PAPER • OPEN ACCESS

Trustworthy AI-based crack-tip segmentation using domain-guided explanations

IOP Publishing Mach. Learn.: Sci. Technol. 7 (2026) 015015 https://doi.org/10.1088/2632-2153/ae3660

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
26 September 2025

REVISED
7 January 2026

ACCEPTED FOR PUBLICATION
9 January 2026

PUBLISHED
21 January 2026

PAPER

Trustworthy AI-based crack-tip segmentation using domain-guided explanations

Jesco Talies, Eric Breitbarth and David Melching*

Institute for Frontier Materials on Earth and in Space, German Aerospace Center (DLR), Linder Hoehe, Cologne 51147, Germany
* Author to whom any correspondence should be addressed.

E-mail: David.Melching@dlr.de

Keywords: deep learning, attention-guided training, explainable AI, quantitative evaluation metrics, domain knowledge, fracture mechanics, digital image correlation

Abstract

Ensuring the trustworthiness and robustness of deep learning models remains a fundamental challenge, particularly in high-stakes scientific applications. In this study, we present a framework called *attention-guided training* that combines explainable artificial intelligence techniques with quantitative evaluation and domain-specific priors to guide model attention. We demonstrate that domain-specific feedback on model explanations during training can enhance the model's generalization capabilities. We validate our approach on the task of semantic crack tip segmentation in digital image correlation data, which is a key application in the fracture mechanical characterization of materials. By aligning model attention with physically meaningful stress fields, such as those described by Williams' analytical solution, attention-guided training ensures that the model focuses on physically relevant regions. This finally leads to improved generalization and more faithful explanations.

1. Introduction

Deep learning (DL) has led to enormous breakthroughs in many scientific fields, from computer vision [1] to biology [2], material science [3–5], computational mechanics [6],failure modeling [7], and fracture mechanics [8] because of its ability to identify patterns in complex, high-dimensional data. DL is based on the training of a highly flexible deep neural network, consisting of millions of trainable parameters, with large amounts of data. While typically improving performance, deep neural networks are trained end-to-end and lack interpretability and explainability. This black-box problem raises critical questions regarding reliability and trustworthiness, in particular in high-risk and high-stakes applications such as autonomous vehicles and robots, healthcare, or maintenance of aircraft systems and components. These considerations have been recently recognized as a legal issue in guidelines issued by the EU AI Act [9] and more specifically for aerospace applications by EASA [10] and NASA [11].

Explainable artificial intelligence (XAI) [12] addresses these issues by proposing processes and methods that provide insights into the decision-making processes of DL models. There are two approaches to XAI: on the one hand, one can aim to achieve intrinsic interpretability by designing inherently transparent models that are human-understandable [13]. On the other hand, post-hoc explainability seeks to clarify model predictions without modifying the internal model structure, for example, through gradient-based sensitivity analysis [14] or model-agnostic feature attribution methods [15, 16]. While models designed for intrinsic interpretability offer clarity, they often lack the complexity needed to capture intricate data patterns, which can result in suboptimal performance. This trade-off between interpretability and accuracy is particularly evident in complex tasks tackled with DL, where simpler models do not suffice.

For **convolutional neural networks** (CNNs), a class of DL models used mainly for spatial, grid-like data, post-hoc explanations in the form of attention heatmaps based on class activation mappings (CAM) [17] have gained wide popularity. While originally only designed for classification tasks, CAM-based methods have recently been extended to semantic segmentation architectures [18] and applied for crack tip segmentation models [19]. Despite their success, post-hoc explainability also faces criticism for producing explanations that can be misleading or unfaithful to the original black-box model or may oversimplify complex relationships [20]. Moreover, even when focusing solely on CAM-based approaches, there is a vast array of methods,including gradient-based variants such as Grad-CAM and its extensions [21–24], as well as perturbation- and decomposition-based approaches [25–27], making it challenging to determine which method is best suited for a given model and task.

To address these concerns, researchers have called for rigorous, standardized metrics to quantitatively evaluate XAI methods. Vilone & Longo [28] propose a hierarchical classification of XAI methods and conceptualize their evaluation, also pointing out that research currently lacks a consensus on how to assess explainability. Nauta *et al* [29] propose conceptual properties such as correctness (i.e. faithfulness), completeness, and compactness to measure, among others, how well explanations align with the true behavior of a model and how effectively they convey relevant information to users.

Moreover, recent work has explored novel approaches that incorporate explanations into the training process. Some methods integrate human-based feedback on explanations, for example, through corrective supervision [30], interactive rationale selection [31], or explanation-informed learning objectives [32]. These approaches demonstrate that interacting with and reflecting upon model explanations can prevent bias, improve accuracy, and reduce required training samples. However, relying on human interaction is tedious and time-consuming, and particularly for specific tasks, it requires experts with domain knowledge. Recently, Stammer *et al* [33] instead sourced explanatory feedback from a secondary critic model. Although their results show improved model generalization and provide more faithful explanations, it does not allow for the explicit inclusion of known concepts and domain knowledge.

Addressing these challenges, this paper introduces a framework that integrates faithful explanations and domain knowledge directly into the model training process. We call this framework **attention-guided training** (AGT). The framework is shown in figure 1 and enables the determination of trustworthy explanations and provides alignment with domain knowledge during training. The framework starts with the identification of relevant expert knowledge for a specific task, which should be tackled by DL. Then, suitable explainability methods are identified and evaluated using objective metrics. Lastly, the core of AGT is a novel training process where the task-specific classical loss function is combined with an additional loss, which assesses the alignment of the explanation with domain-specific expectations.

While AGT serves as a general framework, its development was inspired by the findings of Melching *et al* [19]; hence, we focus on its specific application in the field of fracture mechanics. In [19], the authors employ the explainability method Grad-CAM [21] for the semantic segmentation of fatigue crack tips in digital image correlation (DIC) data. They train models with different architectures, including a U-Net [34] and the so-called ParallelNets approach [19]. The resulting model explanations exhibit distinct semantic characteristics, as illustrated in figure 2. While the U-Net primarily highlights the crack path, the ParallelNets explanations align more closely with the physical crack tip field described by Williams [35]. Despite the improved generalization capabilities of the latter approach, a framework that explicitly controls network attention during training remains absent.

In this work, we present an AGT of a U-Net model for crack tip segmentation in DIC data. First, we adapt existing CAM-based XAI methods to semantic segmentation models. Next, we systematically evaluate their suitability for the given task by assessing correctness, completeness, continuity, and compactness, as proposed by Nauta *et al* [29], and select the most appropriate XAI method accordingly. Finally, as AGT leverages domain knowledge to guide the training process and align model explanations with a predefined target, we calculate the representative von Mises equivalent stress field for the present load case—depicted exemplarily in figure 2(c) – as the desired attention target. This choice is motivated by the findings of [19] and discussions with experts in the field of fracture mechanics.

The paper is structured as follows: In section 2.1, we adapt CAM-based XAI methods to the application of crack tip segmentation in DIC data and present their quantitative evaluation. Building on these evaluations, we apply AGT using physical target explanations motivated by domain knowledge in the form of the crack tip field. We then compare the resulting model performances with models trained using AGT on non-physical target explanations and a reference batch trained conventionally (without AGT) in section 2.2. In section 3, we discuss the results together with possible applications and limitations of the framework. Details regarding the methodology of the fracture mechanical experiments, the machine learning and XAI approach, and details on the AGT framework and application-specific implementations can be found in section 4.

