Deep Learning Unlocks X#ray Microtomography Segmentation of Multiclass Microdamage in Heterogeneous Materials

Deep Learning Unlocks X-ray Microtomography Segmentation of Multiclass Microdamage in Heterogeneous Materials

Reed Kopp, Joshua Joseph, Xinchen Ni, Nicholas Roy, and Brian L. Wardle[*]

Dr. R. Kopp, Dr. X. Ni, Prof. N. Roy, Prof. B. L. Wardle
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139, USA
Email: wardle@mit.edu

Dr. J. Joseph, Prof. N. Roy
MIT Quest for Intelligence
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139, USA

Prof. B. L. Wardle
Department of Mechanical Engineering
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139, USA

Keywords: deep learning, machine learning, heterogeneous materials, 3D multiclass damage, synchrotron radiation computed tomography, material characterization

[*] Corresponding author.

Abstract

Four-dimensional quantitative characterization of heterogeneous materials using *in situ* synchrotron radiation computed tomography can reveal 3D sub-micron features, particularly damage, evolving under load, leading to improved materials. However, dataset size and complexity increasingly require time-intensive and subjective semi-automatic segmentations. Here, we present the first deep learning (DL) convolutional neural network (CNN) segmentation of multiclass microscale damage in heterogeneous bulk materials, teaching on advanced aerospace-grade composite damage using ~65,000 (trained) human-segmented tomograms. The trained CNN machine segments complex and sparse (<<1% of volume) composite damage classes to ~99.99% agreement, unlocking both objectivity and efficiency, with nearly 100% of the human time eliminated, which traditional rule-based algorithms do not approach. The trained machine is found to perform as well or better than the human due to 'machine-discovered' human segmentation error, with machine improvements manifesting primarily as new damage discovery and segmentation augmentation/extension in artifact-rich tomograms. Interrogating a high-level network hyperparametric space on two material configurations, we find DL to be a disruptive approach to quantitative structure-property characterization, enabling high-throughput knowledge creation (accelerated by 2 orders of magnitude) via generalizable, ultra-high-resolution feature segmentation.

1. Introduction

Heterogeneous materials increasingly inhabit indispensable yet historically unpopulated spaces in materials engineering for high-performance structures subjected to extreme loading and/or environment, fostering modern breakthroughs in safety, efficiency, and operational envelope.[1–8] Disparate natural[9,10] and synthetic[11–14] constituent materials discoveries, concomitant with the growing diversity of combinatory processing techniques for design of microstructure[15,16] and interphase[17,18], has expanded both the range and knowledgebase of heterogeneous and composite materials, leading to synergistic mechanical property enhancements in bulk stiffness, strength, and toughness over traditional homogeneous engineering materials[19–21], while often also permitting interdisciplinary strategies for multi-functionality (thermal, electrical, optical, etc.).[22,23] Several cross-cutting research themes have emerged for mechanical property engineering via composite nano- and microstructural heterogeneity and anisotropy, including additive manufacturing[24–26], biomimetics[27–29], and hybrid advanced composites.[30,31] However, incomplete understanding of complex structure-property relationships[32,33], particularly progressive damage in tough heterogeneous systems built from brittle constituents, has emerged as a unifying theme limiting further performance enhancement among the breadth of cutting-edge advanced materials for structural applications. Fundamental knowledge of heterogeneous material mechanics across scales, particularly as related to ‘failure’, strongly limits performance predictive capabilities as well as rational nano/microstructural design toward optimization.[34–36] Motivated by ongoing failure prediction challenges faced globally by academia, government, and industry, aerospace-grade advanced composites exemplify the reality of a costly (~$100M[37] and up to 20 years[38,39] for new materials insertion), experimentally driven qualification campaigns, incurring materials design restrictions and conservative safety margins that undercut theoretical structural efficiency.[40,41]

The next advances in mechanical performance understanding are underpinned by higher-fidelity experimental characterizations (especially temporal data)[33] of complex failure processes comprised of multiscale, multi-modal (multiclass) interacting progressive damage[23,35,36] that inform and validate predictive models for design. Relative to conventional destructive microscale damage characterization techniques[42] like optical and electron microscopy (2D) and acoustic signaling/scanning (low resolution, <3D), synchrotron radiation computed tomography (SRCT) is the most data-rich imaging approach enabling 4D (spatial and temporal) mechanical failure characterization.[43–46] CT can nondestructively visualize interior 3D sub-micron (ultra-high resolution[44]) morphology including processing defects, and when coupled with *in situ* mechanical loading[47,48], establish full-field interplay of complex damage

progression (**Figure 1a**). Compared to modern lab-based X-ray CT, SRCT can provide unparalleled performance in aspects such as resolution and scan speed (at present, greater than tens of scans per second have been demonstrated, such as 'tomoscopy'[46]). Despite access to a complete progression of damage incipience and evolution, objective quantitative mechanistic insights are challenging to extract from the resultant big (~10 GB/mm^3), potentially damage-sparse (<<1% of scan volume) datasets due to differential X-ray attenuation and myriad X-ray imaging artifacts[49] (*e.g.*, rings, interface-enhancing phase contrast, motion blur, noise), computational expense, and particularly important herein, human-based (see Figure 1b, 'Trained human') time-intensive, subjective semi-automatic damage labeling/segmentation.[44,47,50] Human segmentation culminates in a data-to-knowledge bottleneck that significantly hinders new knowledge from the data-rich scans. SRCT datasets can be acquired over a day, but then require years of segmentation and analysis before knowledge is gained and quantified. Furthermore, given sensitive image properties and feature diversity, including irregular damage morphology, generalized automated damage segmentation based on traditional "rule-based" programming (*i.e.*, "hard" computing[51]), which employs digital image processing tools[45,47,50] (*e.g.*, filtering, thresholding, clustering, transforming), that would promote objectivity, speed, and feature flexibility is thus far impossible to codify. To date, limited work has demonstrated only subjectively tuned automated hard algorithms for damage segmentation with narrow applicability[44,45,52,53], preventing translation to other materials and *in situ* configurations.

Recently, mirroring the rise in availability of both computational power and big data, deep learning (DL)[54,55], a branch of machine learning (ML) within artificial intelligence (AI) employing artificial neural network (ANN) models, related to "soft" computing[51], to learn complex input-output mathematical mappings of various data types, has disrupted many scientific fields with leap-ahead computational classification and regression[56], including in materials science.[57–62] Particularly, semantic segmentation, defined as pixel-level image classification, is a proven DL application within computer vision, commonly achieved via a (fully) convolutional neural network (CNN) architecture. CNNs are a spatially invariant class of deep ANNs that especially excel in scalable discriminative ("example-based") learning of big, complex imagery datasets (Figure 1b, 'Trained machine').[63–65] Although commercial[66] and open-source[67–69] image analysis tools have recently integrated ML capabilities, generally, all datasets are extracted from a single 2D or 3D image for the purpose of quickly inferring segmentation of the rest of the image; thus, more generalized CNN learning across numerous scans (as required for damage segmentation) would require a custom algorithmic workflow spanning

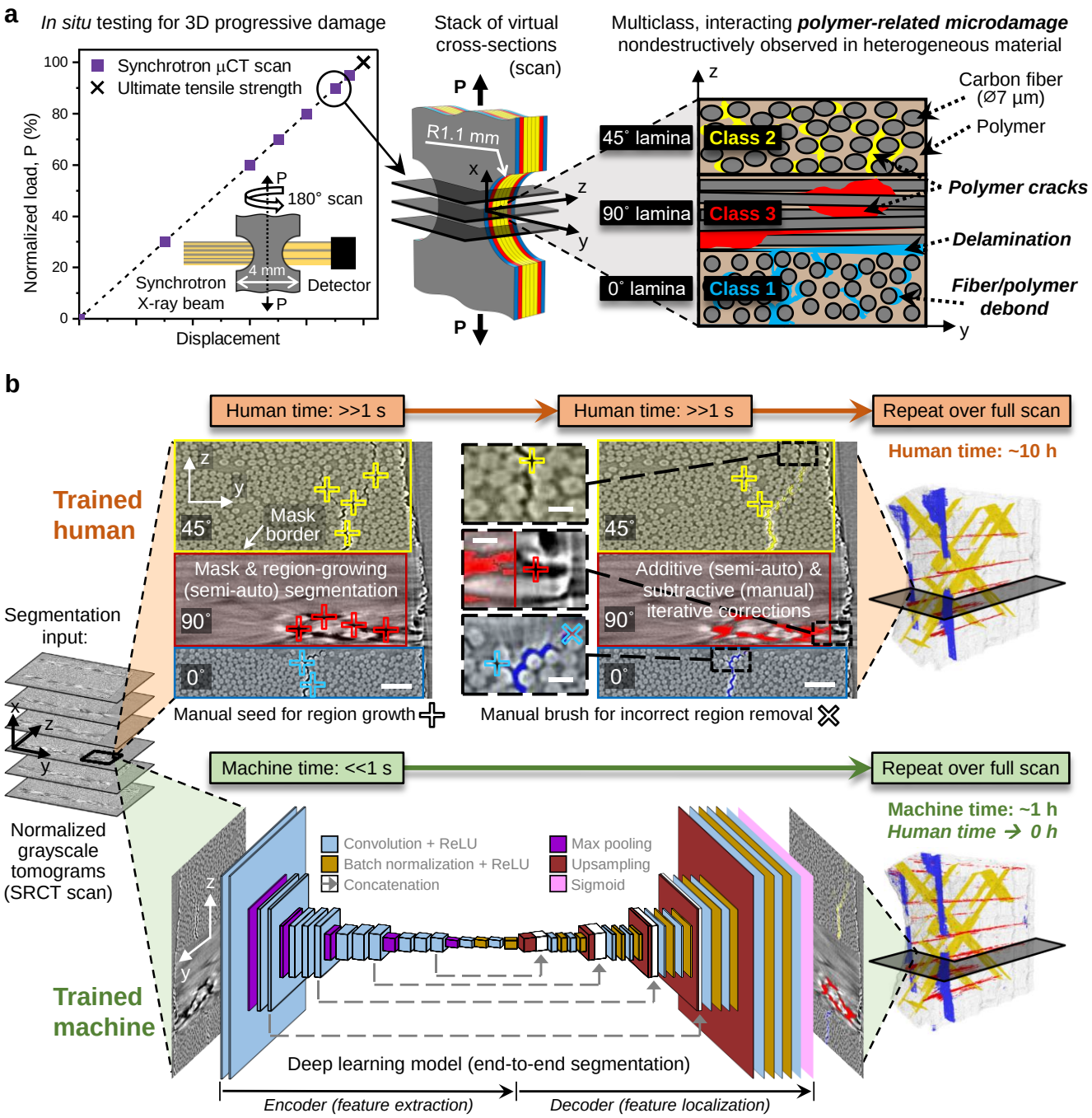

Figure 1. Deep learning unlocks elusive automated damage segmentation of big, complex data. a) *In situ* mechanical testing via synchrotron radiation (X-ray) computed microtomography generates big, rich benchmark datasets of complex (multiclass, interacting microscale modalities) progressive polymer-related damage that underpin heterogeneous materials failure understanding. b) Segmentation input is a SRCT scan, reconstructed as a stack of normalized grayscale tomograms. (top) Trained human pipeline, which includes manual masking and seed-point identification for semi-automatic region growing, and both manual and semi-automatic corrections. Diligent extraction of seminal failure insights from complex data is currently bottlenecked by time-intensive, iterative region growing-based human segmentation analyses that introduce subjectivity, inconsistency, misclassification, etc. and limit the

feasible scope of material types and load step-resolution. Scale bars are 50 μm in zoom-out and 10 μm in zoom-in. (bottom) Trained machine pipeline, which begins from the same starting point as the human, a deep learning network machine performs end-to-end segmentation (no pre- or post-processing) rapidly with negligible human interaction required.

