LUWA Dataset: Learning Lithic Use-Wear Analysis on Microscopic Images

Jing Zhang[1,*], Irving Fang[1,*], Hao Wu[1], Akshat Kaushik[1], Alice Rodriguez[1], Hanwen Zhao[1], Juexiao Zhang[1], Zhuo Zheng[2], Radu Iovita[1,†], Chen Feng[1,†]

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.02129

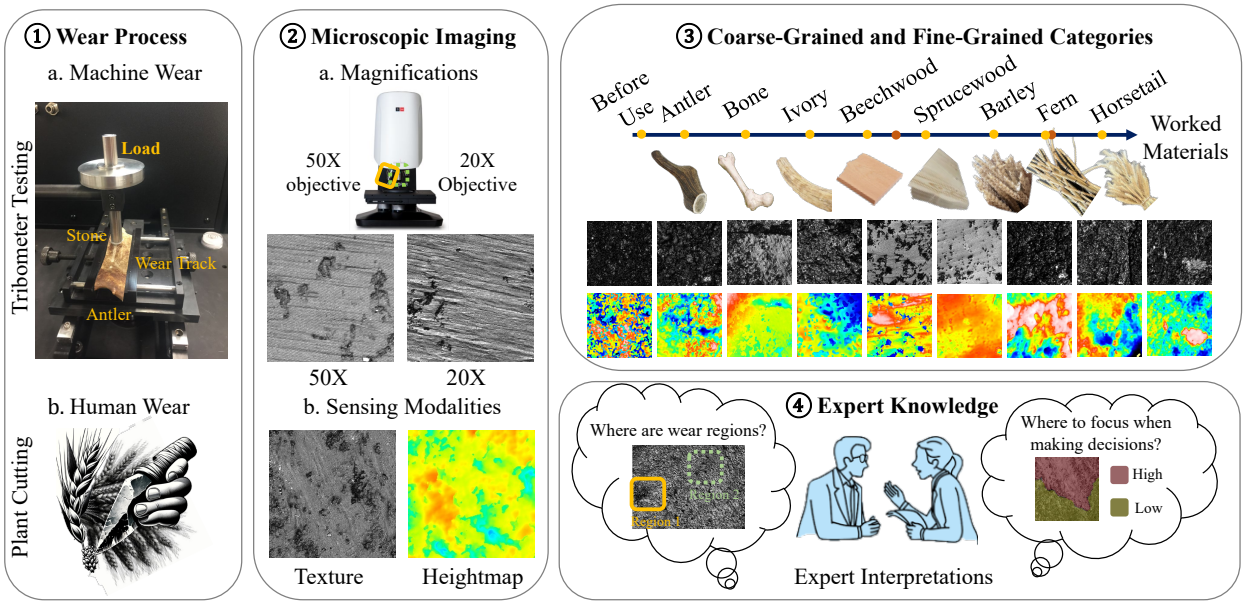


Figure 1. LUWA poses a unique computer vision challenge due to: *its complex wear formation and irregular wear patterns*, *ambiguous sensing modalities and magnifications in microscopic imaging*. Facing these challenges, the LUWA dataset encompasses both texture and heightmap with different magnifications, encouraging the exploration of image classification beyond common objects.

Abstract

Lithic Use-Wear Analysis (LUWA) using microscopic images is an underexplored vision-for-science research area. It seeks to distinguish the worked material, which is critical for understanding archaeological artifacts, material interactions, tool functionalities, and dental records. However, this challenging task goes beyond the well-studied image classification problem for common objects. It is affected by many confounders owing to the complex wear mechanism and microscopic imaging, which makes it difficult even for human experts to identify the worked material successfully. In this paper, we investigate the following three questions on this unique vision task for the first time: ***(i)*** *How well can state-of-the-art pre-trained models (like DINOv2) generalize to the rarely seen domain?* ***(ii)*** *How can few-shot learning be exploited for scarce microscopic images?* ***(iii)*** *How do the ambiguous magnification and sensing modality influence the classification accuracy? To study these, we collaborated with archaeologists and built the first open-source and the largest LUWA dataset containing 23,130 microscopic images with different magnifications and sensing modalities. Extensive experiments show that existing pre-trained models notably outperform human experts but still leave a large gap for improvements. Most importantly, the LUWA dataset provides an underexplored opportunity for vision and learning communities and complements existing image classification problems on common objects.*

1. Introduction

Lithic Use-Wear Analysis (LUWA) is a long-standing scientific problem (see Fig. 1) to identify the functions of stone tools by examining wear traces at the microscopic level on the tool's surface [30, 41]. It seeks to distinguish the worked material (like bone, wood, ivory and antler) us-

*Equal contribution
†Corresponding author: Radu Iovita (iovita@nyu.edu) and Chen Feng (cfeng@nyu.edu)

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.02129

ing microscopic images, creating a classification problem. Investigating this unanswered vision-for-science problem will provide invaluable insights for uncovering the hidden story of ancient tools [38] and advancing the understanding of material interactions [62, 66]. However, few studies have begun to explore advanced learning-based methods, and this field is still in a nascent stage [8, 45].

Besides its potential scientific impact, learning-based LUWA on microscopic images also poses unique computer vision challenges beyond common objects. Unlike common objects with distinct boundaries and typical size, wear traces in microscopic images are *irregular and discontinuous* (see Fig. 1), making it difficult to define clear visual characteristics [40]. In particular, their absence of clear foreground and background might increase the difficulty of feature extraction. Furthermore, recorded wear traces are *a hybrid consequence of the complex wear process and microscopic imaging*, which is affected by many factors [41]. For example, the wear duration and the motion type that generates the wear can create significant intra-class variability. Another crucial aspect is *the ambiguous magnification and sensing modality of microscopic images*. High-sensitivity microscopic imaging allows for capturing complementary sensing modalities and zoom-in details at different magnifications, respectively (see Fig. 1). We found that their modality and magnification differences lead to very different visual features. However, due to the understudied characteristics of wear traces, there has been no conclusive answer to this vision-for-science problem about which magnification and sensing modality is better. In practice, the accessibility of expensive equipment also limits the flexible selection of captured data.

