Corrosion Image Data Set for Automating Scientific Assessment of Materials

Biao Yin[1]
byin@wpi.edu

Nicholas Josselyn[1]
njjosselyn@wpi.edu

Thomas Considine[2]
thomas.a.considine.civ@army.mil

John Kelley[2]
john.v.kelley8.civ@army.mil

Berend Rinderspacher[2]
berend.c.rinderspacher.civ@army.mil

Robert Jensen[3]
robert.e.jensen.civ@army.mil

James Synder[4]
james.f.snyder.civ@army.mil

Ziming Zhang[1]
zzhang15@wpi.edu

Elke Rundensteiner[1]
rundenst@wpi.edu

[1] Data Science Program
Worcester Polytechnic Institute
Worcester, MA

[2] Weapons and Materials Research Directorate
US Army Research Laboratory
Aberdeen Proving Ground, MD, USA

[3] Weapons and Materials Research Directorate
ARL Northeast Regional Extended Site
Burlington, MA, USA

[4] Weapons and Materials Research Directorate
DEVCOM Army Research Laboratory
Adelphi, MD, USA

Abstract

The study of material corrosion is an important research area, with corrosion degradation of metallic structures causing expenses up to 4% of the global domestic product annually along with major safety risks worldwide. Unfortunately, large-scale and timely scientific discovery of materials has been hindered by the lack of standardized corrosion experimental data in the public domain for developing machine learning models. Obtaining such data is challenging due to the expert knowledge and time required to conduct these scientific experiments and assess corrosion levels. We curate a novel data set consisting of 600 images annotated with expert corrosion ratings obtained over 10 years of laboratory corrosion testing by material scientists. Based on this data set, we find that non-experts even when rigorously trained with domain guidelines to rate corrosion fail to match expert ratings. Challenges include limited data, image artifacts, and millimeter-precision corrosion. This motivates us to explore the viability of deep learning approaches to tackle this benchmark classification task. We study (i) convolutional neural networks powered with rich domain-specific image augmentation techniques tuned to our data, and (ii) a recent self-supervised representation learning approach either pretrained on ImageNet or trained on our data. We demonstrate that pretrained ResNet-18 and HRNet models with tuned augmentations can reach up to **0.83** accuracy. With this corrosion

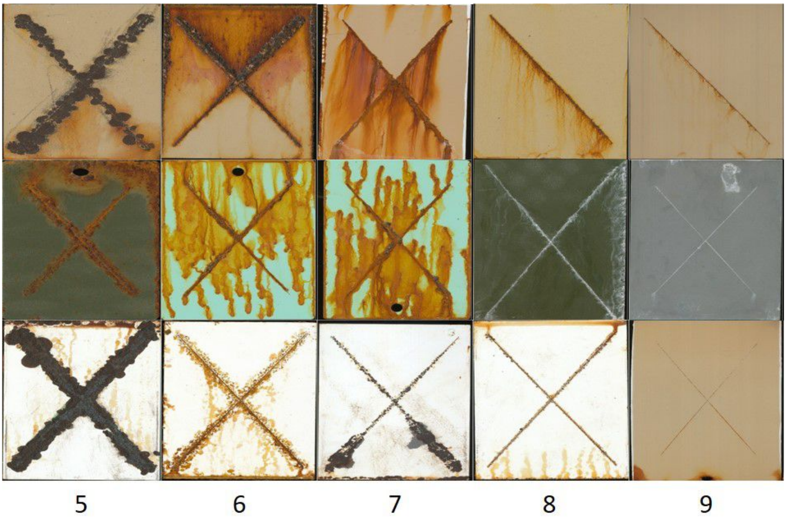

Figure 1: Three sample images per corrosion rating class 5 - 9 from our data set.

data set, we open the door for the design of more advanced deep learning models to support this real-world task, while driving innovative new research to bridge computer vision and material innovation. Our data and code are available at: https://arl.wpi.edu

1 Introduction

Background. Corrosion is defined as the gradual degradation of a metal over time due to chemical interactions with its environment. Corrosion comes with significant economic burden, costing an equivalent of approximately 4% of the gross domestic product (GDP) in losses worldwide – around $2.5 trillion [5, 21]. It results in major safety risks worldwide and negatively impacts the environment, societal health, national infrastructure, manufacturing, and transportation. The study of corrosion is an active area of research in material science focused on the design of new materials to prevent corrosion with corrosion tests conducted by diverse industries, government agencies, and countries [1, 7, 8, 22, 30, 37, 42, 47]. However, it is challenging to assess corrosion progression precisely, quickly, and safely, which impedes the understanding of corrosion and material discovery in general. The reasons include: costs of manufacturing processes and running tests, danger of hazardous chemicals, expert knowledge required to identify corrosion, long periods of observation for corrosion to progress, and inherent biases associated with human analysis and measurements [7, 8, 37]. Although AI and computer vision have increasingly become popular for automation of many critical tasks in science and engineering, there has been limited usage in materials research, possibly due to requiring large amounts of pedigreed, high-quality data [23].

