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

Supplementary Material

Appendix

This document supplements the main paper as follows:

1. Describe dataset fidelity, material properties and human annotations (supplement Section 3.1).
2. More details about the training recipe and reproducibility (supplement section 4.1).
3. More visualizations and detailed tables (supplement section 4.1).
4. More details about the human expert tests(supplement section 4.2).

A. LUWA Dataset

A.1. Dataset Fidelity

Archaeological samples. Archaeologists struggle to reach a consensus on how to identify the worked material on ancient lithic tools because of a lack of ground truth information. LUWA aims to be the first step to building the benchmark and tool that can help archaeologists make more informed decisions as archaeologists believe the underlying physics should remain the same across real-world and lab-made use wear, and models that can work well on lab-made data could be an ancillary input to archaeologists' heuristics.

Worked time. We followed a tightly controlled protocol and "worked time" to reflect various wear degrees.

Impact of aging and conservation status. This is minimized because post-depositional alterations are usually visible under the microscope, and archaeologists can exclude pieces with signs of weathering.

A.2. Material Properties

Existing studies have indicated that both the hardness of materials and their silicon content can have an impact on the visual features of wear traces. This suggests that the properties of materials being worked or worn play a significant role in shaping the wear patterns observed. In machine wear experiments, we listed the hardness of worked materials for further explorations of wear mechanisms (see Tab. I). In human wear experiments, LUWA dataset supports fine-grained analysis on representative plants: horsetail has the highest silicon content, followed by ferns, and then barley.

A.3. Human Annotations

Human experts provide domain-specific knowledge for LUWA dataset in the following aspects (see Fig. I):

- Identification of Wear Traces: Human experts are actively involved in the process of data collection and are responsible for identifying wear traces on objects. Their expertise allows them to recognize and differentiate between various types of wear patterns, such as microwear polish, scratches, and impact marks. This identification is fundamental for understanding the history and use of the objects.

	Ivory	Antler	Bone	Beechwood	Sprucewood
Hardness	3.930±0.025	3.253±0.727	2.961±0.246	2.833±1.672	0.122±0.004

Table I. Hardness of worked materials in machine wear experiments.

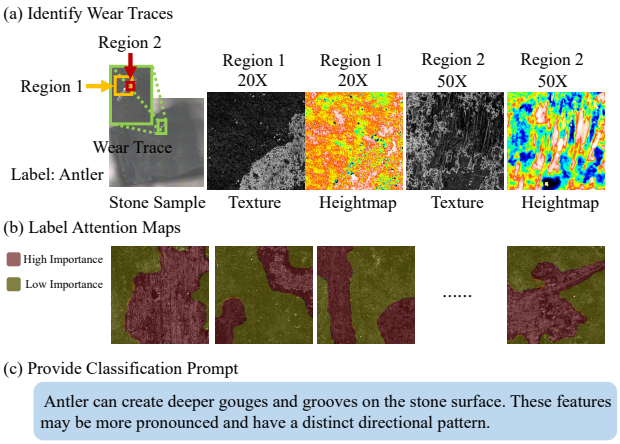


Figure I. Domain-specific expert knowledge: (a) human experts helped to identify wear traces during the process of data collection; (b) human experts labeled the most important region with red and the secondary important region with yellow when making decisions on worked materials; (c) human experts provided classification prompt for GPT-4V.

- Color Labeling for Attention Maps: During the decision-making process regarding worked materials, human experts use a color-coded system to label different regions of the objects. The most important regions are labeled with the color red, while less important regions are labeled with the color yellow. This color-coded system likely helps prioritize the analysis of wear traces and their significance in understanding the function and use of the objects.
- Classification Prompt for GPT-4V: Human experts also contribute by providing a classification prompt for GPT-4V, an AI model. This classification prompt likely guides the AI in recognizing and categorizing wear traces on objects, benefiting from the expertise of human specialists to enhance the accuracy of the AI's analysis.

B. Algorithm Benchmarking

B.1. Training Recipe

The start learning rate, which is also the η_{max} in the linear warmup with cosine annealing scheduler, is set to 0.01. The batch size for the smaller models, such as ResNets and ConvNeXts, is set to 200, while for larger models, such as ViT and DINOv2, it's set to 100 to save VRAM. We do not adjust the learning rate based on changing batch size because we believe our learning rate scheduler will offset the changes. When training from scratch, we train

for 20 epochs. We reduce that to 10 epochs when fine-tuning and linear probing. The dataset is partitioned into portions of 6/2/2 as train/val/test. To ensure fairness, we put all images from the same stone sample in the same set. We report all results based on the checkpoints with the lowest validation error. All experiments are done on a single Nvidia A100 with 80 GB VRAM. All models are trained three times with different random seeds and PyTorch `deterministic=True` and `benchmark=False` to maximize reproducibility. No data augmentation is applied except simple resizing to 224×224 to match pre-trained models' input dimension.