dataset generation to achieve multi-scan inferences. To date, DL for semantic CT segmentation of material degradation has been leveraged by several fields, including medicine[70,71] and materials engineering (*e.g.*, concrete[72] and advanced composites[73,74]), though to the authors' knowledge, multiclass, micron-scale damage (and progression) that underpins failure of advanced composites remains entirely unexplored and a critical open research question in the space of heterogeneous materials generally. Moreover, as a general practice in literature, rule-based automated or even simulated segmentations comprise the learning datasets[73], such that successful CNN replication is highly expected. Alternatively, if manual segmentations were required for learning, then small manually annotated datasets (~100 or fewer images) have been employed[72,74], neglecting the potential advantages of large-scale CNN generalizability. Here, we use DL CNNs to unlock a traditionally impossible image analysis challenge: end-to-end (no pre- or post-processing) automated segmentation of multiclass microdamage progression revealed by *in situ* (sometimes called 4D due to the temporal aspect) SRCT. We study the effects on DL of dataset size and composition, using 30 semi-automatically ('trained human') annotated scans (comprised of 65,000 trained-human-segmented tomograms) for training the CNN, comprising 6 specimens and 2 types of advanced (aerospace-grade carbon fiber) composites, as well as high-level CNN hyperparameters. Following machine downselection for highest performer, we examine a generalizable case study of test set scans, demonstrating machine-inferred segmentation ('trained machine') of composite damage according to host laminae/plies within composite laminates to ~99.99% class binary accuracies, with negligible human time required. Compared with semi-automatic human segmentation (order of ~10 h per scan), the AI segmentation exhibits objectivity in addition to generalizability and high throughput, with numerous examples of exceeding human effectiveness.

2. Results and Discussion

2.1. Trained-Human Segmentation Database of Progressive Damage in Heterogeneous Materials to Create Trained Machines

Deep learning is a form of AI/ML-based machine development wherein a discriminative multi-layer (deep) NN model is trained (generally under supervision) to predict an output signal given an annotated dataset of input-output relations, validated, and tested on independent annotated datasets, as standard

practice. Although it is common in the ML community for a trained CNN to be referred to as a learned model and the broader system of algorithms including data processing and CNN-based prediction to be referred to as a computer, in the context of this study, we refer to both as a trained machine, as we're directly contrasting with a trained human. The performance benefits, in terms of accuracy and generalizability, of DL techniques over traditionally programmed automated algorithms are generally found to increase with increasing training dataset size and diversity that encapsulates the desired feature space.[54] Naturally, *in situ* SRCT testing output, which can generate on the order of 10^3 tomograms (*i.e.,* virtual cross-sections) per SRCT scan and incorporate on the order of 10^1 to 10^3 scans over all load steps per specimen lifetime[44,45], dovetails with the fundamental big data requirements for robust DL. Yet, human-driven annotations are temporally and often financially expensive, and along with the computational expertise and resources required for effective DL management, constitute a high barrier to entry for robust DL investigation; hence, the striking current absence of precedent for such a study. Therefore, compared to DL architecture studies that focus on new algorithms applied to existing benchmark datasets[75], we instead focus here on applying an existing DL architecture proven in computer vision to SRCT-imaged damage progression in two different types of advanced composites (here, carbon fiber reinforced polymer). The trained-human database forms the ground truth and underpins the objective study of hyperparameter effects (*i.e.,* high-level variables in model architecture and training strategy, excluding low-level learned model parameters/weights) on DL performance, which directly informs machine downselection and the subsequent evaluations of sample scan segmentation.

Here, we assess DL for heterogeneous materials damage segmentation in the context of typical data output from a single SRCT experimental campaign[76,77], wherein the number of acquired scans is broadly limited by the properties of the synchrotron light source, data acquisition, and *in situ* test apparatus complexity, among other variables. As depicted in Figure 1a and discussed further in Methods, the *in situ* test considered is double edge-notched tension (DENT) of composite laminates, for which the symmetric specimen geometry is known to promote progressive damage concentration between the notches. The specimens feature common lamina stacking sequences designed to exhibit in-plane elastic isotropy: 0° (aligned with tensile load), ±45°, and 90°. Visualized up to 95% of ultimate tensile strength (UTS), the primary damage mechanisms revealed via SRCT in this configuration are intralaminar polymer cracking and fiber/polymer interfacial debonding, as well as relatively small interlaminar delaminations connected to (and thus classified as) 0° lamina damage — we note that fiber

breakage, normally observed in laminae aligned with tensile loading, was not observed appreciably here due to the polymer-related damage that induces considerable stress redistribution. As shown in **Figure 2a**, while all laminates were ~1 mm thick, two different lamina thicknesses — termed 'Thick' (8 plies per laminate) and 'Thin' (16 plies per laminate), with each consisting of a different fiber/polymer system — were tested to examine the strong effect of ply thickness on intrinsic polymer damage suppression, which is induced by "*in situ*" geometric/size effects on strength and fracture energy release rate.[78] While Thick and Thin specimens exhibit identical damage modalities, their extent and mix of such damage varies significantly, such that the segmentation problem correspondingly varies in terms of damage mechanism morphology (*i.e.,* fracture surface opening displacement, width, and length), as well as distinct gray values of the bulk composite material (elemental-dependent X-ray attenuation properties, which also vary generally with X-ray energy). Therefore, we identify a damage segmentation problem that features two different levels of complexity: (i) relatively simpler Thick composites that exhibit a low quantity of relatively large-volume polymer-related damage mechanisms, and (ii) relatively more complex Thin composites that exhibit a high quantity of relatively small-volume polymer-related damage mechanisms. Accordingly, in view of a generalized damage segmentation tool that learns various manifestations of damage across different composites, we can determine the effects of Thick and Thin data contribution on DL.

In our SRCT experiments, four Thick specimens and two Thin specimens were tested (specimens denoted by color in Figure 2a), with each test comprising several load steps (*i.e.,* SRCT scans executed during pauses in monotonic loading, distinguished by letter) in the range of 60%–95% UTS as well as an unloaded step (~0% UTS). In total, a resultant set of ~65,000 raw (no pre-processing beyond standard grayscale normalization, see Figures S1 and S2, Supporting Information) tomograms, which were semi-automatically human-annotated following a human-seeded region growing algorithm as described in Methods, representing 30 SRCT volumetric scans across 6 specimens, forms a robust database for evaluating DL-based damage segmentation. Prior studies[72–74] of DL damage segmentation of X-ray tomograms considered only a single class of damage, and the annotated learning datasets incorporated on the order of 10^2 or fewer human-annotated tomograms, precluding general assessment of DL capacity. From our ~65,000-tomogram database, the effects of training dataset size, composition, and sequence on learning performance are studied, keeping the test dataset identical for all machine/DL network development to facilitate unbiased performance evaluation. As discussed further in Methods

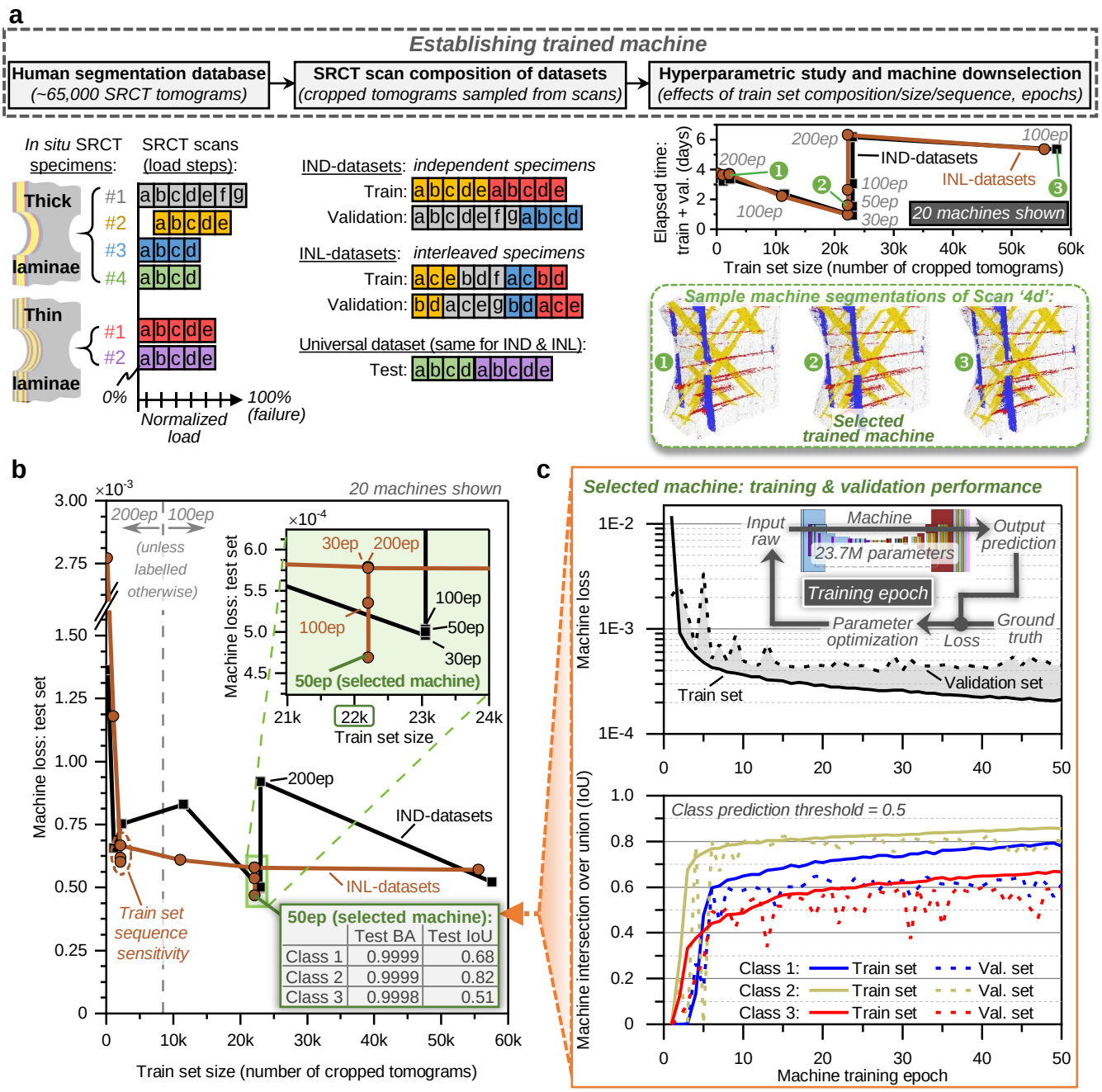

Figure 2. Deep learning machine development and downselection to preferred 'trained machine'. a) SRCT dataset composition (numbers for different specimens; letters for different specimen scans) and hyperparameter spectrum studied (20 different machines), including sample inference comparison (scan Thick 4d), leading to rigorous machine downselection. b) Unbiased downselection to the selected machine (dark green) based strictly on minimization of test loss, comparing the complex influence on machine performance of various hyperparameters: number of training crops, number of training epochs, training and validation dataset composition, and train set sequence sensitivity. The test set class intersection over union (IoU) and binary accuracy (BA) standard metrics are shown for the selected machine following prediction threshold training using the validation set. c) Training and validation learning curves of the downselected machine '5INL-50ep' (top), in which gray shading exhibits

relatively low degree of overfitting, along with training epoch (feedforward and backpropagation loop over dataset) depicted in inset. Class IoU (bottom) reflects the degree of overlap of human vs. selected machine segmentations.