Recent advances in computer vision provide a good opportunity for this challenging scientific problem, especially the emerging paradigm shift with foundation models pre-trained on large-scale data [2, 7]. These foundation models generate task-agnostic visual features and have shown excellent performance on image classification with common datasets like ImageNet-$\{1k, A\}$ [43]. But how much can state-of-the-art (SOTA) pre-trained representations benefit rarely seen domains not extensively covered by uncurated data available on the Internet? The lack of diversity in domain-specific datasets leaves uncertainty about this question, particularly when addressing real-world vision challenges. Unfortunately, existing studies on LUWA focus on private data and the research community faces a deficiency in accessible datasets. The complexity of wear forms, microscopic imaging, equipment limitations, and the need for expert interpretations hinders the collection of an appropriate dataset that can represent the variability of this domain.

To explore these unique challenges and opportunities, we collaborated with anthropological archaeologists and built the first open-source and the largest LUWA dataset containing 23,130 microscopic images. Its major characteristics are summarized as follows: (**i**) **Multi-scale wear patterns.** Lower and higher magnifications allow for an observation of pattern distributions and topographical details, respectively, which increases the scale variations. (**ii**) **Complementary sensing modalities.** Both grayscale microscopic images and corresponding 3D surface profiles are available to provide complementary texture information and geometry cues, helping to identify discriminative features [21]. (**iii**) **Both machine and human wear processes.** Samples of stone tools were collected from both machine and human wear processes. The former is to isolate the effect of worked material by tightly controlling other factors [37] while the latter is to reflect the complexity of real scenes, especially for newly discovered categories without baseline data [44, 52]. We envision that the LUWA dataset will encourage researchers to develop and evaluate algorithms for image classification beyond common objects, as a precursor for downstream tasks like segmentation and detection [3].

Based on the LUWA dataset, we go further in answering the following three problems facing real-world applications: (**i**) How well representative classification models can generalize to the rarely seen domain? (**ii**) How can few-shot learning be exploited when scarce microscopic images are available, especially for newly discovered categories? (**iii**) How do the ambiguous magnification and sensing modality of microscopic images influence the classification accuracy? Our contributions are summarized as follows:

- We introduce the first open-source and the largest LUWA dataset including 23,130 microscopic images with different magnifications and sensing modalities. We collaborated with archaeologists and set up a data collection pipeline considering the influence of wear formation, microscopic imaging, and expert knowledge. The LUWA dataset allows for reproducible investigations on this rarely seen domain and complements existing image classification problems on common objects.
- Facing the image classification problem beyond common objects, we benchmark the generalization capability of state-of-the-art vision models (ResNet, ViT, DINOv2, ConvNeXt) in this specific domain. Experimental results show that DINOv2 has the most stable performance amidst varying levels of granularity, magnification, and sensing modalities in the data. We also observed many trends regarding the impacts that features like magnification and sensing modality have on classification accuracy, and some of them are not consistent with experts' heuristics. In general, state-of-the-art computer vision models display super-human accuracy over domain experts.
- Considering the scarcity of microscopic images in real scenarios, we investigate the performance of few-shot image classification and reasoning on the LUWA dataset. Particularly, we collected prompts from archaeologists

and did case studies on whether GPT-4V(ision) can mimic the experts' reasoning process. Further explorations are required to improve the performance.

2. Related Work

Lithic Use-Wear Analysis. Lithic use-wear analysis was originally developed in the 1950s and aims to answer the scientific problem of how to distinguish the worked material according to wear traces on the stone surface [30, 41, 55]. Low-power and High-power microscopy methods provide useful magnified polishes or micro-fractures via confocal microscopes, tactile profilometers, scanning electron microscopes, and even atomic force microscopes [20, 31, 58]. Existing studies focus on blind tests [4, 19, 39] and quantification methods [4, 5, 20, 28, 59, 60]. However, it remains an insufficiently developed research area due to complex wear patterns and the subjectivity of methods [47]. Existing blind tests demonstrate unreliable identification results on different worked materials. For example, correct identifications of plant and wood are 32.4% and 49.1%, respectively [19]. The identification of worked materials can be regarded as a unique vision problem, but learning-based algorithms are rarely explored in this research area [8]. The deficiency in accessible datasets also limits its research progress heavily [19]. To solve these, we present the first open-source LUWA dataset and investigate the capability of SOTA classification algorithms on this unique vision-for-science problem for the first time.

Image Classification beyond Common Objects. Image classification is a fundamental vision task that categorizes a whole image as a specific label or class based on its visual contents. Outstanding performance can be achieved on the image classification task of common objects with representative frameworks like ConvNeXt using only public training data [67]. However, real-world scenarios, especially scientific applications, involve non-trivial objectives with different physical properties like tiny pollen grains [6], constituent materials [15, 18, 61], texture in the wild [12], and hyperspectral remote sensing images [50]. Facing uncommon objects, many studies focus on specialized models in respective fields. However, we seek to answer whether simpler pipelines can provide comparable performance. Complementing existing datasets, we introduce a new vision challenge to identify irregular and discontinuous wear traces at the microscopic level and build a classification benchmark to explore how far the practical performance of image classification on uncommon objects can be pushed utilizing existing datasets and architectures, especially foundation models.

Microscopic Image Classification. In the context of using microscopic images for classification, a wide variety of applications can be found across different scientific disciplines. This includes the study of biological cells [11, 29], bacteria [65], tissue types [64], and material structures [23]. Each application presents unique challenges, particularly in terms of the high-level detail and focus required in the images. These challenges are often compounded by issues such as ambiguity in magnification and sensing modality [23, 68]. To alleviate these challenges, the LUWA Dataset presents a diverse configuration that covers different magnification levels and sensing modalities.

3. LUWA Dataset

In this section, we describe the LUWA dataset creation process and provide basic statistics. To represent the variability of this domain, in Section 3.1, we present a data collection pipeline, which consists of four key aspects (see Fig. 1). Specifically, ① considering the complexity of wear formation, we introduced both machine and human wear experiments [36] to create stone samples; ② to enrich the dataset diversity and investigate the ambiguous magnification and sensing modality, we utilized an optical 3D profilometer with both $20\times$ and $50\times$ objective lenses to acquire high-quality texture and heightmap; ③ natural materials were selected according to existing blind tests in the literature [19], particularly including fine-grained categories of wood and plants. ④ domain-specific knowledge is twofold including the identification of wear degrees to increase the dataset diversity and expert interpretations for potential explorations on explainability and the application of vision language models. In Section 3.2, we summarized the LUWA dataset and analyzed its diversity in spatial distributions, magnifications, and sensing modalities.