B.2. More Fully-Supervised Image Classification Results

We present more results that cannot fit into the main text.

More Visualization. A larger and clearer visualization is contained in Fig. II and Fig. III. As we can see, the trend described in Sec. 4.1 still holds true.

Quantitative Analysis. We provide quantitative analysis of the distribution overlap in the regions of interest as described in Sec. 4.1. As shown in Tab. II, we select IoU as the quantitative metric for evaluating the distribution overlap.

Table II. IoU for human labeling and Grad-CAM heatmaps.

	Fern	Sprucewood	Ivory	Beechwood	Before Use	Horsetail	Barley	Antler	Bone
IoU	0.9089	0.8577	0.7070	0.6959	0.6165	0.5773	0.4929	0.4535	0.3501

Data Configurations for the Best Performance. Tab. III shows the data configuration to achieve the best performance for each model. We can see the patterns described in Sec. 4.1 are well reflected among the top-performing models. Note that even though the best model for SIFT+FVs can achieve a reasonable performance of 52.88%, most of the other data configurations result in a significant performance downgrade for this method. In fact, this is the only super-human performance ($> 49.5\%$ accuracy) for this specific method.

Models that Achieve Super-Human Performance. Tab. VI contains all the models and their corresponding data configurations that achieve super-human performance. Out of 358 possible data configurations, 79 (22%) are able to achieve super-human performance. Tab. IV contains the count and ratio of different features that appear in super-human models, and we can see that this aligns with the trends described in the main text as well.

More on the Voting Mechanism. For the best performing models, Tab. V shows that when the final voted prediction is correct, how many partitions are predicted correctly before the voting (Corr Consis), and when the final voted prediction is incorrect, how many partitions are correct (Incorr Consis) or the same as the final wrongly-voted result (Incorr Common Consis). As we can see here, the predictions for each partition are relatively consistent before voting.

Model	Granularity	Magnification	Modality	Training Strategy	Accuracy
SIFT+FVs	24	50×	heightmap	N/A	52.88
ResNet50	6	20× + 50×	heightmap	Linear Probing	66.91
ResNet152	24	20× + 50×	heightmap	Linear Probing	67.05
ConvNeXt-tiny	24	20× + 50×	texture	Linear Probing	62.27
ConvNeXt-Large	24	20× + 50×	texture	Linear Probing	66.82
ViT-H	6	20× + 50×	heightmap	Linear Probing	62.5
DINOv2	24	20× + 50×	texture	Linear Probing	66.82

Table III. Best Performing Data Configuration for Each Model

Model Name	Count	Ratio	Training Strategy	Count	Ratio
ResNet50	16	20%	Linear Probing	66	84%
ResNet152	14	18%	From Sratch	8	10%
DINOv2	13	16%	Full-Parameter Fine-Tuning	4	5%
ConvNeXt-tiny	13	16%	**Granularity**	**Count**	**Ratio**
ConvNeXt-large	12	15%	24	37	47%
ViTH	10	13%	6	25	32%
SIFT+FVs	1	1%	1	17	22%
Magnification	**Count**	**Ratio**	**Sensing Modality**	**Count**	**Ratio**
20×	2	3%	Texture	36	46%
50×	38	48%	Heightmap	43	54%
20×+50×	39	49%	-	-	-

Table IV. Count and ratio of different features that appear in super-human models

