Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multilabel Corrosion Morphology Classification Using Deep Learning with Multimodality Fusion

Atira Ferdoushi
University of Texas at El Paso
Daria Bentley
The Ohio State University
Hailong Dai
Colorado School of Mines
Xiaolei Guo
Colorado School of Mines
Jenifer S. Locke
The Ohio State University
Fei Xu
fxu@utep.edu

University of Texas at El Paso

Article

Keywords: Corrosion morphology, ResNet-18, CLIP, Swin-T, Multimodal learning, Class-weighted loss

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11380194/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multilabel Corrosion Morphology Classification Using Deep Learning with Multimodality Fusion

Atira Ferdoushi[1], Daria Bentley[2], Hailong Dai[3], Xiaolei Guo[3], Jenifer S. Locke[2], Fei Xu[1,#]

[1] University of Texas at El Paso, El Paso, TX, United States

[2] The Ohio State University, Columbus, OH, United States

[3] Colorado School of Mines, Golden, CO, United States

[#]Corresponding author

Fei Xu: fxu@utep.edu

Abstract

Accurate characterization of corrosion morphology is critical for understanding degradation mechanisms and supporting predictive maintenance. Corrosion images often exhibit multiple coexisting morphologies, making single-label classification inadequate. This study presents a multilabel classification framework using ResNet-18, Contrastive Language-Image Pretraining (CLIP), and Swin Transformer (Swin-T), with environmental and experimental metadata incorporated through multimodal fusion. A detailed experimental evaluation examines model architecture and training configurations. Comprehensive experiments demonstrate that class-weighted loss functions are critical for enabling metadata fusion and improving macro-F1 performance for minority labels. Transformer-based and vision-language models exhibit strong adaptability to complex corrosion patterns. Ablation studies further quantify the effects of input modality, loss weighting, training epochs, and batch size. The results establish a benchmark for multilabel corrosion morphology classification and highlight the importance of loss-function design in multimodal learning.

Keywords: Corrosion morphology, ResNet-18, CLIP, Swin-T, Multimodal learning, Class-weighted loss

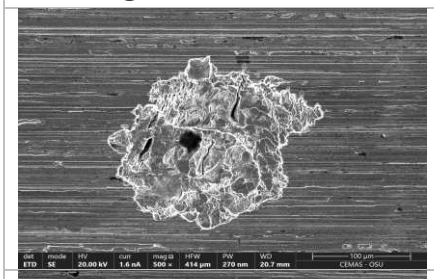
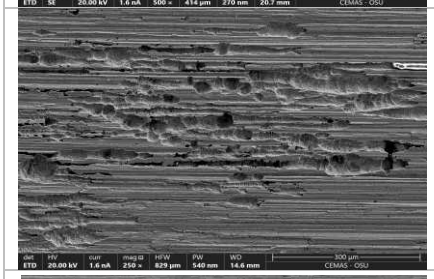
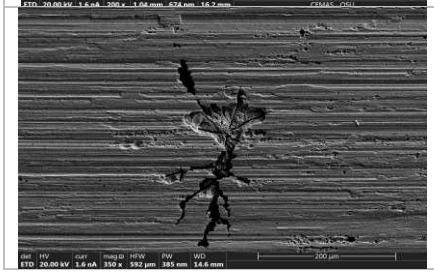
Introduction

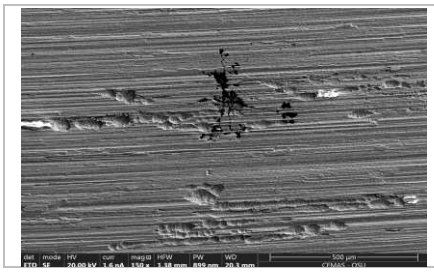
Corrosion is one of the leading causes of structural degradation and failure in engineering systems, resulting in substantial economic losses and significant safety risks [1,2]. Among the various forms of corrosion, localized corrosion is of particular concern because it initiates at discrete sites, often progresses rapidly with limited overall material loss, and can ultimately trigger catastrophic structural failure. In susceptible materials, localized corrosion such as pitting can further interact with applied or residual stresses to initiate stress corrosion cracking (SCC), one of the most aggressive forms of environmentally assisted degradation [3,4].

Reliable lifetime prediction therefore requires accurate identification of pit morphology. To address this need, an automated corrosion assessment framework is proposed to establish a meaningful basis for categorizing corrosion morphologies. In this study, the morphology of pitting corrosion observed in 304H stainless steel (SS304H) samples exposed to seawater droplets was classified into three categories—Complex, Smooth and Cracking—as summarized in Table 1. These morphologies were obtained on mill-finished SS304H samples exposed in a RH-temperature-controlled environmental chamber, where artificial seawater brines equilibrated to the target RH were applied by either pipetting or salt spraying [5]. Specifically, the Smooth morphology refers to corrosion pits with relatively smooth, polished

pit bottoms seen at higher RH conditions. The Complex morphology describes pits exhibiting cross-hatched surface features associated with preferential dissolution along slip planes at lower RHs. Cracking represents a distinct morphological feature also seen at low RH conditions in which small cracks emanate from the boundaries of corrosion pits, which can occur in tandem with the complex morphology. These differences in pit morphology arise from solution chemistry differences that are RH-dependent [3,5,6], as well as surface plastic deformation and the associated residual stress [7,8]. The resulting morphologies have important implications for structural integrity. In particular, complex pits and pits exhibiting crack initiation can act as more deleterious stress concentrators, increasing the susceptibility to SCC [3,6]. Therefore, accurately distinguishing among these morphologies is essential for accurate lifetime predictions.

Table 1 | Illustration of example pit morphologies

Images	Metadata	Labels
	File name: 2TB40_28d, Material: SS304H tensile bar; Temp: 35 °C; RH: 40%; Time: 28 days; Cleaned: Yes	Complex
	File name: TB66_28d, Material: SS304H tensile bar; Temp: 35 °C; RH: 56%; Time: 28 days; Cleaned: Yes	Polished/Smooth
	File name: T2_56RH, Material: SS304H tensile bar; Temp: 35 °C; RH: 56%; Time: 28 days; Cleaned: Yes	Complex, Smooth
	File name: 3CP_56d, Material: SS304H coupon; Temp: 35 °C; RH: 56%; Time: 28 days; Cleaned: Yes	Complex, Cracking

	File name: TB61_28d_56soln, Material: SS304H tensile bar; Temp: 35 °C; RH: 40%; Time: 28 days; Cleaned: No	Complex, Polished/Smooth, Cracking

However, corrosion morphologies observed in the real world rarely fall into a single, well-defined category [9]. The scanning electron microscopy (SEM) image in Fig. 1 illustrates this challenge: although the surface exhibits the broad textural irregularities characteristic of a Complex morphology, a distinct inner crack-like region is also present, highlighted for clarity. Features of this kind introduce additional structural variation that complicates attempts to assign a single dominant label. Such overlapping or nested patterns are common in real-world corrosion, where multiple degradation mechanisms may act simultaneously within the same region of interest. Therefore, multilabel classification is essential for accurately capturing the complexity of real-world corrosion, where multiple degradation mechanisms frequently coexist within a single image.

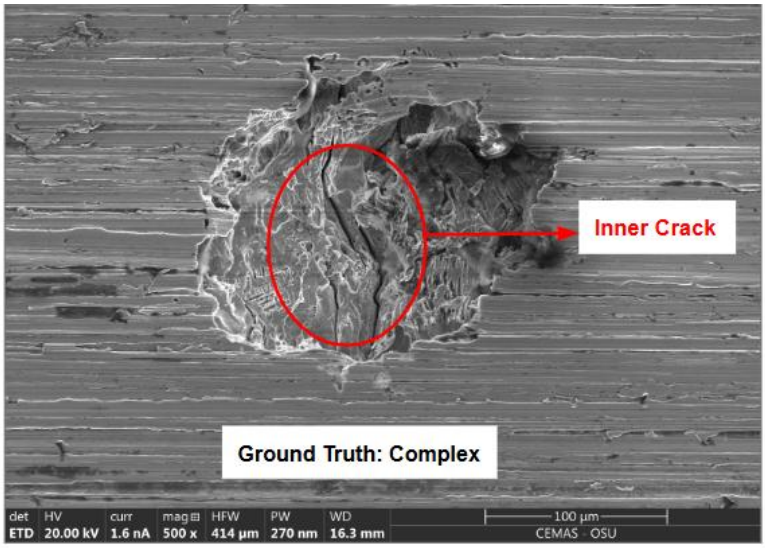


Fig. 1 | SEM image with ground-truth label Complex, showing an inner crack-like region highlighted for visual interpretation.

These challenges are amplified by the dataset itself. Certain morphologies and morphology combinations occur infrequently, leaving only a small number of examples for the model to learn from. As a result, minority classes are consistently underpredicted, even when present. The most problematic cases are the subtle ones—where corrosion signatures blend into the surrounding texture and even expert evaluators may disagree on the correct

classification [10]. These limitations highlight the need for machine learning approaches that can robustly recognize subtle morphological features and effectively learn from imbalanced datasets. Therefore, researchers in materials science have increasingly adopted machine learning tools to analyze the growing volume of image data from microscopy and surface analysis.

Recent studies have demonstrated that computer vision models can segment corrosion damage and classify defect types across a range of microscopy and surface imaging modalities [11]. Cai et al. [12], for example, applied convolutional neural networks to streamline the evaluation of the tri-structural isotropic (TRISO) fuel cross-section, reducing the need for manual inspection. Similar advances have been reported for localized corrosion, where deep learning models can delineate damaged regions in optical and SEM images with high accuracy [13]. More recent work has expanded into corrosion specific feature extraction, including pit detection, crack mapping, and surface condition classification using transformer-based architectures and self-supervised learning [12,13,14]. Despite this progress, many existing image-based materials characterization approaches focus on segmentation or single-label identification [15]. However, these approaches do not explicitly address multilabel morphology classification, where multiple corrosion morphologies may coexist within the same image. This gap is further complicated by class imbalance: only a small number of samples contain certain morphologies or morphology combinations, causing models to underpredict minority classes even when present. These limitations highlight the need for methods that can handle subtle, overlapping, and sparsely represented corrosion features—conditions that remain a central challenge in current corrosion research.

To address the challenge of reliably identifying overlapping corrosion morphologies, a multilabel classification framework is developed using three complementary vision architectures—ResNet-18 [16], CLIP [17], and Swin-Transformer (Swin-T) [18], the tiny variant of Swin), to provide a fair comparison across convolutional, vision–language and hierarchical transformer models when learning corrosion features from SEM images. The study further investigates whether incorporating experimental metadata and applying class-weighted loss can improve performance, particularly for minority labels such as Complex and Cracking coexisting. Model performance is assessed using exact-match accuracy, Hamming loss, micro-F1, and macro-F1, and qualitative evaluation on unseen data helps reveal where each architecture succeeds or fails.

Two key findings emerge. First, metadata fusion does not improve performance under an unweighted loss; the architectures show no consistent gain over the Image Only baseline, with noticeable drops in macro-F1. Second, class-weighted loss fundamentally changes

this behavior: once weighting is applied [19], metadata fusion becomes the best-performing configuration for all models, yielding the highest macro-F1 scores and the most balanced multilabel predictions. These results highlight the importance of multimodal inputs and loss function design for robust corrosion feature classification in real-world materials analysis.

The Results section reports model performance, qualitative behavior, generalization to unseen data and ablation studies; the Discussion addresses label ambiguity and limitations; and the Methods describe the dataset, model architectures and training procedures.