3.1. Dataset Creation

Both Machine and Human Wear Processes. LUWA dataset contains stone samples from both machine and human wear processes. Key factors that affect wear results are material properties, mechanical factors, and environmental conditions [48]. To isolate the effect of worked materials, a tightly controlled protocol was used for machine wear experiments [36, 37]. We utilized a tribometer to simulate cutting actions so as to quantify the load applied to the material (a load of 20N), the type of movement (a straight back and forth motion with the speed of 35 repetitions per minute), and the worked duration (0h, 1h, 3h, 5h, and 12h) (see Fig. 1). To limit the influence of material properties of stone samples, we chose the same flint (Baltic/morainic flint from Denmark) for all experiments [47, 51]. Considering the low classification accuracy of various plants in blind tests [19], we chose the plant cutting process as the human wear experiment.

High-Quality Microscopic Imaging. To capture the high-precision wear traces on stone samples, we utilized an optical 3D profilometer (S neox, Sensorfar Metrology) to collect data with a standardized and reproducible process (see

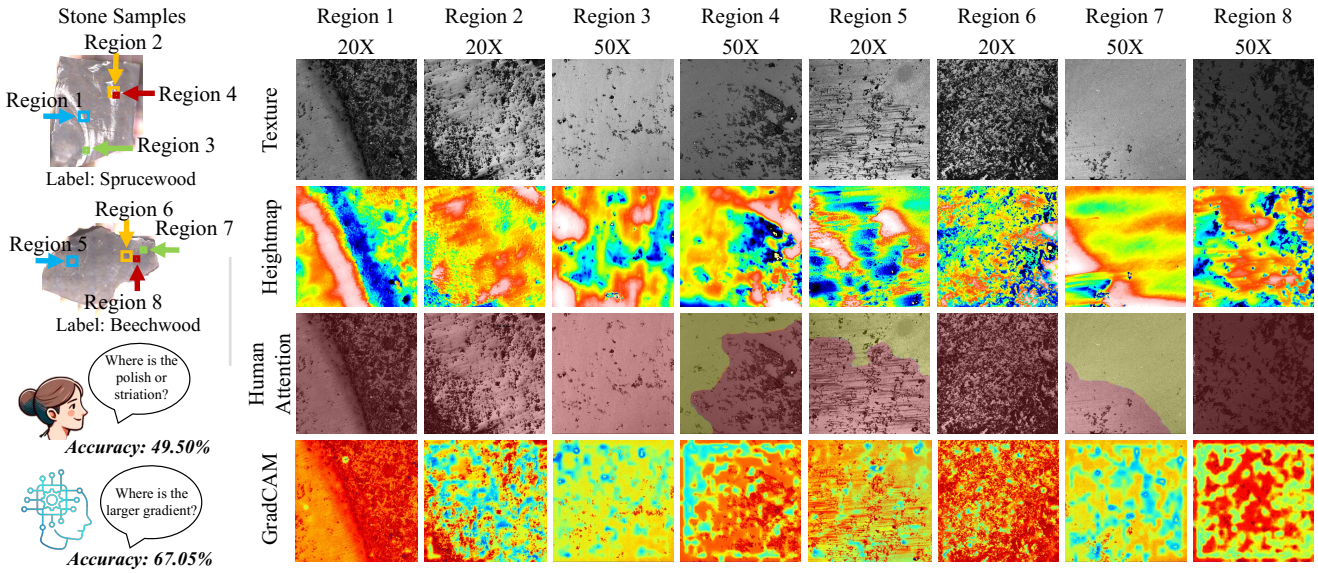


Figure 2. **Image diversity of LUWA dataset and corresponding visual explanations for human and model decision-making processes. (i)** LUWA dataset provides diverse microscopic images associated with spatial distributions (e.g. Regions 1 and 2), magnifications (e.g. Regions 2 and 4) and sensing modalities (texture in the first row and heightmap in the second row); **(ii)** We compared visual explanations in both human (in the third row) and model (in the fourth row) decision-making processes. Human experts labeled the most important region with red and the less important region with yellow when looking at details of microscopic images to distinguish the worked material. Similarly, Grad-CAM [54] heatmaps use red for the highest importance, yellow for lower importance, and blue for the lowest importance. Interestingly, similar areas (e.g. Regions 1, 4 and 6) are labeled with higher importance for both humans and models.

Fig. 1). To test the influence of magnifications for microscopic image classification, both $20\times$ and $50\times$ objective lenses were chosen for measurements. Their spatial resolutions are 0.65 and 0.26 μm/pixel, respectively. Furthermore, complementary grayscale images and corresponding 3D surface profiles are acquired via Sensormap. We applied a standard filtering protocol to extract the worn surface and alleviate the effect of natural flints' surface topography [9].

Domain-Specific Expert Knowledge. Domain-specific knowledge is twofold: (**i**) human experts help to identify microscopic traces with different wear degrees, enriching the dataset diversity; (**ii**) for further investigations on explainability and the application of vision language models, human experts also labeled their attention maps when making decisions on worked material (see Fig 1) and provided classification prompt for GPT-4V [42] (see Fig. 6).

Material Selection and Processing. To benchmark models in this field, we chose representative natural materials (see Fig. 1) according to blind test results in the literature [19] and included fine-grained categories on wood and plants in particular. For the further exploration of wear mechanisms, we analyzed their properties, including the hardness [47] and silicon content [22] in the supplementary material.

3.2. Dataset Analysis

To reflect the variability of this scientific domain, we built the first public and largest LUWA dataset containing 23,130 microscopic images. Specifically, key factors of the complex wear formation and microscopic imaging are considered in the LUWA dataset. As shown in Tab. 1, we report (**i**) the number of microscopic images, (**ii**) the number of stone samples, motion types, worked time, and material classes which are exploited, (**iii**) the number of magnifications and sensing modalities LUWA dataset supports.

Stone Samples	Motion Types	Worked Time	Material Categories
34	2	7	9
Magnifications		**Sensing Modalities**	
2		2	

Table 1. Key factors considered in the LUWA dataset that can reflect the complex wear formation and microscopic imaging.