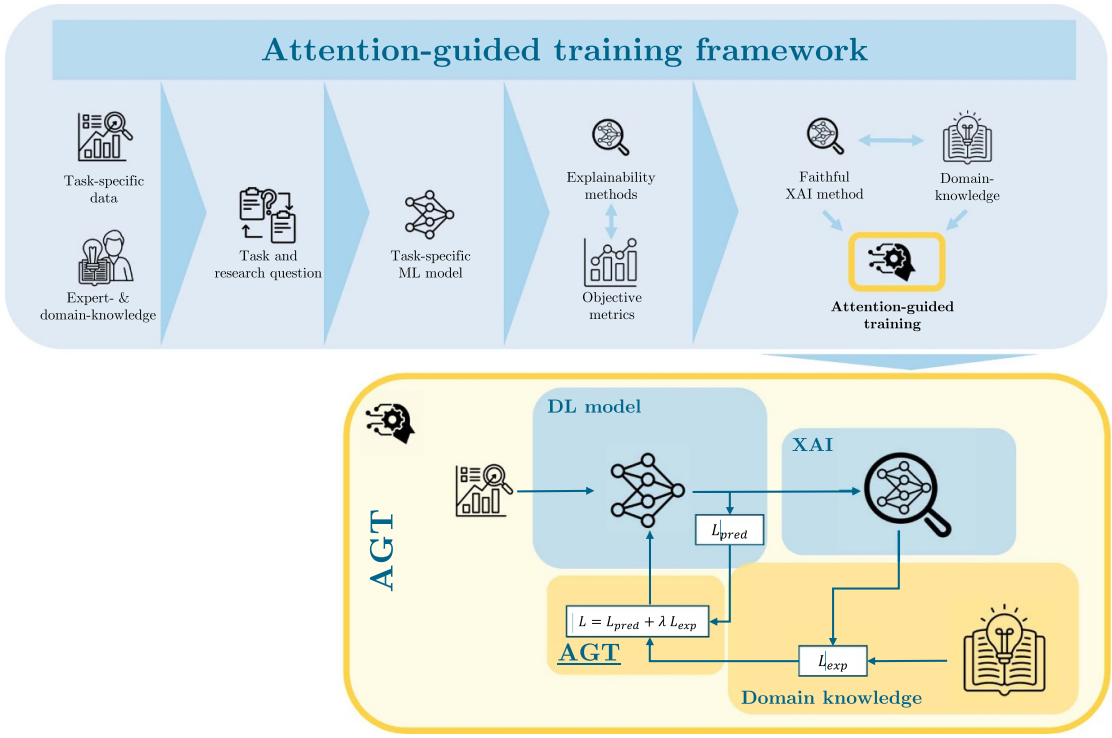

Figure 1. Overview of attention-guided training (AGT); AGT builds upon conventional machine learning principles, providing enhancements through explainable artificial intelligence (XAI) techniques. It incorporates post-hoc explanations and objective XAI evaluation metrics to assess model explainability and integrates domain knowledge through a modified training loop, ensuring better alignment between trained patterns and expert insights.

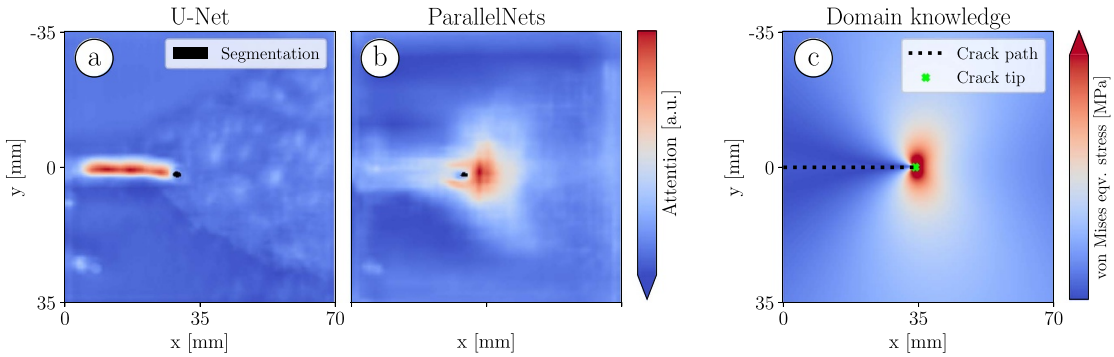

Figure 2. Semantic differences in model explanations, calculated using Grad-CAM, despite similar predictions. (a) The explanation of the U-Net model prediction shows that the model identifies the crack path to predict the crack tip location. It is known that experimental data in this region likely contains artifacts from the digital image correlation evaluation (see section 4.1 for the experimental context). (b) In contrast, the explanation of the ParallelNets [19] model prediction highlights the region ahead of the crack tip as most relevant for its prediction This attention aligns closely with domain knowledge shown in (c), where the crack tip stress field is accurately represented by the widely accepted Williams series (equation (3)), thus relating model behavior directly to theoretical foundations in fracture mechanics.

2. Results

We present the key findings of our study for the task of semantic segmentation of crack tips in DIC data.

Accurately quantifying fatigue crack growth (FCG) is essential for assessing the service life and damage tolerance of critical engineering structures and components exposed to variable service loads [36]. In recent years,DIC has played a crucial role in capturing full-field surface displacements and strains during fatigue crack propagation experiments. The fracture mechanical evaluation of DIC data requires an exact and robust (i.e. reliable) determination of the crack tip positions [37]—an extremely challenging task due to inherent noise and artifacts [38]. Strohmann *et al* [8] created a labeled dataset and trained

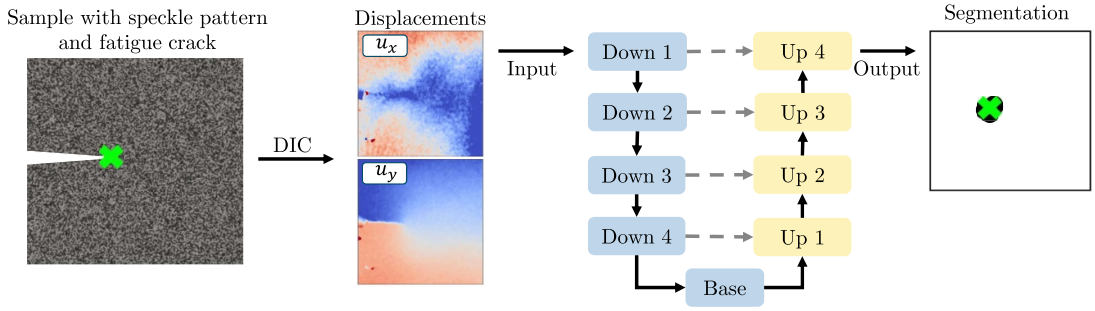

Figure 3. Illustration of the exemplary DL task. We obtain data through full-field DIC analysis during fatigue crack growth experiments (left). The resulting displacement fields are input into a segmentation CNN (here, U-Net architecture), which is trained to segment the most likely crack tip pixels. The centroid of the predicted segmentation yields the crack tip position, which is then translated back into the reference frame of the sample surface.

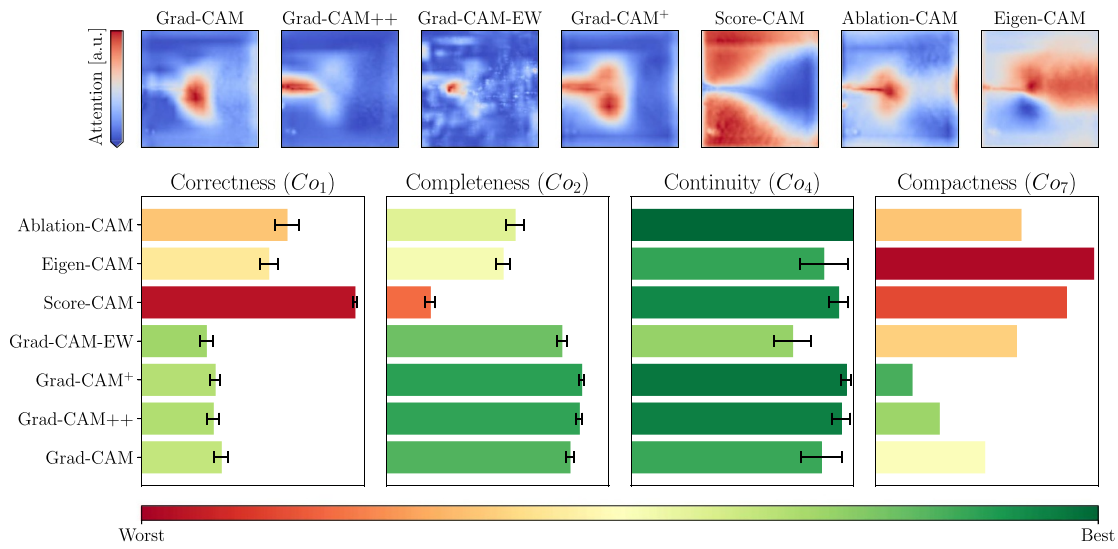

Figure 4. Visualization of explanations and performance comparison: The top row displays attention heatmaps generated using various CAM techniques, each corresponding to the same prediction made by *ParallelNets* on a sample of the $S_{160,4.7,val}$ dataset (see section 4.1). These heatmaps are based on the encoder layers of the U-Net architecture. Given the significant visual differences between methods, the accompanying bar plot quantifies their relative performance based on the four criteria—correctness, completeness, continuity, and compactness.

a U-Net [34] for crack tip segmentation. Melching *et al* [19] refined this model and employed Grad-CAM [21] to generate attention heatmaps, providing explanations that guided the selection of models aligning with domain knowledge. An overview of the deep learning task is given in figure 3.

