SOTA: Self-adaptive Optimal Transport for Zero-Shot Classification with Multiple Foundation Models

Supplementary Material

Contents

A. Optimization

We propose a *joint optimization* framework that simultaneously learns the GMM parameters $\Theta = \{\mu_c, \mathbf{\Sigma}_c, \pi_c\}$ and the transport plan $\mathbf{T}$:

$$\max_{\mathbf{T}\in\Pi(\mathbf{p},\mathbf{q}),\Theta} \sum_{v=1}^{V_1} \langle \mathbf{T}, \mathbf{P}_v(\Theta)\rangle^2 + \sum_{v=1}^{V_2} \langle \mathbf{T}, \hat{\mathbf{P}}_v\rangle^2 + \epsilon\,\mathcal{H}(\mathbf{T}), \tag{1}$$

A.1. Optimization procedure

Theorem 1 *Let $f(x) = x^2$ be a convex function and let $x^{(k)} \in \mathbb{R}$ be a given point. Then the first-order Taylor expansion of $f(x)$ at $x^{(k)}$ yields the global lower bound*

$$x^2 \geq (x^{(k)})^2 + 2x^{(k)}\left(x - x^{(k)}\right) = 2x^{(k)}x - (x^{(k)})^2, \tag{2}$$

with equality if and only if $x = x^{(k)}$. Moreover, this affine function serves as a valid minorizer *of $f(x)$ that can be maximized in iterative optimization schemes.*

Let $\mathbf{P}_v^{(k)} = \mathbf{P}_v(\Theta^{(k)})$ be the posteriors computed from the current GMM parameters. We have

$$a_{P,v}^{(k)} := \langle \mathbf{T}^{(k)}, \mathbf{P}_v^{(k)}\rangle, \qquad a_{\hat{P},v}^{(k)} := \langle \mathbf{T}^{(k)}, \hat{\mathbf{P}}_v\rangle, \tag{3}$$

According to Theorem 1, applying tangent minorization to every quadratic term yields the surrogate (up to additive constants independent of $\mathbf{T}$ and Θ):

$$\begin{aligned} G^{(k)}(\mathbf{T}, \Theta) \;=\; & \sum_{v=1}^{V_1} 2a_{P,v}^{(k)} \langle \mathbf{T}, \mathbf{P}_v(\Theta)\rangle \\ & + \sum_{v=1}^{V_2} 2a_{\hat{P},v}^{(k)} \langle \mathbf{T}, \hat{\mathbf{P}}_v\rangle + \epsilon\,\mathcal{H}(\mathbf{T}). \end{aligned} \tag{4}$$

Hence, at iteration k we maximize the surrogate $G^{(k)}$, which is linear in $\mathbf{T}$ for fixed Θ. In practice we optimize $G^{(k)}$ by alternating updates:

- **(Update $\mathbf{T}$ given $\Theta^{(k)}$).** With $\mathbf{P}_v = \mathbf{P}_v(\Theta^{(k)})$ fixed, define adaptive weights

$$\lambda_v^{(k)} := 2a_{P,v}^{(k)}, \qquad \mu_v^{(k)} := 2a_{\hat{P},v}^{(k)}. \tag{5}$$

  The $\mathbf{T}$-subproblem becomes

$$\max_{\mathbf{T}\in\Pi(\mathbf{p},\mathbf{q})} \left\langle \mathbf{T},\ \mathbf{S}^{(k)} \right\rangle + \epsilon\,\mathcal{H}(\mathbf{T}), \tag{6}$$

  where

$$\mathbf{S}^{(k)} := \sum_{v=1}^{V_1} \lambda_v^{(k)} \mathbf{P}_v + \sum_{v=1}^{V_2} \mu_v^{(k)} \hat{\mathbf{P}}_v. \tag{7}$$