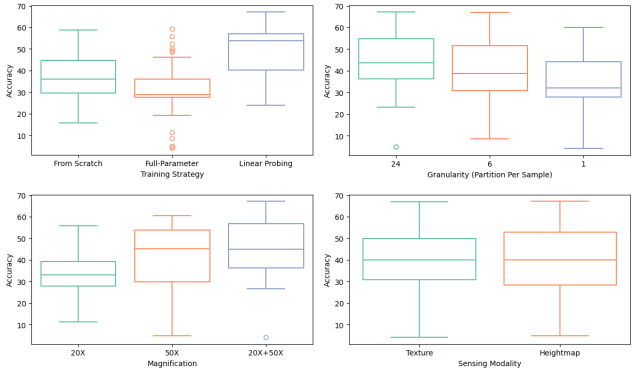


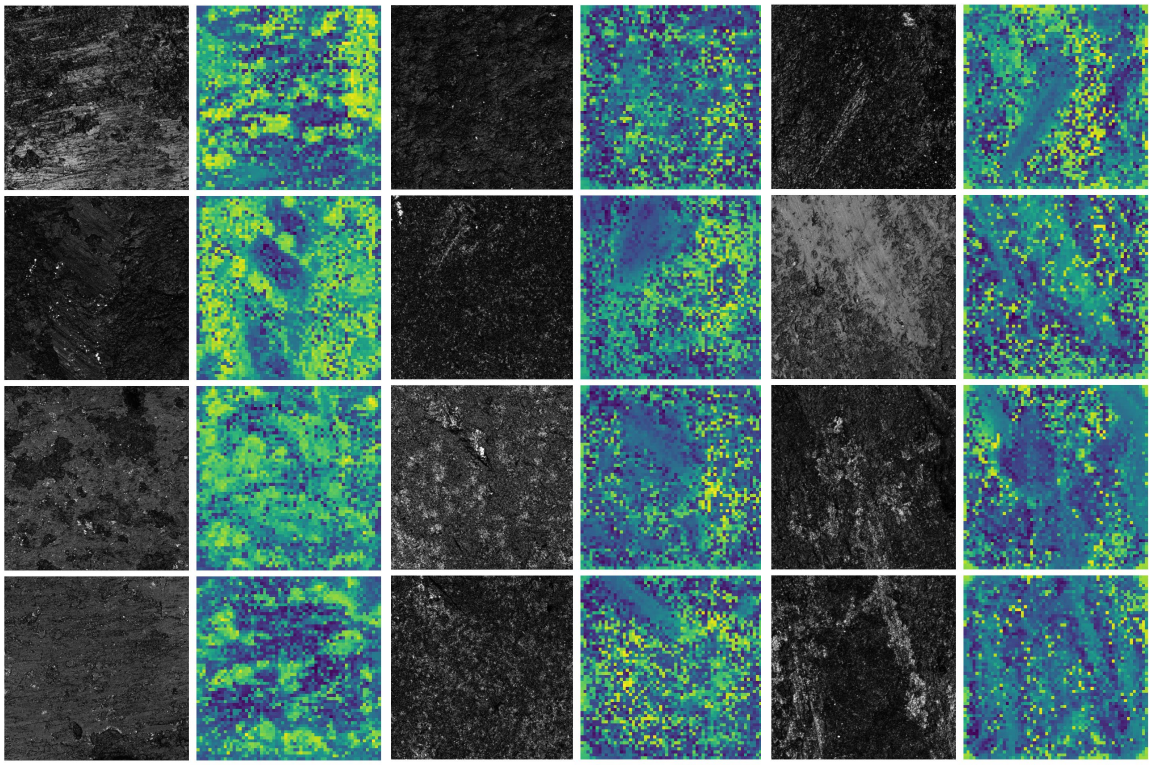
Figure II. The impact of the training strategy, granularity, magnification, and sensing modality on top-1 classification accuracy in %: Larger numbers in granularity mean more detailed information about a use-wear is fed into the model.