2.1. Evaluation of explainability methods

In order to answer the question of which explainability method to choose, we fix the model choice to a single trained instance, namely the open-source *ParallelNets* from Melching *et al* [19], truncated to the U-Net architecture for inference. We generate explanations for various methods based on CAM. We refer to appendix A below for details on these methods and their adaptation to semantic segmentation. The resulting explanations in the form of attention heatmaps for the different methods are shown in figure 4(top) and show significant differences, both regarding semantic concepts and intuitive quality. While minor differences are expected and can be observed in explanations generated for classification CNNs as well, see, e.g. [25, 39], the adaptation to semantic segmentation appears to have a significant impact. Specifically, the explanations differ considerably across methods, in contrast to the more consistent patterns observed for classification in [40]. We further observed that the choice of considered neural block(s) and score (see figure 1 and appendix A, respectively) impact the resulting attention heatmaps even within the context of a single CAM method. This is related to the different fidelity and nature of features learned by each model block. Furthermore, we hypothesize that the skip connections in the U-Net architecture, along with the associated feature propagation, further contribute to a more complex and less localized feature representation [41]. As a result, depending on the explanation method,

this complexity may not be adequately captured in the final explanations. Since each of the considered methods aims to explain the predictions of the underlying model, our objective is to quantitatively assess which method is best suited for this task. To achieve this, we employ a subset of the *Co*-12 criteria proposed by Nauta *et al* [29]. We objectively evaluate which of the provided methods is the most faithful, referred to, in the spirit of *Co*-12, as correctness (Co_1), complete (Co_2), continuous (Co_4), in the sense that small changes in the input data lead to small changes in the explanations, and compact (Co_7), referring to the size of the explanation. Using data obfuscation strategies, we implemented metrics for each of these four criteria and refer to section 4.4 for details. Figure 4(bottom) presents a concise, comprehensive comparison of the method performance among these criteria.

For the chosen *ParallelNets* model and the task of crack tip segmentation, the above metrics conclusively indicate that the gradient-based methods (Grad-CAM [21], Grad-CAM++ [22], Grad-CAM$^+$ [23], and elementwise-GradCAM [24]) are significantly more *correct, complete,* and *compact*, but are similarly *continuous*, compared to the gradient-free approaches (Score-CAM [25], Eigen-CAM [26], and Ablation-CAM [27]). While, e.g. Score-CAM appears unsuitable for this model and task, particularly due to its poor performance in terms of correctness and completeness, Grad-CAM+ and Grad-CAM++ demonstrate strong overall evaluation results. Ultimately the results found in figure 4 indicate that it is entirely possible to explain segmentation models using the CAM techniques originally intended for classification CNNs.

2.2. AGT

Considering the vast possibilities of model attention patterns, as presented in figure 2, we argue that it is beneficial to guide these patterns by aligning them with domain-specific, theory-guided target explanations. The alignment is carried out by training a model with a total loss function L_{total}, combining a prediction loss L_{pred} and an explanation loss L_{exp} between the current and target explanation. We provide the target explanation $\hat{\mathbf{\Phi}}$ using the relevant domain knowledge. The total loss is expressed as:

$$L_{\text{total}} = L_{\text{pred}}\left(\mathbf{y}, \hat{\mathbf{y}}\right) + \lambda L_{\text{exp}}\left(\mathbf{\Phi}, \hat{\mathbf{\Phi}}\right), \tag{1}$$

where $\lambda \geqslant 0$ is a hyperparameter balancing both loss contributions. Here, $\mathbf{y}$ and $\hat{\mathbf{y}}$ denote the actual and target predictions, respectively, while $\mathbf{\Phi}$ and $\hat{\mathbf{\Phi}}$ denote the current and domain-guided target explanations, respectively. Thus, the explanation loss term ensures coherence (Co_{11}) of explanations with expert knowledge (see section 4.4).

For the case of crack tip segmentation, this total loss consists of the Dice loss [42] between current and target predictions and the cosine similarity (S_{C}) between current and target explanations. The loss then reads:

$$L_{\text{total}} = \text{Dice}\left(\mathbf{y}, \hat{\mathbf{y}}\right) + \lambda S_{\text{C}}\left(\mathbf{\Phi}, \hat{\mathbf{\Phi}}\right), \tag{2}$$

where

$$\text{Dice}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = 1 - \frac{2\sum_{ij} y_{ij}\hat{y}_{ij}}{\sum_{ij}\left(y_{ij} + \hat{y}_{ij}\right)},$$

and

$$S_{\text{C}}\left(\mathbf{\Phi}, \hat{\mathbf{\Phi}}\right) = 1 - \frac{\sum_{ij}\phi_{ij}\hat{\phi}_{ij}}{\left(\sum_{ij}\phi_{ij}^2\right)^{1/2}\left(\sum_{ij}\hat{\phi}_{ij}^2\right)^{1/2}}.$$

We further evaluate the model's generalization capabilities in terms of reliability. Reliability scores serve as a quantitative measure of generalization performance and are defined as the fraction of samples with valid segmentations (i.e. exactly one contiguous patch of segmented pixels; section 4.2) for any given dataset.

2.2.1. Training phases

To ensure successful training, we divide the process into two phases. The first phase, an initial pretraining,follows a conventional DL approach using only the prediction loss (i.e. $\lambda = 0$), which serves to prime the explanatory component of AGT and allows the model to produce meaningful explanations. In the second phase, a finite $\lambda > 0$ is introduced to refine the learned behavior towards the theory-guided target explanations $\hat{\mathbf{\Phi}}$. To preserve the model's predictive performance throughout the alignment stage,

the hyperparameter λ must remain sufficiently small, preventing over-steering, which would result in non-salient explanations. For this particular use case, the useful range $\lambda \in (0.5, 3)$ has been empirically estimated.

2.2.2. Domain knowledge

For crack tip segmentation, our domain knowledge is based on analytical expressions describing the displacement and stress fields near the crack tip of an open crack. More precisely, we use the analytical derivation by Williams [35]:

In planar linear-elastic fracture mechanics, the stress and displacement fields induced by a single open crack with traction-free crack faces can be described in polar coordinates (r, θ) by the Williams series expansion [43]

$$\sigma_{ij}(r,\theta) = \sum_n r^{\frac{n}{2}-1} \left(A_n f_{\mathrm{I},ij}(\theta, n) + B_n f_{\mathrm{II},ij}(\theta, n)\right), \tag{3}$$

$$u_{ij}(r,\theta) = \sum_n \frac{r^{\frac{n}{2}}}{2\mu} \left(A_n g_{\mathrm{I},ij}(\theta, n) + B_n g_{\mathrm{II},ij}(\theta, n)\right). \tag{4}$$

To estimate the stress field σ_{ij} in equation (3) for experimental DIC data, we optimize the parameters $A_n, B_n \in \mathbb{R}$, called Williams coefficients, by fitting the theoretical field (4) to the DIC displacement data using the over-deterministic method implemented in CrackPy [44]. From this, the von Mises equivalent stress is computed as $\sigma_{\mathrm{VM}} = \sqrt{\sigma_{11}^2 + \sigma_{22}^2 - \sigma_{11}\sigma_{22} + 3\sigma_{12}^2}$.

2.2.3. Effectiveness of AGT

Figure 5 shows an example result of AGT for crack tip segmentation, illustrating the initial unaligned attention after 30 epochs of pretraining (see figures 5(a)–(c)) and the final explanation, aligned to the theory-guided target (see figure 5(e)). The number of epochs used for pretraining was empirically determined. To avoid overfitting, we used the trained weights saved from the epoch with the lowest total validation loss (see figures 5(b)–(d)) as the final model to compute predictions and corresponding explanations. In this example, the target explanation is based on the corresponding stress field using the strategy Gradual Williams (GW) introduced below. This strategy steers the attention towards regions of high von Mises equivalent stress larger than 75 MPa (see figure 5(e)). For details on the used processing of our target explanations, we refer to section 4.5 and appendix C.