Corrosion image data set. In this paper, we provide a unique data set for automating scientific assessment of materials, aiming at end-to-end learning the corrosion ratings process using computer vision techniques. This data set is composed of 600 images of tested material panels and expert-confirmed annotations of corrosion assessment scores under standardized laboratory environments and procedures [7, 14, 47]. To our knowledge, this is the first-ever open data set with images and expert annotations from standardized corrosion tests for the use of discovering new materials. Examples of this data set are depicted in Figure 1. As we see, images can contain a single or double scribe across the image, have differing background colors for the same rating, and can contain noisy areas around the actual corrosion. Sometimes corrosion can even get too thin to be measured visually by an untrained eye.

Challenges. From the computer vision perspective, challenges on our data set include:

- The amount of training images for corrosion ratings is very limited. This leads us to the following questions: Can deep learning based models be trained well from scratch to achieve comparable performance to human experts? Can data augmentation help improve the performance? Manually or automatically?
- Corrosion on the collected images shows difficult patterns, tiny mm-level scales, and similar colors and textures with non-corrosion objects such as water staining. Such appearances are far away from natural image sets such as ImageNet [11]. This observation leads us to the following questions: Can pretrained models on ImageNet be transferred to our task? Which models work better, pretrained or trained from scratch?

Empirical study. We demonstrate the automation of corrosion assessment (in fact, a classification task) on our data set using convolutional neural networks (CNNs). Specifically,

- *Non-expert human study:* Two raters, instructed by a corrosion scientist, worked together to identify and measure areas of corrosion with the help of computer tools. Following the corrosion rating guidelines, they were required to specify the location of corrosion, measure its length, and then assign a rating to each image. The whole procedure can take up to 5 minutes per image. After several trials, the best test accuracy is only 0.38. This clearly indicates that our corrosion rating task is non-trivial especially for non-experts.
- *Learning with manual data augmentation:* ResNet-18, ResNet-50, DenseNet, and HRNet [17, 18, 44] are trained from scratch with extensive tuning with 9 different data augmentation approaches, achieving 0.81, 0.77, 0.80, and 0.77 best test accuracy using 10-fold cross-validation, respectively.
- *Self-supervised learning:* We investigate the potential of a recent self-supervised learning approach, PIRL [27], which takes advantage of data augmentation automatically for representation learning. We compare the pretrained PIRL on ImageNet and the PIRL trained from scratch on our data, leading to 0.75 and 0.72 test accuracy, respectively.

Contributions. In summary, our key contributions in the paper are listed as follows:

- We demonstrate that deep learning has great potential in automating scientific assessment of materials, and works significantly better than non-experts.
- We conduct comprehensive experiments on our data set, and identify key issues that mislead our deep learning classifiers.
- We provide rich research opportunities with our data set in material science, machine learning, and computer vision via the first meticulously collected corrosion data set with high-quality images and expert annotations.

2 Related Work

Deep learning & data sets in material science. While studies with traditional machine learning approaches in the material science domain exist [12, 15, 32, 39, 41, 45, 46, 48], the use of deep learning techniques has been limited. Over the last few years, initial work *w.r.t.* the detection of material defects has emerged, such as LEDNet [25], Faster Region-based

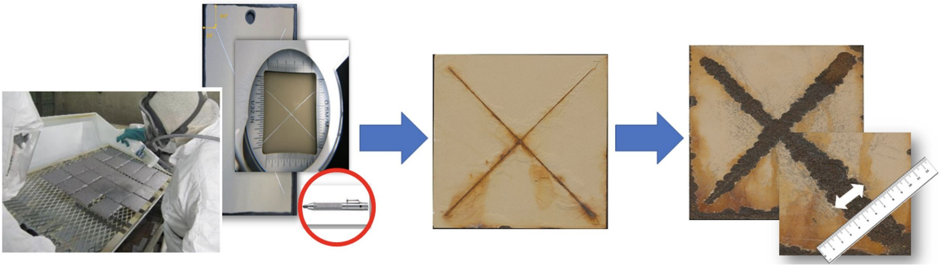

Figure 2: Material experimentation and corrosion rating process. *Left:* Application of coatings to test panels. Test panels are coated with a pretreatment, primer, and topcoat and then scribed with tungsten carbide (circled in red). *Middle:* Panel is placed in accelerated corrosion chamber to initiate corrosion. *Right:* Final time point. Corrosion width is measured at 12 evenly distributed locations along the scribes and assigned an appropriate rating.

CNN (Faster R-CNN) [6, 35], fully connected networks [3], and Texture CNN (T-CNN) working on a data set of 150 raw corroded pipe images and 3 imprecise rating categories (*i.e.,* non-defective, medium corrosion, and aggravated corrosion) [43]. These prior works are done with the purpose of monitoring or detecting defects on corroded pipes, bridges, or synthetic data. None are for material discovery as our work, and thus our proposed data set is unique due to being intentionally prepared for scientific study over a period of ten years.

Data augmentation. In this work we leverage the benefit of data augmentation techniques to improve corrosion assessment with our relatively small data. Manual data augmentation such as rotation and random crop is widely used in training deep models. Other advanced techniques include, for instance, TANDA [33], AutoAugment [9], Fast AutoAugment [24], RandAugment [10], and SelfAugment [34]. These approaches work with unlabeled data or large-scale data sets, which cannot be applied to our case. Instead, we employ Pretext-Invariant Representation Learning (PIRL) [27], a self-supervised approach involving automatic data augmentation that can be easily fine-tuned on small data sets, and compare it with manual data augmentation.