Table V. Consistency Analysis of the Voting Mechanism

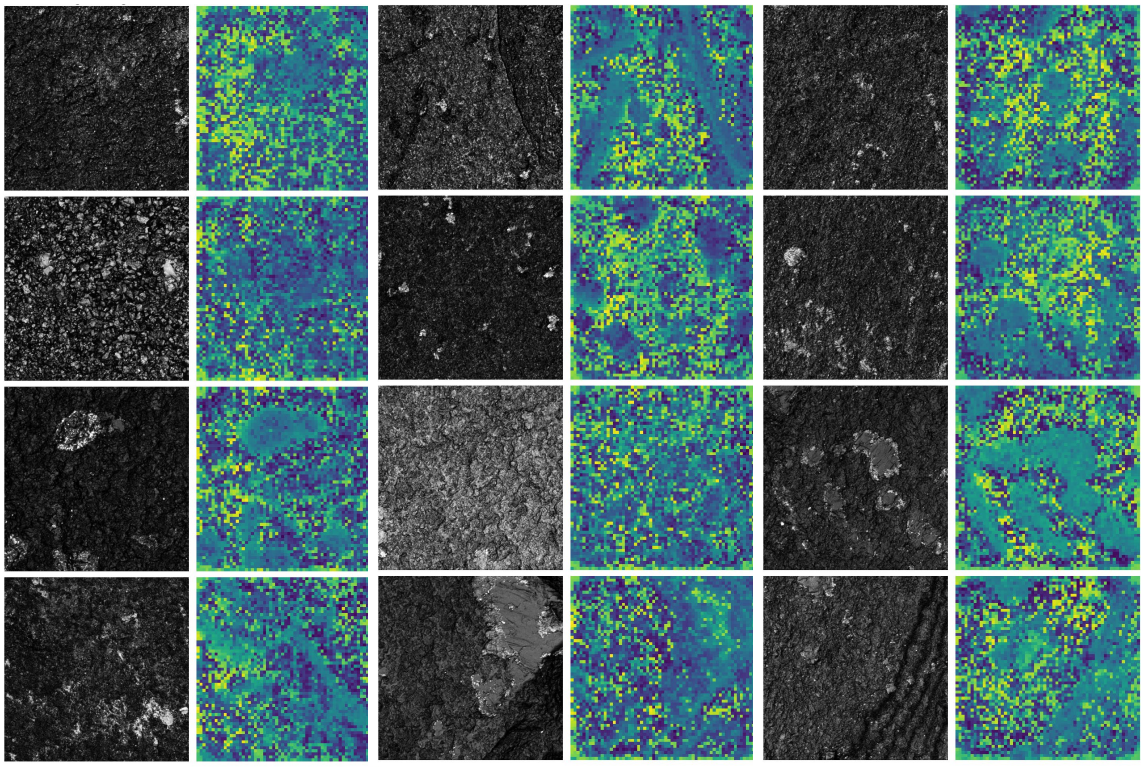
Model	Corr Consis	Incorr Consis	Incorr Common Consis
ResNet50	86.30%	8.15%	78.52%
ResNet152	78.85%	11.59%	62.14%
ConvNext-Tiny	82.48%	12.27%	60.84%
ConvNext-Large	78.57%	9.79%	66.55%
ViT-H	89.80%	9.33%	72.00%
DINOv2	86.34%	7.34%	66.90%

B.3. More Few-Shot Image Classification Details

In a test scenario where new categories of wear traces were identified, we provided identical support and query sets to both GPT-4V and two anthropologists. These anthropologists had no prior exposure to the samples in the sets, and we selected their best results for analysis.



(a) Ivory (b) Before Use (c) Bone



(d) Barley (e) Fern (f) Horsetail

Figure III. More feature visualization of LUWA dataset using frozen pre-trained DINOv2.