Preliminary experiments indicated that it is beneficial to avoid large corrective updates; large λ can cause over-correction of the model weights within the first few AGT epochs, which causes the predictive performance to deteriorate, leading to non-salient explanations. Models found in this state were rarely able to recover the intended training. Similarly, small λ had insufficient effect on the resulting explanations. To mitigate these effects, loss scaling was applied, and the present experiment was conducted using a weight factor of $\lambda = 2$. We observe that both training phases, pretraining and attention-guided, successfully converge. In this example, both validation and explanation loss exhibit significant variance, underpinning our approach of selecting the epoch with the lowest validation loss. For this model, the attention (figure 5(d)) aligns with the target GW attention (figure 5(e)), demonstrating the effectiveness of AGT.

2.2.4. Comparison of attention strategies

To investigate whether models guided by physically meaningful explanations exhibit improved generalization and trustworthiness, we conduct a series of experiments. For this, we introduce different target attention maps—two *physical* strategies that build on expert knowledge and intuition, i.e.

- *Binary Williams (BW)*: The target explanation is obtained by binarizing the Williams stress field using a fixed threshold: regions of elevated mechanical stress were assigned a target attention of 1, all others 0,
- *GW*: The target explanation is derived by truncating the Williams stress field at a specified threshold and rescaling the resulting values to the range $[0, 1]$, yielding a continuous, non-binary attention map with gradually fading intensity,

and two *non-physical* strategies that are intentionally designed to steer attention towards allegedly less informative regions, i.e.

- *Binary Misleading (BM)*: The target attention is set to 1 within a small square located in the bottom-right corner of the domain and 0 elsewhere.

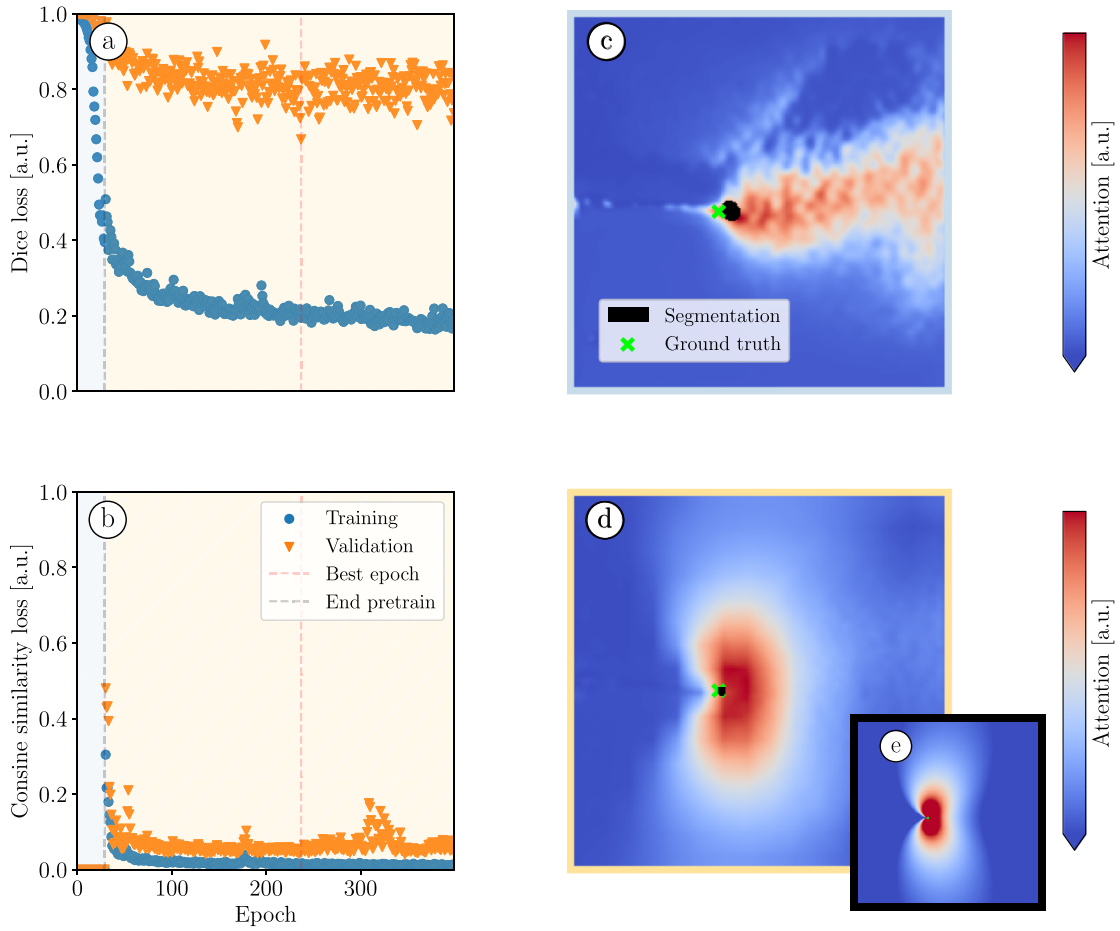

Figure 5. Example of an AGT for crack tip segmentation. The Dice loss provides feedback on the segmentation quality, while the cosine similarity loss quantifies coherence with the target explanation (e). Plots (a) and (b) show the loss curves of the segmentation quality and explanation coherence, respectively. Plot (c) shows the attention heatmap after the pretraining phase of 30 epochs. Plot (d) depicts the heatmap of the model with the smallest validation loss during the AGT phase (indicated by a dashed orange line in (a) and (b)). All attention heatmaps were obtained using the Grad-CAM++ method on the encoder layers of the U-Net.

- *Multi-gradual misleading (MGM)*: The target attention is set to 1 at the top- and bottom-right corners and gradually fades to 0.

The results presented in figure 6 show similar features as discussed in more detail in figure 5. Although the models can be qualitatively aligned with each of the considered explanation strategies without compromising predictive performance, the quantitative alignment of the explanation loss term is markedly lower for the non-physical target attentions compared to those guided by more salient, physically informed strategies.

Considering the large volatility in the validation loss and following up on previous work [19], we evaluate the models trained with different AGT strategies regarding their robustness and trustworthiness. For this, we compare the validation (Dice) loss, the reliability of the model (see section 4.2), and the explanation correctness (see section 4.4), ensuring that the learned explanations are faithful. By calculating both validation loss and reliability on in-distribution and reliability on out-of-distribution datasets (see section 4.1), we can qualitatively estimate the generalization capabilities of the respective models.

For each attention strategy, we trained 10 randomly initialized models. All experiments were performed with $\lambda = 2$, which was determined empirically and worked consistently for the present task. In addition, we trained 10 models without AGT as an independent baseline (Reference (R); $\lambda = 0$).

The results of this study are presented visually in figure 7 and quantitatively in table 1, with statistical significance assessed using Mann–Whitney-U (MWU) tests (appendix D).