3 Data Set Description

In this section, we introduce our corrosion image data set, its real-world applications, and the manufacturing and assessment processes involved in obtaining it. All experiments follow standard material science procedures. An example process can be viewed in Figure 2.

Corrosion panel samples. In the field of coatings and corrosion research, test articles are typically assessed using a layered stack up of materials consisting of a topcoat, primer, pretreatment, surface profile, and substrate layer. This can be seen in the supplementary document. The five constituent layers of the coating stack up for each sample are reinforced by multiple replicates of the same stack up to provide an adequate statistical sample for material performance testing. When panels with the coatings stack up of materials are assessed, corrosion scientists consider two main elements: the panel surface condition and the composition of the five layers in the stack up. Any commercial names of materials are omitted in this paper and data set for proprietary reasons.

Real-world data applications. Corrosion panels such as those in our work are used through-

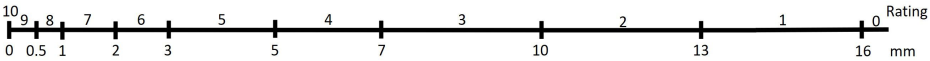

Figure 3: Scribe corrosion rating scale. *Top:* Discrete corrosion rating assigned for each mm measurement range, i.e. rating 10 is 0 mm, rating 0 is 16+ mm (higher ratings mean less corrosion). *Bottom:* mm measurements of average scribe corrosion width for a panel.

out various industrial domains, government branches, and countries. Corrosion tests following these standards are conducted anywhere a paint is applied to a surface to prevent corrosion. This includes industries from automotive, aerospace, chemical, construction, healthcare, to mining [4, 31]; government branches such as the US Army, Navy, Air Force, Marines, Department of Defense (DoD), Department of Energy, NASA, and even the United States Postal Service [16, 29], and similarly in all US-allied countries. Similar tests are even conducted for developing household items such as air conditioning units. These corrosion tests represent the simplest form of testing criteria for industries and companies and thus are ubiquitous in use. The application of these standardized tests are described across countless studies in the DoD-Allied Nations Technical Corrosion Conference [13, 19, 20, 26, 38].

Experimental corrosion testing procedure. Our data set comes from standardized laboratory corrosion tests that are primarily used as a quality control element for production. Each panel is assessed on daily or weekly timescales. The process of obtaining this data is extremely expensive due to the skilled labor required, time it takes to assess the panels, and the costs to process the panels and operate the machines used to conduct corrosion testing.

The laboratory experiments that are conducted come from two possible experimental methods. The first is ASTM B117, a static salt-fog corrosion experiment that consists of a continuous 5% salt-fog (NaCl) atomized into the test chamber containing the panel at 35°C. The second laboratory experiment that is conducted, cyclic corrosion, is an extension from the static salt-fog experiment and is regarded as a more realistic analogue to outdoor environmental conditions. Each experiment cycle consists of a period of ambient dwell, high humidity at an increased temperature, and a high temperature, low humidity dry cycle event. The ambient phase of the test includes 4 spray events where a solution of NaCl, $NaHCO_3$, and $CaCl_2$ (an approximate to seawater) is sprayed across the surface of the panels. Cyclic experiments are observed and rated by investigators at intervals of 10 cycles [42]. For the purpose of this work, the two indoor tests are considered together as one data set, as individually they do not have a significant number of representative samples.

Panel rating procedure. The assessment and ratings done by corrosion scientists are done in accordance with standards defined in ASTM D1654 [47]. These standards define the standard practice to visually evaluate the amount of scribe corrosion for a panel. Scribe corrosion is referred to as corrosion creep, which is emanating out from a deliberately cut area in a panel. Analysis is done using an optical magnifying tool to measure the amount of corrosion emanating from the scribe at 12 equally spaced points along the scribe; 6 points along one scribe direction and 6 points along the other (6 points are used if the panel only has a single scribe), as seen in Figure 2. The 12 measurements are then averaged, divided by two, and correlated to a discrete rating label between 0 and 10 as seen in Figure 3. A rating of 0 signifies a large amount of corrosion, whereas a rating of 10 signifies no corrosion, and is generally only observed at the very start of the testing process. Due to the earliest corrosion assessments typically having high ratings, this leads to imbalance in data collected.

Data set details. Our data set includes 600 images of corroded panels (physical size 4 inches

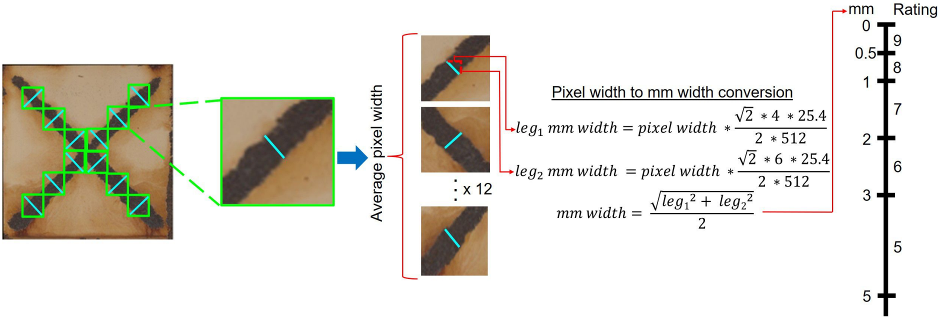
Average pixel width
x 12
Pixel width to mm width conversion
$leg_1\ mm\ width = pixel\ width * \frac{\sqrt{2} * 4 * 25.4}{2 * 512}$
$leg_2\ mm\ width = pixel\ width * \frac{\sqrt{2} * 6 * 25.4}{2 * 512}$
$mm\ width = \frac{\sqrt{leg_1^2 + leg_2^2}}{2}$
mm
Rating
0
0.5
1
2
3
5
9
8
7
6
5
5