Results

This section summarizes the performance of the three models on the multilabel corrosion-classification task. Four evaluation metrics are reported: exact-match accuracy, Hamming loss, micro-F1 and macro-F1. The results also allow a comparison of how different input configurations influence performance across the three architectures.

Evaluation metrics

To analyze model performance, four standard multilabel metrics were used: exact-match accuracy, Hamming loss, micro-F1 and macro-F1. Let *N* be the number of test images and *L* = 3 the number of labels (Complex, Smooth and Cracking). For each image i and label j, the ground-truth value is $y_{ij} \in \{0, 1\}$, and the model's prediction is $\hat{y}_{ij}$.

1. **Exact-match accuracy** checks whether the entire predicted label set for an image matches the ground truth. It is defined as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} I(y_i = \widehat{y_i})$$

where *I(·)* returns 1 when all labels for image *i* are correctly predicted and 0 otherwise. This metric is easy to interpret but very strict—missing even one label counts the whole image as incorrect.

2. **Hamming loss** looks at errors on individual labels rather than the full label set. Lower values indicate better performance.

$$\text{Hamming Loss} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} I\left(y_{ij} \neq \widehat{y_{ij}}\right)$$

This metric is less strict than exact-match accuracy because partially correct predictions are credited.

To define the F1-based metrics, standard per-label precision and recall are used:

$$\text{Precision}_j = \frac{TP_j}{TP_j + FP_j}, \qquad \text{Recall}_j = \frac{TP_j}{TP_j + FN_j}$$

For each label j, TP_j , FP_j, and FN_j correspond to the true-positive, false-positive, and false-negative counts. Precision reflects the correctness of positive predictions, while recall captures the model's ability to recover actual positive instances.

3. **micro-F1** aggregates true positives, false positives, and false negatives across all labels. Since frequent labels dominate the counts, micro-F1 tends to reflect performance on the common classes.

$$\text{micro-F1} = \frac{2\sum_{j=1}^{L} TP_j}{2\sum_{j=1}^{L} TP_j + \sum_{j=1}^{L} FP_j + \sum_{j=1}^{L} FN_j}$$

4. **macro-F1** computes the F1-score for each label separately and then averages them. This metric gives equal weight to every label, regardless of how often it appears.

$$\text{macro-F1} = \frac{1}{L}\sum_{j=1}^{L} \frac{2 \cdot \text{Precision}_j \cdot \text{Recall}_j}{\text{Precision}_j + \text{Recall}_j}$$

The rarity of Cracking compared to the other labels leads us to use macro-F1 for model selection and early stopping. This metric penalizes models that perform well on common labels but ignore rare ones, which is exactly what class-weighted loss helps prevent.

Overall performance

Fig. 2 shows each model at its best batch size (8 for ResNet-18, 16 for CLIP and 8 for Swin-T) for three configurations (see Methods, Input configurations). These configurations—Image Only, Image with Text, and Image with Text (Weighted)—allow evaluation of how metadata influences multilabel corrosion classification.

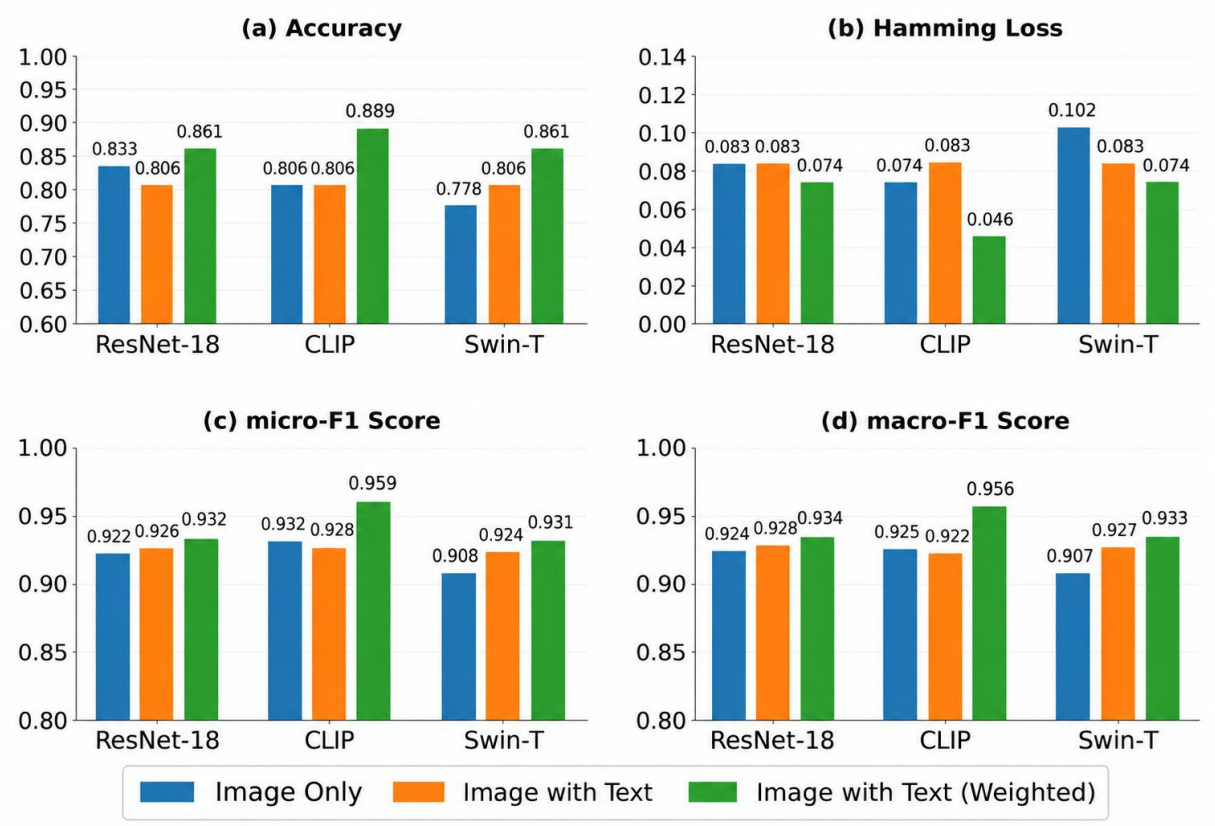


Fig. 2 | Overall Performance

Two observations emerge from the figure.

1. Adding metadata without loss weighting does not consistently improve performance. In several cases, it even reduces it. Under an unweighted loss, all three architectures show either a drop or only marginal gains when metadata is fused with the image. For ResNet-18, macro-F1 increases slightly from 0.924 to 0.928, and Hamming loss remains unchanged, indicating that unweighted metadata provides only minor gains. For CLIP, macro-F1 decreases from 0.925 to 0.922, while Hamming loss improves from 0.083 to 0.074. This shows that metadata without weighting does not reliably enhance CLIP's already strong baseline. For Swin-T, macro-F1 increases from 0.907 to 0.927, and Hamming loss improves from 0.102 to 0.083, suggesting modest benefits. A plausible explanation is that increasing the input dimensionality makes the model more prone to

overfitting and that the metadata do not provide additional morphological cues as the essential information is already present in the image.

2. Under weighted loss, all three architectures improve across nearly all metrics as shown in Fig. 2. Metadata clearly helps when the model is encouraged to pay more attention to minority labels. CLIP's weighted fusion configuration reaches the highest accuracy (0.889), 0.956 macro-F1 and the lowest Hamming loss. ResNet-18 achieves its highest macro-F1 of 0.934 with a Hamming loss of 0.074. Swin-T follows the same trend: weighted fusion outperforms both the Image Only and unweighted fusion variants across accuracy, micro-F1, macro-F1, and Hamming loss.

Metadata becomes more helpful when paired with class-weighted loss: across all three architectures, the weighted fusion model surpasses the Image Only baseline on every metric. Among the three models, CLIP stands out as the strongest overall performer: it reaches the highest accuracy (0.889), achieves the highest macro-F1 (0.956), and achieves consistently low Hamming loss. CLIP's strong Image Only baseline, combined with its additional gains under weighted loss, makes it the most robust architecture across configurations.

Per-label analysis

Per-label F1 scores reveal distinct strengths among the three architectures (Table 2). ResNet-18 achieved its best results with weighted fusion, reaching high F1-scores for Complex and Smooth. Its Cracking performance remained between 0.87 and 0.88 across all configurations, indicating that class weighting improves texture-based classification but provides limited benefit for crack detection.

Table 2 | Per-label F1 performance on the test set.

Models	Experiments	Complex F1	Smooth F1	Cracking F1
ResNet-18	Image Only	0.93	0.97	0.87
	Image with Text	0.94	0.97	0.88
	Image with Text (Weighted)	0.96	0.97	0.88
CLIP	Image Only	1.00	0.86	0.92
	Image with Text	0.98	0.86	0.93
	Image with Text (Weighted)	1.00	0.94	0.93
Swin-T	Image Only	0.93	0.91	0.88
	Image with Text	0.94	0.97	0.88
	Image with Text (Weighted)	0.95	0.97	0.88

CLIP exhibited the strongest Cracking performance, achieving an F1-score of 0.93, higher than that of the other models. Metadata integration improved Smooth (0.86 to 0.94) and

restored perfect performance on Complex, whereas Cracking changed little (0.92 to 0.93), indicating that the pretrained visual encoder alone captures crack-related features most effectively. In contrast, Swin-T consistently benefited from metadata for Complex and Smooth, but its Cracking score also remained fixed at 0.88.

Fig. 3 shows confusion matrices for the best-performing configuration of each model. For the Complex class, CLIP achieves perfect classification, while ResNet-18 makes one error and Swin-T produces two false negatives. Smooth shows a similar pattern of high accuracy: ResNet-18 and Swin-T each miss one positive sample, and CLIP adds one false positive to its single false negative. The largest difference occurs for Cracking: CLIP achieves perfect recall with only three false positives, whereas ResNet-18 and Swin-T each miss two Cracking samples and generate three false positives. These results indicate that CLIP is particularly sensitive to crack-like features, while ResNet-18 and Swin-T are more effective at distinguishing subtle surface textures.

Despite similar overall performance, the models achieved their results through different mechanisms. CLIP performs strongly even without metadata, reflecting the benefits of large-scale pretraining and a frozen backbone with a lightweight classification head. Under weighted fusion, CLIP attains the highest macro-F1 of 0.956 and the highest accuracy of 0.889, while ResNet-18 reaches a competitive macro-F1 of 0.934. These findings suggest that ResNet-18 provides stronger per-label discrimination, while CLIP is more reliable for predicting complete label sets.

