID, and RSI-CB. Unless otherwise specified, we directly use the pre-trained models without fine-tuning and evaluate their transfer performance. The brackets "()" indicate the effect changes of vanilla baselines after introducing HCS. More details are shown in Appendix E.

Model	NWPU-RESISC45		AID		RSI-CB	
	Top-1 ACC	Top-5 ACC	Top-1 ACC	Top-5 ACC	Top-1 ACC	Top-5 ACC
ViT-B/32	1.84	4.61	0.97	4.63	6.15	15.33
ViT-B/32+HCS	17.15 (+15.31)	28.56 (+23.95)	12.17 (+11.20)	20.61 (+15.98)	19.51 (+13.36)	27.64 (+12.31)
ViT-L/14	2.23	6.85	1.37	5.53	8.89	17.26
ViT-L/14+HCS	19.23 (+17.00)	29.00 (+22.15)	15.12 (+13.75)	21.59 (+16.06)	21.32 (+12.43)	31.88 (+14.62)
CLIP	3.41	12.19	2.16	7.97	12.32	29.17
CLIP+HCS	21.36 (+17.95)	36.78 (+24.59)	18.22 (+16.06)	27.91 (+19.94)	31.39 (+19.07)	47.05 (+17.88)
ContextCLIP	1.48	11.05	1.01	6.25	11.32	25.88
ContextCLIP+HCS	18.56 (+17.08)	33.12 (+22.07)	16.01 (+15.00)	24.37 (+18.12)	29.46 (+18.14)	44.95 (+19.07)
LLaVA-hf/llava-v1.6-mistral-7b-hf	41.20	52.15	35.13	49.36	52.17	59.30
LLaVA-hf/llava-v1.6-mistral-7b-hf+HCS	44.15 (+2.95)	54.48 (+2.33)	40.05 (+4.92)	54.13 (+4.77)	55.66 (+3.49)	65.59 (+6.29)
LLaVA-hf/llama3-llava-next-8b-hf	53.12	64.84	39.65	51.06	58.33	64.89
LLaVA-hf/llama3-llava-next-8b-hf+HCS	58.42 (+5.30)	69.15 (+4.31)	42.05 (+2.40)	54.89 (+3.83)	60.50 (+2.17)	68.36 (+3.47)
Qwen/Qwen2-VL-7B-Instruct	61.15	75.56	46.26	62.03	68.63	76.69
Qwen/Qwen2-VL-7B-Instruct+HCS	66.12 (+4.97)	80.38 (+4.82)	51.04 (+4.78)	68.11 (+6.08)	76.19 (+7.56)	80.97 (+4.28)

Coreset theory

Coreset theory provides a principled framework for data compression by constructing small, weighted subsets of large-scale datasets that approximate the original data with respect to a given optimization objective. By preserving essential geometric or statistical properties, coresets significantly reduce computational complexity, making them particularly effective for large-scale or high-dimensional learning tasks.

References

[1] James A Storer. *Data compression: methods and theory*. Computer Science Press, Inc., 1987.

[2] Jeff M Phillips. Coresets and sketches. In *Handbook of discrete and computational geometry*, pages 1269–1288. Chapman and Hall/CRC, 2017.

[3] Satyan L Devadoss and Joseph O'Rourke. *Discrete and computational geometry*. Princeton university press, 2025.