Figure 4: Non-expert rating procedure. Corrosion areas are identified in 12 locations on an image, measured on the computer in a pixel width, averaged across the 12 boxes, converted to a mm width, and then assigned the appropriate scribe corrosion rating 0-10.

x 6 inches) that underwent laboratory experiments, with resolution of 512x512 pixels. Each panel received a ground-truth rating based on the ASTM standard assigned by a corrosion scientist. To avoid heavy imbalance in corrosion ratings, and also in accordance with the most domain-relevant corrosion ratings obtained, we only assess panels of ratings in the range 5-9. We do this because any rating less than a 5 is generally considered a failed test, often leading to the sample being removed from testing to be discarded. A rating of 10 is generally only observed at the very start of testing (prior to the initiation of corrosion). Across each of these 5 rating classes 5-9, we provide a balanced data set of 120 images for each rating class. Present in the data set exists a mix of both single and double scribe panels.

4 Experiments

Given the challenges in obtaining the data and ratings in the corrosion science domain, we are motivated to develop automated techniques that can reliably classify levels of corrosion. The first approach we take is a non-expert corrosion rating study that establishes the need for expert knowledge in corrosion assessment. Through this study, we find that individuals who have been trained and are supervised by expert raters are not able to rate corrosion effectively. Therefore, we explore deep learning models (ResNet-18, ResNet-50, DenseNet, and HRNet) [17, 18, 44] and tune these models with various image augmentation methods. After determining that tuning augmentation methods can improve performance significantly, we explore the use of PIRL [27] to demonstrate that automatic representation learning approaches are able to achieve similar performance to those based on heavily tuned deep models.

4.1 Non-Expert Rating Study

We first establish a baseline rating performance by conducting a non-expert corrosion rating study that follows domain rating rules as defined in the ASTM D1654 standards. Two raters, who have undergone several sessions of review and training by a corrosion expert, working together, identified where corrosion was present on a set of 60 image panels (our test set) and measured and averaged the corrosion width at 12 equally spaced locations (6 for single scribe panels) with the help of an interactive segmentation tool, called Grabcut [36]. They measure the width at each point in terms of a pixel length, average all widths, then convert that averaged pixel length to a mm length by calculating the legs of a right triangle formed by the diagonal pixel length measurement, then scaling pixel lengths to the physical panel size of 4 inches x 6 inches, converting from inches to mm using the conversion 1 inch = 25.4 mm, getting the diagonal mm width using the Pythagorean theorem, then dividing by two as required by domain standards to only measure corrosion emanating out along the scribe in one direction, and finally associating that mm length with one of the 11 (0 - 10) possible corrosion ratings. See this process in Figure 4.

Using this method, the well-trained non-material scientists are only able to obtain a 0.38 classification accuracy. This confirms that rating corrosion is a non-trivial task requiring domain expertise and experience typically gained by years of working with materials. Strictly following measurement rules from the ASTM standard, without any other in-depth knowledge of corrosion, leads to the inability to make judgement decisions on identifying corrosion for difficult panels. A corrosion scientist is more experienced and can make better decisions on panels that are borderline between two rating classes. We also observe that in this non-expert study there is also a 0.75 classification accuracy when considering a relaxation of ± 1 rating class, meaning there exist small discrepancies between what a non-expert and an expert sees when reviewing panels. We can visually see some of these difficulties in Figure 1 where for each rating 5-9, a variety of panel appearances for a single rating exist, there are difficult distinctions between close rating categories such as ratings 6 and 7, and challenging areas of corrosion to identify due to water staining and corrosion not perforating the topcoat. Ultimately, with this classification accuracy of 0.38 we conclude that pursuing the avenue of deep learning with our data is a valuable avenue as it gives more potential for learning domain specific information corrosion scientists use in determining corrosion ratings.

4.2 Convolutional Neural Networks and Data Augmentation

After determining the inability for non-experts to assess corrosion, we moved to using deep learning classification approaches to more closely learn how an expert determines corrosion ratings. Not only do experts follow prescribed rules, but they better identify areas of corrosion based on their expertise on scenarios where a panel is close between ratings.

We use our 600 corrosion images and split our data into 10-folds of training (0.80) and validation (0.10) data sets with a held out test set (0.10) of 60 images (same 60 used in the non-expert study). We survey a variety of image augmentation methods and their parameters to improve test accuracy. All augmentations experimented with are seen in Figure 5. Details on each augmentation method and the parameters surveyed can be found in the supplementary materials and tuned parameters for each method can be found in Table 1. We present classification results using ResNet-18, ResNet-50, DenseNet, and HRNet [17, 18, 44] with tuned data augmentation methods in Table 1. Key tuned hyperparameters for: ResNet-18 and ResNet-50 are a base learning rate of $1x10^{-3}$ and weight decay of $5x10^{-2}$, DenseNet is a base learning rate of $1x10^{-4}$ and weight decay of $5x10^{-2}$, and HRNet is a base learning rate of $1x10^{-3}$ and weight decay of $1x10^{-4}$. For ResNet-18, a batch size of 64 is used whereas for all other models a batch size of 32 is used. For all four models, a cosine learning rate scheduler with exponential warmup was used and were all trained for 2000 epochs. Further, all images for all models are resized to 256x256 pixels prior to training. All ex-