Choosing among the three models depends on the features that matter most. CLIP is the strongest choice for detecting crack-like patterns, ResNet-18 performs best on smoother textures, and Swin-T offers balanced gains when metadata is included. No single model

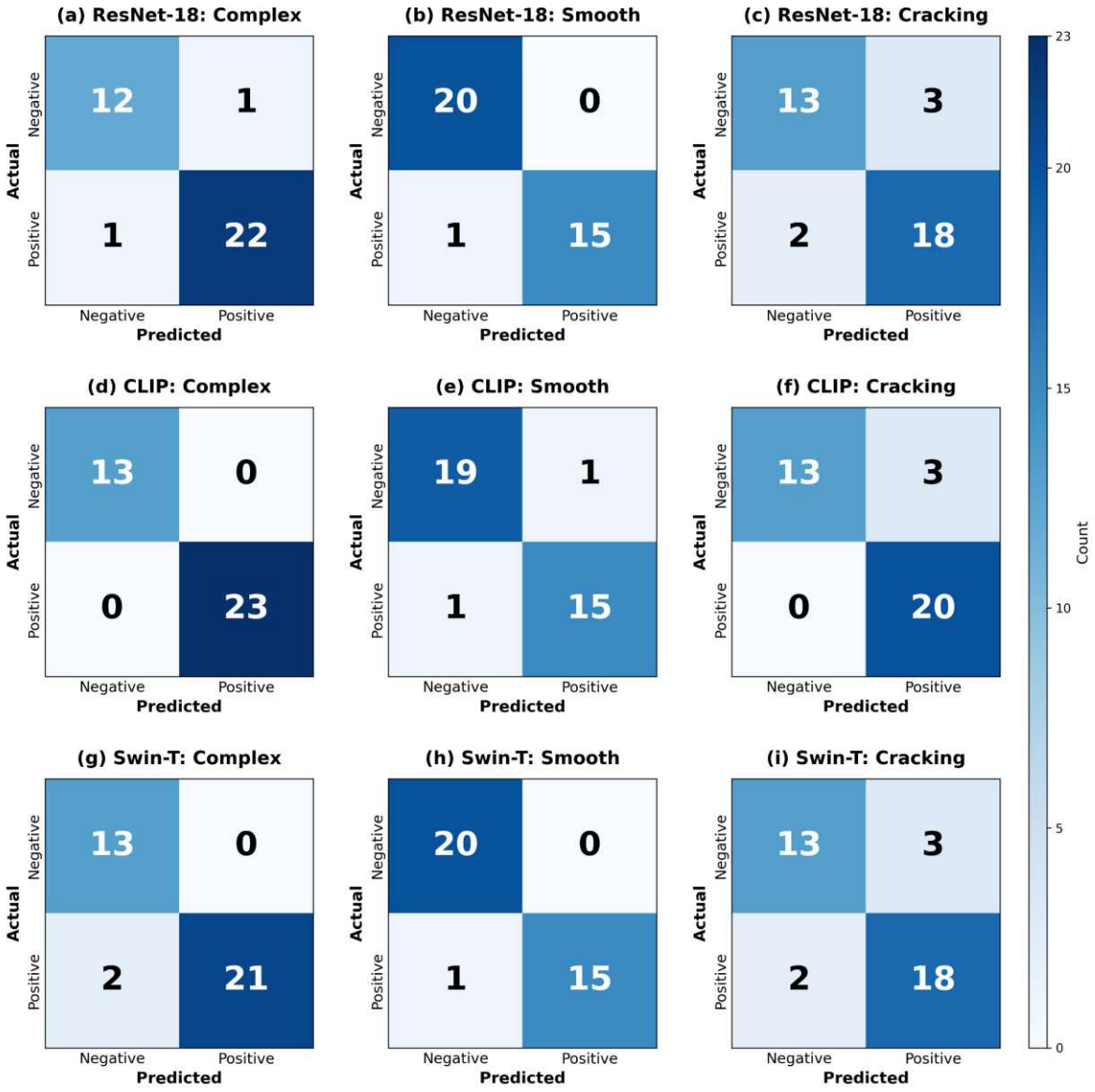


Fig. 3 | Per-label confusion matrices for ResNet-18, CLIP and Swin-T under the Image with Text (Weighted) configuration on the test set. Columns show the three morphology labels: (a, d, g) Complex, (b, e, h) Smooth, and (c, f, i) Cracking.

excels at every morphology, but CLIP provides the most reliable performance for the patterns that are hardest to capture.

Training curve analysis

Fig. 4 shows the training and validation loss curves for all nine input setups. Across all configurations, the loss steadily declines, which confirms that the models are learning as training progresses.

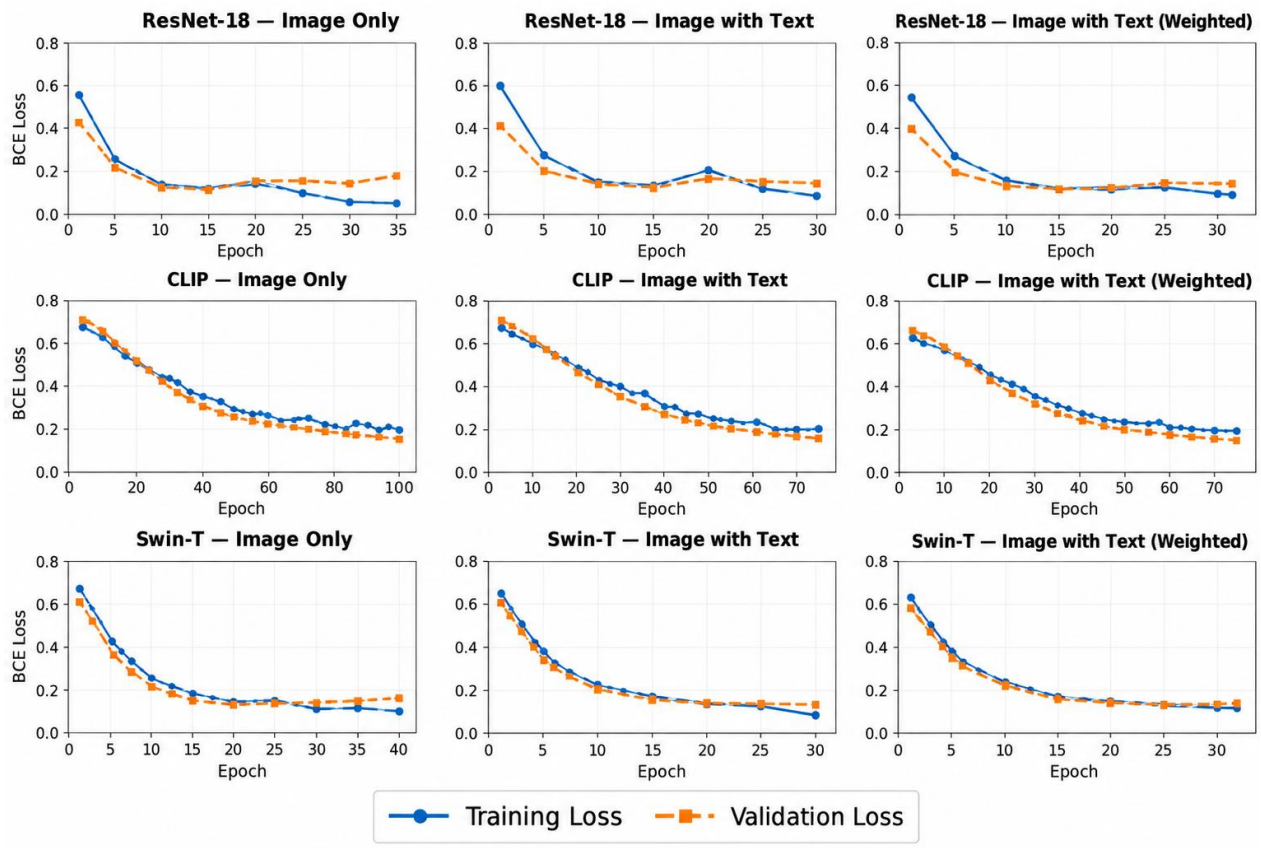


Fig. 4 | Training and validation loss curves for ResNet-18, CLIP and Swin-T across the three input configurations.

For ResNet-18, the Image Only run develops a clear gap between the training and validation curves toward the end. The training loss keeps dropping, but the validation loss starts to rise, which signals mild overfitting. When metadata and class-weighted loss are added, the gap between the training and validation curves shrinks, and the two curves remain closer throughout training. This suggests that both additions help the model generalize. The ResNet-18 curves also look slightly uneven, including a small bump around epoch 20 in the Image with Text run. That bump is most likely due to the learning rate scheduler adjusting the rate rather than any real instability in training.

CLIP shows a different pattern. The Image Only configuration has the closest match between training and validation loss across all nine plots, which fits with the strength of pretrained features. The fusion setups introduce more variation, and in the later epochs the validation loss even dips slightly below the training loss. This can happen when dropout is active during training but disabled during evaluation, so it is not a sign of overfitting.

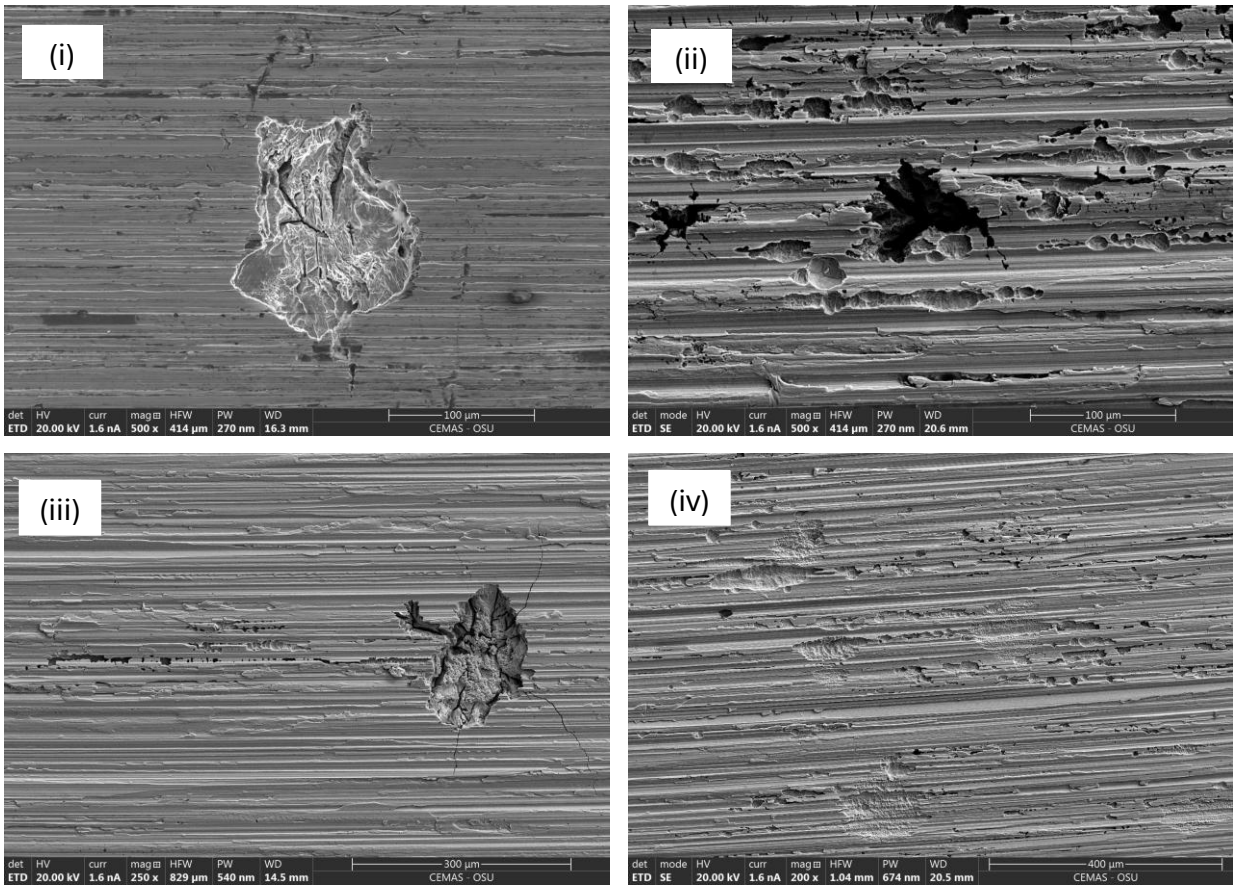
Swin-T exhibits steady and stable curves across all three setups. The training and validation loss both decline smoothly, and the two lines stay close to each other throughout training.

