Material Open-Vocabulary Recognition

Margarita Kichik[a], Dmitry Yudin[a,b]

[a]*Moscow Institute of Physics and Technology, 9 Institutsky per., Dolgoprudny, 141701, Moscow Region, Russia*
[b]*Artificial Intelligence Research Institute (AIRI), 32 Kutuzovsky Ave., Moscow, 121170, Russia*

Abstract

Material recognition is a critical capability for robotics, enabling applications from object manipulation to environmental interaction. However, existing methods remain limited in scalability, often relying on predefined material categories or struggling with real-world complexity. To address these challenges, we introduce MatLVIS, a large-scale, open-vocabulary dataset for material recognition, constructed by augmenting the LVIS benchmark with fine-grained material annotations generated via vision-language models. We further propose an efficient CLIP-based architecture optimized for real-time material recognition, enhancing its ability to resolve small-scale objects (10px) through feature fusion.

Keywords: kw1, kw2

1. Introduction

Material recognition and property detection are critical for robotics, where inferring object characteristics directly impacts manipulation, navigation, and task planning. While existing methods often rely on semantic segmentation, current datasets are constrained by two dominant paradigms: synthetic data from simulators, which lack real-world complexity, and limited annotated real-world imagery. Notably, real-world datasets predominantly emphasize outdoor scenes, creating a significant gap in indoor material recognition due to sparse, contextually narrow data.

To address these limitations, we introduce MatLVIS, a large-scale, automatically annotated dataset that augments existing LVIS object labels with fine-grained material descriptions. MatLVIS supports diverse tasks, including open-vocabulary

recognition and material classification, while bridging the indoor-outdoor data disparity. Furthermore, we propose a novel CLIP-based architecture enhanced through feature fusion, which integrates multi-scale visual cues to resolve ambiguities in small or textureless objects. We also introduce a generalizable training protocol that combines contrastive learning with adversarial robustness, achieving state-of-the-art performance on zero-shot material recognition without compromising inference speed. Together, these contributions advance robotic perception by enabling precise, real-time material understanding in cluttered, real-world environments.

2. Related work

Datasets.

Datasets for material segmentation in real-world imagery include MINC [1], DMS [2], and MCubeS [3]. MINC is a large-scale dataset containing approximately 440,000 images, though only 7,000 are paired with binary masks across 23 material categories. DMS, a more recent benchmark, provides over 44,000 images annotated with pixel-level semantic masks for 46 material categories. MCubeS focuses on outdoor scenes, offering multi-modal data (RGB, reflectance imaging [RI], and polarization parameters [angle and degree]) alongside semantic masks for 20 material classes, enabling cross-modal analysis of surface properties.

In addition to real-world datasets, synthetic simulators and indoor-focused datasets have gained traction. 3D-FUTURE [4] features 15 material categories for furniture-centric scenes, while 3DSSG [5], built using the 3R-Scan simulator [6], similarly supports 15 categories. The 3D Scene Graph dataset [7], constructed in the Gibson environment, extends material annotations derived from MINC to indoor robotic navigation contexts.

Methods.

Specialized methods for material recognition remain limited in scope and diversity. In [2], the authors propose a convolutional neural network (CNN) augmented with a pyramidal feature aggregation layer, designed to capture multi-scale texture patterns critical for distinguishing materials. Similarly, [3] introduces a multi-modal transformer architecture, employing four parallel encoders (one per modality: RGB, reflectance, polarization angle, and polarization degree) followed by a cross-modal feature filtering mechanism to suppress irrelevant or noisy signals.

Recent work has also emphasized efficiency for edge deployment: [8] trains a multi-gate mixture-of-experts (MoE) transformer on the DMS dataset, dynamically routing computations to balance accuracy and latency on resource-constrained devices. Meanwhile, vision-language models (VLMs) like CLIP have been adapted for material recognition—[9] inserts a lightweight feature fusion network between CLIP's image and text encoders to refine material-specific embeddings, though their evaluation is limited to a narrow 5-class taxonomy.