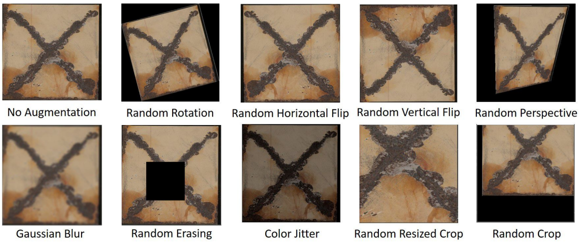

Figure 5: Illustration of data augmentation in our experiments for corrosion classification.

Augmentation	Parameters	ResNet-18	ResNet-50	DenseNet	HRNet
None	N/A	0.78 ± 0.03	0.72 ± 0.03	0.79 ± 0.01	0.76 ± 0.04
Color Jitter	Prob. 25%, Brightness (1.5, 2), Contrast (0.5, 1.5), Hue 0.5, Saturation (0.5, 1.5)	0.79 ± 0.03	0.74 ± 0.03	0.76 ± 0.02	0.68 ± 0.04
Gaussian Blur	Prob. 75%, Kernel 11, Sigma 5	0.75 ± 0.04	0.71 ± 0.05	0.74 ± 0.03	0.75 ± 0.02
Horiz. Flip	Prob. 25%	0.74 ± 0.03	0.70 ± 0.03	0.78 ± 0.03	**0.77** ± **0.04**
Rand. Erasing	Prob. 25%, Max Attempt 5, Area Ratio (0, 0.05)	0.78 ± 0.04	0.74 ± 0.04	**0.80** ± **0.02**	0.75 ± 0.02
Rand. Perspective	Prob. 75%, Distortion Scale 25%	0.76 ± 0.03	0.74 ± 0.04	0.74 ± 0.02	0.76 ± 0.05
Rand. Resized Crop	Prob. 25%, Scale (0.3, 0.7)	0.77 ± 0.03	0.76 ± 0.04	0.77 ± 0.03	**0.77** ± **0.04**
Rand. Rotation	Prob. 75%, Degrees (-25, 25)	0.77 ± 0.03	0.69 ± 0.03	0.74 ± 0.04	0.75 ± 0.03
Vert. Flip	Prob. 50%	0.75 ± 0.03	0.72 ± 0.04	**0.80** ± **0.02**	0.76 ± 0.04
Rand. Crop	Prob. 50%, Padding 4, Padding Mode Constant	**0.81** ± 0.04	0.74 ± 0.02	0.79 ± 0.02	**0.77** ± **0.03**
Combination	Same settings	**0.81** ± 0.04	**0.77** ± **0.03**	**0.80** ± **0.03**	0.76 ± 0.04
Pretrained + Combination	Same settings	**0.83** ± **0.01**	0.76 ± 0.02	0.79 ± 0.04	**0.83** ± **0.03**

Table 1: Test accuracy comparison using 10-fold cross-validation.

periments make use of the Apex package to improve computational efficiency [28] with our Pytorch codes running on GeForce RTX 2080 Ti GPUs.

We establish baseline performance for all four models by first using no augmentations. ResNet-18, ResNet-50, DenseNet, and HRNet, averaged over the 10 trained models, achieved 0.78 ± 0.03, 0.72 ± 0.03, 0.79 ± 0.01, and 0.76 ± 0.04 test accuracies, respectively. Table 1 demonstrates that different augmentation methods can yield higher classification accuracy. The best performance for ResNet-18 uses random cropping or a combination of random cropping and color jitter and achieves 0.81 ± 0.04 test accuracy. For ResNet-50, using a combination of color jitter, random erasing, random perspective, random resized crop, and random crop achieves 0.77 ± 0.03 test accuracy. DenseNet uses a combination of vertical flip and random erasing, or each one individually, and achieves 0.80 ± 0.03 or 0.02 test accuracy. HRNet uses horizontal flip, random resized crop, or random crop individually and achieves 0.77 ± 0.04 or 0.03 test accuracy. Combining these three augmentations for HRNet decreases performance by 1% back to baseline. Combinations of augmentations were chosen based on if individually they achieved higher than baseline no augmentation accuracy.

All rows in Table 1 are trained from scratch on our corrosion data, except the final row where each model is first pretrained on ImageNet then finetuned on our corrosion data set with augmentation. We see performances increase for ResNet-18 and HRNet. From this, we conclude that using pretrained models we can boost our best test accuracy by 0.2 to 0.83.