  This is a classical entropic-regularized optimal transport problem, which can be efficiently solved using the Sinkhorn algorithm [19]. Concretely, form the kernel $\mathbf{K} = \exp(\mathbf{S}^{(k)}/\epsilon)$ and compute

$$\mathbf{T}^{(k+1)} = \mathrm{Diag}(\mathbf{p}) \cdot \exp\left(\frac{\mathbf{S}^{(k)}}{\epsilon}\right) \cdot \mathrm{Diag}(\mathbf{q}), \tag{8}$$

  where $\mathbf{p} \in \mathbb{R}^K$ and $\mathbf{q} \in \mathbb{R}^N$ are scaled to satisfy the marginals by the usual Sinkhorn updates

$$\mathbf{p}^{(s+1)} = \frac{\mathbf{1}_K}{\mathbf{K}\mathbf{q}^{(s)}}, \tag{9}$$

$$\mathbf{q}^{(s+1)} = \frac{\mathbf{1}_N}{\mathbf{K}^\top \mathbf{p}^{(s+1)}}, \tag{10}$$

  It is worth noting that once $\mathbf{T}$ is updated, we re-update the weight parameters according to Eq. (5).
- **(Update GMM parameters Θ given $\mathbf{T}^{(k+1)}$).** We treat $\mathbf{T}^{(k+1)}$ as soft assignments of samples to classes and perform an M-step update for the GMM parameters. Specifically,

$$\mu_c^{(k+1)} = \frac{\sum_i T_{ic}^{(k+1)} \mathbf{v}_i}{\sum_i Q_{ic}^{(t)}}, \tag{11}$$

$$\mathbf{\Sigma}^{(k+1)} = \frac{1}{N} \sum_i \sum_c T_{ic}^{(k+1)} (\mathbf{v}_i - \mu_c^{(k+1)})(\mathbf{v}_i - \mu_c^{(k+1)})^\top, \tag{12}$$

  where we omit the GMM index for brevity, though these updates are independently applied to each visual model's GMM. After the M-step we recompute the posterior matrices $\mathbf{P}_v(\Theta^{(k+1)})$ (E-step).

The full iterative procedure alternates the two blocks above (See Algorithm 1). Because each $\mathbf{T}$-update maximizes the surrogate $G^{(k)}(\cdot, \Theta^{(k)})$ and the construction of $G^{(k)}$ is a valid minorizer of the original objective 1, the algorithm yields a non-decreasing sequence of objective values and converges to a stationary point.

Algorithm 1 Optimization of **SOTA** via MM

Require: Cost matrices $\{\mathbf{C}_v\}_{v=1}^{V_1+V_2}$, marginal distributions $\mathbf{p}, \mathbf{q}$, visual features $\{\mathbf{v}_v\}_{v=1}^{V_1+V_2}$, entropy weight ϵ, and the number of maximum iteration T.
Ensure: Optimal transport plan $\mathbf{T}^*$

Initialize transport plan $\mathbf{T}^{(0)}$.
Initialize $\Theta^{(0)}, \lambda_v^{(0)}, \mu_v^{(0)}$.
for $k = 1$ to T **do**
  # Step 1: Compute posterior probability.
  for $v = 1$ to $V_1 + V_2$ **do**
    $\mathbf{P}_v^{(k)}(\Theta) \leftarrow (\Theta^{(k)}, \mathbf{v}_v)$ ▷ GMM E-step
  end for
  # Step 2: Update transport plan.
  $\mathbf{T}^{(k+1)} = \mathrm{Diag}(\mathbf{p}) \cdot \exp\left(\frac{\mathbf{S}^{(k)}}{\epsilon}\right) \cdot \mathrm{Diag}(\mathbf{q})$
  ▷ Eq.(8)
  # Step 3: Update adaptive weights.
  $\lambda_v^{(k+1)} \leftarrow (\mathbf{T}^{(k+1)}, \mathbf{P}_v^{(k)})$
  $\mu_v^{(k+1)} \leftarrow (\mathbf{T}^{(k+1)}, \hat{\mathbf{P}}_v^{(k)})$ ▷ Eq.(5)
  # Step 4: Update GMM.
  $\Theta^{(k+1)} \leftarrow (\mathbf{T}^{(k+1)}, \{\mathbf{v}_v\}_{v=1}^{V_1+V_2})$ ▷ Eq.(11) and Eq.(12)
end for
return $\mathbf{T}^* \leftarrow \mathbf{T}^{(k)}$