3. Approach

MatLVIS dataset.

Our annotation pipeline for LVIS proceeds in three stages to ensure high-quality material descriptions. First, we generate object-centric captions using a hybrid approach: Set-of-Marks (SoM) [10] localizes objects via spatial prompts, while Qwen2.5-VL-7B [11] generates free-form captions describing object attributes (e.g., "a metallic spoon on a wooden table"). Next, we validate captions using the DeepSeek API, categorizing results into (see these classes distribution in Table 1.):

- High-confidence captions (directly retained);
- Auto-correctable captions (refined via context-aware material insertion or artifact removal);
- Human-reviewed captions.

Finally, we extract material mentions from validated captions, ensuring future alignment with a unified taxonomy (e.g., mapping "steel"→ "metal"). The resulting MatLVIS dataset contains over 1M object-caption pairs, with detailed splits and material distributions provided in Table **??**.

To generate object-centric embeddings (rather than whole-image representations), we extract region-of-interest (RoI) crops centered on each object. However, naively cropping objects can truncate critical context or introduce edge artifacts. Our experiments reveal two key optimizations:

- Adaptive Padding: Apply a 10% padding margin relative to the object's bounding box. For objects smaller than 50px in any dimension, we iteratively increase padding until reaching a 50px minimum, preserving structural integrity;

Таблица 1: Statistics over labelled MatLVIS dataset.

Split	Images (total)	Objects (total)	High-confidence captions	Auto-correctable captions	Human-reviewed captions
Train	118287	-	-	-	-
Val	5000	-	-	-	-

- Background Suppression: Blur the background outside the padded RoI using a Gaussian kernel (σ=15). This mitigates interference from irrelevant visual cues while retaining material-relevant context.

Model

The core idea is to train an image model to produce high-quality embeddings that enable robust material classification or recognition of object materials.

We began with a standard vanilla CLIP model. However, we observed that fine-tuning primarily caused the model to "ignore"material-related information. To address this, we introduced a reconstruction loss from a separate classification MLP, resulting in the following combined loss function:

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{CE}}$$

The next improvement involved incorporating features extracted from the entire image using encoders like RADIO and DINOv2. We hypothesized that leveraging global contextual information would enhance the model's ability to generate more discriminative embeddings. To this end, we introduced a transformer-based fusion module where CLIP-derived features serve as queries, while global image features act as keys and values. The architecture of our model is illustrated in Fig. 1.

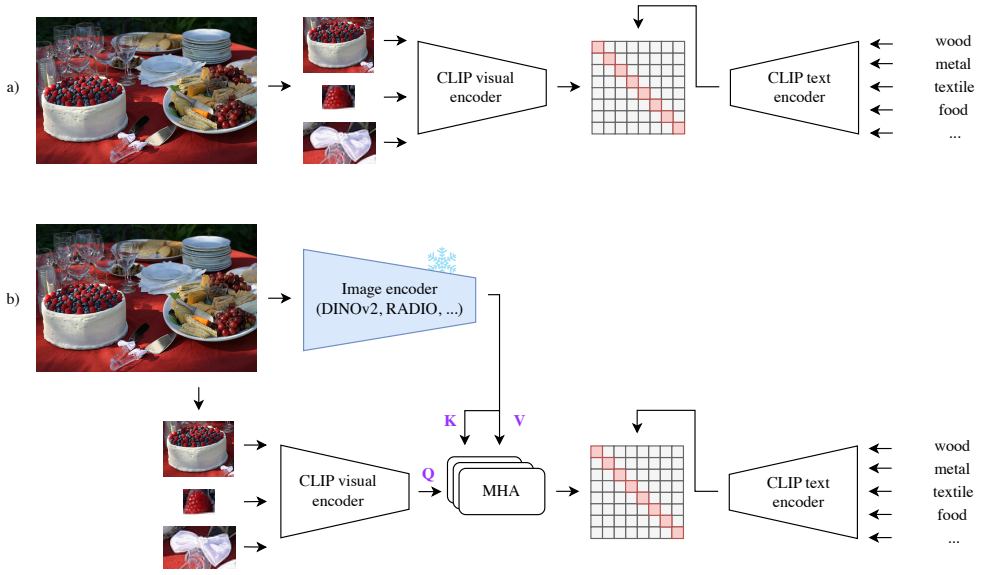

Рис. 1: Proposed architecture of CLIP: a) OpenAI CLIP; b) Ours CLIP with fusion module.