On average, models trained with AGT using physical target explanations (BW, GW) achieve lower validation loss compared to the non-physical (BM, MGM) and the unguided reference (R). This difference is statistically significant when comparing physical strategies against both misleading targets and the

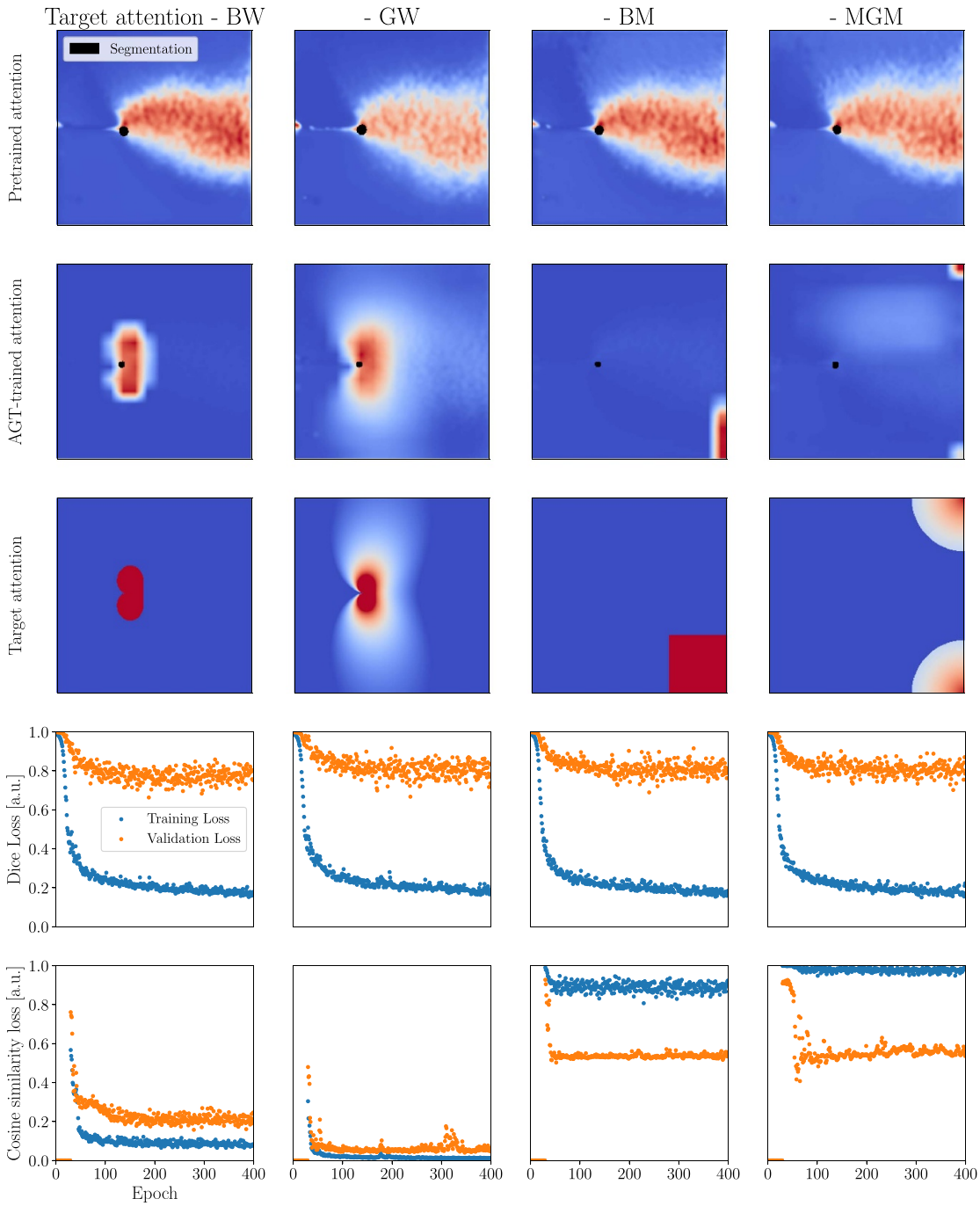

Figure 6. Comparison of different attention target strategies for exemplary training runs: *Binary Williams* (BW) and *Gradual Williams* (GW) are adaptations of the Williams stress field, whereas *BM* and *Multi-Gradual Misleading* (MGM) are intentionally steering attention towards allegedly less informative regions. All attention heatmaps were obtained using the Grad-CAM++ method on the encoder layers of the U-Net.

reference baseline (see table 2; appendix D). Among all strategies, the BW explanation yields the best single model in terms of validation loss.

In terms of reliability, on the in- or near-distribution datasets ($S_{160,4.7}, S_{160,2.0}$), all strategies exhibit saturated behavior, whereas for the further out-of-distribution datasets ($S_{950,1.6}$) the BW strategy shows significantly higher reliability values compared to every other strategy with statistical significance (see appendix D for details).

As far as correctness is concerned, the physical strategies (BW and GW) improve the correctness of their explanation with AGT, while non-physical strategies (BM and MGM) deteriorate correctness, as indicated by decreased and increased area under the curve (AUC) values, respectively. Moreover, the physical strategy BW resulted in the overall best model in terms of validation loss, reliability, and correctness of the explanations.

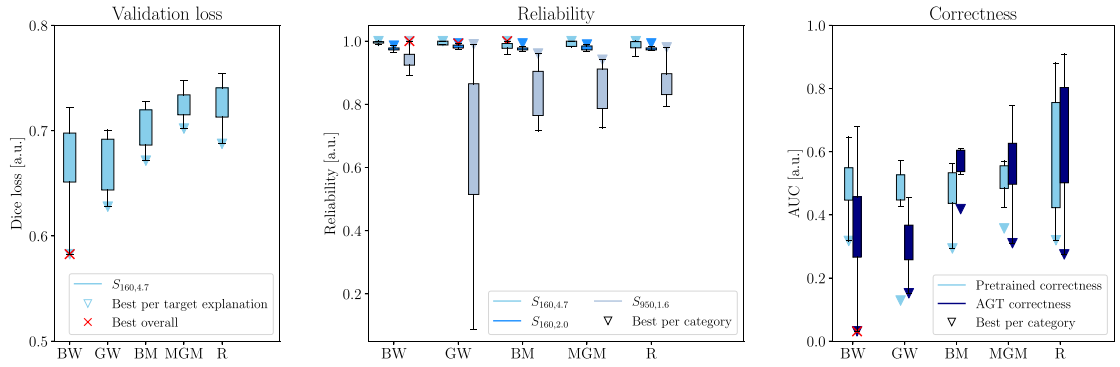

Figure 7. Quantitative evaluation of attention strategy experiments. Left: Validation (Dice) loss (lower is better) to measure predictive performance. Middle: Reliability of detection (larger is better) as a metric for generalization capacity (see section 4.2). Right: Explanation correctness, measured by area under the curve (AUC) (lower is better) to characterize faithfulness of explanations (see section 4.4). Both in-distribution ($S_{160,4.7}$) and out-of-distribution datasets ($S_{160,2.0}, S_{950,1.6}$) are considered (see section 4.1). Numeric representation can be found in table 1.

Table 1. Performance comparison across target explanation types. Reported values are mean ± standard deviation and 95% confidence intervals over 10 independent training runs. The evaluate model instances were selected based on the lowest-achieved validation loss, representing their respective peak performance potential. Sub-tables show (a) validation loss on the in-distribution dataset $S_{160,4.7}$; (b) reliability on in-distribution and out-of-distribution datasets; and (c) explanation correctness before and after attention-guided training (AGT). Bold values indicate the best-performing method within each category. For correctness, additionally the relative change induced by AGT is evaluated per method. Colored arrows indicate whether explanation trustworthiness improved (↓), deteriorated (↑), or remained unchanged (–).

(a) Validation loss (lower is better)

	$S_{160,4.7}$	
	Mean ± Std	95% CI
BW	**0.67 ± 0.04**	[0.64, 0.69]
GW	**0.67 ± 0.03**	[0.65, 0.68]
BM	0.70 ± 0.02	[0.69, 0.71]
MGM	0.72 ± 0.01	[0.72, 0.73]
R	0.72 ± 0.02	[0.71, 0.74]

(b) Reliability (higher is better)

	$S_{160,4.7}$		$S_{160,2.0}$		$S_{950,1.6}$	
	Mean ± Std	95% CI	Mean ± Std	95% CI	Mean ± Std	95% CI
BW	**0.99 ± 0.01**	[0.99, 1.00]	0.97 ± 0.01	[0.97, 0.98]	**0.94 ± 0.05**	[0.90, 0.96]
GW	**0.99 ± 0.02**	[0.97, 1.00]	**0.98 ± 0.01**	[0.98, 0.99]	0.67 ± 0.28	[0.50, 0.82]
BM	0.98 ± 0.01	[0.98, 0.99]	**0.98 ± 0.01**	[0.97, 0.98]	0.78 ± 0.22	[0.63, 0.88]
MGM	0.98 ± 0.03	[0.97, 1.00]	**0.98 ± 0.01**	[0.97, 0.98]	0.85 ± 0.08	[0.80, 0.90]
R	**0.99 ± 0.02**	[0.98, 0.99]	**0.98 ± 0.01**	[0.97, 0.98]	0.86 ± 0.12	[0.79, 0.92]

(c) Correctness (lower is better)