and illustrated in Figure 1b, the input and output image dimensions of the DL machine are smaller than the output SRCT tomogram dimensions in our database, requiring the typical practice of patchwise sampling[54] of cropped sub-tomograms to populate the learning datasets (cropped sub-tomogram area is ~5–7% of tomogram area). Patchwise dataset sampling enables efficient learning, and as observed in Figure 1b for a Thick tomogram, the current patch size importantly can enable simultaneous visualization of the three damage classes. Moreover, since polymer-related damage in aggregate constitutes <<1% of the scan volume and is dominate by ±45° laminae damage, efficient and generalized dataset crop sampling of a given tomogram was enforced via both random crop positioning and presence (or not) of each class, in which the three possible classes ('Class 1, 2, and 3', see Figure 1a) refer to damage in either 0° ±45°, or 90° lamina, respectively. Class-based crop sampling was found to be far more effective than purely random sampling (Figure S6, Supporting Information). With the mechanisms of sub-tomogram dataset sampling established, two different strategies for SRCT scan composition of train and validation datasets were assessed, organized by distribution of specimens and their load steps: (i) independent datasets where each specimen appears in only one dataset ('IND-datasets'), and (ii) interleaved datasets where each specimen has load steps mixed across the datasets ('INL-datasets'). These strategies for controlling representation (or not) of Thick and Thin in both train and validation datasets may have broader implications for *in situ* experimental design to minimize specimen count (*e.g.*, prioritization of load steps and envelope). For example, the INL training dataset only contains 20% Thin scans, whereas the IND dataset contains 50% Thin scans. Despite train and validation dataset variation, we note that the test dataset is split nearly evenly between Thick and Thin specimens that intentionally do not appear in any other datasets, facilitating a proper and balanced performance evaluation. Finally, dataset size, which was varied primarily for training here, was controlled by identically linearly downsampling each of its constituent scans, followed by the class-based sub-tomogram sampling algorithm being applied identically to each downsampled tomogram. More detailed discussion of dataset size and composition and data augmentation, as well as the primary software and hardware used in this study, accompanies Tables S1 and S2, Supporting Information.

2.2. Deep Learning to Train Machine Semantic Damage Segmentation

Semantic segmentation is a particularly challenging application of DL for pixel-/voxel-level image classifications, inherently demanding a relatively complex, many-layered model architecture to encode feature classes and then decode feature locations.[75] Recently, among various architectures for semantic segmentation in self-driving vehicle and biomedical applications, the U-Net[64,67] fully convolutional encoder-decoder NN has demonstrated great success in biomedical microscopy segmentation and inspired numerous derivative models.[75] As another seminal architecture, though for large-scale image recognition, VGG16[79] is an early application of a relatively deep CNN for enhanced feature classification. Combining U-Net as the encoder-decoder high-level structure with VGG16 as the deep CNN backbone, we employ here their hybridized model (38 trainable layers, 23.7M trainable parameters) for multiclass semantic segmentation, with the architecture and layer definitions overviewed in Figure 1b and detailed in Figures S4 and S5, Supporting Information. Counterintuitive observations leading to the identification of universally effective hyperparameters were discovered in preliminary work and uniformly applied to all machines during subsequent refined hyperparametric investigations (*i.e.*, machine downselection), as discussed further in Section S1, Supporting Information. First, although the current segmentation problem is defined (based on physics) by multiclass classification (*i.e.*, mutually exclusive pixel-level damage classes), which functions via a softmax output layer in coordination with the categorical cross entropy loss function, the severe imbalance of background (composite bulk, air, or SRCT reconstruction mask) over damage pixels skewed learning toward only uninteresting undamaged regions, despite implementation of background-class-weighted predictions. Therefore, generalization to multi-label classification (*i.e.*, independent pixel-level damage classes, though inference selects only the highest prediction value relative to its class threshold) was employed, which functions via the sigmoid output layer in coordination with the binary cross entropy loss function. Second, a negligible impact of background class inclusion during multi-label classification learning was found, such that the learning datasets generated here exclude background class annotations.

Beyond the aforementioned universal hyperparameter identification, additional hyperparameter investigations were conducted to study DL performance effects from several factors including train set size, composition (IND vs. INL), and sequence sensitivity, as well as the usual number of training epochs (*i.e.*, the feedforward-backpropagation iterations over the entire training dataset, with major steps delineated in Figure 2b inset). As shown in Figure 2a, 20 different machines (encompassing 12 different train/validation datasets) developed across this hyperparametric space are plotted to study combinations

of these hyperparameters on training and validation time, with the general goal of informing minimization of required machine development time toward increased focus on machine inference. Additionally, for comparison, sample inferences from three different machines, spanning a large training dataset size range, of the same Thick test dataset scan (90% UTS) are shown below the plot, with machine 2 (‘2’ inside green circle) highlighted since it is found to be highest performing (referred to as ‘5INL-50ep’ later). Qualitatively, the 3D-arranged segmentations (*i.e.*, vertical stacking of 2D tomograms) appear very similar, suggesting that small training datasets may potentially be counterbalanced by increased training epochs for a similar cost in training and validation time, though clearer differences in segmentation performance can be distinguished in sample 2D tomograms (see discussion and Tables S7 and S8, Supporting Information).

For an objective quantitative comparison of performance of different machines/networks for unbiased downselection, we track the performance of each machine/network through training, validation and testing processes. Here, a unitless logarithmic loss function, which directly affects parameter optimization/learning during training, as well as class binary accuracy (BA) and class intersection over union (IoU), which are standard semantic segmentation performance metrics that measure agreement of prediction and ground truth (here, trained-human segmentation), computed over the training and validation datasets are important for understanding general learning history. Though, the most straightforward comparison of machine performance focuses simply on the loss function, which is defined between the annotated label and the machine/network output for each pixel, computed over the test dataset (Figure 2b). Interestingly, we find a complex relationship between test loss function and the hyperparameters of training dataset size and composition and number of training epochs (labeled as ‘ep’). For both IND- and INL-datasets, machine performance generally improves with increasing training dataset size, as expected, but only through dataset sizes of ~20,000 cropped tomograms; larger training dataset sizes exhibit unchanged or worse performance, depending on number of epochs. This is the known general tradeoff between underdeveloped feature learning (*i.e.*, underfitting) and overfitting, which will be examined later. Overall, as clearly depicted in the Figure 2b inset, the highest performing (and thus downselected for further examination) machine is found to feature the following hyperparameters: INL-datasets, ~22,000 cropped tomograms in train set, and 50 training epochs. Additionally, the class-based performance on the test dataset is listed for the selected trained machine, following class prediction threshold training on the IoU metric, with macro-averages of 0.67 and >0.9998 for class IoU and BA, respectively, on the test set (see Section S2, Supporting Information, for

class prediction thresholds and additional performance metrics discussion, including a validation set macro-average BA of 0.9999), motivating detailed qualitative and quantitative examination in the next section. For qualitative insight into hyperparameter effects on learning of the overall database, sample tomogram inferences for similarly performing machines (relative to the downselected trained machine) are included in Section S3, Supporting Information. Finally, shown in Figure 2b is that machine performance is relatively insensitive to the sequence of training dataset tomograms during training, observed as minimal deviation in test dataset loss.

Focusing on the downselected machine (5INL-50ep) only, representative training and validation results are shown in Figure 2c, capturing the general trends in loss function and class IoU, which are computed on both train and validation datasets after each training epoch. As is typical, the training dataset loss continually decreases during training and asymptotes at an imprecise value based on the intrinsic ground truth error, which will be inspected subsequently; the validation dataset loss decreases in an increasingly stable manner during training initially, but remains relatively constant as training concludes. The gray-shaded region between loss curves indicates the degree of overfitting to the training dataset, and despite its increase through training conclusion at 50 epochs, better performance is found here than for identically configured machines trained for only 30 epochs, attributed likely to underdeveloped feature learning with fewer epochs (see Figure S29, Supporting Information, for training and validation results of other machines). Class IoU is an alternative significant metric to track during learning, particularly because it decomposes machine performance into its individual class contributions, which is particularly valuable since different classes have different ground truth error and are also notionally more challenging to learn than others. It is important to note that common image analysis metrics like IoU, F_1 score, Precision, and Recall require criteria definition for transforming the three output class prediction scores for each pixel into a single criterion score. Whereas class prediction thresholds are optimized eventually over a desired metric, this so-called class prediction threshold training is not possible until after the machine training has converged; thus, a default class prediction threshold of 0.5 is assumed during training and validation, and importantly, it has no effect on parameter learning (loss function definition is independent of prediction threshold). For reference, recall that a 0.5 threshold is exactly correct in an unbiased purely binary classification problem. Similar to the loss curves, the class IoU curves reflect steep performance increases followed by asymptotic regions. Here, we observe in Figure 2b that Classes 1 and 2 are easiest to learn, followed by Class 3 (attributed to

inherent physics-related challenges with SRCT imaging of Class 3, see Figures S25–S27, Supporting Information).