Adding metadata lowers the validation loss and keeps the curves tighter, especially in the Image with Text (Weighted) runs. Even with the backbone frozen, the head still learns to use the extra information. Swin-T does not show the widening gap seen in ResNet-18 or the slight dip seen in CLIP, giving it a calm and predictable training pattern.

Overall, ResNet-18 becomes more stable when metadata and class weighting are added, CLIP keeps smooth curves because its backbone is frozen, and Swin-T shows steady learning with clear gains from metadata even though the backbone stays fixed. CLIP also shows the most consistent training behavior, with training and validation loss staying closely aligned throughout.

1.1 Qualitative results

(a) Input Images



(b) Ground truth and Predicted labels

Images	Ground truth	ResNet-18	CLIP	Swin-T
(i)	Complex	Complex, Cracking	Complex, Smooth, Cracking	Complex, Cracking
(ii)	Complex, Smooth, Cracking	Smooth	Complex, Cracking	Smooth
(iii)	Complex, Cracking	Complex, Cracking	Complex, Cracking	Complex, Cracking
(iv)	Smooth	Smooth	Smooth	Smooth

Fig. 5 | (a) Input SEM images from the test set. (b) Predicted multilabel outputs for each model compared with the ground-truth labels across the images for Image with Text (Weighted) configuration.

Fig. 5 presents the classification outputs of the three models on four representative test images, alongside the corresponding ground truth labels. Although all three models reliably identify the Complex morphology when it is present, their behavior differs noticeably when additional classes such as Smooth or Cracking appear.

Fig. 5 (i) and (ii) illustrate typical mismatch cases. In 5(i), the ground truth is Complex, but the image contains a small inner crack, which leads all models to predict Cracking as well. CLIP predicts Complex, Smooth, and Cracking, showing that it sometimes interprets fine-scale texture as belonging to multiple classes. In 5(ii), the ground truth includes Complex, Smooth and Cracking. ResNet-18 and Swin-T predict only Smooth, missing both the complex texture and the crack. CLIP correctly identifies Complex and Cracking, capturing the most important features even though it does not include Smooth.

Fig. 5 (iii) and (iv) represent cases where all models predict the labels correctly. In 5(iii), the ground truth includes Complex and Cracking, and all three models correctly recover the full label set. When cracks are visually prominent, the models detect them consistently without introducing additional labels. In 5(iv), the ground truth contains only Smooth, and all models correctly identify this single morphology. This demonstrates that the models can accurately identify simple morphologies when no overlapping features are present.

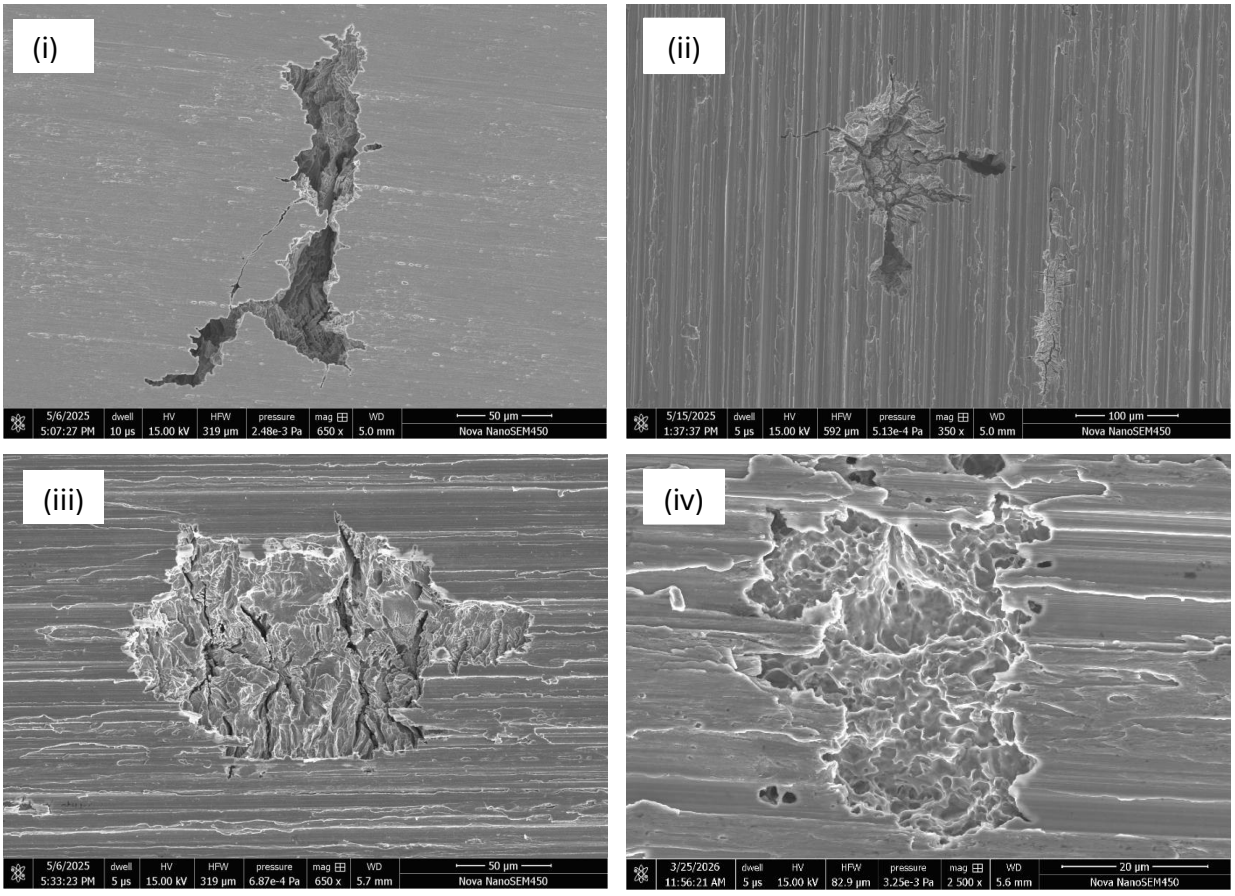
These examples match the trends seen in the quantitative results. CLIP is the most consistent when cracks are present, ResNet-18 handles Smooth well but sometimes adds extra labels, and Swin-T leans toward predicting Cracking even when the evidence is weak. These sample cases help illustrate how each model behaves on real images and where their strengths and weaknesses show up.

Generalization on unseen data

The trained ResNet-18, CLIP, and Swin-T models were evaluated on 24 unseen SEM images obtained from additional physical experiments outside the original training and test sets. This set included 14 images labeled Complex, 12 Smooth and 12 Cracking. All three models were evaluated using the Image with Text (Weighted) configuration. Fig. 6 presents selected unseen images with their ground-truth and predicted labels.

In Fig. 6(i), all three models correctly assign the labels Complex and Cracking, which match the rough surface and the clear crack-like lines visible in the image. Fig. 6(ii) contains cracks and uneven texture but also includes a noticeably smoother region. ResNet-18 and Swin-T classify it as Complex and Cracking, whereas CLIP adds Smooth, reflecting the presence of that smooth surface area. In Fig. 6(iii), the ground truth remains Complex and Cracking. ResNet-18 and CLIP correctly identify both labels, while Swin-T predicts only Complex, missing the cracking feature.

(a) Input Images



(b) Ground truth and Predicted labels

Images	Ground truth	ResNet-18	CLIP	Swin-T
(i)	Complex, Cracking	Complex, Cracking	Complex, Cracking	Complex, Cracking
(ii)	Complex, Cracking	Complex, Cracking	Complex, Smooth, Cracking	Complex, Cracking
(iii)	Complex, Cracking	Complex, Cracking	Complex, Cracking	Complex
(iv)	Smooth	Smooth	Smooth, Cracking	Complex, Smooth

Fig. 6 | Predicted labels for selected unseen samples. (a) Input SEM images. (b) Ground-truth

In Fig. 6(iv), all three models predicted Smooth, but only ResNet-18 matched the complete ground-truth label set; CLIP additionally predicted Cracking and Swin-T predicted Complex. Fig. 8(c) offers a clearer look at why these extra labels appeared.

Table 3 summarizes performance on the full unseen dataset. Swin-T obtained the highest exact-match accuracy, micro-F1, and macro-F1, along with the lowest Hamming loss. Its predictions matched the complete ground-truth label set for 20 images, compared with 17 for ResNet-18 and 15 for CLIP.

Table 3 | Performance on unseen SEM images.

Model	Accuracy	Hamming loss	micro-F1	macro-F1
ResNet-18	70.83%	0.1667	0.8500	0.8489
CLIP	62.50%	0.1806	0.8354	0.8346
Swin-T	83.33%	0.1111	0.8974	0.8974

The per-label F1 scores in Fig. 7 highlight differences across morphologies. Swin-T obtained the highest F1 scores for both Complex and Smooth, while ResNet-18 achieved the strongest performance on Cracking. Although Swin-T performed best overall on this unseen set, no model achieved the highest F1 score for every label.

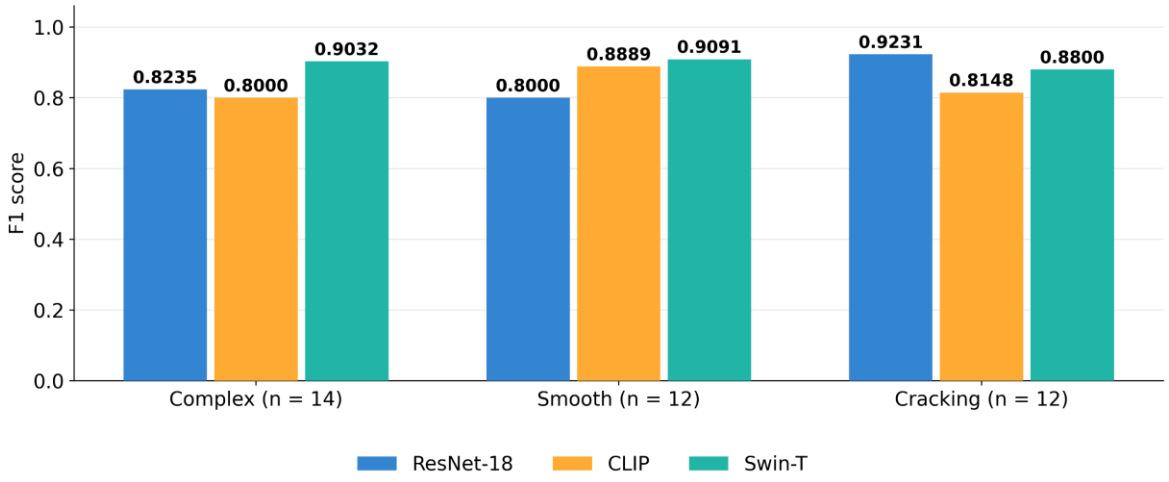


Fig. 7 | Per-label F1 scores on the unseen dataset

The small size of the unseen dataset makes accuracy highly sensitive to individual errors. Some images also contain low-contrast or weakly resolved features, which can lead to missed labels or extra predictions. CLIP performed best on the original test set and Swin-T performed best on the unseen set under the weighted configuration. These results do not establish the general superiority of either model. Evaluation on a larger and more diverse independent dataset is needed to more reliably assess their generalization performance.