	Pretrained—$S_{160,4.7}$		AGT-trained—$S_{160,4.7}$		
	Mean ± Std	95% CI	Mean ± Std	95% CI	
BW	0.51 ± 0.1	[0.45, 0.57]	0.36 ± 0.18	[0.25, 0.46]	↓
GW	0.47 ± 0.14	[0.38, 0.54]	0.31 ± 0.09	[0.25, 0.37]	↓
BM	0.47 ± 0.10	[0.41, 0.52]	0.59 ± 0.12	[0.52, 0.66]	↑
MGM	0.52 ± 0.09	[0.46, 0.57]	0.55 ± 0.13	[0.47, 0.62]	↑
R	0.60 ± 0.22	[0.49, 0.72]	0.60 ± 0.22	[0.49, 0.72]	–

3. Discussion

Utilizing DL methods in a faithful and trustworthy manner remains a challenge, especially in scientific domains, as highlighted by recent work on robustness and generalization [45] and on the faithfulness of learned representations [46]. In this work, we show that integrating XAI techniques with evaluation metrics and domain knowledge to guide model attention can enhance both generalization capabilities and trustworthiness of DL models. We validate this claim in the context of machine-learned crack tip

segmentation in full-field displacement fields obtained by DIC during FCG, which is a critical task in fracture mechanics where model interpretability and robustness are paramount.

We build upon state-of-the-art methodology by extending techniques based on CAM to semantic segmentation tasks, similar to [18]. These methods can be applied to a variety of neural layers of the network, or a combination of such, and provide meaningful insight into the internal decision-making process of our DL model, as illustrated in figures 2 and 4. Adapting these methods has proven to be non-trivial. Even upon fixing the CAM target layers, the visualizations in figure 4 reveal substantial variation among different CAM-based methods with respect to attribution shape, spatial alignment, and total relevance. These differences underscore the heuristic and often inconsistent nature of post-hoc XAI techniques, exposing them to warranted critique [20] and necessitating cautious application, especially in contexts where explanation fidelity (correctness [29]) is critical.

Therefore, it is imperative to complement the visualizations with objective and quantitative metrics, evaluating, among others, the fidelity between the model and provided explanations. To that end, we address four of the twelve XAI evaluation criteria of Nauta *et al* [29], tailored to the task of crack tip segmentation in DIC data. Evaluation of these metrics allows us to systematically determine the faithfulness and quality of different methods, as depicted in figure 4, identifying the gradient-based methods Grad-CAM [21] and Grad-CAM++ [22] as most effective for this task and data. While these objective criteria provide valuable insights, further investigation is required to refine the implementation of the metrics and their applicability to other tasks and explanation methods. In general, it should be mentioned that the evaluation results depend on multiple parameters, among others the chosen model architecture, task, and data. Therefore, one cannot provide general guidance or static performance values, as each use case and model has to be evaluated separately.

To leverage explanations during training, we adapt the learning by self-explaining (LSX) approach recently proposed by Stammer *et al* [33]. LSX introduces a novel training paradigm where a learner model is optimized not only for the primary predictive task but also through feedback from a critic model that evaluates the quality of the determined explanations. However, a limitation of the LSX framework is that it relies on explanations that have not been externally validated.

In our AGT approach, we address these issues by choosing a correct, complete, continuous, and compact explanation method (see figure 4) and incorporating domain-specific knowledge to inform and evaluate explanations, thereby providing more reliable feedback and enhancing the overall trust in the model's predictions and explanations. This idea is motivated by earlier work conducted by Melching *et al* [19], where the authors observed that different model architectures lead to distinct attention patterns. Specifically, while some models learn to focus attention on physically relevant regions like the crack path or crack tip field, others display attention in presumably less meaningful areas, indicating that not all trained models inherently learn the same physical features. However, the combination of physical features and the inherent flexibility of deep learning often makes these models more powerful in terms of generalization capabilities. To guide model attention towards regions of physical significance, i.e. areas of high mechanical stress, we utilize the well-known von Mises stress field calculated using a Williams series expansion fitted to displacement data as the physical attention prior.

Our findings show that models guided away from the natural crack path attention—including the non-physical strategies (BM and MGM) – exhibit predictive performance comparable to, or slightly exceeding, that of the unguided reference model. In this study, predictive performance is assessed using a qualitative combination of the validation Dice coefficient, computed on the spatially and statistically distinct validation side of our labeled dataset $S_{160,4.7}$, and a task-specific reliability metric that enables performance estimation on unlabeled, out-of-distribution datasets $S_{160,2.0}$, $S_{950,1.6}$.

Due to the severe class imbalance of the segmentation task and the resulting stochastic fluctuations during training, model instances are selected based on their peak validation Dice performance. Consequently, the quantitative results reported in figure 7 and table 1 represent the maximum achievable performance of each training strategy rather than an unbiased estimate of expected deployment behavior. While this model selection procedure constitutes a known form of optimization bias, it is applied uniformly across all training regimes. As a result, relative comparisons between strategies remain meaningful and allow a fair assessment of their achievable performance.

Within this evaluation protocol, physically inspired attention strategies (BW and GW) consistently outperform both misleading and unguided baselines. This is reflected in lower mean validation Dice losses ($\approx$0.67) compared to misleading and reference strategies ($\geqslant$0.7; table 1). One-sided non-parametric comparisons using the MWU test confirm that physically guided strategies yield significantly lower Dice losses than misleading baselines ($p < 10^{-5}$) and the unguided reference ($p < 10^{-3}$); full test statistics are reported in appendix D.

We attribute the improved performance of physically guided models to the induction of information-dense, crack-tip-field-dependent decision patterns in the model's internal representation. The marginal performance gains of the misled models over the unguided reference are likely attributed to the avoidance of experimental artifacts inherent to DIC measurements near the crack path. All guided strategies—physical and non-physical alike—systematically suppress attention in this noise-dominated region and instead emphasize areas containing crack-tip field information (see figure 6), which appears sufficient to recover modest segmentation gains even in the absence of physically meaningful guidance.

Beyond in-distribution performance, generalization to unlabeled out-of-distribution data can be assessed using the task-specific reliability metric. Here, physically guided models exhibit markedly improved robustness, most prominently on the non-saturated OOD dataset $S_{950,1.6}$. In this regime, the BW strategy achieves mean reliability scores of approximately 94%, whereas all other strategies do not exceed 86%. Corresponding hypothesis tests confirm that BW yields significantly higher reliability scores than each other attention strategy ($p < 7 \cdot 10^{-3}$; appendix D). Evidently, providing a binarized, physically grounded representation of the crack-tip field as an attention prior is particularly effective in promoting reliable and robust model behavior under distributional shift.

Importantly, even under this favorable peak-performance evaluation protocol, which affords all models their best possible chance, physically guided strategies retain a clear and statistically supported advantage. This demonstrates that aligning model attention with domain-consistent stress-field priors raises the attainable performance beyond what can be achieved through unguided or arbitrarily guided training.

With respect to explanation correctness, physically guided strategies (BW and GW) exhibit a consistent improvement after attention-guided training, whereas non-physical strategies (BM and MGM) show a deterioration in correctness, as indicated by decreasing and increasing AUC values, respectively (see table 1(c)). Importantly, this trend cannot be attributed to longer training alone: although all AGT-trained models undergo an extended optimization phase, the correctness of the unguided reference model ($\lambda = 0$) remains unchanged throughout the training, while correctness for misleading strategies degrades. This demonstrates that the observed correctness changes are not an artifact of additional training epochs but are specifically induced by attention guidance.

The pretrained correctness values of all strategies exhibit noticeable variability; however, these initial values lie comfortably within the standard deviation of the reference baseline (0.60 ± 0.22) and can therefore be attributed to the inherent stochasticity of deep learning training rather than systematic differences between strategies. The subsequent divergence in correctness-improvement for physically guided models and deterioration for misleading ones thus reflects a genuine effect of the respective attention strategies.

While, in principle, predictive accuracy and explanation quality need not be coupled, our results indicate that under attention-guided training, aligning optimization with physically meaningful explanations can positively influence explanation faithfulness. In particular, models guided by physical attention priors consistently produce more faithful explanations, suggesting a non-trivial interplay between prediction learning and explanation alignment. This interaction represents a promising direction for future investigation into training paradigms that jointly optimize predictive performance and interpretability.