2.3. Generalizable Segmentation of Multiclass Damage in Layered Composite

Following the trained machine downselection, a detailed qualitative and quantitative inference-based evaluation of 5INL-50ep performance is presented using a salient subset of sample scans from the test dataset: one Thick and one Thin specimen, each with load steps of 70%, and 90% of DENT UTS (corresponding 0% UTS results are presented in Figure S28, Supporting Information). This subset enables a clear assessment of machine intelligence capacity, emphasizing objective and consistent segmentation across different composite types (Thick and Thin) and their corresponding damage mixes as well as sequential progressive damage states, thus facilitating an overall assessment of generalizability, which is arguably as critical as segmentation accuracy for adoption in other similar problems.

The full 3D segmented damage states are the primary desired qualitative outcome of *in situ* CT testing, and are shown in **Figure 3**. In the first two columns, 3D renderings of human- and machine-segmented damage are shown, as are the textured specimen exterior surfaces (standard skin feature caused by surface film during autoclave curing) and rough water-jetted notched edge, which contextualize the location of damage in the DENT specimens. Additionally, the schematic at the right portrays the lamina orientation sequence (simply colored to match the resident damage class) of each composite type to assist in situating features. Overall, the sparse 3D-connected damage segmentations show excellent agreement between human and machine across various damage classes and scales. For a more detailed discernment of human vs. machine performance, the intersecting (column 3) and non-intersecting (column 4, *i.e.*, difference sets) components of each class segmentation are shown, with the specimen exterior skin rendering removed. Taken together, these columns illustrate the quantitative IoU metric measuring overlap. Thus, for each scan, the corresponding scan-level IoU and Recall (*i.e.*, portion of human segmentation that is positively predicted by machine) scores are included for quantitative context. In column 3, we observe strong agreement by comparing the intersection components to the human segmentation, which is summarized quantitatively by the all-around relatively high Recall scores. In column 4, noting application of a different color legend to decompose human vs. machine segmentation for each class, where for a given class ‘human only’ aggregates false negative (background) or false positive for other classes, we visualize a breadth of non-intersecting components that appear to mirror the major intersecting component trends. However, as will be recognized in 2D

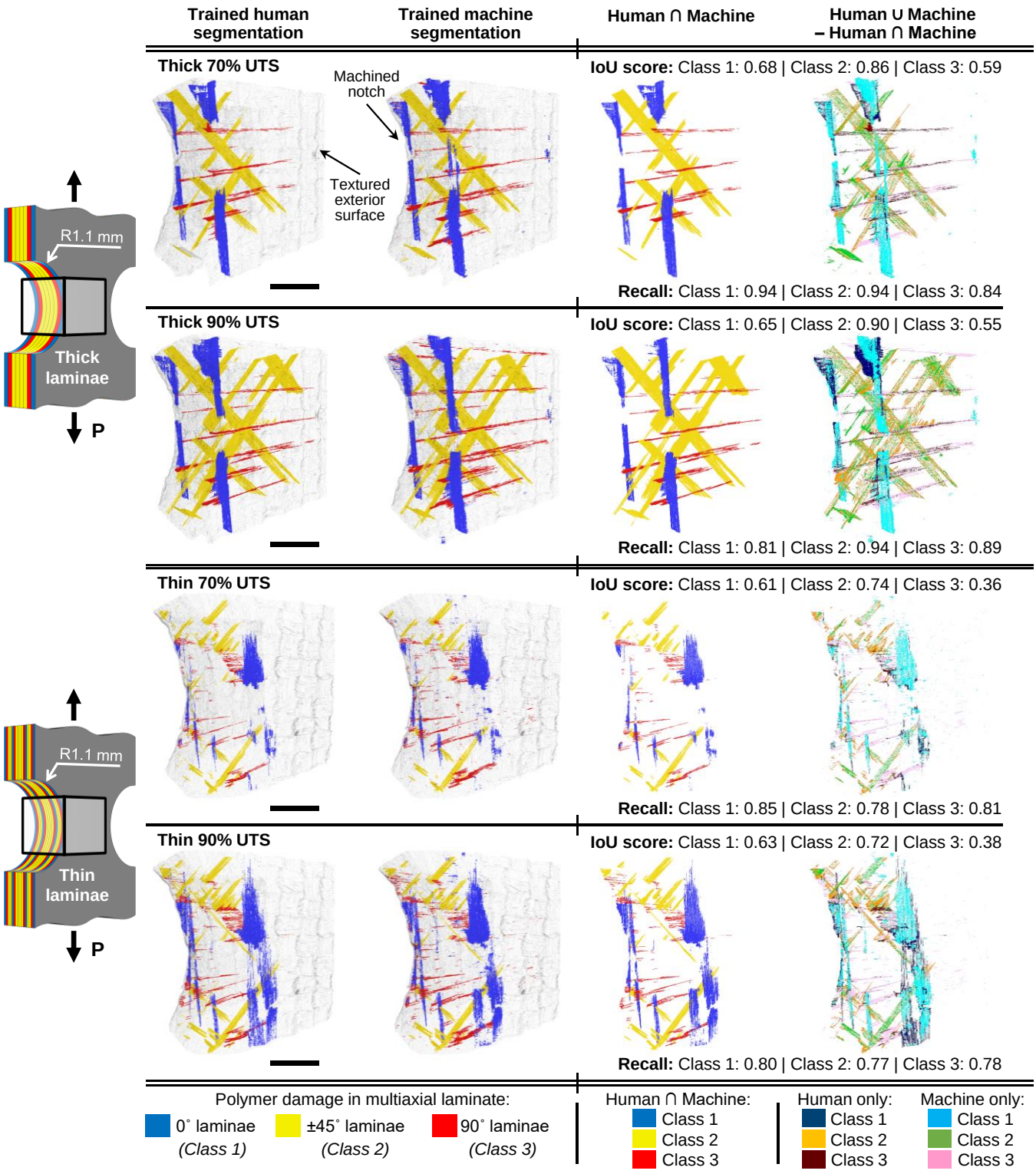

Figure 3. Comparison of 3D segmentations of selected test set scans. Thick and Thin material type specimens at 70% and 90% of ultimate tensile strength (UTS) are segmented by a human and the selected trained DL machine (columns 1 and 2). Column 3 visualizes the 3D intersection of human and machine results, showing good agreement via solidly filled-in damage instances, whereas column 4

visualizes the non-intersecting components, which generally form thin films over intersection components. Corresponding intersection over union (IoU) and Recall scores quantify excellent human and machine agreement. The schematic (right) visualizes the lamina stacking sequence in each specimen. Scale bar for all images, 500 μm.

examination of segmentations, the unfilled non-intersecting components exaggerate the actual non-intersecting volume of damage, since it will be seen that generally the non-intersecting components form a thin-film over the intersecting components, which can be interpreted generally as the machine improving on the ground truth with an objective/consistent selection of the extent of damage. Quantitatively, the combination of columns 3 and 4 represent the standard IoU metric components. In analyzing the class IoU scores, it is critical to note their potential to be misleading in the presence of nontrivial ground truth error, as any superior performance by the machine is reflected in a IoU penalty, which motivates our subsequent 2D analysis. Thus, considering both Recall and IoU, the overall qualitative 3D assessment of machine segmentation is strikingly positive, effectively replicating human performance. Of particular import, concentrating on problematic scan regions that normally preclude simple gray value thresholding and traditional hard programming, the machine consistently avoids segmentation of the artifact-prone notch edge and exterior surface, which are common regions where semi-automatic segmentation schemes would fail.

For deeper quantitative insight into trained-machine segmentation performance, the test dataset sample scan segmentations are examined via 3D object analysis (see Methods section), focusing on three key comparative perspectives: spatial distribution (**Figure 4a**), volume (Figure 4b), and grayscale intensity (Figure 4c). Regarding trained-human segmentations, note that each 3D-connected object is the result of one application of region growing, though iterative manual subtractive corrections are generally required for each such object. First, in Figure 4a, aggregating the barycenter locations for 3D-isolated damage segmentation instances, we map the volume-weighted barycenter location of each class (color) in each scan (shape) within the corresponding linear 3D image space (axes normalized by scan dimensions). For all segmentation scan/class combinations, we observe relatively small (<5%) or even negligible shifts in normalized barycenter, suggesting good agreement in global spatial distribution. Next, in Figure 4b, the segmented class volumes for each scan, which are also the pointwise volumes represented in Figure 4a, are shown relative to a general measure of scan sparsity (<<1% volume), underscoring an inherent impediment to segmentation learning in such CT scans of damage even though the notches concentrate damage in the scanned regions. Noting the difference in scale between Thick