When using data augmentation methods, the goal is to supplement our data set to increase sample size and data variance. While applying any of these methods will theoretically achieve this, certain augmentation methods are more fitting for some data sets. We observe that color jitter can boost accuracy for ResNet models. Corrosion panels vary in background and corrosion color and therefore color invariance should be learned. We observe that gaussian blur never improves performance. We suspect blurring degrades image quality such that distinguishing between corrosion and background is blurred. Horizontal flip only slightly improves HRNet and vertical flip only slightly improves DenseNet. Flipping an image results in an "identical" image, allowing the model to see more examples. Random erasing tends to consistently improve performance. Removing image patches allows models to become robust to image occlusion. It thus leads to learning with information loss with overall image structure still being preserved [49]. In general, random perspective and random rotation are not helpful. All corrosion images are well centered and do not present distortions nor re-orientations given by these methods. Random resized crop does improve ResNet-50 and HRNet. However, cropping and resizing effectively changes scribe ratings by expanding corrosion areas. Finally, random crop is observed as one of the most consistent augmentations that improves accuracy. This is not surprising, given random cropping allows for occluding parts of images, makes small shifts in positions of the images, and preserves overall image structure and relative corrosion size.

From these experiments, we see that we are able to significantly improve classification performance over our non-expert rating study (0.38) using deep learning with tuned data augmentation methods. We identify that a smaller model, ResNet-18, tends to work better. We expect that this is due to fewer parameters in the model that allow better generalizability to unseen test data and the unique augmentation methods selected for our data set.

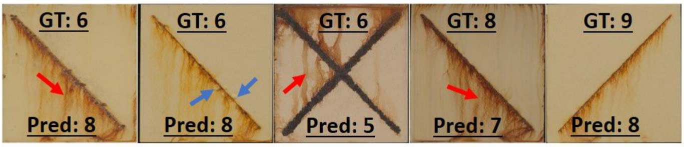

Figure 6: *Misclassified* samples with ground truth (GT) and predicted (Pred) ratings noted. Red arrows identify water stain, blue arrows identify raised blistering under the topcoat.

In further investigation of the CNN performance, we analyze instances of misclassification, as seen in Figure 6. We observe a fairly uneven pattern of corrosion across the single scribe along with water staining shown in red leading to potentially varied measurements across the scribe. The blue arrows point at raised blistering that have not yet perforated the topcoat. These areas would be identified and measured by an expert, but may not be recognized by current models. In panels 1 and 2, our model predicts a less harsh corrosion rating (Pred=8) when the ground truth is actually a lower rating (GT=6). This means that an expert accounted for the blistering and this resulted in a lower rating (more corrosion). In panels 3 and 4, we see more water staining identified with red arrows, and lower predicted ratings indicate that the similar colored water stain is confused with the prominent and dark corrosion. Finally, in panel 5, we observe very thin corrosion with a small amount of water staining around the thinly corroded area, potentially leading to the predicted lower corrosion rating (Pred=8) than ground truth (GT=9). These observations provide us with many hints about how to improve the accuracy of deep models, which we will explore in our future work.

4.3 Grad-CAM Visualization

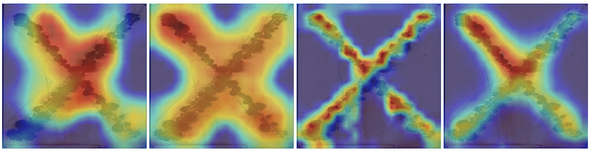

Figure 7: Grad-CAM visualization for correct predictions. Left to right: ResNet-18, ResNet-50, DenseNet, HRNet. High activation in red, low activation in blue.

For each of the four models, we select the best trained from scratch model and visualize a coarse localization map of important regions using the gradients flowing into the final convolutional layers via Gradient-weighted Class Activation Mapping (Grad-CAM) [40]. Grad-CAM produces an attention heatmap that can be applied to any neural network architecture. We illustrate our results in Figure 7 for an example image from our test set; with additional examples placed in the supplementary document. The example in Figure 7 with a ground truth rating of 5 has all four models predict this image correctly. We observe that for DenseNet and HRNet, the scribe area is distinctly highlighted with minimal activation of the background. This provides good evidence that these models can learn to focus on corrosion. For ResNet-18, the central area of the image is mainly focused on, leading us to believe that a smaller area of the panel may only need to be observed for it to make an accurate prediction (ResNet-18 had the highest test accuracy, 0.81). Finally, for ResNet-50, the scribe area is fully activated, but there is a large amount of activation beyond the scribe area. In this case, ResNet-50 appears to not be learning as well (0.77 test accuracy) as other models to focus on corrosion areas and thus does not generalize well to unseen images.

4.4 Self-Supervised Representation Learning: PIRL

PIRL	Backbone	Classifier	Test Acc.
Pretrained	ResNet-50	MLP	0.75 ± 0.03
Pretrained	ResNet-50	Linear	0.68 ± 0.02
Trained	ResNet-50	Linear	0.72 ± 0.04
Trained	ResNet-18	Linear	0.70 ± 0.03

Table 2: Pretrained and trained from scratch PIRL + classification results.