Ablation experiments

To evaluate the contribution of individual design choices, a series of ablation experiments was conducted. In each experiment, only one component of the framework was modified while all other settings were held constant, allowing its impact to be assessed independently. The factors investigated included the loss function, training duration (number of epochs), and batch size. Comparative analyses were performed for ResNet-18, CLIP and Swin-T to examine how these factors influenced model performance.

Effect of class-weighted loss

A clear improvement in performance was observed with the use of class-weighted loss, as shown in Table 4. Without weighting, ResNet-18, CLIP, and Swin-T tended to favor majority classes, resulting in reduced performance on underrepresented morphologies. Introducing class weights increased the penalty for minority-class errors, encouraging the models to learn more discriminative features for these labels.

These benefits were reflected in lower Hamming loss and higher micro- and macro-F1 scores across all three architectures. CLIP exhibited the largest gain, with macro-F1 increasing from 0.9220 to 0.9559 and accuracy rising from 0.8056 to 0.8889, likely because the weighted loss provided stronger supervision for its lightweight classification head. ResNet-18 also improved in F1 performance, with macro-F1 increasing from 0.9282 to 0.9341 and accuracy increasing from 0.8056 to 0.8611. Swin-T showed more modest gains, with macro-F1 rising from 0.9273 to 0.9334 and accuracy improving from 0.8056 to 0.8611. These results suggest that class-weighted loss is particularly beneficial for small, imbalanced corrosion datasets, and enables more effective utilization of metadata information.

Table 4 | Performance of models with and without class-weighted loss across different input settings

Models	Batch Size	Experiments	Accuracy	Hamming Loss	micro-F1	macro-F1
ResNet-18	8	Image with Text (Unweighted)	0.8056	0.0833	0.9256	0.9282
		Image with Text (Weighted)	0.8611	0.0741	0.9322	0.9341
		Image with Text (Unweighted, Without Comments)	0.7500	0.1296	0.8833	0.8801

		Image with Text (Weighted, Without Comments)	0.8611	0.0741	0.9322	0.9300
CLIP	16	Image with Text (Unweighted)	0.8056	0.0833	0.9280	0.9220
		Image with Text (Weighted)	0.8889	0.0463	0.9587	0.9559
		Image with Text (Unweighted, Without Comments)	0.8611	0.0556	0.9508	0.9488
		Image with Text (Weighted, Without Comments)	0.8611	0.0463	0.9587	0.9562
Swin-T	8	Image with Text (Unweighted)	0.8056	0.0833	0.9244	0.9273
		Image with Text (Weighted)	0.8611	0.0741	0.9310	0.9334
		Image with Text (Unweighted, Without Comments)	0.8333	0.0833	0.9231	0.9234
		Image with Text (Weighted, Without Comments)	0.8333	0.0833	0.9231	0.9234

Effect of epochs and early stopping

Training duration is a critical factor when working with small datasets. Insufficient training may prevent a model from fully learning relevant feature representations, whereas excessive training can lead to overfitting and reduced generalization performance. To balance these competing effects, early stopping with a patience of 25 epochs was employed. Training was terminated when the validation macro-F1 score failed to improve for 25 consecutive epochs, allowing the models sufficient opportunity to converge while limiting overfitting.

To investigate the effect of training duration, four maximum epoch limits (10, 20, 50, and 100) were evaluated for ResNet-18, CLIP and Swin-T using the Image with Text (Weighted) configuration. The results in Table 5 show that the three architectures responded differently as the allowable training window increased, reflecting their distinct optimization characteristics and learning dynamics. CLIP requires substantially longer optimization. Performance increased steadily from a macro-F1 of 0.8696 at 10 epochs to 0.9131 at 20 epochs, reaching its best result (macro-F1 = 0.9559, accuracy = 0.8889) when training continued to 50 epochs. Increasing the limit to 100 epochs yielded identical results because

training stopped early at epoch 73 and the same best checkpoint was selected. These results suggest that CLIP converges more gradually than ResNet-18. Swin-T showed the most stable behavior. Performance was unchanged between 10 and 20 epochs and then plateaued, with a macro-F1 of 0.9264 maintained across all longer training schedules. Early stopping consistently terminated training at epoch 34, indicating that additional epochs provided little benefit.

Overall, early stopping effectively prevented unnecessary training while selecting the best-performing checkpoints. ResNet-18 converged quickly, CLIP required a longer optimization period to achieve peak performance, and Swin-T demonstrated stable performance with minimal sensitivity to extended training.

Table 5 | Comparison of models' performance at different maximum epochs.

Model	Batch Size	Max Epochs	Early Stop	Accuracy	Hamming Loss	micro-F1	macro-F1
ResNet-18	16	10	—	0.8333	0.0926	0.9153	0.9196
		20	—	0.8333	0.0833	0.9244	0.9240
		50	37	0.8333	0.0833	0.9244	0.9240
		100	37	0.8333	0.0833	0.9244	0.9240
CLIP	16	10	—	0.6667	0.1574	0.8661	0.8696
		20	—	0.7778	0.0926	0.9138	0.9131
		50	—	0.8889	0.0463	0.9587	0.9559
		100	73	0.8889	0.0463	0.9587	0.9559
Swin-T	16	10	—	0.8333	0.0833	0.9231	0.9264
		20	—	0.8333	0.0833	0.9231	0.9264
		50	34	0.8333	0.0833	0.9231	0.9264
		100	34	0.8333	0.0833	0.9231	0.9264

Effect of batch size

Batch size influenced each model differently, and the results in Table 6 highlight how the three architectures respond under the Image with Text (Weighted) configuration with a maximum of 100 epochs. The observed trends reflect how each model handles changes in gradient stability, feature variability, and optimization dynamics.

ResNet-18 is trained end-to-end, so the batch size directly affects how noisy or stable the gradient updates are. Smaller batches introduce more variation, which can help the model learn subtle corrosion patterns in a small and imbalanced dataset. Batch size 8 gave the strongest results, with the lowest Hamming loss (0.0741) and the highest macro-F1 (0.9341). Batch sizes 16 and 32 showed weaker performance, especially on minority corrosion types.

This suggests that ResNet-18 benefits from the extra gradient noise provided by small batches.

CLIP behaves differently because its backbone is frozen and only the classification head is trained. Larger batches give smoother and more stable updates, which helps CLIP converge. Batch size 8 performed worse than the larger batches. Batch sizes 16 and 32 reached the same macro-F1 (0.9559), showing that CLIP improves when moving beyond very small batches but does not gain anything from increasing the batch size past 16. Batch size 16 appears to be the most efficient choice.

Swin-T shows only small changes across batch sizes. Batch sizes 8 and 32 produce nearly identical results, while batch size 16 shows a slight drop in accuracy and F1 scores. The early-stopping epoch increases with batch size, but the longer training does not lead to better performance. Swin-T handles different batch sizes well, with batch sizes 8 and 32 giving the most consistent results.

Table 6 | Effect of batch sizes on the evaluated models.

Model	Batch Size	Early Stop	Accuracy	Hamming Loss	micro-F1	macro-F1
ResNet-18	8	31	0.8611	0.0741	0.9322	0.9341
	16	37	0.8333	0.0833	0.9244	0.9240
	32	37	0.8333	0.0926	0.9138	0.9183
CLIP	8	50	0.8611	0.0648	0.9431	0.9418
	16	73	0.8889	0.0463	0.9587	0.9559
	32	100	0.8889	0.0463	0.9587	0.9559
Swin-T	8	31	0.8611	0.0741	0.9310	0.9334
	16	34	0.8333	0.0833	0.9231	0.9264
	32	39	0.8611	0.0741	0.9310	0.9334

Discussion

This study addressed three major challenges in corrosion morphology classification: a limited dataset size (234 SEM images), significant label imbalance, and the frequent coexistence of multiple corrosion morphologies within a single image. In addition, some image annotations were inherently uncertain because morphological boundaries were not always clearly distinguishable. Representative examples are shown in Fig. 8.

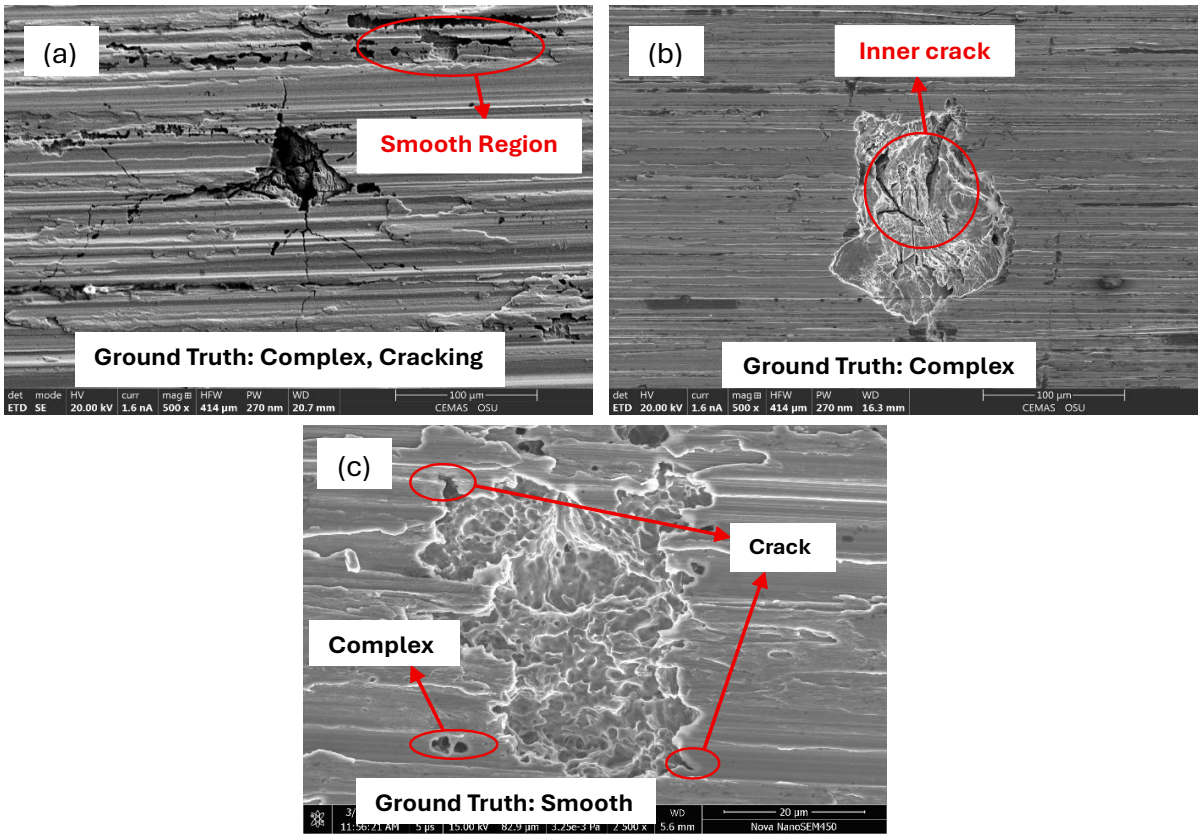


Fig. 8 | Examples of morphological ambiguity in SEM-based corrosion classification.

In Fig. 8(a), the ground-truth labels are Complex and Cracking, as the image contains an irregular damaged region with cracks propagating from the center. However, the model also identified the relatively smooth surface region at the top of the image. This prediction illustrates the difficulty of analyzing SEM images containing multiple texture patterns, where localized features may resemble characteristics associated with additional morphology classes. Such mixed visual cues can lead to ambiguity in both annotation and model prediction.