Image diversity of LUWA dataset brings challenges for algorithm robustness. It is associated with the spatial distribution of the region collected, the selection of the magnification, and the sensing modality. Greater distances between sampled areas typically result in more pronounced variations in their surface distributions (see Regions 1 and 2 in Fig. 2). Even collected from the same wear trace, the selection of the magnification also contributes to scale difference, which causes totally different wear patterns (see Regions 2 and 4 in Fig. 2). Moreover, LUWA dataset provides both the texture and heightmap, helping to identify discriminative features. We explore the semantic diversity of LUWA dataset on the magnification and sensing modality (see Fig. 2) [3]. We selected VGG [56], ResNet [26],

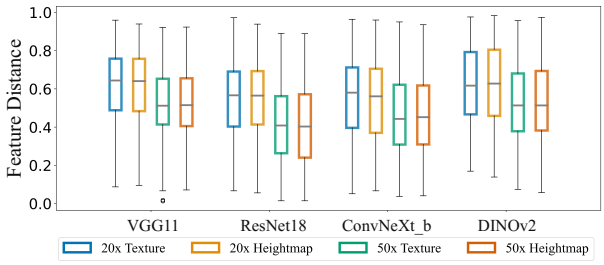


Figure 3. Cosine similarity distribution of LUWA dataset on different magnifications and sensing modalities.

ConvNeXt [34], and DINOv2 [10, 14, 43] as feature extractors. Then we compute the mean cosine distance of images with different magnifications and sensing modalities, respectively. Scale difference leads to obvious diversity of semantic information and visual descriptions (see Fig. 3).

4. Algorithm Benchmarking

By benchmarking a wide range of image classification methods, both classic and state-of-the-art, on this unique vision-for-science dataset, we explore how different features of this dataset affect model performance and hope that we can provide some useful analysis that future work can build on. Specifically, we divide our experiments into two major segments: (1) fully-supervised image classification and (2) few-shot image classification, with specific motivations explained in their corresponding sections below.

4.1. Fully-Supervised Image Classification

Unlike datasets crawled from the internet, such as ImageNet-1k [16] or LAION-5B [53], LUWA contains niche microscopic images with irregular and discontinuous wear traces that often lack obvious foreground or background. In this experiment, we investigate how well classic and state-of-the-art image classification algorithms generalize to the LUWA dataset and seek to find compelling patterns affecting different models' performance. We want to see how well these patterns align with domain experts' knowledge. We also aim to position SOTA methods from the computer vision community with respect to classification performance achieved by human experts.

Experimental Settings. We deploy classic methods such as SIFT+FVs [49], ResNets[24] and cutting-edge models such as ViT [17], ConvNeXts [34] and DINOv2 [13, 43] to our LUWA dataset. We believe the time-tested classic methods can serve as a lower-bound benchmark, while the more recent advancements such as DINOv2, which can often be characterized by intensive scale-up in parameter count, can be used as a reference comparable to human experts' performance on the same task.

Another major reason that propels us to deploy this wide range of models is to see if there are consistent trends across different model architectures and parameter counts and if these trends align with domain experts' knowledge. Specifically, we study the impact of image granularity, magnification, and sensing modality on image classification performance. We also compare different training strategies. By combining the following configurations, we have 324 total experiment results.

Granularity refers to how many pictures one single use-wear is partitioned into. The use-wear is first captured as an image at 865×865 resolution. Because many pre-trained models resize the input, which results in pixelated images and loss of fine-grained details that many experts believe are crucial to such a classification task, we partition the original image into 24 or 6 patches and feed all the patches to the model. Importantly, to make our results comparable to those of human experts, we adopt a *voting mechanism* during test time. If the majority of the 24 patches are classified as class 1, then all the patches of the same use-wear will be classified as class 1 regardless of their actual test results. We believe this method most resembles how archaeologists perform classification when given an 865×865 image, as they do not assign a label to each partition.

Magnification represents the magnification multiplier on our microscopic imaging equipment. Our dataset comprises images after $20\times$ and $50\times$ magnification. We also mix $20\times$ and $50\times$ data together without any indicator of magnification added to the data to see if mixing magnification will cause any confusion in our models. Note that the magnification multipliers are fixed when the images are taken, and $50\times$ images will not look like $20\times$ images, even if we significantly downsize them.

Sensing Modality refers to whether the picture of the use-wear is stored as texture scans or heightmaps. Texture scans have no depth information, although depth cues are still present. Meanwhile, heightmaps explicitly store the depth information and only the depth information. We want to see if different ways to represent the use-wear will affect computer vision models.

Training Strategy is also varied in our experiments. Some models are initialized with state-of-the-art initialization methods [25] and then trained from scratch. We also apply full-parameter fine-tuning and linear probing [1] with unfrozen and frozen pre-trained weights, respectively.

Implementation Details. We deploy ResNet50 (25.6M parameters), ResNet152 (60.4M), ConvNeXt-tiny (28.6M), ConvNeXt-large (197.7M), ViT-H (632M), and DINOv2-ViT-g/14 with registers [13] (1.1B). All models are trained with Adam optimizer [32] on the default setting in PyTorch. We employ linear warmup with cosine annealing [35] as the learning rate scheduler strategy. No data augmentation technique is applied during pre-processing. We defer more details to our supplementary material.