A.2. Discussion

The proposed optimization framework leverages the MM principle to address the nonlinear coupling between the transport plan $\mathbf{T}$ and the GMM parameters Θ in the joint objective (1). Importantly, the MM reformulation induces a *self-adaptive weighting* of different foundation models. Specifically, the coefficients

$$\lambda_{P,v}^{(k)} = 2\langle \mathbf{T}^{(k)}, \mathbf{P}_v^{(k)} \rangle, \quad \mu_{\hat{P},v}^{(k)} = 2\langle \mathbf{T}^{(k)}, \hat{\mathbf{P}}_v^{(k)} \rangle$$

are updated at each iteration based on the current transport cost for each model. Models with lower transport cost, indicating stronger semantic alignment, receive larger weights in the subsequent OT step, thereby exerting greater influence on the updated transport plan. This adaptive mechanism eliminates the need for manual weight tuning. Furthermore, the coupling between $\mathbf{T}$ and Θ forms a closed-loop refinement: the updated $\mathbf{T}$ yields semantically informed soft assignments that guide the GMM parameter updates, while the refined Θ produces posterior matrices $\mathbf{P}_v(\Theta)$ that reshape the cost structure in the OT problem. Over iterations, this synergy progressively improves alignment quality across heterogeneous foundation models.

B. Experimental Setup.

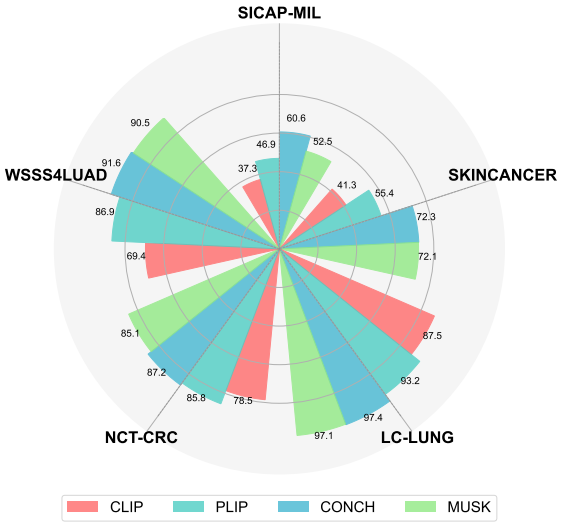
B.1. Datasets.

The natural datasets include: `ImageNet` [5], `SUN397` [29], `Aircraft` [16], `EuroSat` [8], `StanfordCars` [11], `Food101` [2], `Pets` [18], `Flower102` [17], `Caltech101` [6], `DTD` [4], and `UCF101` [23]. The remote sensing datasets include: `AID` [27], `EuroSAT` [8], `MLRSNet` [20], `OPTIMAL31` [24], `PatternNet` [32], `RESISC45` [3], `RSC11` [31], `RSICB128` [13], `RSICB256` [13], and `WHURS19` [26]. The medical datasets include: `SICAP-MIL` [22], `SKINCANCER` [12], `LC-LUNG` [1], `NCT-CRC` [10], `WSSS4LUDA` [7]. These datasets encompass a broad spectrum of image classification, allowing us to comprehensively assess the robustness and generalization capability of our method across distinct domains. The specific dataset templates can be found in the code.

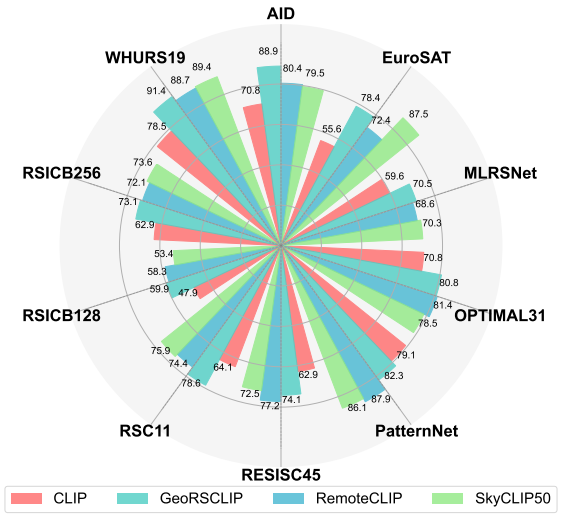
B.2. Implementation details.