In Fig. 8(b), the ground-truth label is Complex, and the image clearly contains an inner crack within the damaged zone. The model detects this crack as well. Depending on how the observer interprets the feature, it may be viewed as a distinct crack or simply part of the

broader complex deformation. Because these boundaries are gradual rather than sharply defined, different annotators may reasonably disagree on how to classify such regions.

Fig. 8(c) shows another case where the labeling can be unclear. The ground-truth label is Smooth, but the surface contains small regions with complex texture and some faint cracks. Because the crack is very minor, some annotators may not consider it a true crack, but the model still detects it. The complex features are also subtle, which makes this image easy to misinterpret.

The errors shown in Fig. 5(i) and (ii) largely follow the same pattern. In many cases, the model predicts additional morphologies that are not included in the ground-truth annotations. Consequently, the dominant error type is false positives rather than false negatives, resulting in a larger reduction in precision than recall. Importantly, several of these predictions occur in borderline cases where the visual evidence is ambiguous and alternative annotations could be reasonably justified. The inconsistency of metadata descriptions and image comments further contributes to this uncertainty. As a result, the model tends to identify potentially relevant additional morphologies that warrant further inspection, while generally maintaining sensitivity to the primary corrosion features present in the image.

Several limitations should be considered when interpreting these results. First, the test set contained only 36 images, making small differences in macro-F1 scores difficult to interpret with confidence. The dataset was limited to experiments conducted at 35 °C and at relative humidities of 40%, 56%, 61%, 66%, and 76%, with most samples collected after 28 days of exposure. Because these conditions represent only a narrow portion of the broader environmental space, the findings should not be generalized to other temperatures, humidity levels, or exposure durations without additional validation. Although the proposed framework performed well on the unseen dataset (see Results, Generalization on unseen data), further validation is needed before extending the findings to other alloys, corrosion environments, or degradation mechanisms. Second, inconsistencies in metadata descriptions and annotator interpretations likely introduced label noise, contributing to some of the borderline classification errors observed in the qualitative results. These limitations highlight the need for larger datasets, standardized annotation procedures, and broader validation to improve model robustness and generalizability.

This study presents a multilabel corrosion classification framework that combines class-weighted loss with metadata, comparing three models—ResNet-18, CLIP, and Swin-T—on this task. For each model, the weighted fusion setup achieved the highest macro-F1 score, reaching 0.956 for CLIP. The ablation studies demonstrate that class weighting is essential for learning less common morphologies, and that metadata improves performance only when comments and labels are reasonably consistent. The three models also exhibit

distinct behavior: CLIP performs better on Cracking, while ResNet-18 is more consistent on Smooth. Model selection may therefore depend on which morphology is most critical for a given application. The overall goal of this work was to establish a practical approach for learning multiple corrosion morphologies from a small, imbalanced SEM dataset, and the experiments specifically examined how metadata consistency, class weighting, and model architecture jointly affect classification performance.

It is worth noting that corrosion morphology labeling carries inherent subjectivity, even among human experts. During dataset annotation, disagreements occasionally arose between annotators regarding borderline cases—for instance, distinguishing "Complex" from "Smooth" morphologies is not always definitive, even upon close visual inspection. This suggests that some fraction of model misclassifications may reflect genuine ambiguity in the underlying labels rather than model error, and that inter-annotator variability should be considered when interpreting model performance in future work.

For materials science applications, this framework can reduce the burden of manual labeling, which is typically time-consuming and dependent on expert availability. The model can provide more consistent classifications and help identify subtle surface features that might be overlooked during rapid visual inspection. Such a tool could support corrosion monitoring in settings where stainless steel components require regular inspection, including chemical plants, marine equipment, and nuclear systems. Future work should focus on expanding the dataset to include additional alloys and humidity conditions, as well as refining metadata quality through more consistent, expert-verified labeling. A larger and cleaner dataset would allow further evaluation of the method's robustness across a broader range of corrosion scenarios.

Methods

The corrosion dataset was first curated to ensure consistency between image annotations and associated metadata. Three deep learning models, ResNet-18, CLIP, and Swin-T, were then trained and evaluated for multilabel corrosion morphology classification. The subsequent sections detail the experimental data collection, dataset preparation process, input representation, and the training and evaluation methodologies adopted in this study.

Experimental data collection

Type 304H SS specimens with a mill finish (~120 grit) were used to generate the training data. The sample surface was cleaned with ethanol, and arrays of 2 μL droplets of an equilibrium brine solution were then pipetted onto the surface. Equilibrium RH solutions of ASTM artificial seawater were used to represent the salt composition of the solution at a given RH. These equilibrium solutions were calculated by using the solubility code EQ3/6 and the

Pitzer thermodynamic database and were developed by Sandia National Laboratories [20,21,22]. The specimens were exposed in an Allied Environmental Systems RH-temperature chamber or in a desiccator with a saturated salt solution at 35 °C and the target RH for up to 60 days. After exposure, samples were sonicated for 60 minutes in 10% nitric acid and imaged by SEM.

An independent experimental dataset collected by a separate research group was used to validate the model developed in this study. In their study, SS304H specimens from the same vendor were prepared using three surface-finishing methods (hand grinding, wheel grinding, and milling) to introduce varying degrees of subsurface deformation and residual stress. Arrays of droplets (2 μL) containing corrosive brines in equilibrium with 40% relative humidity were then deposited onto the specimen surfaces. The specimens were subsequently exposed in an environmental chamber maintained at 35 °C and 40% RH for exposure periods of up to 60 days. At predetermined intervals, the specimens were removed, rinsed with deionized water, dried, and characterized by SEM to evaluate the resulting corrosion morphologies.

Dataset preparation

The dataset comprises 234 SEM images of stainless steel corrosion. Each image is tagged with one or more morphology types: Complex, Smooth and Cracking. Many samples contain two or even three morphologies, which makes the task a multilabel rather than a single-label classification problem.

Fig. 9 illustrates the distribution of morphology labels within the dataset. Most images contain two active labels, with the Complex and Cracking combination representing the most frequently observed class. In contrast, the Smooth morphology is more commonly observed as a standalone label, while only a limited number of samples exhibit all three morphologies simultaneously. This uneven label distribution introduces a class imbalance that can bias model training toward the majority classes. To mitigate this issue, a class-weighted loss function was employed, assigning greater importance to underrepresented labels and encouraging the models to improve their recognition of minority morphology classes.

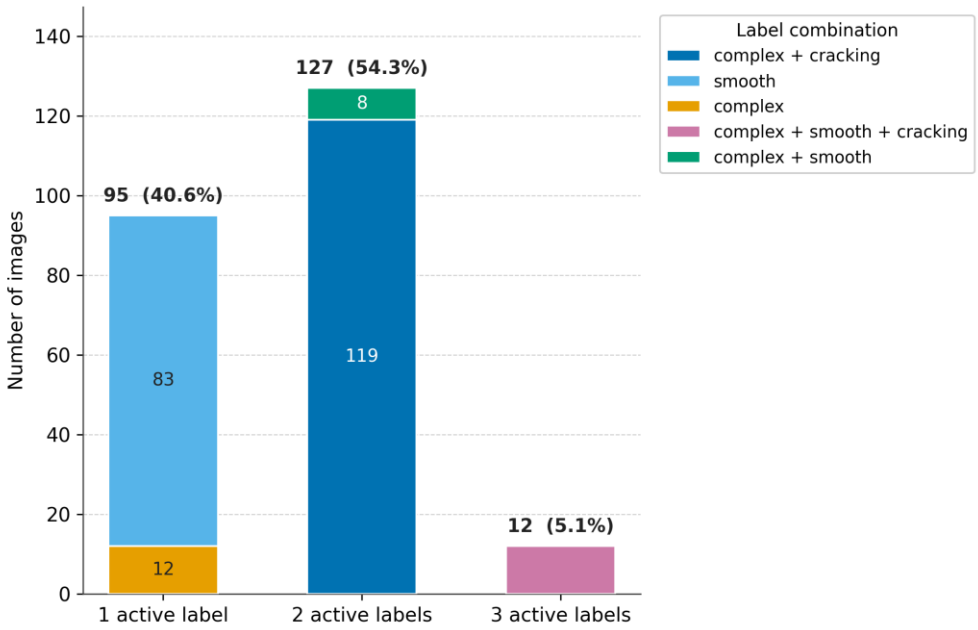


Fig. 9 | Distribution of morphology label combinations in the dataset.

The raw data come from five Excel spreadsheets, each corresponding to a different relative humidity condition (40%, 56%, 61%, 66%, and 76% RH). Each row includes the image filename, its morphology labels, and several pieces of experimental metadata. These metadata fields cover material type, temperature, humidity, exposure time, cleaning condition, and additional comments from the experimenter. Representative examples are given in Table 1.

Image preprocessing and metadata encoding

Each architecture requires a dedicated preprocessing pipeline. ResNet-18 operates on 512×512 images that are normalized using ImageNet [23] statistics. To improve model generalization and reduce overfitting, data augmentation was applied to the training set, including random horizontal and vertical flips, rotations of up to ±15°, and minor adjustments to image brightness and contrast, as illustrated in Fig. 10. Validation and test images underwent only resizing and normalization to ensure consistent evaluation conditions. CLIP and Swin-T utilized 224×224-pixel inputs and followed their respective preprocessing protocols. CLIP used its native normalization parameters, whereas Swin-T used the standard ImageNet normalization scheme.

Metadata were transformed into a structured feature vector to facilitate model training. Categorical variables, including material type and cleaning status, were represented using one-hot encoding, while temperature, relative humidity, and exposure time were retained as

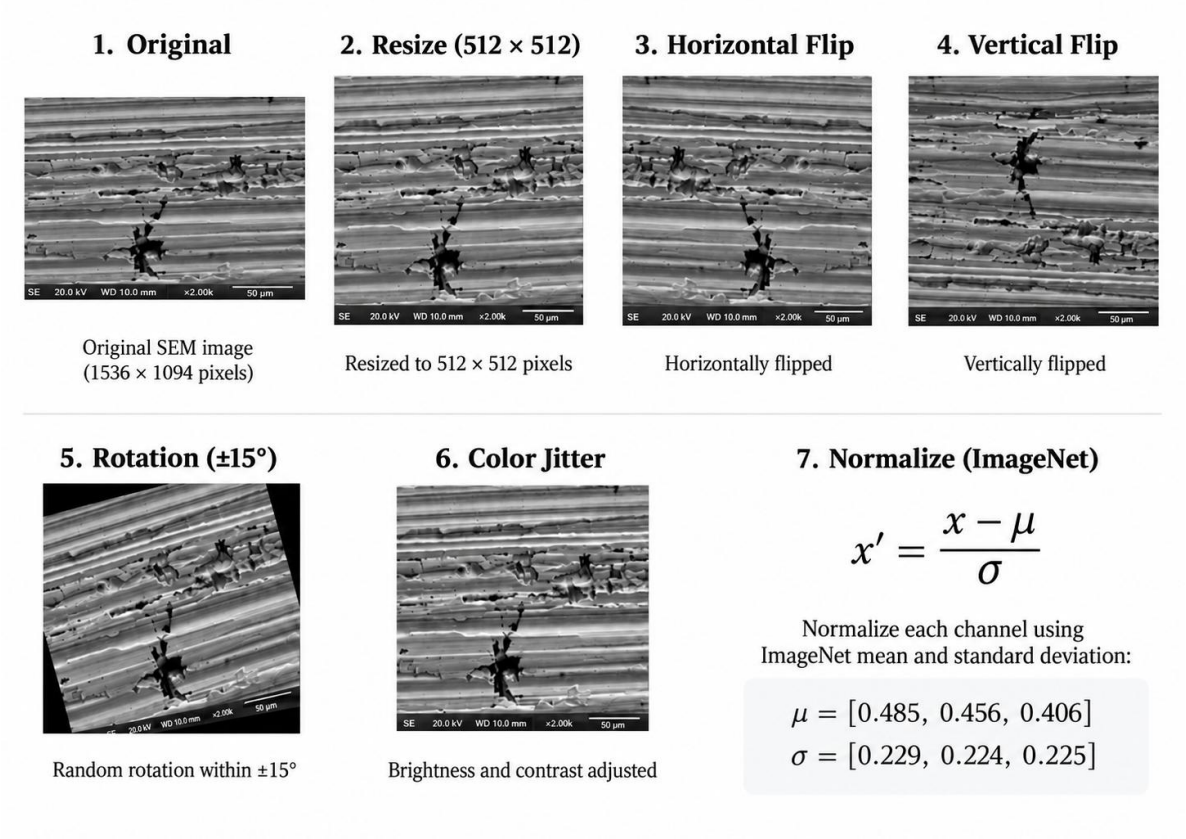


Fig. 10 | ResNet-18's image preprocessing and augmentation pipeline, including resizing, flipping, rotation, color jitter, and ImageNet normalization.