Results and Analysis. The overall experiment results can

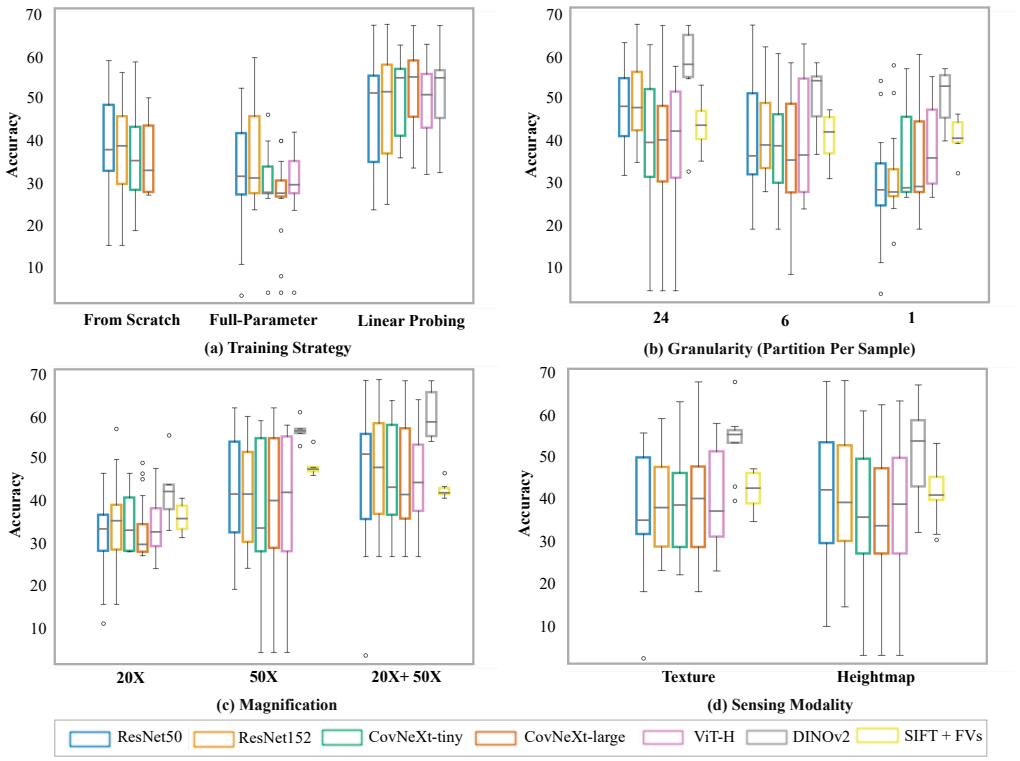


Figure 4. The impact of the training strategy, granularity, magnification, and sensing modality on top-1 classification accuracy in %: (a) Due to their huge parameter counts, the experiments do not include full-parameter fine-tuned DINOv2, and ViT-H and DINOv2 trained from scratch. (b) Larger numbers in granularity mean more detailed information about a use-wear is fed into the model.

be found in Fig. 4. Fig. 4 (a) demonstrated that linear probing yields the most stable and optimal performance across the broad. These results align with our expectation as fine-tuning on uncommon domain-specific datasets may cause catastrophic forgetting [33]. On the other hand, the LUWA dataset itself is too small to train generalizable models from scratch. From the perspective of granularity (Fig. 4 (b)), the more granular partition of the images tends to result in better outcomes, although a diminishing marginal return can be observed. This aligns with our speculation that keeping the original larger image's information as intact as possible is beneficial. It is also possible that the introduction of a voting mechanism brought about a positive regularization effect. More discussion on the voting mechanism can be found in the supplementary material. Considering the selection of magnification, results in Fig. 4 (c) indicated that a higher magnification multiplier is beneficial, which aligns with some human experts' opinions. Notably, mixing data with different magnifications does not confuse the models, and they are able to reap the benefit of abundant data. The same cannot be said for humans, as images with different magnifications can cause confusion. For the sensing modality, we observed that while the best results are usually trained with heightmaps, larger models tend to favor texture. However, the discrepancy is small in general, and the overall performances of the two modalities are comparable as in Fig. 4 (d). More visualizations and detailed tables that follow the trends described above can be found in the supplementary material.

The best performance of 67.05% top-1 accuracy is achieved by linear probing on ImageNet-1K pre-trained ResNet152 with the heightmap data of 24 partitions and $20\times$+$50\times$ magnification. Overall, DINOv2 excels across all aspects, demonstrating the most stable performance amidst varying levels of granularity, magnification, and sensing modalities. The worst performance of the traditional baseline SIFT+FVs is better than the worst configurations of other deep learning methods, but its better configurations are significantly worse than those of the deep learning methods that achieved upper-echelon performance.

Explainability and Comparison with Human Experts. Currently, archaeologists have about 49.5% accuracy in a double-blind test with a similar setup [19]. Our in-house testing with two professional archaeologists yields an accuracy of 43.75% in a few-shot setting (Tab. 2). All tested models, except SIFT+FVs, are able to achieve far better accuracy (over 59.5%) under several configurations of the dataset. Notably, DINOv2 with linear probing is able to achieve superior or comparable performance under almost all possible configurations.

However, feature visualization demonstrated that DINOv2 (with registers for better visualization) sometimes

PreTr	20X TEX		50X TEX		20X HM		50X HM		20X+50X TEX		20X+50X HM	
	9w5s	9w20s	9w5s	9w20s	9w5s	9w20s	9w5s	9w20s	9w5s	9w20s	9w5s	9w20s
ResNet18	54.54	61.97	54.43	62.48	31.19	38.79	35.27	42.34	42.11	49.60	26.43	31.80
ResNet50	54.13	59.20	55.08	62.18	32.67	38.97	36.71	43.95	45.37	51.46	28.91	34.56
ResNet152	52.92	59.14	57.59	64.26	30.83	38.74	34.12	41.40	44.32	51.40	26.39	31.89
ConvNeXt-tiny	46.27	52.44	52.74	59.23	32.25	39.72	36.43	43.46	42.64	49.43	27.33	33.04
ConvNeXt-base	48.04	54.45	54.74	62.48	31.62	39.70	35.26	43.46	41.91	48.56	26.35	32.12
ConvNeXt-large	50.89	57.00	56.65	63.51	30.15	37.70	35.20	42.91	43.80	50.67	25.46	30.79
ViT-base	41.00	48.80	43.89	50.99	20.60	24.98	22.68	27.59	35.65	42.31	19.32	22.64
DINO-small	**58.85**	**66.10**	**59.50**	**67.35**	**33.55**	**41.27**	**42.52**	**51.11**	**46.94**	**53.93**	**28.02**	**33.49**
DINO-base	**57.28**	**65.33**	**61.39**	**69.67**	**33.07**	**41.83**	**42.39**	**51.23**	**47.52**	**55.34**	**28.23**	**34.00**
DeiT-small	47.00	55.99	52.08	60.64	29.36	36.70	35.45	44.43	39.93	47.68	26.18	32.14
DeiT-base	53.70	61.48	55.12	63.81	32.71	40.68	37.67	46.80	44.21	52.57	27.20	34.13
CLIP-base	42.98	51.30	46.52	55.01	29.75	36.92	36.91	44.51	34.45	41.29	27.81	34.02
GPT-4V	37.78	-	31.11	-	20.00	-	20.00	-	21.11	-	23.33	-
Human Expert	35.00	-	43.75	-	20.00	-	18.75	-	33.33	-	19.44	-

Table 2. Few-shot image classification performance on LUWA dataset is associated with the magnification and sensing modality. 'PreTr' denotes pre-trained models we used; '20X' and '50X' denote microscopic images at $20\times$ and $50\times$ magnifications; 'TEX' and 'HM' denote texture and heightmap; '9w5s' and '9w20s' denote 9-way-5-shot and 9-way-20-shot, respectively.

recognizes important polished regions in microscopic images of beechwood (see Fig. 5) as the foreground, but recognizes the same polish of different categories (sprucewood, bone, and antler, see Fig. 5) as the background. More explorations are needed to explain this unwanted behavior. Interestingly, we found that the regions recognized as highly important for classification are similar for human experts and our best model ResNet152, under the same data configuration (see Regions 1, 4, and 6 in Fig. 2). We visualize the results for ResNet152 using Grad-CAM [54].