This work presents a framework for objectively selecting suitable XAI methods for specific semantic segmentation tasks and introduces a novel approach to guide model attention by domain knowledge. It serves as a compelling example of how XAI can move beyond post-hoc interpretation to actively inform and guide model development. Importantly, the proposed framework is not limited to specific data types, physical models, or explanation formats. In principle, any expert knowledge, independent of its origin, that can be used to determine the coherence of explanations may be used to guide model attention. In practice, however, we recognize the significant difficulty of finding a suitable setup that would benefit from AGT. Partly due to the sparsity of meaningful domain priors that can be related to model explanations without requiring manual generation or annotation of such.

Future research should aim to develop more principled and robust XAI methods tailored specifically for segmentation tasks. Establishing standardized evaluation protocols and benchmarks will be essential to assess explanation quality and model alignment across applications. Ultimately, we aim for a more interactive and iterative approach to scientific machine learning, where explanations are not just used after the fact but become an integral part of training, evaluation, and scientific understanding.

4. Methods

To demonstrate AGT, we applied it to crack tip segmentation in fatigue experiments. Figure 8 shows the task-specific implementations of the general framework.

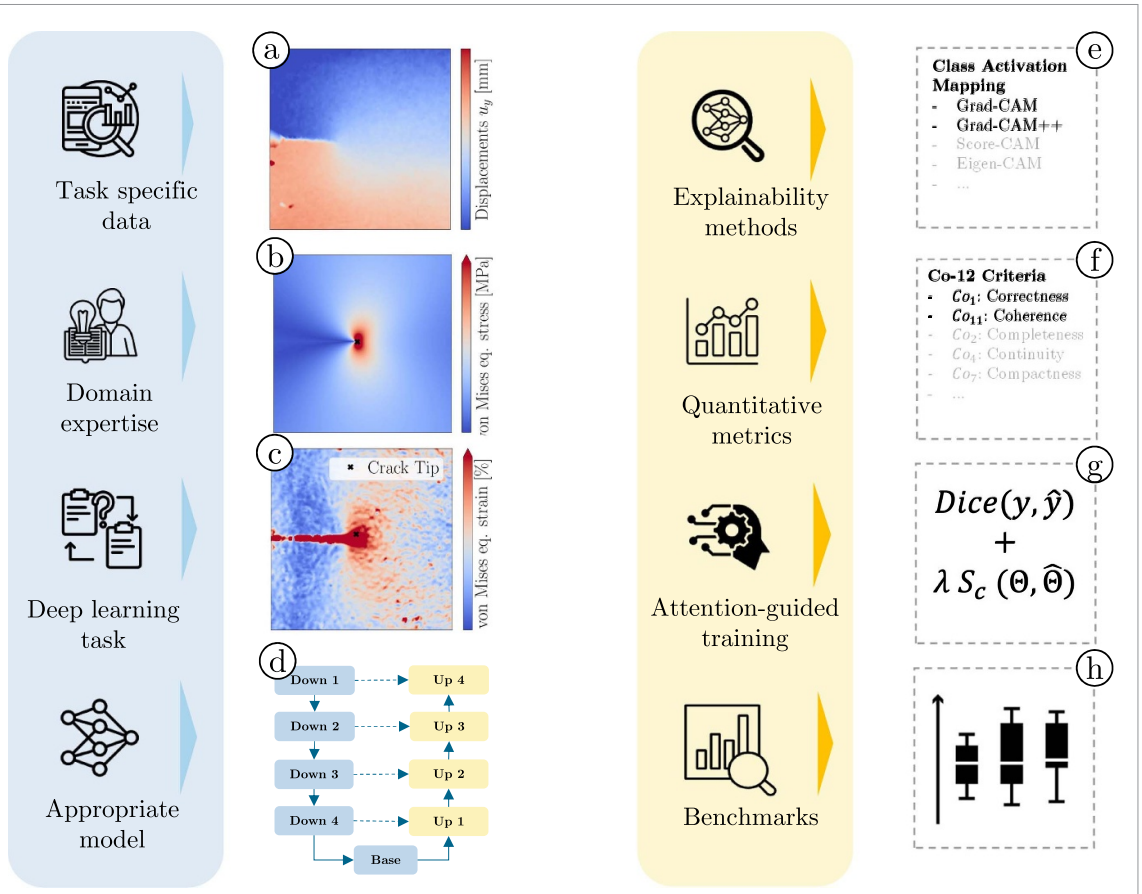

Figure 8. Visual summary of the task-specific AGT framework. (a) Experimental DIC data (section 4.1). (b) Crack tip specific domain knowledge via Williams series (section 4.5). (c) and (d) Crack tip segmentation task and U-Net architecture (section 4.2). (e), (f) CAM explanations and evaluation with *Co*12 criteria (sections 4.3 and 4.4). (g) Joint loss combining segmentation and explanation supervision (section 4.5). (h) Metrics to estimate model performance and robustness during generalization (section 4.2).

4.1. Experimental data

We used previously published DIC data from three FCG experiments on AA2024-T3 middle-crack tension specimens [8, 47, 48]. Displacement fields were interpolated on a 256×256 grid using CrackPy [44], yielding two-channel inputs (u_x, u_y) for deep learning. No new experiments were performed.

The dataset $S_{160,4.7}$ (166 samples at maximum force per side of the crack) was labeled using additionally obtained images of the polished sample surface. The data is split into training and validation based on sample side, where the crack growing to the right is used for training and the left for validation. The left side inputs are mirrored such that they match the right. Two additional datasets, $S_{160,2.0}$ (280 samples) and $S_{950,1.6}$ (102 samples), provided out-of-distribution test cases with different geometries and loading conditions. Further details on the data and the generation of ground-truth labels are provided in our previous work [8].

4.2. Machine learning

4.2.1. Task and architecture

Crack tip detection was formulated as a binary semantic segmentation task [19]. A U-Net [34] with LeakyReLU activations was used, mapping two-channel displacements to a single-channel crack-tip-probability map. The encoder consists of four convolutional blocks and a bottleneck, mirrored by a decoder with linear upsampling. A visual representation of this structure is depicted in figure 3.

4.2.2. Data preparation

Displacements were channel- and sample-wise normalized using the mean and standard deviation of each sample. Annotated single-pixel crack tips were expanded to 5×5 (train) and 3×3 (validation) masks. Augmentations included random crops (130–180 px), rotations ($\pm 10°$), and flips. Validation and test data were not augmented.

4.2.3. Training

Each training uses a randomly initialized U-Net with LeakyReLu activation functions [19]. Training used PyTorch with Adam (5×10^{-4}, AMSGrad, batch size 16) and dropout ($p = 0.3$) at the bottleneck. A two-stage strategy was applied: (i) 30 epochs of initial pretraining with segmentation loss only and (ii) 370 epochs with additional AGT supervision. Reference models were trained for 400 epochs without AGT.

4.2.4. Loss and metrics

Segmentation loss: Dice loss addressed the strong class imbalance.

Explanation loss: Cosine similarity (CSI) measured alignment of model attention Φ with target attentions $\hat{\Phi}$.

Reliability: A prediction was valid if it contained exactly one connected region; reliability was the fraction of valid predictions:

$$\mathrm{Rel} = \frac{\#\text{samples with exactly one segmentation}}{\#\text{total samples}}. \tag{5}$$

4.3. Explainability

CAM was used to estimate model attention w.r.t. input features. Because segmentation outputs are spatial, we replaced the class score with the global average of output logits (following the approach from [18, 19]). Explanations were generated from encoder and bottleneck layers, where features retain higher spatial fidelity. We evaluated gradient-based (Grad-CAM, Grad-CAM++, LayerCAM, etc) and gradient-free methods (EigenCAM, Score-CAM, Ablation-CAM), normalizing the resulting explanations to $[0, 1]$. For AGT, we selected Grad-CAM++ with encoder-layer aggregation. We refer to appendix A for further details.

4.4. Objective metrics for explainability

To assess explanation quality, we used five criteria from Nauta *et al* [29]: **Correctness** (here via incremental deletion), **Completeness** (here via incremental insertion), **Continuity** (here via structural similarity index measure across time steps), **Compactness** (here via the minimal input features required to recover a Dice coefficient of $\geqslant 0.8$), and **Coherence** (here via agreement with domain targets measured using S_c [49]). These metrics provided a quantitative basis for comparing CAM variants beyond visual inspection. We refer to appendix B for further details on the equations and calibration details.