continuous numerical features. Additionally, two features were derived from the comments field: the comment length and a binary indicator denoting the presence or absence of a comment. To ensure consistent feature scaling and prevent data leakage, all numerical metadata features were standardized using a scaler fitted exclusively on the training set and subsequently applied to the validation and test sets.

Input configurations

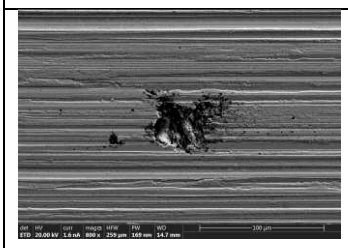
Three input configurations were evaluated to investigate the contribution of image and metadata information to model performance, as shown in Fig. 11. This comparison enabled a systematic assessment of the relative importance of visual and contextual information for multilabel corrosion morphology classification.

- Image Only uses the SEM image alone, without any accompanying information. This configuration serves as the baseline and evaluates how well visual information alone can classify corrosion morphology.
- Image with Text combines the image with all metadata fields, including the experimental conditions and the experimenter's comments. This represents the most complete

configuration and tests whether textual cues—information experts often rely on during manual labeling—improve the model's classification accuracy.

- Image with Text (without Comments) pairs the image with only the objective experimental conditions (temperature, humidity, exposure time, material type, and cleaning status), excluding comments. This configuration isolates the contribution of structured experimental parameters to classification performance.

(a) Example raw data

SEM image	Experimental conditions	Expert comment
	File Name: 3CP_56d; Material: SS304H Tensile bar; Temperature (Celsius): 35 C; Relative Humidity (%): 56; Time period (days): 28; Cleaned? Yes	"Slightly dirty"

(b) Five model input configurations

No	Configuration	Image	Experimental conditions	Expert comment	Weighted fusion
1	Image only	✓	—	—	—
2	Image with Text	✓	✓	✓	No
3	Image with Text (Weighted)	✓	✓	✓	Yes
4	Image with Text (Without Comments)	✓	✓	—	No
5	Image with Text (Without Comments, Weighted)	✓	✓	—	Yes

Fig. *11* | Example SEM data with associated experimental metadata (a) and the model-input configurations evaluated in this study (b).

Model architectures

This study evaluates three complementary architectures—ResNet-18, CLIP, and Swin-T—to capture different aspects of corrosion morphology. These models represent convolutional, vision–language, and transformer-based approaches, offering a balanced set of baselines for analyzing SEM images. Each model is adapted to the limited dataset size and the multimodal nature of the inputs.

ResNet-18 model

ResNet-18 is a compact convolutional network built around residual blocks, which help maintain stable gradient flow in deeper networks [16]. In this study, a ResNet-18 model pretrained on ImageNet is fine-tuned for corrosion-morphology classification. Although SEM images differ from natural photographs, low-level ImageNet features—such as edges, contours and fine textures—transfer effectively, and the model's relatively small capacity helps reduce overfitting on a limited dataset [24].

To adapt the network for this task, the original classification head is replaced with a custom module consisting of a dropout layer ($p = 0.3$), a fully connected layer with 128 units and ReLU activation, a second dropout layer ($p = 0.2$), and a final linear layer that produces three logits corresponding to the corrosion morphology classes. All parameters, including those in the pretrained backbone, are updated during training, allowing the model to learn the specific visual characteristics present in the SEM images.

Input images are resized to 512×512 pixels to preserve the fine-scale surface details needed to distinguish Smooth from Complex morphologies; lower-resolution experiments indicated that downsampling tended to obscure these subtle textural cues. Image normalization follows standard ImageNet mean and standard deviation values. The complete architecture is shown in Fig. 12.

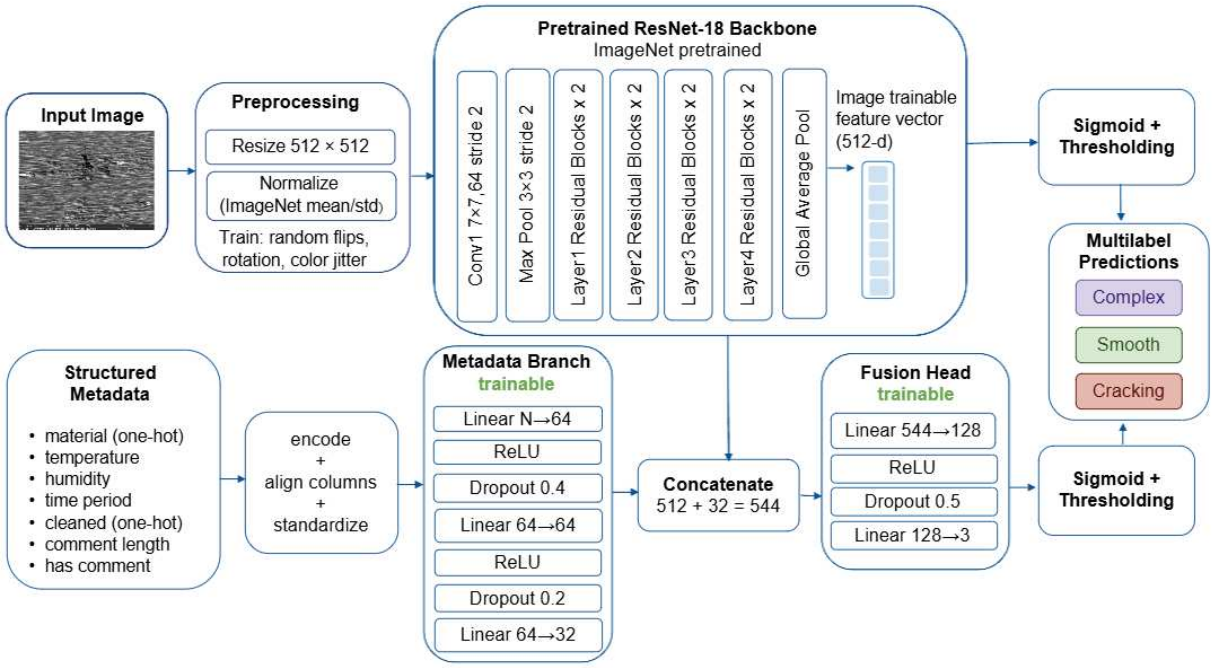


Fig. 12 | Architecture of the ResNet-18 multilabel corrosion classification model.

CLIP-based model

CLIP is a large vision–language model trained on 400 million image-text pairs collected from the web [17]. Instead of learning from images in isolation, the model is optimized to align images with their corresponding textual descriptions. Through contrastive learning, semantically related image–text pairs are drawn closer together in a shared embedding space, while unrelated pairs are pushed apart. This training strategy produces visual representations that generalize effectively to domains that differ substantially from the original pretraining distribution.

The ViT-B/32 variant of CLIP is employed for this work. This model uses a Vision Transformer backbone [24], which divides each image into fixed-size patches and processes them as a sequence of tokens rather than relying on convolutional operations. The pretrained encoder generates a 512-dimensional feature vector that summarizes the semantic content of the input image.

Given the limited size of the training set (163 images), fine-tuning the full CLIP model would likely lead to overfitting. To mitigate this risk, the pretrained backbone is kept frozen, and only a lightweight classification head is trained on top of the extracted features. This transfer learning approach leverages CLIP's general-purpose visual representations while keeping the number of trainable parameters low [25]. To ensure a fair comparison, the classification head mirrors the architecture used for ResNet-18: a dropout layer, a fully connected layer with 128 units and ReLU activation, a second dropout layer, and a final output layer producing three logits corresponding to the corrosion morphology classes. All images fed into CLIP are resized to 224 × 224 pixels and normalized using CLIP's standard preprocessing. Fig. 13 presents the full architecture.

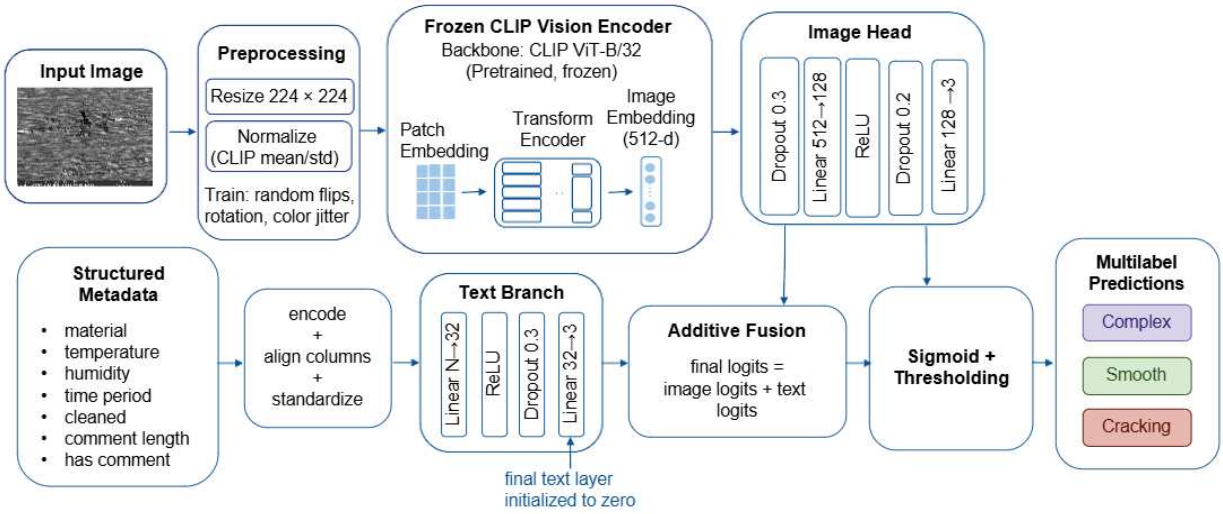


Fig. 13 | Architecture of the CLIP multilabel corrosion classification model.