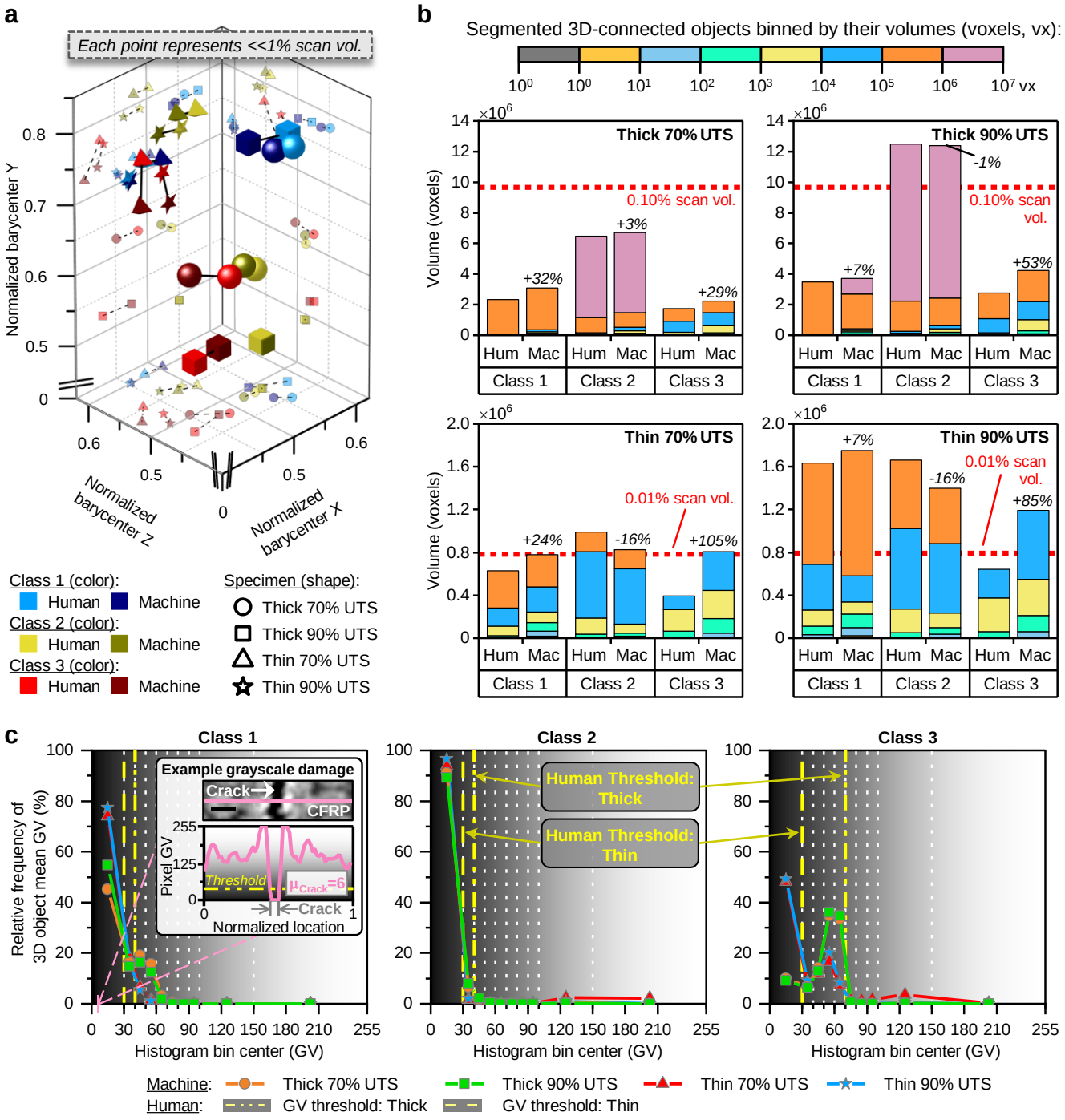

Figure 4. Quantitative comparison of 3D segmentations of selected test set scans. a) Mapping of segmentation barycenters (centers of volume normalized by scan dimensions) translation reveals agreement (shifts <5%) of global spatial distribution of segmented damage. b) Quantification of sparse (<<1% scan volume) segmentation volumes (Human: 'Hum' vs. Machine: 'Mac') decomposed into subgroup volumes contributed by 3D objects within selected volume ranges, revealing the wide 3D object volume range and degree of damage connectedness. Good agreement is demonstrated in the overall class volumes segmented, as well as the decomposition of total volume into contributions by objects within given volume ranges. c) Gray value (GV) intensity-based histograms exhibit the mean

gray values of machine-segmented 3D objects (via analysis of stack-arranged segmented slices). Dotted white lines reflect the histogram bin boundaries, and grayscale backdrop contextualizes the mean GV (see inset for example grayscale damage). The machine (5INL-50ep) appears to sufficiently learn the human region growing GV thresholds. Scale bar (inset), 5 μm.

and Thin scans, the listed percent differences in total class volume segmented indicate strong agreement for several scan/class combinations, while considerable disagreement is noted for others. Diving deeper, the total segmented volume is then decomposed into the aggregate contributions of all objects falling into log-scale-separated volume ranges, facilitating understanding of object segmentation discrepancies in the context of 3D-connectedness: noise (small object volumes on the order of 10 voxels or less) vs. ostensibly 3D-connected damage (larger object volumes). Representative depictions of these damage volume classes are shown in Figure S13, Supporting Information. Generally, the total volume discrepancies between human and machine correspond to discrepancies in larger-volume objects, suggesting that discrepancies arise due to machine augmentation/extension or at the level of entire 3D-connected damage instances (as also reflected in Figure 4a), requiring the higher resolution investigation discussed later. An example is the thin-film of machine vs. human disagreement in column 4 of Figure 3, as discussed earlier. Additionally, it is recognized that linear scale of Figure 4b prevents clear resolution of small volume object presence; thus, a companion log-scale version of the plot is found in Figure S12, Supporting Information, and supports the aforementioned explanation. Interestingly, smaller-volume objects are characterized in the human segmentations as well, due to subtractive manual operations that fragment larger region-grown objects, suggesting a source of ground truth error since the maximum tomographic resolution is 2–3× voxel size, setting the theoretical minimum acceptable human-segmented connected object volume to be on the order of 10 voxels. Finally, in Figure 4c, we report 3D object analysis histograms per class focusing on grayscale intensity, with the plot backdrops contextually given according to the 8-bit grayscale (gray values: 0–255) of the tomograms, showing distribution of 3D object mean relative to the aforementioned gray value thresholds (which vary according to composite type and damage class) applied during human region growing. In the inset, we illustrate the tomographic material representation along a line plot of grayscale, exemplifying a sample crack segmentation with sub-threshold gray values (here, mean object gray value of 6). Clearly, a great majority of machine-segmented objects, including those of Class 3, across both laminate types and load steps, exhibit mean gray values within the human-defined thresholds, suggesting flexible ML of the (usefully variable) human segmentation strategies for different composite systems. An expansive set of

histograms representing basic 3D object statistics, broken down by object volume range, can be found in Figures S14–S22, Supporting Information, particularly addressing numerical outliers (recall, no post-processing here at all).

To better understand the trained human vs. selected trained machine (5INL-50ep) performance, we consider comparative segmentation themes generally observed in 2D tomograms, which are direct inputs and outputs of the machine, while focusing on the more damage-prevalent (90% UTS) Thick test dataset scan segmentation, shown in **Figure 5** (corresponding results for Thin presented in companion Figure S23, Supporting Information). The following discussion applies to both Thick and Thin figures. Applying the same color legend as in Figure 3, we illustrate the general themes in representative Thick sub-tomograms (vertical location along specimen marked in bottom schematic; location selection not based on any rules), with the organizing principle being qualitative comparative performance, *i.e.*, equal, superior, or inferior performance of the machine relative to human. Though, practically, examples of all three distinctions are actually interspersed throughout each sample sub-tomogram. In each tomographic image and inset, we identify several factors/artifacts that complicate generalized CT image analysis of damage: damage sparsity, multiple scales of isolated damage instances (ranging from single fiber/polymer debonds to multi-lamina polymer cracks), highly irregular damage morphologies and orientations, morphological, spatial, and grayscale intensity differences in Thick vs. Thin damage, jagged notch edges, phase contrast fringes at interfaces due to X-ray detection in the near-field Fresnel region[44], unclear reconstruction of 90° laminae (*e.g.*, Figure S25, Supporting Information), and noise, among others. Nonetheless, equivalent performance (row 1) is demonstrated by virtually complete segmentation of matrix damage instances, which is observed in the vast majority of tomograms across all scales and classes, consistent with the high validation and test dataset macro-averages of class BA scores of 99.99% and 99.98%, respectively. Interestingly, although the machine can only access the yz-plane during training, without the aid of human-generated masking, it demonstrates adequate performance on 90° lamina damage despite its lower quality reconstruction; in comparison, human-driven Class 3 segmentation relies also on the xz-plane. Machine superiority (row 2) is demonstrated where the machine discovers entirely new damage instances missed by the human, augments existing diffuse segmentations (caused partly by local grayscale noise in tomogram near human-defined thresholds, and also includes the thin film feature from Figure 3 column 4), or correctly extends segmentations to artifact-prone specimen edges. Note that the rate of damage-positive machine

superiority is infeasible to calculate here given the nontrivial extent of human error. Occurring with least frequency, machine inferiority (row 3) is demonstrated where the

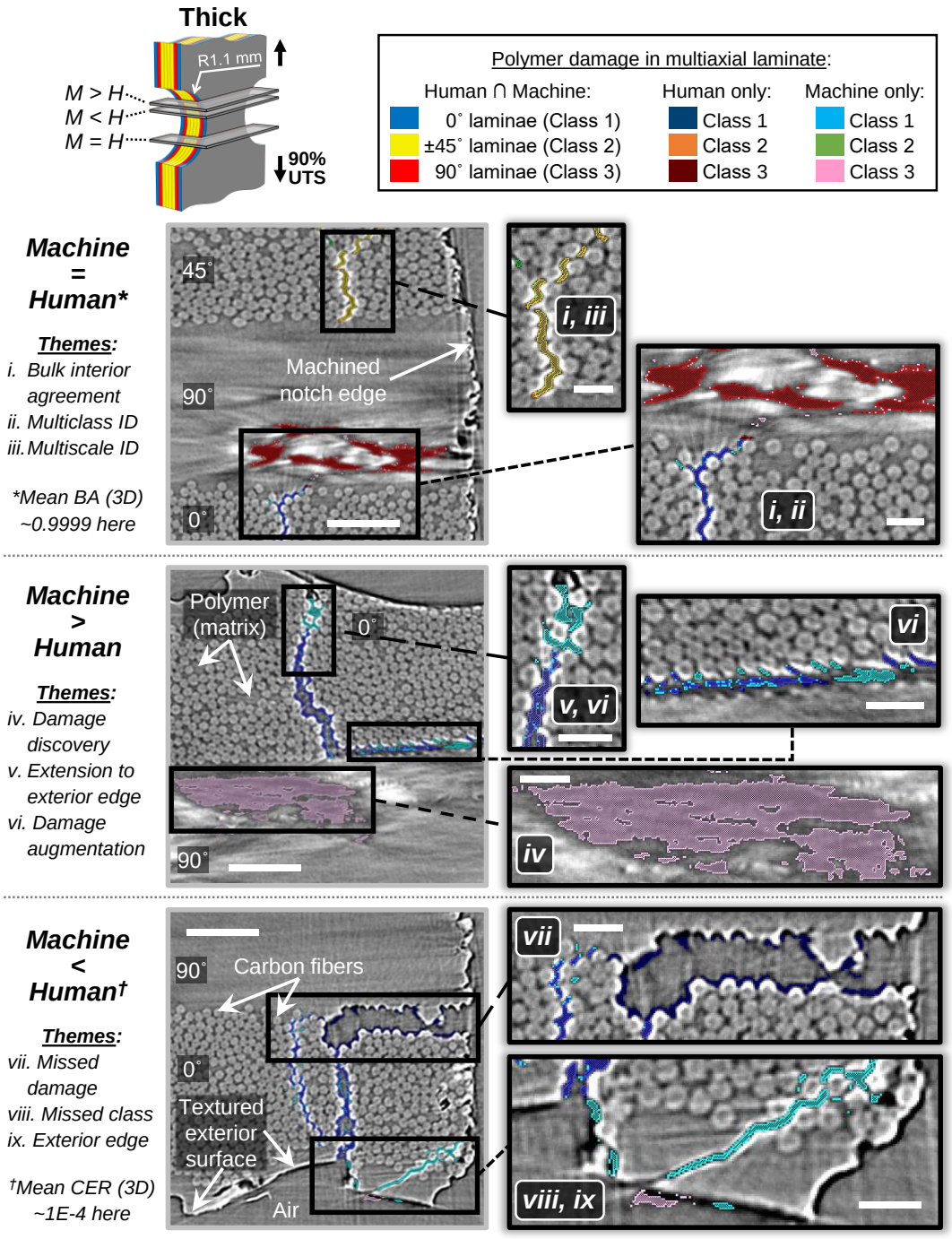

Figure 5. Examples of machine vs. human segmentations where: (top) the machine performs similar to the human, (middle) the machine outperforms the human, and (bottom) the machine underperforms the human. The quality of the selected trained machine (5INL-50ep) segmentations of Thick 90% UTS is evaluated based on a (subjective) scale with 3D locations of cropped tomograms shown in a schematic. The top row characterizes regions (i–iii) of human-repeated performance by the machine (majority case

noted by mean class binary accuracy (BA)) as the bulk of multiscale, diffuse damage is segmented with the correct class, differing only by a thin outer film/layer. The middle row characterizes observed causes (iv–vi) of machine superiority as discovering new damage, augmenting interior human segmentation, and extending near-edge damage closer to the edge. The bottom row characterizes the observed causes (vii–ix) of machine inferiority (minority case noted by mean classification error rate (CER)) as damage misclassification (false positive), absent classification (typically only in complex edge regions, false negative), and exterior edge overrun. Such quality characterizations appear interspersed in actuality throughout each representative row. Scale bar is 50 μm in zoomed-out views, and 20 μm in zoomed-in (inset) views.

model misclassifies or entirely misses damage, or segments regions outside of the specimen interior. Examples of machine inferiority in 2D are typically closely located to, and significantly outnumbered by, larger regions of machine equivalence or even superiority. Note that machine inferiority is the minority case due to the test set macro-average classification error rate (CER, equivalent to one minus the class BA) of 10^{-4}, which includes nontrivial human error and can be considered as a proxy for damage-positive misclassification (false positives) rate in this class-independent prediction framework (wherein multi-label classification predictions are class-thresholded to a single class or background for each pixel).