Using pretext tasks, self-supervised learning develops representations that are semantically meaningful without training on a large data set. In this way, learned representations are covariant with the pretext tasks, which involves redundant use of information. Recently, Pretext-Invariant Representation Learning (PIRL) was developed to learn invariant representations with pretext tasks. Solving jigsaw puzzles as the pretext task, PIRL with ResNet-50 achieves 0.64 accuracy on ImageNet comparing to 0.76 accuracy entirely supervised with ResNet-50 [27]. Following common practice in self-supervised learning, we applied the pretrained PIRL representation as an encoder and fine-tuned on our data to test our downstream corrosion classification task. Instead of a linear layer as the downstream classifier, we also tried a simple MLP layer with one relu activation layer to add complexity for this transfer learning process. We then trained a customized PIRL encoder on our training set entirely and tested the learned representations to classify corrosion. In Table 2, this empirical evaluation shows that: (i) the MLP layer rather than the linear layer improves the transfer learning performance from ImageNet to our data, (ii) with the linear layer classifier, using the PIRL representation trained on ImageNet (0.68) is worse than using the representation trained on our data with ResNet-50 (0.72) or ResNet-18 (0.70) backbone, and (iii) comparing with results in Table 1, PIRL pretrained on ImageNet with ResNet-50 backbone and MLP classifier (0.75) outperforms the baseline supervised ResNet-50 (0.72), but does not beat supervised ResNet-50 with tuned data augmentation (0.77).

5 Conclusion

In this work, we introduce a corrosion image data set with expert annotated ratings derived from standardized experiments used for materials research. Such data comes with several challenges, including the domain knowledge required to assess panels, the time to prepare panels, run experiments, and analyze each panel, and the cost to operate laboratory facilities. With our data set, we demonstrate that we can leverage deep learning techniques to automate corrosion assessment.We demonstrate that image augmentation methods can be tuned to our data achieving 0.83 classification accuracy in corrosion assessment.

Our longer-term goal is to build quality assessment models and integrate the assessment model with standardized experimental procedures for speeding up experimental workflows. Our data set will drive innovation and development of deep learning techniques such as generative models for corrosion progression prediction and new representation learning techniques for small data sets – over time bridging computer vision and material innovation.

Acknowledgements

This research was supported by the US ARMY, ACC-APG-RTP, Cooperative Agreement W911NF-19-2-0112, US. CDCC Army Research Lab. Ziming Zhang was partially supported by NSF grant CCF-2006738 as well. We also thank the Data Science Research Group at Worcester Polytechnic Institute.

References

[1] Brahim Aïssa and Maha Mohamed Khayyat. Self-healing materials systems as a way for damage mitigation in composites structures caused by orbital space debris. In *Handbook of Research on Nanoscience, Nanotechnology, and Advanced Materials*, pages 1–25. IGI Global, 2014.

[2] ASTM B117. Standard practice for operating salt spray (fog) apparatus. *ASTM International (1997 Edition)*, 2011.

[3] Yongsheng Bai, Halil Sezen, and Alper Yilmaz. End-to-end deep learning methods for automated damage detection in extreme events at various scales. *arXiv preprint arXiv:2011.03098*, 2020.

[4] David Banis, Arthur Marceau, and Michael Mohaghegh. Design for corrosion. *Boeing*, Jun 1999. URL https://www.boeing.com/commercial/aeromagazine/aero_07/corrosn.html.

[5] David M. Bastidas. Corrosion and protection of metals. *Metals*, 10(4), 2020. ISSN 2075-4701. doi: 10.3390/met10040458. URL https://www.mdpi.com/2075-4701/10/4/458.

[6] Young-Jin Cha, Wooram Choi, Gahyun Suh, Sadegh Mahmoudkhani, and Oral Büyüköztürk. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, 33(9):731–747, 2018.

[7] Noelle Easter C Co and James T Burns. Effects of macro-scale corrosion damage feature on fatigue crack initiation and fatigue behavior. *International Journal of Fatigue*, 103:234–247, 2017.

[8] Thomas Considine, Daniel Braconnier, John Kelley, Thomas Braswell, Christopher Miller, Brian Placzankis, and Robert Jensen. Data analytic prediction and correlation visualization of corrosion assessment for DoD sustainment. *Technical Report of US. CDCC Army Research Lab*, 2018, Unpublished.

[9] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2020.

[11] 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, 2009. doi: 10.1109/CVPR.2009.5206848.

[12] Yupeng Diao, Luchun Yan, and Kewei Gao. Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features. *Materials & Design*, 198:109326, 2021.

[13] Jeannine Elliot and Ronald Cook. Flexible chromate-free aerospace primer. *2019 Department of Defense Allied Nations Technical Corrosion Conference*, 2019. URL https://sspc.org/.

[14] American Society for Testing and Materials. Standard test methods for cyclic (reversed) load test for shear resistance of vertical elements of the lateral force resisting systems for buildings. *ASTM International (1997 Edition)*, 2011.

[15] Xiaoyu Gong, Chaofang Dong, Jiajin Xu, Li Wang, and Xiaogang Li. Machine learning assistance for electrochemical curve simulation of corrosion and its application. *Materials and Corrosion*, 71(3):474–484, 2020.

[16] Angel Green. New aircraft pretreatment and wash primer system. In *NACE CORROSION*, March 2010. URL https://onepetro.org/NACECORR/proceedings-pdf/CORR10/All-CORR10/NACE-10088/1717299/nace-10088.pdf.

[17] 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.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[19] Jiheon Jun, Adrian S Sabau, Zach Burns, and Mike Stephens. Corrosion performance of mil-prf-23377 primer coated of laser-interference structured aluminum alloy 2024. *2019 Department of Defense Allied Nations Technical Corrosion Conference*, 2019. URL https://sspc.org/.

[20] Jesse C Kelly, Adrian Goff, Jeier Yang, Charles Sprinkle, and Sarah E Galyon Dorman. Effect of inhibitor leaching from several aircraft primers on the corrosion protection of aa7075-t651. *2019 Department of Defense Allied Nations Technical Corrosion Conference*, 2019. URL https://sspc.org/.