Models. Our framework is built upon publicly available implementations of both vision-language models (VLMs) and vision foundation models (VFMs). For natural datasets, we employ CLIP ViT-B/16 and CLIP Resnet-50, released by OpenAI, as the primary VLMs without any additional fine-tuning. For extracting visual priors, we adopt DINOv2 ViT-L/14 and DINOv3 ViT-L/16, chosen for their strong clustering capabilities. For remote sensing datasets, we use CLIP [21], RemoteCLIP [14], SkyCLIP [25] and GeoRSCLIP [30] as VLMs. For medical pathology datasets, we use CLIP [21], CONCH [15], PLIP [9], and MUSK [28] as VLMs. Additionally, the image encoder from VLMs is reused as an auxiliary visual encoder.

Process. During inference, each image is processed using a single center-cropped view of size 224×224 to reduce computational overhead. Class names are embedded using a fixed prompt template, with no prompt ensembling or optimization. We set the temperature parameter $\epsilon = 0.01$ to control entropy regularization. During the initialization phase, we initialize both for V_1 and V_2 with equal values, subject to the normalization constraint $\sum_{v=1}^{V_1} \lambda_v = 1$ and $\sum_{v=1}^{V_2} \mu_v = 1$. In the **Transductive** setting, our method is applied directly to test data. In the inductive setting, we first execute our algorithm on the training data to learn both the GMM parameters corresponding to different visual models and weight coefficients. The test data are then processed by obtaining the posterior probabilities of GMM and inte-



(a) Medical datasets



(b) Remote sensing datasets

Figure 1. Clustering accuracy comparison of visual features extracted from different foundation models on medical pathology and remote sensing datasets.

grating multiple outputs using the learned weight coefficients to produce the final prediction. All models operate in inference-only mode without gradient updates. Experiments are conducted on a single NVIDIA RTX 4090 GPU with 24GB of memory.

C. Extended Discussions

C.1. How important is VFM introduction?

To further highlight the importance of visual features, we conduct analyses on remote sensing and medical image datasets by extracting features from the visual branches of different vision-language models. As shown in Fig. 1, the clustering accuracy of visual features is consistently and substantially higher than the corresponding baseline models' zeroshot performance. In particular, although CONCH achieves the highest clustering accuracy across all datasets, it is nevertheless outperformed by MUSK on the zero-shot task. This observation reveals an inherent modality gap in vision-language models: a naïve strategy that assigns categories solely based on the similarity between sample visual features and candidate class prototypes proves ineffective. Therefore, incorporating high-quality visual features as a complementary source of information is crucial for improving overall performance.

C.2. Qualitative Visualization and Analysis

To intuitively compare the differences between visual features from VFMs and outputs from VLMs, we employed t-SNE to visualize output of different models on the `EuroSat` and `Food101`. As shown in Fig. 2, the experiment included four types of output sources: **nCLIP** (Visual representations from CLIP visual encoder), **DINOv3** (Visual representations from DINOv3), **nLogits** (Outputs from raw CLIP), and **SOTA** (Outputs from raw CLIP optimized with SOTA). The results reveal that while the visual features (e.g., from CLIP and DINOv3) show clearer and more compact clusters, the zero-shot logits from CLIP remain poorly separated due to the strong modality misalignment between visual and textual spaces. Among the visual encoders, DINOv3 exhibits superior feature organization and stronger inter-class separation than CLIP, highlighting its advantage in capturing discriminative structures. These observations confirm that our adaptive optimal transport alignment effectively exploits the structured visual information from VFMs to calibrate and refine the multimodal logits, thereby enhancing the zero-shot classification performance of VLMs.