3. Conclusion

Materials insertion for high-performance, safety-critical structural applications has historically been slowed by uncertainty related to damage initiation and growth, which affects strength and toughness, particularly in heterogeneous advanced fiber composites. Data-rich emergent *in situ* SRCT studies for high-fidelity damage characterization are underpinning mechanistic understanding and allow microstructural-optimization-validated predictive modeling. However, despite the conceptual completeness of 4D damage state visualization via SRCT, objective mechanistic insights are slowed by large quantities of artifact-prone tomograms that are indeterminable by conventional “rule-based” automated segmentation, necessitating reliance on subjective, tedious human-driven (manual or semi-automatic) segmentation techniques. Altogether, these factors constitute a choked big data bottleneck on advancing understanding via CT studies.

We present a foundational investigation of DL capacity to classify sparse, multiclass polymer-related microdamage in advanced composite laminates via *in situ* SRCT, exemplary materials that feature notoriously complex failure behavior of great interest to several research communities, including aerospace, automotive, and renewable energy. Recognizing that DL performance scales with training dataset size and diversity, we robustly study generalizability of novel feature learning across 30 SRCT

scans (~65,000 tomograms, totaling ~3×10^{11} voxels) comprising 6 specimens and 2 different laminate types, necessary for maximizing the speed and objectivity benefits of automation. Following a high-level hyperparametric optimization study involving 20 different machines featuring a fully convolutional neural network in an encoder-decoder architecture, proven in semantic segmentation within other domains, the selected trained machine (5INL-50ep) was found via 2D and 3D quantitative and qualitative analyses to have excellent agreement (~99.99% class binary accuracies on validation and test datasets) with the ground truth via time-intensive, subjective human-driven semi-automatic segmentation methods, while concurrently introducing significant improvements in efficiency and consistency, and in some cases improving upon the trained-human ground truth. The DL/AI segmentation approach accelerates materials knowledge creation by 2 orders of magnitude, *e.g.,* the 65,000 tomograms were segmented by the trained human in ~60 working days, or ~0.23 years, while a single trained machine would take ~2 full days (~0.005 years). This corresponds to 1 GPU as considered here for the trained machine; further acceleration is achievable with multi-GPU machines, or multiple machines working in parallel.

Looking forward, while transfer learning techniques that can accelerate ML were not used here, future machines may be initialized with the presented model to aid learning efficacy, though this likelihood needs to be proven with new human-generated ground truth labels in future work, particularly on noisier, lower-resolution (more challenging) lab-based μCT datasets, which are becoming common in materials engineering, as well as similar high-resolution SRCT datasets which continually enlarge in scale as scan rate and detector size increase at all global beamlines. DL-based segmentation can successfully characterize sparse, extremely complex damage within vast SRCT datasets, establishing such tools as presently unmatched candidates to accelerate understanding of basic structure-property relationships, underpinned by failure mechanisms, across a spectrum of heterogeneous materials (*e.g.,* biological and biomimetic). Future straightforward extensions include microstructural morphological segmentation that elucidates constituent interplay with damage progression and trends for advanced and other types of heterogeneous and composite materials, as well as machines capable of 3D and 4D segmentation to deepen learning via the full-field and temporal (*in situ*) nature of CT data.

4. Methods

In situ synchrotron radiation computed tomography testing of advanced composite laminates: In this study, two different types of aerospace-grade carbon fiber/epoxy advanced composite laminates comprised of unidirectional prepreg laminae are examined: standard-thickness-ply (termed ‘Thick’) and

thin-ply (termed ‘Thin’), as illustrated in Figure S1, Supporting Information. The Thick laminate material system is comprised of Hexcel AS4/8552 (130 μm nominal cured lamina thickness) and employs a quasi-isotropic lamina stacking sequence of [0°/90°/±45°/∓45°/90°/0°] ($[0°/90°/±45°]_S$ in composite shorthand); the Thin laminate material system is comprised of Toho Tenax HTS40/Q-1112 (54 μm nominal lamina thickness) and employs a similar quasi-isotropic lamina stacking sequence of [0°/90°/±45°/0°/90°/±45°/∓45°/90°/0°/∓45°/90°/0°] ($[0°/90°/±45°]_{2S}$ in composite shorthand). Both laminates were cured in an autoclave according to their respective (different) manufacturer cure schedules, which are documented in refs. [76,77], producing a ~1 mm laminate thickness for both types. Based on refs. [80,81], double edge-notched tension (DENT) specimens (length of 70 mm and grip section width of 4 mm, with two 1.1 mm-radius edge notches centered lengthwise) were manufactured with a high-precision waterjet (Omax 2652 Jet Machining Center; 0.01-in tool offset). Aluminum tabs with 1.5-mm thickness were adhered to each specimen end to aid load transfer from the loading rig grips, resulting in 50 mm of specimen length not being tabbed. *Ex situ* DENT testing was performed next, using a Zwick/Roell Z010 uniaxial loading rig equipped with a 10-kN load cell to plan *in situ* DENT ultimate tensile strength (UTS, equivalent to the maximum applied tensile load divided by the grip section cross-sectional area) load steps as a percentage of *ex situ* UTS. Next, using beamline ID19 at the European Synchrotron Radiation Facility (ESRF) in Grenoble, France, *in situ* SRCT testing was performed in displacement control (1 mm/minute, 2 Hz sample rate), where DENT specimens were monotonically loaded in a Deben electromechanical (screw-driven) loading rig (parallel alignment of loading axis, 0° lamina fibers, and tomography stage rotational axis, all of which were orthogonal to the X-ray beam plane) and paused via fixed displacement at various load steps (0%, 60%–95% of the mean *ex situ* UTS for each laminate type) to accommodate SRCT scanning of one of the two notch edge regions (see Figure 1a for idealized specimen loading curve). Each scan took an average of 7 minutes, including time for displacement application, stress relaxation (typically less than 10 MPa nominal stress over ~1 minute) during fixed grip displacement that mitigates blurring artifacts caused by specimen motion, field-of-view (FoV) positioning, and scan acquisition. The ultra-high-resolution SRCT scans and scan reconstructions were performed using the following parameters: 20 keV X-ray energy (monochromatic), 50 ms exposure, 2996 radiographic projections (180° angular range), 0.65 μm isotropic voxel size, and 1.66 mm × 1.66 mm × 1.40 mm FoV. An established ID19 imaging protocol employing a 60-mm propagation (sample to detector) distance was used to allow enhanced visualization of individual fibers and micron-scale crack opening displacements via edge-enhancing propagation-

based phase contrast that appears in tomograms as black/white fringes at interfaces due to differential X-ray refraction[80], facilitated by positioning the detector in the near-field Fresnel region.[44] Following radiograph acquisition, tomographic scan reconstruction was performed using an algorithm based on the inverse radon transform via filtered back projection on either a tomogram-by-tomogram basis[82], as well as a generalized 3D basis for arbitrary slice monitoring during testing[83], with both approaches including ring artifact correction and center shift determination, and then finished by grayscale normalization over all tomograms in a single scan/stack. Following tomographic reconstruction and normalization, 32-bit floating-point grayscale raw volumetric images (2,560 pixels × 2,560 pixels × 2,160 pixels) were generated, which were subsequently adjusted identically in brightness/contrast (histogram adjusted to -60 to +60 gray value range, reflecting local linear X-ray attenuation coefficients for each specimen FoV, for 32-bit real-valued tomograms), downsampled to 8-bit grayscale images without loss of generality, and cropped to remove broad air and SRCT mask regions using Fiji ImageJ. In total, a set of thirty SRCT scans (~65,000 tomograms) encompassing six specimens (four Thick specimens and two Thin specimens) were acquired, as shown in Figure 2a. Additional methodology details for this study are provided elsewhere.[84]

Traditional (trained-human) damage segmentation for X-ray microtomography of advanced composites as ground truth: Since simple gray value thresholding (Figure S2, Supporting Information) and even more sophisticated rule-based programming approaches involving digital image processing tools (*e.g.,* ref. [85]) are unsuitable/inaccurate for automation of the present multiclass 3D segmentation problem, a semi-automatic seeded region growing technique was implemented to create ground truth (*i.e.,* baseline)-labeled damage segmentation results, which collectively comprise the DL model development database. The state-of-the-art trained human-driven segmentation approach of iterative seeding then region growing, which prioritizes accuracy over labor and objectivity, iteratively employs in FEI Avizo Version 9.4 a blend of manual lamina masking (via the brush tool) to limit segmentation propagation to within the laminae and the magic wand tool to segment 3D-connected regions that both contain a manual seed point and possess gray values less than a human-defined threshold, as exemplified in Figure 1b. Consequently, features of interest segmented by region growing remain within masked regions (laminae) and comprise gray values that are identified manually as representing damage (much darker in grayscale intensity than bulk composite, air, or SRCT reconstruction mask). Particularly, associated with slight differences in the X-ray attenuation properties of their fiber/matrix systems, different gray value threshold ranges were selected for Thick and Thin specimens: [0, 40] for 0° or ±45° lamina damage and