4.2. Few-Shot Image Classfication

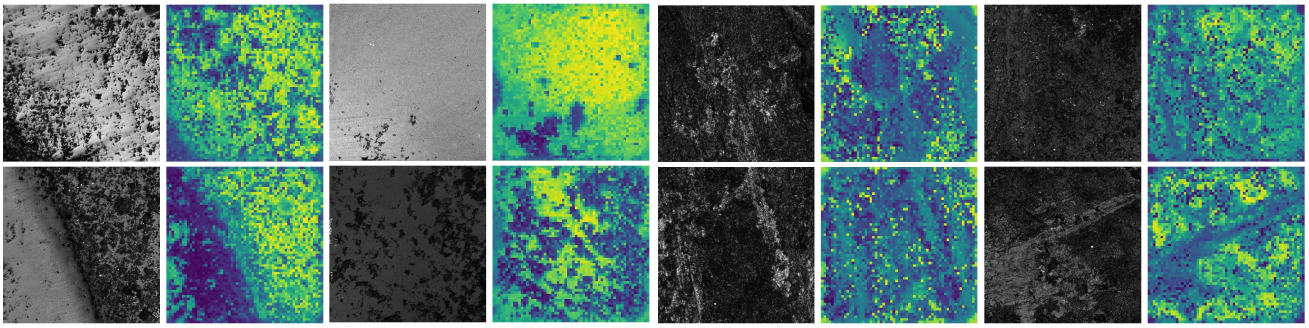
In practice, LUWA faces a scarcity of microscopic images due to limited stone tools, expertise requirements, and specialized equipment, especially when discovering new categories. Human experts can identify new classes of wear traces with a few examples. To emulate this, we investigate whether few-shot learning can be utilized and how microscopic image magnifications and sensing modalities influence the model's performance.

Experimental Settings. We designed two main experiments: **(i)** Few-shot image classification with a simple but effective pre-train + ProtoNet pipeline [27]. We evaluate the performance of powerful pre-trained models (including ResNet, ViT, DINO [10], ConvNeXt, CLIP [46], DEIT [63]) and popular meta-learners ProtoNet [57] on LUWA dataset. We simulated 600 episodes/tasks and results are demonstrated under 9-way-5/20-shot settings. **(ii)** GPT-4V [42] experiments: few-shot image classification and reasoning following instructions from human experts. To explore the potential mode of AI-human collaboration facing scientific domains in the advent of large multi-modal models, we collected prompts from three archaeologists and conducted case studies on whether the latest GPT-4V can follow and mimic the experts' reasoning process when analyzing the samples. Then we summarized key points that matter during the experts' analysis and used that to prompt the GPT-4V for few-shot image classification and reasoning. The experiments are illustrated in Fig. 6. Additionally, we included human experts' test results to reflect the difficulty for humans to distinguish these discovered categories with just a few examples. Results are reported under 9-way-5-shot settings in Tab. 2.

Results and Analysis. Experimental results in Tab. 2 demonstrated that DINO excels at few-shot learning classification. Note that in the case of limited microscopic images, classification results of textures at $50\times$ magnification setting yield notably superior results compared to others, which provides valuable guidance for few-shot learning tasks in our domain. Moreover, the number of parameters in pre-trained models has a limited impact on this few-shot learning task. We found that GPT-4V can effectively follow the experts' analysis as highlighted in Fig. 6. It learns to emphasize the same points that the experts pay attention to. However, the analysis doesn't always lead to the correct answer. GPT-4V did poorly on few-shot classification. We hypothesize that our data is very different from the web data that GPT models are trained on, and the vision module in GPT-4V still struggles to efficiently present detailed vision information to the language module, especially in a long context such as our multi-image few-shot classification scenario. This means the vision ability of SOTA multi-modal language models still needs improvement before they can be used in scientific tasks with domain-specific data like ours.

5. Impact and Limitations of LUWA Dataset

Scientific Impacts. AI-expert collaboration can provide invaluable insights for scientific research. To tackle the



(a) Sprucewood (b) Beechwood (c) Bone (d) Antler

Figure 5. Feature visualization of LUWA dataset using frozen pre-trained DINOv2.