Model Name	Granularity	Magnification	Sensing Modality	Training Strategy	Accuracy
ResNet152	24	20× + 50×	heightmap	Linear Probing	67.05
ResNet50	6	20× + 50×	heightmap	Linear Probing	66.91
ConvNeXt-large	24	20× + 50×	texture	Linear Probing	66.82
DINOv2	24	20× + 50×	texture	Linear Probing	66.82
DINOv2	24	20× + 50×	heightmap	Linear Probing	66.14
ResNet50	24	20× + 50×	heightmap	Linear Probing	62.73
ViTH	6	20× + 50×	heightmap	Linear Probing	62.50
ConvNeXt-tiny	24	20× + 50×	texture	Linear Probing	62.27
ResNet152	6	20× + 50×	heightmap	Linear Probing	61.76
ConvNeXt-large	24	20× + 50×	heightmap	Linear Probing	61.59
ResNet50	24	50×	heightmap	Linear Probing	60.58
ConvNeXt-large	24	50×	heightmap	Linear Probing	60.58
ConvNeXt-tiny	6	20× + 50×	heightmap	Linear Probing	60.25
ConvNeXt-large	1	20× + 50×	heightmap	Linear Probing	60.00
DINOv2	24	50×	heightmap	Linear Probing	59.62
ResNet152	24	20× + 50×	heightmap	Full-Parameter Fine-Tuning	59.32
ResNet152	24	50×	heightmap	Linear Probing	58.65
ConvNeXt-large	1	50×	heightmap	Linear Probing	58.65
ResNet50	24	20× + 50×	heightmap	From Scratch	58.64
ResNet152	6	20× + 50×	texture	Linear Probing	58.50
ConvNeXt-tiny	24	20× + 50×	heightmap	From Scratch	58.41
ConvNeXt-tiny	6	20× + 50×	texture	Linear Probing	58.09
ConvNeXt-large	6	20× + 50×	heightmap	Linear Probing	58.09
DINOv2	6	20× + 50×	heightmap	Linear Probing	58.09
ConvNeXt-tiny	6	50×	heightmap	Linear Probing	57.69
ResNet152	24	50×	texture	Linear Probing	57.69
ConvNeXt-large	6	50×	heightmap	Linear Probing	57.69
ResNet152	1	20× + 50×	heightmap	Linear Probing	57.50
ViTH	6	20× + 50×	texture	Linear Probing	57.35
ViTH	24	20× + 50×	texture	Linear Probing	57.27
ViTH	24	50×	heightmap	Linear Probing	56.73
ConvNeXt-tiny	24	50×	texture	Linear Probing	56.73
ConvNeXt-tiny	1	20× + 50×	texture	Linear Probing	56.67
ConvNeXt-tiny	1	20× + 50×	heightmap	Linear Probing	56.67
DINOv2	1	20× + 50×	texture	Linear Probing	56.67
ResNet152	24	20× + 50×	heightmap	From Scratch	55.91
ConvNeXt-large	6	20× + 50×	texture	Linear Probing	55.88
ResNet152	24	20×	texture	Full-Parameter Fine-Tuning	55.82
DINOv2	1	50×	texture	Linear Probing	55.77
DINOv2	24	50×	texture	Linear Probing	55.77
ViTH	6	50×	heightmap	Linear Probing	55.77
ConvNeXt-tiny	24	50×	heightmap	Linear Probing	55.77
ResNet50	6	50×	heightmap	Linear Probing	55.77
ResNet50	24	20× + 50×	texture	Linear Probing	55.23
ResNet50	6	20× + 50×	heightmap	From Scratch	55.15
ResNet50	24	50×	texture	Linear Probing	54.81
ViTH	1	50×	heightmap	Linear Probing	54.81
DINOv2	6	50×	texture	Linear Probing	54.81
ConvNeXt-tiny	6	50×	texture	Linear Probing	54.81
DINOv2	6	50×	heightmap	Linear Probing	54.81
ConvNeXt-large	24	50×	texture	Linear Probing	54.81
ResNet152	24	20× + 50×	texture	Linear Probing	54.77
ConvNeXt-tiny	24	20× + 50×	heightmap	Linear Probing	54.55
DINOv2	24	20×	texture	Linear Probing	54.39
ResNet152	6	50×	heightmap	Linear Probing	53.85

ConvNeXt-tiny	1	$50\times$	texture	Linear Probing	53.85
ConvNeXt-large	1	$50\times$	texture	Linear Probing	53.85
ResNet50	6	$50\times$	texture	Linear Probing	53.85
ViTH	24	$50\times$	texture	Linear Probing	53.85
ResNet50	1	$50\times$	heightmap	Linear Probing	53.85
ViTH	6	$50\times$	texture	Linear Probing	53.85
DINOv2	1	$20\times + 50\times$	heightmap	Linear Probing	53.33
ResNet50	24	$20\times + 50\times$	texture	From Scratch	53.18
DINOv2	6	$20\times + 50\times$	texture	Linear Probing	52.94
SIFT+FVs	24	$50\times$	heightmap	NaN	52.88
ConvNeXt-large	6	$50\times$	texture	Linear Probing	52.88
ConvNeXt-tiny	1	$50\times$	heightmap	Linear Probing	52.88
ResNet50	24	$20\times + 50\times$	heightmap	Full-Parameter Fine-Tuning	52.27
ResNet152	6	$50\times$	texture	Linear Probing	51.92
DINOv2	1	$50\times$	heightmap	Linear Probing	51.92
ResNet50	6	$20\times + 50\times$	texture	Linear Probing	51.47
ViTH	1	$50\times$	texture	Linear Probing	50.96
ResNet152	1	$50\times$	heightmap	Linear Probing	50.96
ResNet152	24	$20\times + 50\times$	texture	From Scratch	50.91
ResNet50	1	$20\times + 50\times$	heightmap	Linear Probing	50.83
ViTH	24	$20\times + 50\times$	heightmap	Linear Probing	50.45
ConvNeXt-large	24	$50\times$	heightmap	From Scratch	50.00
ResNet50	24	$50\times$	texture	From Scratch	50.00
ResNet50	24	$50\times$	texture	Full-Parameter Fine-Tuning	50.00

Table VI. All the models and their data configuration that achieve super-human performance (accuracy $> 49.5\%$)