[0, 70] for 90° lamina damage in Thick specimens, and [0, 30] for 0° or ±45° lamina damage and [0, 30] for 90° lamina damage in Thin specimens. The larger threshold range selected for 90° lamina damage in Thick specimens was driven by the presence of local reconstruction clarity artifacts (blurring), as the 90° fibers are disadvantageously arranged parallel to the X-ray beam plane. In contrast, no such region growing threshold range expansion was needed to improve 90° lamina damage segmentation in Thin specimens, since they exhibited an observed relative suppression of overall polymer damage extent. In practice, the trained human (baseline) segmentation process initiates by examining a relatively small region of a lamina-masked tomogram and manually setting seed points for region growing in human-identified damage sub-regions, as demonstrated in the baseline segmentation pipeline in Figure 1b. Note that when using this semi-automatic method, a trained human has access to all three mutually orthogonal image planes captured by volumetric SRCT, providing a strong analytical advantage as 0° and ±45° laminae damage are more clearly visualized in the yz-plane and 90° lamina damage is more clearly visualized in the xz-plane (planes normal to lamina/fiber direction are preferred, as cracks are more evident there). For comparison, as discussed previously, the current trained-machine segmentation only accesses the yz-plane, as it is 2D slice-based. Once all human-identified damage regions in the tomogram have been segmented via region growing, manual pixel-level corrections are typically needed and performed using the brush tool to subtract erroneous selections from segmentation label sets or refine masks as needed, particularly in interlaminar (lamina/lamina) regions where different damage class instances interact and near specimen edges and lamina interfaces due to expected and observed masking errors associated with their rough/wavy morphologies, which display dark/bright imaging artifacts caused by inherent interference/fringe effects of propagation-based phase contrast in SRCT, underpinning masking of the notch edge region as especially salient due to its susceptibility to false-positive segmentations. Overall, polymer damage present in interior regions located at least five fibers or ~30 μm away from the rough notch exterior edge were examined, to promote feasible and consistent region-growing results without compromising accuracy. Once a tomogram sub-region (or sub-stack) was segmented relatively accurately (based on trained human assessment), neighboring tomographic regions were processed with the same baseline procedure. In total, iterating with 2D-based inspections over a single entire 3D tomographic image (stack of tomograms), the semi-automatic region growing segmentation method generally required on the order of ~10 h (up to greater than 20 h) of trained human labor that comprises considerable levels of subjectivity. Following segmentation, the quantification module (label analysis) in Avizo was used for class-level analyses.

Deep learning-based (trained-machine) damage segmentation of X-ray microtomography of advanced composites: The DL machine development workflow as executed here, including model and data inputs and outputs, is presented in Figure S3, Supporting Information. The workflow, comprising the six steps of data ingestion, dataset preparation, machine training and validation, prediction threshold training, machine testing, and machine inference, are organized such that depending on the development stage of the machine as well as the segmentation needs, different combinations of these steps may be warranted, which can be further generalized to include DL hierarchies comprised of developing desired broader-scoped machines from the combined output of numerous more tractable narrower-scoped machines. For example, developing a new model (without any transfer learning/pre-training) necessitates performing all six steps of the workflow. In contrast, once a model has been fully developed (trained, validated, and tested), if new, relatively similar SRCT data (in terms of lamina stacking sequence, specimen orientation relative to X-ray beam, damage mechanisms, etc. based on similar or nearly identical experimental setups) needs to be segmented, then only steps 1 (*i.e.*, data ingestion) and 6 (*i.e.*, machine inference) need to be executed. Overall, compared to the trained human, the trained machine begins from a similar starting point (here, yz-plane sub-tomogram; note that the machine cannot learn from xz- or xy-planes) and performs end-to-end multiclass segmentation (no pre- or post-processing) in a much shorter timeframe and with (presumed) greater levels of consistency (*n.b.*, machine segmentation is repeatable). Repeating over the full tomogram stack (complete scan), the machine requires ~1 h of computational time, which is strongly dependent on computational resources. The human time required to operate a trained machine pipeline is negligible. Public access to our code repository is discussed below in Supporting Information.

The DL model development was performed with Python 3.5 using Keras/Google Tensorflow Version 2.1 in combination with the publicly available DL model library Segmentation Models[86] (GitHub public repository). Regarding hardware, virtual machines facilitated by Google Cloud Platform were utilized, with each featuring one NVIDIA Tesla P100 GPU (additional GPUs would be expected to reduce machine time). The primary details and hyperparameters used in the current DL model/network architecture are presented in Tables S1 and S2, Supporting Information. We note that from the full set of semi-automatically labeled 2D tomograms (*n.b.*, region growing masks excluded from labeled learning datasets) spanning thirty 3D SRCT scans (~65,000 tomograms) of both Thick and Thin specimens, class-based crop sampling was employed to more efficiently train the model. Note that no transfer learning was used in this study; models/networks were randomly initialized as discussed in Section S1,

Supporting Information. Class-based cropping involves patchwise sampling of a human-specified number of sub-tomograms (512 pixels × 512 pixels) from each full-sized tomogram (~2,000 pixels × ~2,000 pixels here, following minor initial cropping of air regions) based on the presence (or not) of a given class of polymer damage within the sub-tomogram. Three classes of (polymer) damage were annotated semi-automatically by a trained human, as previously discussed, and used for model training: Class 1 comprises polymer damage located in 0° laminae, Class 2 comprises polymer damage located in ±45° laminae, and Class 3 comprises polymer damage located in 90° laminae. Additionally, we note that to reinforce more general learning of damage features, class-sampled crops/sub-tomograms were subsequently rotated randomly by integer multiples of 90°, a form of a technique known as data augmentation. Finally, as discussed previously, we note that while each pixel was treated here in a multi-label context, the final inference workflow step enforced selection of only the single highest (relative to each respective class prediction threshold) probability class, or no class at all since the background was not included explicitly as a learned class here. Further details regarding multi-label classification and background class exclusion are presented in Section S1, Supporting Information.

Quantitative segmentation analysis in two and three dimensions: Three-dimensional object analysis of each segmentation class was conducted via 3D ImageJ Suite[87], a Fiji ImageJ plug-in.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgments

This work was supported by Airbus, ANSYS, Embraer, Lockheed Martin, Saab AB, and Teijin Carbon America through MIT's Nano-Engineered Composite aerospace STructures (NECST) Consortium, as well as partially supported by the NASA Space Technology Research Institute (STRI) for Ultra-Strong Composites by Computational Design (US-COMP), grant NNX17AJ32G. We are also grateful for the technical support from the MIT Quest for Intelligence that made this work possible. This study was supported by the Google Cloud Research Credits program with the award GCP823539415517. This work made use of facilities supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office through the Institute for Soldier Nanotechnologies, under contract number W911NF-13-D-0001, the facilities at the U.S. Army Natick Combat Capabilities Development Command - Soldier Center (CCDC-SC), and was carried out in part through the use of MIT's Microsystems Technology Laboratories (MTL). The SRCT experiments were performed on beamline

ID19 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. We are grateful to Lukas Helfen and Elodie Boller at the ESRF for providing assistance in using beamline ID19, as well as Jeonyoon Lee and Estelle Kalfon-Cohen (MIT necstlab), Albertino Arteiro (University of Porto, Portugal), and Gregor Borstnar and Mark N. Mavrogordato (University of Southampton, United Kingdom) for supporting the SRCT experiments. For the first author, this material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All authors thank the entire necstlab at MIT and faculty and lab members at μ-VIS X-ray Imaging Center at the University of Southampton for valuable discussion and input. The authors thank Saab AB for the donation of 'Thick' prepreg materials and Teijin Carbon America for the donation of 'Thin' prepreg materials.

Conflict of Interest

The authors declare no conflict of interests.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author (or Reed Kopp, rkopp@alum.mit.edu) upon reasonable request. The semantic segmentation code repository can be open-source accessed at https://github.com/mit-quest/necstlab-damage-segmentation.

Author Contributions

R.K.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Visualization, Data Curation, Writing – original draft. J.J. Conceptualization, Methodology, Software, Investigation, Resources, Writing – review & editing, Supervision. X.N.: Investigation, Resources. N.R.: Supervision, Funding acquisition, Writing – review & editing. B.L.W.: Resources, Supervision, Funding acquisition, Writing – review & editing.

References

[1] I. Levchenko, K. Bazaka, T. Belmonte, M. Keidar, S. Xu, *Advanced Materials* **2018**, *30*, 1802201.

[2] T. Ghidini, *Nature Materials* **2018**, *17*, 846.

[3] A. Taub, E. De Moor, A. Luo, D. K. Matlock, J. G. Speer, U. Vaidya, *Annual Review of Materials*

Research **2019**, *49*, 327.

[4] A. K. Naskar, J. K. Keum, R. G. Boeman, *Nature Nanotechnology* **2016**, *11*, 1026.

[5] N. P. Padture, *Nature Materials* **2016**, *15*, 804.

[6] X. Wu, Y. Zhu, *Materials Research Letters* **2017**, *5*, 527.

[7] *Materials Research to Meet 21st Century Defense Needs*, National Academies Press, Washington, D.C., **2003**.

[8] M. Ashby, H. Shercliff, D. Cebon, *Materials: Engineering, Science, Processing and Design*, **2007**.

[9] A. Bourmaud, J. Beaugrand, D. U. Shah, V. Placet, C. Baley, *Progress in Materials Science* **2018**, *97*, 347.

[10] V. Dhand, G. Mittal, K. Y. Rhee, S.-J. Park, D. Hui, *Composites Part B: Engineering* **2015**, *73*, 166.

[11] H. G. Chae, S. Kumar, *Science* **2008**, *319*, 908.

[12] Y. Bai, R. Zhang, X. Ye, Z. Zhu, H. Xie, B. Shen, D. Cai, B. Liu, C. Zhang, Z. Jia, S. Zhang, X. Li, F. Wei, *Nature Nanotechnology* **2018**, *13*, 589.

[13] X. Liao, M. Dulle, J. M. de Souza e Silva, R. B. Wehrspohn, S. Agarwal, S. Förster, H. Hou, P. Smith, A. Greiner, *Science* **2019**, *366*, 1376.

[14] I. A. Kinloch, J. Suhr, J. Lou, R. J. Young, P. M. Ajayan, *Science* **2018**, *362*, 547.

[15] S. Torquato, *Annual Review of Materials Research* **2010**, *40*, 101.

[16] M. Bechthold, J. C. Weaver, *Nature Reviews Materials* **2017**, *2*, 17082.

[17] F. Barthelat, Z. Yin, M. J. Buehler, *Nature Reviews Materials* **2016**, *1*, 16007.

[18] J. Karger-Kocsis, H. Mahmood, A. Pegoretti, *Progress in Materials Science* **2015**, *73*, 1.

[19] R. O. Ritchie, *Nature Materials* **2011**, *10*, 817.

[20] X. Zhang, N. Zhao, C. He, *Progress in Materials Science* **2020**, *113*, 100672.

[21] L. Hu, M. O'Neil, V. Erturun, R. Benitez, G. Proust, I. Karaman, M. Radovic, *Scientific Reports* **2016**, *6*, 35523.

[22] A. Mirabedini, A. Ang, M. Nikzad, B. Fox, K. Lau, N. Hameed, *Advanced Science* **2020**, *7*, 1903501.

[23] C. González, J. J. Vilatela, J. M. Molina-Aldareguía, C. S. Lopes, J. LLorca, *Progress in Materials Science* **2017**, *89*, 194.