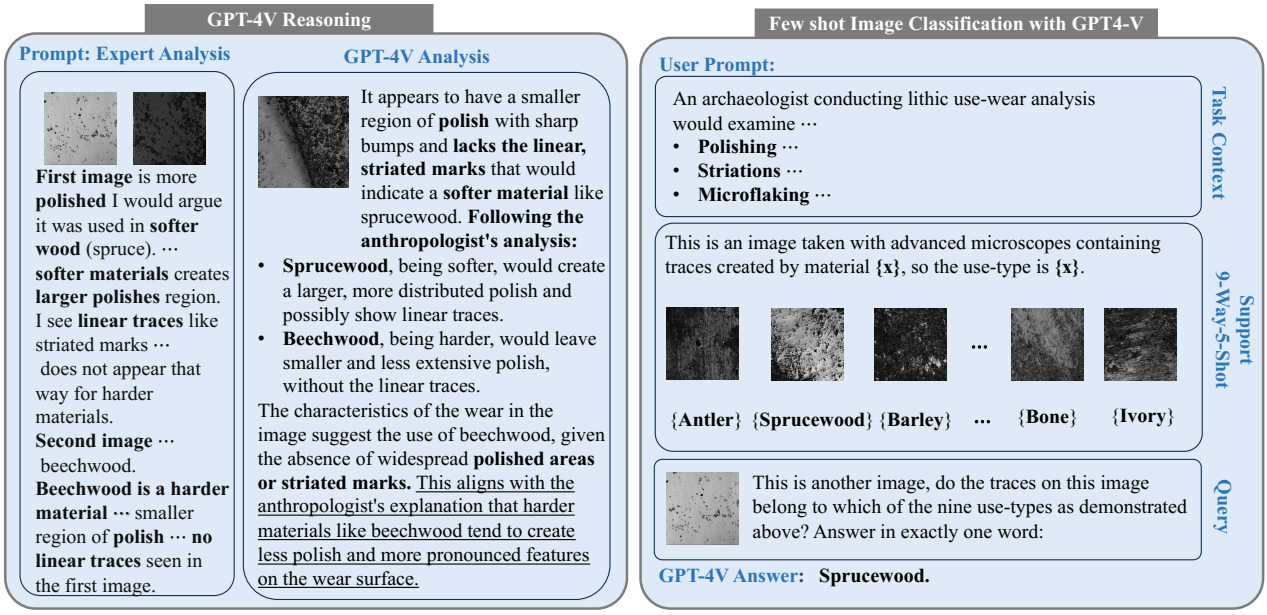


Figure 6. GPT-4V few-shot image classification and reasoning following instructions from human experts.

long-standing problem of stone tool use, we make the first attempt to collaborate with archaeologists and utilize advanced learning-based methods for worked material inference. The LUWA dataset allows for further investigations to advance our understanding of ancient tool use and material processing techniques.

Limitations and Future Directions. We will enrich the LUWA dataset from the following three aspects: **(i)** supplement microscopic images with both worn and unworn regions using lower-magnification objectives, allowing wear trace segmentation and detection tasks; **(ii)** collect images including wear traces caused by different worked material; **(iii)** increase categories of man-made materials for comprehensive analysis on wear features and material properties.

6. Conclusion

We collaborate with anthropological archaeologists and present the first public and the largest Lithic Use-Wear Analysis (LUWA) dataset benefiting both vision and science domains. The LUWA dataset serves as a benchmark to evaluate the generalization capabilities of advanced models on image classification tasks beyond common objects. Addressing specific challenges of wear formation and microscopic imaging, LUWA offers vital guidance on selecting suitable magnifications and sensing modalities facing different scenarios. Our analysis reveals that SOTA models encounter distinct difficulties when facing these specific challenges. Despite DINOv2's superior performance relative to other methods, it overlooks visual features that archaeologists identify as indicative of wear. We anticipate that the LUWA dataset will stimulate further research into enhancing the adaptability of large-scale models to specialized domains within computer vision.

Acknowledgment. This work is supported by NSF Grant 2152565, and by NYU IT High-Performance Computing resources, services, and staff expertise. We gratefully acknowledge Dr. Rakesh Behera for the tribometer hardware, and thank Sara Borsodi, Felix Devis Kisena, Kat Liu, Eugenia Ochoa, Vita Jackman Kuwabara, Alice Jiang, Meiyu Zhang, and Sriram Koushik for their valuable assistance in collecting the microscopic images.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *ArXiv*, abs/1610.01644, 2016. 5

[2] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*, 2023. 2

[3] Reza Akbarian Bafghi and Danna Gurari. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16261–16270, 2023. 2, 4

[4] Douglas B Bamforth. Investigating microwear polishes with blind tests: the institute results in context. *Journal of Archaeological Science*, 15(1):11–23, 1988. 3

[5] Tomasz Bartkowiak and Christopher A Brown. Multiscale 3d curvature analysis of processed surface textures of aluminum alloy 6061 t6. *Materials*, 12(2):257, 2019. 3

[6] Sebastiano Battiato, Alessandro Ortis, Francesca Trenta, Lorenzo Ascari, Mara Politi, and Consolata Siniscalco. Detection and classification of pollen grain microscope images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 980–981, 2020. 3

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[8] Wonmin Byeon, Manuel Dominguez-Rodrigo, Georgios Arampatzis, Enrique Baquedano, José Yravedra, Miguel Angel Maté-González, and Petros Koumoutsakos. Automated identification and deep classification of cut marks on bones and its paleoanthropological implications. *Journal of Computational Science*, 32:36–43, 2019. 2, 3

[9] Ivan Calandra, Lisa Schunk, Alice Rodriguez, Walter Gneisinger, Antonella Pedergnana, Eduardo Paixao, Telmo Pereira, Radu Iovita, and Joao Marreiros. Back to the edge: relative coordinate system for use-wear analysis. *Archaeological and Anthropological Sciences*, 11:5937–5948, 2019. 4

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5, 7

[11] Claire Lifan Chen, Ata Mahjoubfar, Li-Chia Tai, Ian K Blaby, Allen Huang, Kayvan Reza Niazi, and Bahram Jalali. Deep learning in label-free cell classification. *Scientific reports*, 6(1):21471, 2016. 3

[12] 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*, pages 3606–3613, 2014. 3

[13] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 5

[14] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 5

[15] Aniket Dashpute, Vishwanath Saragadam, Emma Alexander, Florian Willomitzer, Aggelos Katsaggelos, Ashok Veeraraghavan, and Oliver Cossairt. Thermal spread functions (tsf): Physics-guided material classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1641–1650, 2023. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[18] Manuel S Drehwald, Sagi Eppel, Jolina Li, Han Hao, and Alan Aspuru-Guzik. One-shot recognition of any material anywhere using contrastive learning with physics-based rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23524–23533, 2023. 3

[19] Adrian Anthony Evans. On the importance of blind testing in archaeological science: the example from lithic functional studies. *Journal of Archaeological Science*, 48:5–14, 2014. 3, 4, 6

[20] Adrian A Evans and Randolph E Donahue. Laser scanning confocal microscopy: a potential technique for the study of lithic microwear. *Journal of Archaeological Science*, 35(8): 2223–2230, 2008. 3

[21] Andrea Ferreri, Silvia Bucci, and Tatiana Tommasi. Multimodal rgb-d scene recognition across domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2199–2208, 2021. 2

[22] Richard LK Fullagar. The role of silica in polish formation. *Journal of Archaeological Science*, 18(1):1–24, 1991. 4

[23] Eric Hayman, Barbara Caputo, Mario Fritz, and Jan-Olof Eklundh. On the significance of real-world conditions for material classification. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8*, pages 253–266. Springer, 2004. 3