C.3. Computational efficiency analysis

Our ensemble inference is iterative, alternating between GMM parameter updates and OT. With diagonal covariances, one EM update costs $O(VNKD)$, where V is the number of VFMs, N is the number of target samples, K is the number of Gaussian components, and D is the feature dimension. The entropic OT is solved by Sinkhorn with cost $O(T_{\text{SK}}NK)$, where T_{SK} is the number of Sinkhorn iterations. Overall, the per-iteration computational complexity is $O(VNKD + T_{\text{SK}}NK)$. Tab. 1 compares representative methods. Compared with graph-based approaches such as ECALP and TransCLIP, which typically incur higher computation and memory overhead due to graph construction and inference, our method achieves competitive efficiency while delivering stronger accuracy. In contrast, ADAPT is more efficient but exhibits inferior accuracy than ours,

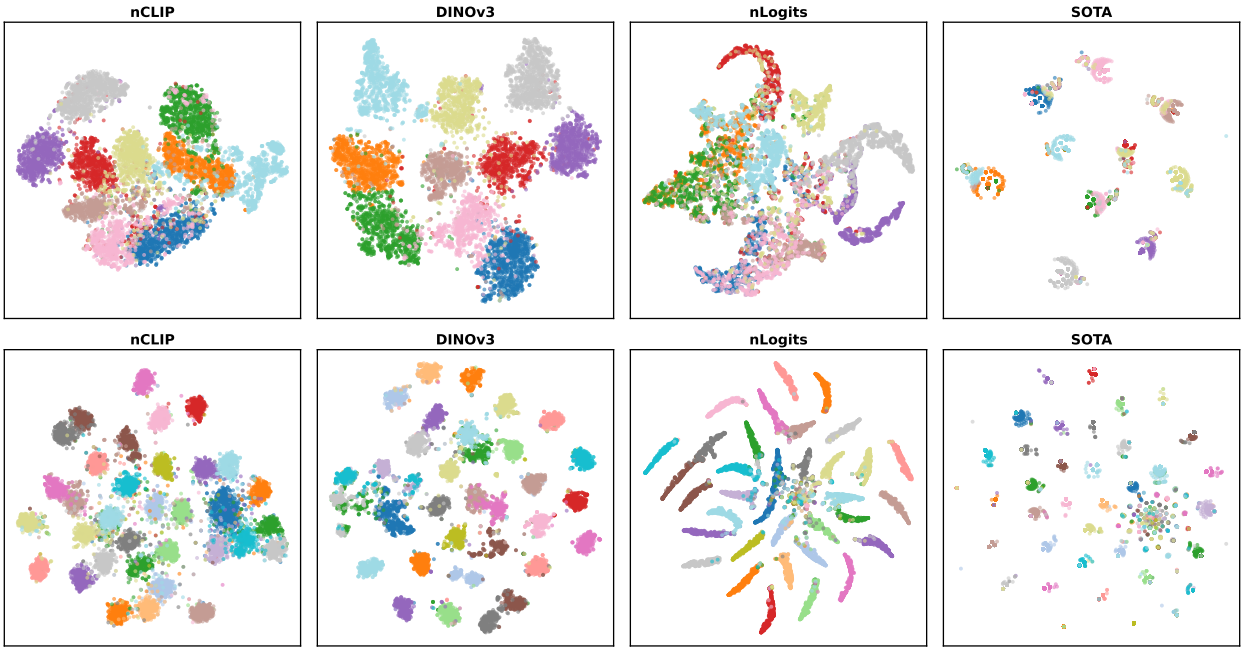


Figure 2. t-SNE visualization of predicted clusters on the `EuroSat`(first row) and `Food101`(second row) dataset. SOTA significantly improves cluster compactness and separation over CLIP, highlighting superior integration of visual and semantic cues.

highlighting that our design offers a favorable balance between effectiveness and efficiency.

C.4. Model subset analysis

For remote-sensing and medical datasets, we further conduct model-subset experiments (Tab. 2). Notably, although CLIP exhibits the weakest cross-domain performance among all VLMs, integrating it with any target-domain-specific model via our framework consistently leads to performance gains, validating the effectiveness of our approach under diverse model subset configurations. This behavior suggests that our method does not rely on a single dominant model; instead, it can leverage diverse model outputs in a cooperative manner.

References

[1] Andrew A. Borkowski, Marilyn M. Bui, L. Brannon Thomas, Catherine P. Wilson, Lauren A. DeLand, and Stephen M. Mastorides. Lung and colon cancer histopathological image dataset (LC25000). *arXiv preprint arXiv:1912.12142*, 2019. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pages 446–461. Springer, 2014. 2

[3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 2

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 2

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, page 178, 2004. 2

[7] Chu Han, Jiatai Lin, Jinhai Mai, Yi Wang, Qingling Zhang, Bingchao Zhao, Xin Chen, Xipeng Pan, Zhenwei Shi, Zeyan Xu, et al. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis*, 80:102487, 2022. 2

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2

[9] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using

Table 1. Comparison on remote sensing and medical pathology (mean over datasets). “max”/“mean” are ensemble rules; adaptive methods (e.g., TransCLIP and ADAPT) are applied per vanilla model before ensembling.