4. Experimental results

To evaluate model performance during training, we employ three key metrics: mean cosine similarity (MCS), mean reciprocal rank (MRR), and recall@k (k$\in\{1, 5\}$). These metrics quantify the alignment between predicted material embeddings and ground-truth textual descriptions in the CLIP latent space. As shown in Table 2, we benchmark our models against the pretrained CLIP baseline (first row) to establish a comprehensive performance comparison. Subsequent rows detail the improvements achieved by our architecture, highlighting gains in both retrieval accuracy (via MRR) and fine-grained material discrimination (via MCS).

To rigorously benchmark our models against pretrained CLIP variants, we evaluate their zero-shot material classification capabilities across three datasets: MINC, DMS, and MatLVIS. For cross-dataset compatibility, we map material categories in MatLVIS to MINC's taxonomy. On DMS, we generate object masks using a pretrained YOLOv8 segmentation model to isolate materials for fair comparison. Table 3 shows the results of our benchmarking.

Таблица 2: Metrics for our main experiments (calculated with batch of 64 items).

Model	MCS	MRR	Recall@1	Recall@5	Recall@10
OpenAI CLIP (before fine-tuning)	-	-	-	-	-
CLIP	-	-	-	-	-
SigLIP	-	-	-	-	-
CLIP with reconstruction loss	-	-	-	-	-
FusionCLIP	-	-	-	-	-

Таблица 3: Zero-shot classification

Model	Dataset	Accuracy@1	Precision@1	Recall@1
CLIP	-	-	-	-
SigLIP2	-	-	-	-
EVA-CLIP(?)	-	-	-	-
CLIP (ours)	-	-	-	-
SigLIP (ours)	-	-	-	-
FusionCLIP (ours)	-	-	-	-

5. Conclusion

Список литературы

[1] S. Bell, P. Upchurch, N. Snavely, K. Bala, Material recognition in the wild with the materials in context database, Computer Vision and Pattern Recognition (CVPR) (2015).

[2] P. Upchurch, R. Niu, A dense material segmentation dataset for indoor and outdoor scene parsing, in: European Conference on Computer Vision (ECCV), 2022.

[3] Y. Liang, R. Wakaki, S. Nobuhara, K. Nishino, Multimodal material segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19800–19808.

[4] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. Maybank, D. Tao, 3d-future: 3d furniture shape with texture, International Journal of Computer Vision (2021) 1–25.

[5] J. Wald, H. Dhamo, N. Navab, F. Tombari, Learning 3d semantic scene graphs from 3d indoor reconstructions, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[6] N. N. F. T. M. N. Johanna Wald, Armen Avetisyan, Rio: 3d object instance re-localization in changing indoor environments, in: Proceedings IEEE International Conference on Computer Vision (ICCV), 2019.

[7] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, S. Savarese, 3d scene graph: A structure for unified semantics, 3d space, and camera, in: Proceedings of the IEEE International Conference on Computer Vision, 2019.

[8] J. Zheng, J. Zhang, K. Yang, K. Peng, R. Stiefelhagen, Materobot: Material recognition in wearable robotics for people with visual impairments, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 2303–2309.

[9] Y. Sun, Z. Gu, S. B. Yang, Probing vision and language models for construction waste material recognition, Automation in Construction 166 (2024) 105629.

[10] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, J. Gao, Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, arXiv preprint arXiv:2310.11441 (2023).

[11] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-vl technical report, arXiv preprint arXiv:2502.13923 (2025).