[24] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 5

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. 5

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[27] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022. 7

[28] Juan José Ibáñez, Talía Lazuen, and J González-Urquijo. Identifying experimental tool use through confocal microscopy. *Journal of Archaeological Method and Theory*, 26:1176–1215, 2019. 3

[29] Hartland W Jackson, Jana R Fischer, Vito RT Zanotelli, H Raza Ali, Robert Mechera, Savas D Soysal, Holger Moch, Simone Muenst, Zsuzsanna Varga, Walter P Weber, et al. The single-cell pathology landscape of breast cancer. *Nature*, 578 (7796):615–620, 2020. 3

[30] L.H. Keeley. *Experimental Determination of Stone Tool Uses: A Microwear Analysis*. University of Chicago Press, 1980. 1, 3

[31] Larry R Kimball, John F Kimball, and Patricia E Allen. Microwear polishes as viewed through the atomic force microscope. *Lithic Technology*, pages 6–28, 1995. 3

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[33] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 6

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 5

[35] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 5

[36] João Marreiros, Ivan Calandra, Walter Gneisinger, Eduardo Paixão, Antonella Pedergnana, and Lisa Schunk. Rethinking use-wear analysis and experimentation as applied to the study of past hominin tool use. *Journal of Paleolithic Archaeology*, 3:475–502, 2020. 3

[37] João Marreiros, Telmo Pereira, and Radu Iovita. Controlled experiments in lithic technology and function, 2020. 2, 3

[38] Shannon P McPherron, Zeresenay Alemseged, Curtis W Marean, Jonathan G Wynn, Denné Reed, Denis Geraads, René Bobe, and Hamdallah A Béarat. Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at dikika, ethiopia. *Nature*, 466(7308):857–860, 2010. 2

[39] Mark Newcomer, Roger Grace, and Romana Unger-Hamilton. Investigating microwear polishes with blind tests. *Journal of Archaeological Science*, 13(3):203–217, 1986. 3

[40] George H Odell. Stone tool research at the end of the millennium: classification, function, and behavior. *Journal of Archaeological Research*, 9:45–100, 2001. 2

[41] George Hamley Odell and Frieda Odell-Vereecken. Verifying the reliability of lithic use-wear assessments by 'blind tests': the low-power approach. *Journal of field Archaeology*, 7(1):87–120, 1980. 1, 2, 3

[42] OpenAI. Chatgpt can now see, hear, and speak, 2023. Accessed: 2023-11-16. 4, 7

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 5

[44] Antonella Pedergnana and Andreu Olle. Monitoring and interpreting the use-wear formation processes on quartzite flakes through sequential experiments. *Quaternary International*, 427:35–65, 2017. 2

[45] Marcos Pizarro-Monzo, Jordi Rosell, Anna Rufà, Florent Rivals, and Ruth Blasco. A deep learning-based taphonomical approach to distinguish the modifying agent in the late pleistocene site of toll cave (barcelona, spain). *Historical Biology*, pages 1–10, 2023. 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[47] Alice Rodriguez, Kaushik Yanamandra, Lukasz Witek, Zhong Wang, Rakesh K Behera, and Radu Iovita. The effect of worked material hardness on stone tool wear. *Plos one*, 17(10):e0276166, 2022. 3, 4

[48] Z Rymuza. Tribology of polymers. *Archives of civil and mechanical engineering*, 7(4):177–184, 2007. 3

[49] Jorge Sanchez and Florent Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR 2011*, pages 1665–1672, 2011. 5

[50] Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2165–2175, 2023. 3

[51] Patrick Schmidt, Alice Rodriguez, Kaushik Yanamandra, Rakesh K Behera, and Radu Iovita. The mineralogy and structure of use-wear polish on chert. *Scientific reports*, 10 (1):21512, 2020. 3

[52] Benjamin J Schoville, Kyle S Brown, Jacob A Harris, and Jayne Wilkins. New experiments and a model-driven approach for interpreting middle stone age lithic point function using the edge damage distribution method. *PloS one*, 11 (10):e0164088, 2016. 2

[53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo

Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pages 25278–25294. Curran Associates, Inc., 2022. 5

[54] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 4, 7

[55] S.A. SEMENOV. *Prehistoric Technology*. 1964. 3

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[57] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 7

[58] W James Stemp and Michael Stemp. Ubm laser profilometry and lithic use-wear analysis: a variable length scale investigation of surface topography. *Journal of Archaeological Science*, 28(1):81–88, 2001. 3

[59] W James Stemp, Adam S Watson, and Adrian A Evans. Surface analysis of stone and bone tools. *Surface Topography: Metrology and Properties*, 4(1):013001, 2015. 3

[60] Nathan E Stevens, Douglas R Harro, and Alan Hicklin. Practical quantitative lithic use-wear analysis using multiple classifiers. *Journal of Archaeological Science*, 37(10):2671–2678, 2010. 3

[61] Kenichiro Tanaka, Yasuhiro Mukaigawa, Takuya Funatomi, Hiroyuki Kubo, Yasuyuki Matsushita, and Yasushi Yagi. Material classification using frequency- and depth-dependent time-of-flight distortion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[62] SH Teoh. Fatigue of biomaterials: a review. *International journal of fatigue*, 22(10):825–837, 2000. 2

[63] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 7

[64] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, page 53, 2019. 3

[65] Hongda Wang, Hatice Ceylan Koydemir, Yunzhe Qiu, Bijie Bai, Yibo Zhang, Yiyin Jin, Sabiha Tok, Enis Cagatay Yilmaz, Esin Gumustekin, Yair Rivenson, et al. Early detection and classification of live bacteria using time-lapse coherent imaging and deep learning. *Light: Science & Applications*, 9(1):118, 2020. 3

[66] Ran Wang, Yuanjing Zhu, Chengxin Chen, Yu Han, and Hongbo Zhou. Tooth wear and tribological investigations in dentistry. *Applied Bionics and Biomechanics*, 2022, 2022. 2

[67] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 3

[68] Samuel J Yang, Marc Berndl, D Michael Ando, Mariya Barch, Arunachalam Narayanaswamy, Eric Christiansen, Stephan Hoyer, Chris Roat, Jane Hung, Curtis T Rueden, et al. Assessing microscope image focus quality with deep learning. *BMC bioinformatics*, 19:1–9, 2018. 3