[21] Gerhardus Koch. Cost of corrosion. *Trends in oil and gas corrosion research and technologies*, pages 3–30, 2017.

[22] R.A. Lane, C Fink, C Grethlein, and N Rome. Analysis of alternatives to hexavalent chromium: A program management guide to minimize the use of crvi in military systems. *Rome, NY: Advanced Materials, Manufacturing and Testing Information Analysis Center*, 2012.

[23] Jiali Li, Kaizhuo Lim, Haitao Yang, Zekun Ren, Shreyaa Raghavan, Po-Yen Chen, Tonio Buonassisi, and Xiaonan Wang. Ai applications through the whole life cycle of material discovery. *Matter*, 3(2):393–432, 2020.

[24] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment, 2019.

[25] Hui Lin, Bin Li, Xinggang Wang, Yufeng Shu, and Shuanglong Niu. Automated defect inspection of led chip using deep convolutional neural network. *Journal of Intelligent Manufacturing*, 30(6):2525–2534, 2019.

[26] Jeremy Mattison and Diane Kleinschmidt. Corrosion test method evaluation for aerospace sealants. *2019 Department of Defense Allied Nations Technical Corrosion Conference*, 2019. URL https://sspc.org/.

[27] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] Sharan Narang, Gregory Diamos, Erich Elsen, Paulius Micikevicius, Jonah Alben, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *Int. Conf. on Learning Representation*, 2017.

[29] NASA. Corrosion engineering technology. URL https://corrosion.ksc.nasa.gov/Coatings/CoatingsData.

[30] ASTM Committee D-1 on Paint, Materials Related Coatings, and Applications. *Standard Test Method for Evaluation of Painted or Coated Specimens Subjected to Corrosive Environments*. ASTM International, 2008.

[31] Manjula Panyam and Hariharan Venkatraman. Corrosion simulation tests: Analysis and improvement of corrosion resistance for automotive components. *SAE International Journal of Materials and Manufacturing*, 6(2):154–156, 2013. doi: 10.4271/2013-01-0335.

[32] Zibo Pei, Dawei Zhang, Yuanjie Zhi, Tao Yang, Lulu Jin, Dongmei Fu, Xuequn Cheng, Herman A Terryn, Johannes MC Mol, and Xiaogang Li. Towards understanding and prediction of atmospheric corrosion of an fe/cu corrosion sensor via machine learning. *Corrosion Science*, 170:108697, 2020.

[33] Alexander J. Ratner, Henry R. Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation, 2017.

[34] Colorado J Reed, Sean Metzger, Aravind Srinivas, Trevor Darrell, and Kurt Keutzer. Selfaugment: Automatic augmentation policies for self-supervised learning, 2021.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[36] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23 (3):309–314, 2004.

[37] Christina Rudén and Sven Ove Hansson. Registration, evaluation, and authorization of chemicals (reach) is but the first step–how far will it take us? six further steps to improve the european chemicals legislation. *Environmental health perspectives*, 118 (1):6–10, 2010.

[38] Adrian S Sabau, Jiheon Jun, Zach Burns, and Mike Stephens. Corrosion resistance of laser-interference structured aluminum alloy 2024 coated with mil-prf-85582 primer. *2019 Department of Defense Allied Nations Technical Corrosion Conference*, 2019. URL https://sspc.org/.

[39] S Samudrala, K Rajan, and B Ganapathysubramanian. Data dimensionality reduction in materials science. In *Informatics for Materials Science and Engineering*, pages 97–119. Elsevier, 2013.

[40] 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.

[41] Cher Tian Ser, Petar Žuvela, and Ming Wah Wong. Prediction of corrosion inhibition efficiency of pyridines and quinolines on an iron surface using machine learning-powered quantitative structure-property relationships. *Applied Surface Science*, 512: 145612, 2020.

[42] Jan Ivar Skar and Darryl Albright. Corrosion behavior of die-cast magnesium in astm b117 salt spray and gm9540p cyclic corrosion test. *2003 Magnesium Technology*, 1: 59, 2003.

[43] Daniel Vriesman, Alceu Britto Junior, Alessandro Zimmer, and Alessandro Lameiras Koerich. Texture cnn for thermoelectric metal pipe image classification. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 569–574. IEEE, 2019.

[44] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[45] Luchun Yan, Yupeng Diao, Zhaoyang Lang, and Kewei Gao. Corrosion rate prediction and influencing factors evaluation of low-alloy steels in marine atmosphere using machine learning approach. *Science and Technology of Advanced Materials*, 21(1): 359–370, 2020.

[46] Biao Yin, Thomas A Considine, Fatemeh Emdad, John V Kelley, Robert E Jensen, and Elke A Rundensteiner. Corrosion assessment: Data mining for quantifying associations between indoor accelerated and outdoor natural tests. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 2929–2936. IEEE, 2020.

[47] L YING-YU and W QUI-DONG. Astm D1654: standard test method for evaluation of painted or coated specimens subjected to corrosive environments. *Annual Book of ASTM Standards. West Conshohocken*, 1992.

[48] Wen Yu Zhang. Artificial neural networks in materials science application. In *Applied Mechanics and Materials*, volume 20, pages 1211–1216. Trans Tech Publ, 2010.

[49] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.