Swin-T model

Swin Transformer (Swin-T) is a hierarchical vision transformer architecture that employs shifted-window self-attention rather than global self-attention, enabling efficient modeling of both local and long-range visual dependencies [18]. This design allows the network to capture fine-grained surface textures as well as broader structural patterns, making it particularly suitable for corrosion image analysis, where both localized defects and overall morphological characteristics are important. In this study, the Swin-Tiny variant pretrained on ImageNet-1K was adopted, and the backbone network was frozen during training to reduce the risk of overfitting given the limited dataset size.

For feature extraction, each image was resized to 224 × 224 pixels and partitioned into 4 × 4 patches. The Swin-T backbone first embeds these patches into 96-dimensional tokens and processes them through four hierarchical stages of shifted-window attention. Across these

stages, the spatial resolution is progressively reduced while the feature dimension increases from 96 to 192, 384, and finally 768 channels [18]. Following the final stage, global average pooling is applied to generate a compact 768-dimensional image representation.

To incorporate contextual information, a lightweight multilayer perceptron (MLP) branch was introduced to encode structured metadata, including material type, temperature, humidity, exposure duration, and inspection comments. Categorical variables were one-hot encoded, while numerical variables were standardized prior to processing. The MLP compresses the metadata into a 32-dimensional feature vector. The image and metadata representations were then concatenated to form an 800-dimensional multimodal feature vector (768 + 32). This fused representation was passed through a lightweight classification head consisting of a fully connected layer with 128 hidden units, a ReLU activation function, dropout regularization, and a final output layer producing three logits corresponding to the corrosion classes. This architecture maintains a relatively small number of trainable parameters while effectively integrating visual and contextual information. Fig. 14 illustrates the full architecture.

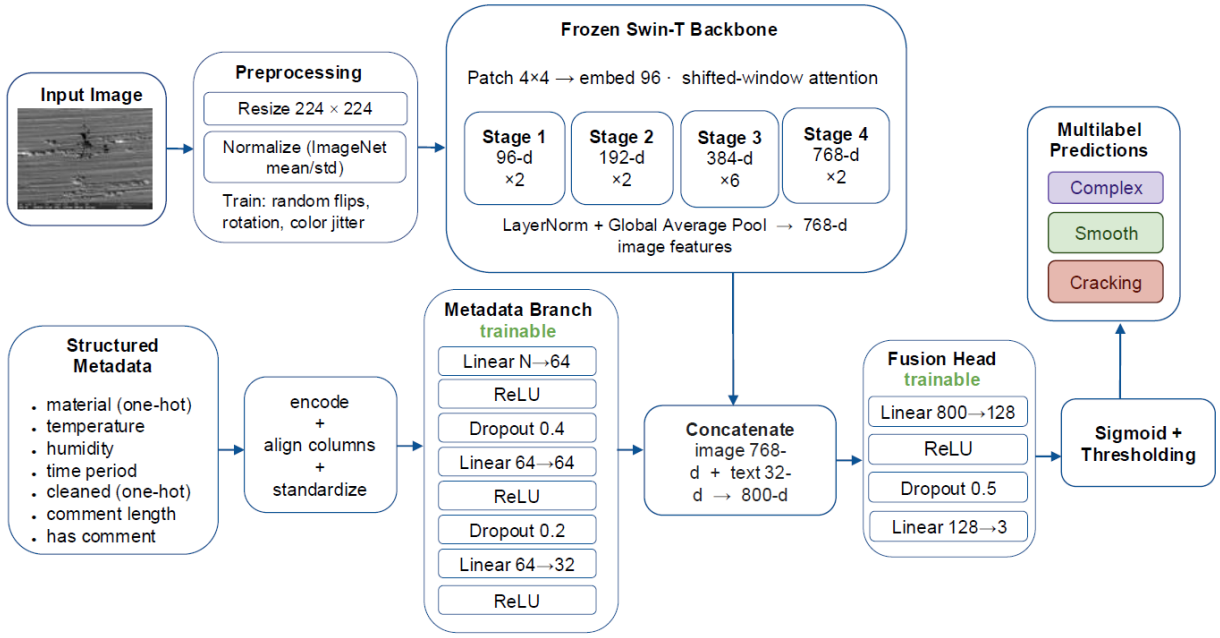


Fig. 14 | Architecture of the Swin-T multilabel corrosion classification model.

Model training

All experiments were conducted in a cloud computing environment using Google Colab with an NVIDIA T4 GPU. Model training was implemented using the PyTorch deep learning framework [26]. To ensure reproducibility, a fixed random seed (42) was applied across Python, NumPy, PyTorch, and CUDA/cuDNN operations under deterministic settings.

To ensure reliable evaluation, the dataset was divided into training, validation and test sets (163/35/36 images) using an approximately 70/15/15 ratio. A stratified multilabel split was applied to preserve the distribution of corrosion morphology combinations across all subsets [27], which is necessary because several label combinations occur fewer than ten times. Following stratification, the test set contains 23 Complex labels, 16 Smooth labels, and 20 Cracking labels. The training configuration is summarized in Table 7.

Table 7 | Overview of the training configurations for ResNet-18, CLIP and Swin-T.

Category	**Parameter**	**Setting**
Common training settings	Optimizer	Adam
	Loss function	BCEWithLogitsLoss (weighted and unweighted)
	Learning rate	1×10^{-4}
	Weight decay	1×10^{-4}
	Batch sizes	8, 16, 32
	LR scheduler	ReduceLROnPlateau (patience = 5, factor = 0.5)
	Early stopping	Patience = 25; validation macro-F1 monitored
	Maximum epochs	10, 20, 50, 100
	Evaluation metrics	Accuracy, Hamming Loss, micro-F1, macro-F1
Model-specific setting	ResNet-18 input size	512 × 512 pixels
	CLIP input size	224 × 224 pixels
	Swin-T input size	224 × 224 pixels

Data availability

The SEM images, morphology labels, environmental metadata and training/validation/test split definitions supporting this study, including the independent unseen test set, will be available at https://github.com/feixu735/corrosion-morphology-multimodal.git following a 6-month embargo from the date of publication. Any additional data is available from the corresponding author upon reasonable request.

Code availability

The code used to preprocess the data and to train and evaluate the ResNet-18, CLIP and Swin-T models, together with configuration files and the fixed data splits, is available at https://github.com/feixu735/corrosion-morphology-multimodal.

Acknowledgements

We acknowledge the financial support from the U.S. Department of Energy, MFC (MFC) of the Nuclear Technology Research and Development program in the office of Nuclear Energy, as well as the Nuclear Energy University Program through Award No. NE0009388. Some of

this work was performed at the following core facility, which is part of Colorado School of Mines' Shared Instrumentation Facility: Electron Microscopy: RRID:SCR_022048. We acknowledge SEM imaging conducted at the Center for Electron Microscopy and Analysis (CEMAS) at The Ohio State University.

Author contributions

Atira Ferdoushi: Writing – original draft, Data curation, Methodology, Formal analysis, Investigation, Software, Validation. **Daria Bentley**: Data curation, Formal analysis, Investigation, Validation, Writing – review and editing. **Hailong Dai**: Data curation, Formal analysis, Investigation, Validation, Writing – review. **Xiaolei Guo**: Funding acquisition, Formal analysis, Investigation, Validation, Supervision, Writing – original draft. **Jenifer S. Locke**: Funding acquisition, Conceptualization, Supervision, Investigation, Writing – review and editing. **Fei Xu**: Funding acquisition, Conceptualization, Supervision, Formal analysis, Methodology, Validation, Writing – review and editing.

Competing interests

The authors declare no competing interests.

References

1 Koch, G. et al. International measures of prevention, application, and economics of corrosion technologies study. NACE International (2016).

2 Cortest Columbus Technologies, Inc. Corrosion costs and preventive strategies in the United States. Federal Highway Administration (2002).

3 Srinivasan, J. et al. Correlation of stainless steel pit morphology to humidity-specific sea salt brine constituents. Corrosion 79, 255–267 (2023).

4 Pinoski, L., Antony Jose, S. & Menezes, P. L. Stress corrosion cracking: mechanisms, materials challenges, and engineering solutions. Materials 19, 898 (2026).

5 Weirich, T. D. et al. Humidity effects on pitting of ground stainless steel exposed to sea salt particles. J. Electrochem. Soc. 166, C3477–C3487 (2019).

6 Katona, R. M., Karasz, E. K. & Schaller, R. F. A review of the governing factors in pit-to-crack transitions of metallic structures. Corrosion 79, 72–96 (2023).

7 Montoya, T. et al. Influence of realistic, cyclic atmospheric cycles on the pitting corrosion of austenitic stainless steels. J. Electrochem. Soc. 170, 041502 (2023).

8 Dai, H. et al. Linking pitting morphologies of stainless steel 304H across electrochemical polarization, laboratory atmospheric exposure, and field exposure. (under review).

9 Wang, Y., Tian, Z. & Hu, S. Multiscale statistical analysis of massive corrosion pits based on image recognition of high resolution and large field-of-view images. Materials 13, 4695 (2020).

10 Yin, B. et al. Corrosion image data set for automating scientific assessment of materials. In Proc. British Machine Vision Conference 102 (2021).

11 Rajendran, M. & Subbian, D. Deep learning in corrosion assessment and control: a critical review of techniques and challenges. Corros. Rev. 44, 20240060 (2026).

12 Cai, L. et al. Automatic feature segmentation of TRISO fuel cross-sections through machine learning. Trans. Am. Nucl. Soc. 132, 661 (2025).

13 Zhao, L., Locke, J., Xu, F., Yao, T. & Guo, X. Accurate segmentation of localized corrosion in structural alloys via deep learning. npj Mater. Degrad. 9, 80 (2025).

14 Arias, F. et al. Automated assessment of marine steel corrosion using visible–near-infrared hyperspectral imaging. Coatings 15, 645 (2025).

15 DeCost, B. L., Francis, T. & Holm, E. A. Exploring the microstructure manifold: image texture representations applied to ultrahigh carbon steel microstructures. Acta Mater. 133, 30–40 (2017).

16 He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In Proc. IEEE Conference on Computer Vision and Pattern Recognition 770–778 (2016).

17 Radford, A. et al. Learning transferable visual models from natural language supervision. In Proc. International Conference on Machine Learning 8748–8763 (2021).
18 Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. In Proc. IEEE/CVF International Conference on Computer Vision 9992–10002 (2021).
19 Cui, Y., Jia, M., Lin, T.-Y., Song, Y. & Belongie, S. Class-balanced loss based on effective number of samples. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition 9260–9269 (2019).
20 Wolery, T. J. EQ3/6, a software package for geochemical modeling of aqueous systems: package overview and installation guide (version 7.0) (Lawrence Livermore National Laboratory, 1992).
21 Wolery, T. J. & Jarek, R. L. EQ3/6, version 8.0: software user's manual. Report 10813-UM-8.0-00 (Sandia National Laboratories, 2003).
22 Benbow, S., Metcalfe, R. & Wilson, J. Pitzer databases for use in thermodynamic modeling (Nuclear Waste Management Organization Technical Memorandum QRS-3021A-TM1, 2008).
23 Deng, J. et al. ImageNet: a large-scale hierarchical image database. In Proc. IEEE Conference on Computer Vision and Pattern Recognition 248–255 (2009).
24 Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. Preprint at https://arxiv.org/abs/2010.11929 (2020).
25 Pan, S. J. & Yang, Q. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359 (2010).
26 Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8024–8035 (2019).
27 Szymański, P. & Kajdanowicz, T. A network perspective on stratification of multi-label data. In Proc. First International Workshop on Learning with Imbalanced Domains: Theory and Applications 22–35 (2017).