Method	Remote sensing			Medical Pathology		
	ACC/%	Time/s	Memory/MiB	ACC/%	Time/s	Memory/MiB
Vanilla(max)	63.50	–	–	65.32	–	–
Vanilla(mean)	69.80	–	–	69.52	–	–
ECALP(ICLR'25)	72.69	260.2535	3330	69.82	72.9642	1012
TransCLIP(NeurIPS'24)	76.99	1.7839	911	75.29	0.7548	1067
ADAPT(NeurIPS'25)	77.46	0.2587	1089	78.07	0.2132	970
Ours	81.45	0.9488	986	83.9	0.3120	1435

Table 2. The results of combining different models. The first line is the results of the baseline model.

Remote Sensing					Medical Pathology				
CLIP	Geo	Remote	Sky	Acc/%	CLIP	CONCH	MUSK	PLIP	Acc/%
56.1	64.5	61.0	64.4	–	30.8	62.9	58.2	66.3	–
✓	✓			76.05	✓	✓			68.25
✓		✓		77.18	✓		✓		73.06
✓			✓	75.52	✓			✓	68.58
	✓		✓	80.41		✓	✓		83.07
	✓		✓	77.58		✓		✓	80.01
		✓	✓	81.16			✓	✓	81.75
✓	✓	✓		78.90	✓	✓	✓		79.87
✓	✓		✓	77.09	✓	✓		✓	79.32
	✓	✓	✓	82.40		✓	✓	✓	84.30
✓	✓	✓	✓	81.50	✓	✓	✓	✓	83.90

medical twitter. *Nature Medicine*, 29(9):2307–2316, 2023. 2

[10] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo*, 2018. 2

[11] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 554–561, 2013. 2

[12] Katharina Kriegsmann, Frithjof Löbers, Christiane Zgorzelski, Joerg Kriegsmann, Charlotte Janssen, Rolf Rudinger Meliß, Thomas Muley, Ulrich Sack, Georg Steinbuss, and Mark Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12:1022967, 2022. 2

[13] Haifeng Li, Xin Dou, Chao Tao, Zhixiang Wu, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data. *Sensors*, 20(6):1594, 2020. 2

[14] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision-language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 2

[15] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 2

[16] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[17] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 2

[18] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR), pages 3498–3505, 2012. 2

[19] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5–6):355–607, 2019. 1

[20] Xiaoman Qi, Panpan Zhu, Yuebin Wang, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P. Takis Mathiopoulos. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 2020. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 2

[22] Julio Silva-Rodríguez, Arne Schmidt, María A. Sales, Rafael Molina, and Valery Naranjo. Proportion constrained weakly supervised histopathology image classification. *Computers in Biology and Medicine*, 147: 105714, 2022. 2

[23] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[24] Qi Wang, Shaoteng Liu, Jocelyn Chanussot, and Xuelong Li. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1155–1167, 2018. 2

[25] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. SkyScript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 5805–5813, 2024. 2

[26] Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, and Henri Maître. Structural high-resolution satellite image indexing. In *ISPRS TC VII Symposium – 100 Years ISPRS*, pages 298–303, 2010. 2

[27] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 2

[28] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Jeffrey J. Nirschl, Sierra Willens, Francesca Maria Olguin, Joel Neal, Maximilian Diehn, Sen Yang, Kun-Hsing Yu, and Ruijiang Li. A vision–language foundation model for precision oncology. *Nature*, 638:769–778, 2025. 2

[29] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010. 2

[30] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024. 2

[31] Lijun Zhao, Ping Tang, and Lianzhi Huo. Feature significance-based multi-bag-of-visual-words model for remote sensing image scene classification. *Journal of Applied Remote Sensing*, 10(3):035004, 2016. 2

[32] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:197–209, 2018. 2