[24] M. A. Skylar-Scott, J. Mueller, C. W. Visser, J. A. Lewis, *Nature* **2019**, *575*, 330.

[25] S. Bose, D. Ke, H. Sahasrabudhe, A. Bandyopadhyay, *Progress in Materials Science* **2018**, *93*, 45.

[26] J. Frketic, T. Dickens, S. Ramakrishnan, *Additive Manufacturing* **2017**, *14*, 69.

[27] C. Sanchez, H. Arribart, M. M. Giraud Guille, *Nature Materials* **2005**, *4*, 277.

[28] Y. Yang, X. Song, X. Li, Z. Chen, C. Zhou, Q. Zhou, Y. Chen, *Advanced Materials* **2018**, *30*, 1706539.

[29] U. G. K. Wegst, H. Bai, E. Saiz, A. P. Tomsia, R. O. Ritchie, *Nature Materials* **2015**, *14*, 23.

[30] J. V. Anguita, C. T. G. Smith, T. Stute, M. Funke, M. Delkowski, S. R. P. Silva, *Nature Materials* **2020**, *19*, 317.

[31] V. P. Veedu, A. Cao, X. Li, K. Ma, C. Soldano, S. Kar, P. M. Ajayan, M. N. Ghasemi-Nejhad, *Nature Materials* **2006**, *5*, 457.

[32] D. D. L. Chung, *Materials Science and Engineering: R: Reports* **2017**, *113*, 1.

[33] T. L. Burnett, P. J. Withers, *Nature Materials* **2019**, *18*, 1041.

[34] E. Maine, P. Seegopaul, *Nature Materials* **2016**, *15*, 487.

[35] B. Cox, Q. Yang, *Science* **2006**, *314*, 1102.

[36] J. LLorca, C. González, J. M. Molina-Aldareguía, J. Segurado, R. Seltzer, F. Sket, M. Rodríguez, S. Sádaba, R. Muñoz, L. P. Canal, *Advanced Materials* **2011**, *23*, 5130.

[37] J. Cunningham, *Engineering Materials* **2015**, 32.

[38] R. R. Boyer, J. D. Cotton, M. Mohaghegh, R. E. Schafrik, *MRS Bulletin* **2015**, *40*, 1055.

[39] *Accelerating Technology Transition: Bridging the Valley of Death for Materials and Processes in Defense Systems*, The National Academies Press, Washington, DC, **2004**.

[40] *Nature Materials* **2016**, *15*, 803.

[41] S. Z. H. Shah, S. Karuppanan, P. S. M. Megat-Yusoff, Z. Sajid, *Composite Structures* **2019**, *217*, 100.

[42] M. E. Ibrahim, *Composites Part A: Applied Science and Manufacturing* **2014**, *64*, 36.

[43] S. C. Wu, T. Q. Xiao, P. J. Withers, *Engineering Fracture Mechanics* **2017**, *182*, 127.

[44] S. C. Garcea, Y. Wang, P. J. Withers, *Composites Science and Technology* **2018**, *156*, 305.

[45] E. Maire, P. J. Withers, *International Materials Reviews* **2014**, *59*, 1.

[46] F. García-Moreno, P. H. Kamm, T. R. Neu, F. Bülk, R. Mokso, C. M. Schlepütz, M. Stampanoni, J. Banhart, *Nature Communications* **2019**, *10*, 3762.

[47] B. M. Patterson, N. L. Cordes, K. Henderson, X. Xiao, N. Chawla, in *Materials Discovery and Design* (Eds.: T. Lookman, S. Eidenbenz, F. Alexander, C. Barnes), Springer International Publishing, Cham, **2018**, pp. 129–165.

[48] H. Proudhon, M. Pelerin, A. King, W. Ludwig, *Current Opinion in Solid State and Materials Science* **2020**, 100834.

[49] G. R. Davis, J. C. Elliott, *Materials Science and Technology* **2006**, *22*, 1011.

[50] P. Iassonov, T. Gebrenegus, M. Tuller, *Water Resources Research* **2009**, *45*, DOI 10.1029/2009WR008087.

[51] D. Ibrahim, *Procedia Computer Science* **2016**, *102*, 34.

[52] S. Rosini, M. N. Mavrogordato, O. Egorova, E. S. Matthews, S. E. Jackson, S. Mark Spearing, I. Sinclair, *Composites Part A: Applied Science and Manufacturing* **2019**, *125*, 105543.

[53] F. Sket, A. Enfedaque, C. Alton, C. González, J. M. Molina-Aldareguia, J. Llorca, *Composites Science and Technology* **2014**, *90*, 129.

[54] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, **2016**.

[55] T. J. Sejnowski, *Proceedings of the National Academy of Sciences* **2020**, 201907373.

[56] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, V. Kumar, *IEEE Transactions on Knowledge and Data Engineering* **2017**, *29*, 2318.

[57] W. Nash, T. Drummond, N. Birbilis, *npj Materials Degradation* **2018**, *2*, 37.

[58] C. Suh, C. Fare, J. A. Warren, E. O. Pyzer-Knapp, *Annual Review of Materials Research* **2020**, *50*, 1.

[59] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.

[60] T. D. Sparks, S. K. Kauwe, M. E. Parry, A. M. Tehrani, J. Brgoch, *Annual Review of Materials Research* **2020**, *50*, 27.

[61] M. Umehara, H. S. Stein, D. Guevarra, P. F. Newhouse, D. A. Boyd, J. M. Gregoire, *npj Computational Materials* **2019**, *5*, 34.

[62] K. Guo, Z. Yang, C.-H. Yu, M. J. Buehler, *Materials Horizons* **2021**, DOI 10.1039/D0MH01451F.

[63] J. Long, E. Shelhamer, T. Darrell, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, **2015**, pp. 3431–3440.

[64] O. Ronneberger, P. Fischer, T. Brox, in *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015* (Eds.: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi), Springer International Publishing, Cham, **2015**, pp. 234–241.

[65] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, **2017**, pp. 1175–1183.

[66] B. Provencher, N. Piché, M. Marsh, *Microscopy and Microanalysis* **2019**, *25*, 402.

[67] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, A. Dovzhenko, O. Tietz, C. Dal Bosco, S. Walsh, D. Saltukoglu, T. L. Tay, M. Prinz, K. Palme, M. Simons, I. Diester, T. Brox, O. Ronneberger, *Nature Methods* **2019**, *16*, 67.

[68] M. G. Haberl, C. Churas, L. Tindall, D. Boassa, S. Phan, E. A. Bushong, M. Madany, R. Akay, T. J. Deerinck, S. T. Peltier, M. H. Ellisman, *Nature Methods* **2018**, *15*, 677.

[69] I. Arganda-Carreras, V. Kaynig, C. Rueden, K. W. Eliceiri, J. Schindelin, A. Cardona, H. Sebastian Seung, *Bioinformatics* **2017**, *33*, 2424.

[70] H. Tang, X. Chen, Y. Liu, Z. Lu, J. You, M. Yang, S. Yao, G. Zhao, Y. Xu, T. Chen, Y. Liu, X. Xie, *Nature Machine Intelligence* **2019**, *1*, 480.

[71] B. J. Antony, B.-J. Kim, A. Lang, A. Carass, J. L. Prince, D. J. Zack, *PLOS ONE* **2017**, *12*, e0181059.

[72] Y. Dong, C. Su, P. Qiao, L. Sun, *Construction and Building Materials* **2020**, *253*, 119185.

[73] D. Sammons, W. P. Winfree, E. Burke, S. Ji, in *AIP Conference Proceedings*, **2016**, p. 110014.

[74] A. Badran, D. Marshall, Z. Legault, R. Makovetsky, B. Provencher, N. Piché, M. Marsh, *Journal of Materials Science* **2020**, 1.

[75] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, G. Hamarneh, *Artificial Intelligence Review* **2020**, 1.

[76] R. Kopp, X. Ni, E. Kalfon-Cohen, C. Furtado, A. Arteiro, G. Borstnar, M. Mavrogordato, L. Helfen, I. Sinclair, S. M. Spearing, P. Camanho, B. L. Wardle, in *AIAA Scitech 2019 Forum*, American Institute Of Aeronautics And Astronautics, Reston, Virginia, **2019**.

[77] X. Ni, R. Kopp, E. Kalfon-Cohen, C. Furtado, J. Lee, A. Arteiro, G. Borstnar, M. N. Mavrogordato, L. Helfen, I. Sinclair, S. M. Spearing, P. P. Camanho, B. L. Wardle, *Composites Part B: Engineering* **2021**, 108623.

[78] A. Arteiro, G. Catalanotti, J. Reinoso, P. Linde, P. P. Camanho, *Archives of Computational Methods in Engineering* **2019**, *26*, 1445.

[79] K. Simonyan, A. Zisserman, in *ICLR 2015*, **2014**.

[80] P. Wright, A. Moffat, I. Sinclair, S. M. Spearing, *Composites Science and Technology* **2010**, *70*, 1444.

[81] A. J. Moffat, P. Wright, J. Y. Buffière, I. Sinclair, S. M. Spearing, *Scripta Materialia* **2008**, *59*,

1043.

[82] A. Mirone, E. Brun, E. Gouillart, P. Tafforeau, J. Kieffer, *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* **2014**, *324*, 41.

[83] M. Vogelgesang, T. Farago, T. F. Morgeneyer, L. Helfen, T. dos Santos Rolo, A. Myagotin, T. Baumbach, *Journal of Synchrotron Radiation* **2016**, *23*, 1254.

[84] R. Kopp, X-Ray Micro-Computed Tomography and Deep Learning Segmentation of Progressive Damage in Hierarchical Nanoengineered Carbon Fiber Composites, Massachusetts Institute of Technology, **2021**.

[85] N. K. Fritz, R. Kopp, A. K. Nason, X. Ni, J. Lee, I. Y. Stein, E. Kalfon-Cohen, I. Sinclair, S. M. Spearing, P. P. Camanho, B. L. Wardle, *Composites Science and Technology* **2020**, *193*, 108132.

[86] P. Yakubovskiy, *GitHub repository* **2019**.

[87] J. Ollion, J. Cochennec, F. Loll, C. Escudé, T. Boudier, *Bioinformatics* **2013**, *29*, 1840.

Table of Contents

A deep learning machine is created to segment sparse, multiclass microdamage in advanced composite laminates that feature complex microstructures and failure behavior, with ~99.99% binary accuracy using 65,000 tomograms in 3D X-ray datasets. The deep learning approach accelerates heterogeneous materials knowledge creation by 2 orders of magnitude in a generalizable way, breaking a longstanding bottleneck while additionally introducing objectivity.

Reed Kopp, Joshua Joseph, Xinchen Ni, Nicholas Roy, and Brian L. Wardle*

Deep Learning Unlocks X-ray Microtomography Segmentation of Multiclass Microdamage in Heterogeneous Materials

Table of Contents figure submitted as a separate file.