4.5. Attention-guided training with domain knowledge

Target attentions were derived from the near-tip stress field using the Williams series expansion. Two physical (binary, gradual Williams) and two unphysical baselines (binary and multi-gradual misleading) were considered as attention targets (see appendix C for details). AGT integrates prediction and explanation supervision into a joint loss:

$$L_{\text{total}} = (1 - \mathrm{Dice}(\mathbf{y}, \hat{\mathbf{y}})) + \lambda \cdot \mathrm{CSI}\left(\mathbf{\Phi}, \hat{\mathbf{\Phi}}\right), \tag{6}$$

where $\mathbf{y}$ is the model output, $\hat{\mathbf{y}}$ the ground truth mask, Φ the Grad-CAM++ explanations, and $\hat{\Phi}$ the target attention maps. We used $\lambda = 2$, chosen empirically. Training followed a two-stage scheme: initial segmentation pretraining, followed by AGT with attention supervision. We refer to appendix C for further details on the training procedure and target attentions.

Data availability statement

The code for training and evaluation is published on Github (https://github.com/dlr-wf/attention-guided-training).

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.5740216, https://doi.org/10.5281/zenodo.16902960 [54].

Acknowledgments

Acknowledgments We acknowledge the financial support of the DLR-Directorate Aeronautics.

Conflicts of interest

All authors are named inventors on the patent application DE102025113325.5 (pending), filed by the German Aerospace Center (DLR), which is related to the algorithm described in this work.

Author contributions

J T and D M conceived the attention-guided training framework. J T implemented the framework. E B provided the fracture mechanical domain knowledge. All Authors discussed, analyzed, and interpreted the results and wrote the manuscript.

Appendix A. Explainability

The model behavior can be approximately described using class activation mapping [17]. The CAM methods provide explanations in the form of spatial attention heatmaps indicating the importance of different input regions to the models prediction for a certain output class. These explanations are calculated as a linear combination of the spatial feature maps $\mathbf{A}^k$ of a convolutional layer l with a set of relevance weights ω_k^c associated with a class c, yielding:

$$L_{\text{CAM}}^{c,l} = \sum_k \omega_k^c \mathbf{A}^k, \tag{7}$$

which is postprocessed by applying ReLU and upsampled to the spatial dimensions of the input to obtain a class-specific attention map $L_{\text{method}}^{c,l}$.

The original CAM method is restricted to specific architectures, containing convolutional layers proceeded by a fully connected layer and applicable only to classification tasks [21]. Grad-CAM [21] was proposed to generalize this approach to more complex architectures. Many novel CAM techniques in the literature are slight variants of Grad-CAM. These typically operate on the final convolutional layer and assume a single class score as output. Further changes have to be considered when applying these approaches to semantic segmentation [18].

Moreover, directly applying CAM to the final decoder layers of U-Net proved uninformative due to spatial sparsity and reduced feature variation in those layers. Approaches proposed by [18, 19, 39] consider earlier convolutional blocks-specifically, the encoder and bottleneck layers-where richer spatial structure is retained. To support explanation at multiple levels, we define block-wise outputs `Down1` through `Down4`, `Base`, and `Up1` through `Up4`, as described in section 4.2. The multi-layer aggregation is defined here as the average over explanations calculated for multiple blocks:

$$L_{\text{method}}^{c,[l_1,\ldots,l_n]} = \sum_{i=1}^{n} \beta_i L_{\text{method}}^{c,l_i}, \quad \beta_i = 1. \tag{8}$$

Adapting CAM for segmentation

Since semantic segmentation tasks produce spatial outputs, a scalar class score S^c is defined instead. S^c is calculated by aggregating logits over a subset of output pixels. Using the entire output mask retains the maximal amount of information:

$$S_{\text{GAP}}^c = \frac{1}{N} \sum_{i,j}^{m,n} f(\mathbf{X})_{ij}^c, \quad N = m \cdot n \tag{9}$$

This score replaces the scalar output used in original CAM methods. Other score definitions were discussed in [18].

Gradient-based CAM methods

Several gradient-based CAM methods were considered, all using the above score function (9):

- Grad-CAM [21]
- Grad-CAM++ [22]
- Grad-CAM+ [23]
- LayerCAM [39]
- Elementwise-Grad-CAM [24]

Gradient-free CAM methods
Alternatively, gradient-free CAM methods were evaluated. In particular:

- EigenCAM [26]
- Score-CAM [25]
- Ablation-CAM [27]

Eigen-CAM can be directly applied to intermediate feature maps without adaptation. Score-CAM and Ablation-CAM typically rely on confidence drops in classification. For segmentation, the degradation of a suitable prediction metric (e.g. logit aggregation, Dice score) is used as an alternative measure. All CAM outputs are normalized to $[0,1]$ and evaluated using objective criteria described in appendix B.

Appendix B. Objective metrics for explainability methods

To ensure that model explanations are meaningful, trustworthy, and unbiased, a set of objective criteria is applied to evaluate and compare different CAM-based explainability methods. A variety of individual criteria to measure explanation quality can be found in literature, including perturbation-based evaluation measures [50], axiomatic or theoretical frameworks for explanation properties [51], and human-aligned faithfulness metrics [52]. To our knowledge, no universal consensus has been reached so far. Following the *Co*-12 framework introduced by Nauta *et al* [29], we adopt a principled subset of explanation quality criteria that are meaningful and operational for CAM-based explanations in crack-tip segmentation. While *Co*-12 provides a versatile and unifying taxonomy, the application domain and practical feasibility determines the emphasis on and selection of the different criteria. Thus, we restrict our evaluation to a task-appropriate subset of criteria, as explicitly encouraged to preserve objectivity and comparability. In our case, criteria that depend on class contrastivity, user interaction, uncertainty quantification, or subjective concept decomposition cannot be meaningfully defined or are not sufficiently relevant for the problem at hand. We therefore focus on correctness, completeness, continuity, compactness, and coherence, as these criteria capture complementary aspects of explanation quality while avoiding ill-posed or subjective measures, enabling a robust comparison and application of explainability methods for our crack tip segmentation task.

Correctness (Co_1)
Correctness (or faithfulness) measures how well an explanation reflects the true decision process of the model and thus is of utmost importance. To quantify correctness, incremental deletion was used, where input pixels are obfuscated in order of decreasing relevance as predicted by the corresponding explanation [50]. A fast degradation of the agreement between predictions made on obfuscated samples with respect to the original (as measured here by the Dice coefficient) is interpreted as evidence of explanation correctness.

The Gaussian noise intensity is calibrated by randomized incremental deletion using:

$$\left(\tilde{\mathbf{X}}_{\mathrm{p}}\right)_{ij} = (\mathbf{X})_{ij} + \begin{cases} \alpha\mathcal{N}(\mu=0,\sigma=1) & \text{if} \quad i,j \in M_{\mathrm{p}}, \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

where $\mathcal{N}$ denotes Gaussian noise, p is the percentage of obfuscated pixels, M_{p} denotes a set of p percent of pixels, and $\tilde{\mathbf{X}}_{\mathrm{p}}$ is the obfuscated input with p percent obfuscation. The noise scale α is calibrated per model, averaging over the results of multiple inputs. A well-calibrated obfuscation should cause an approximately linear decrease in prediction performance as the obfuscated set of pixels M_{p} increases with p. An example of this approach is depicted in figure 9(b).

The relevance estimates obtained through the CAM methods in combination with the calibrated obfuscation yield the incremental deletion technique. The inputs are deleted in order of relevance, where the area under the curve (AUC) for deletion pixel percentages $p \in [0,P]$ defines the correctness score:

$$Co_1 = \int_0^P \mathrm{Dice}\left(\sigma\left(f\left(\tilde{\mathbf{X}}_{\mathrm{p}}\right)\right), \hat{\mathbf{y}}\right) \mathrm{d}p, \tag{11}$$

where $\tilde{\mathbf{X}}_{\mathrm{p}}$ denotes the obfuscated input and $\hat{\mathbf{y}}$ is the original binarized prediction. Co_1 is usually averaged over several inputs, e.g. a whole dataset, and several random Gaussian noises. Lower AUC indicates higher correctness of the explanation.

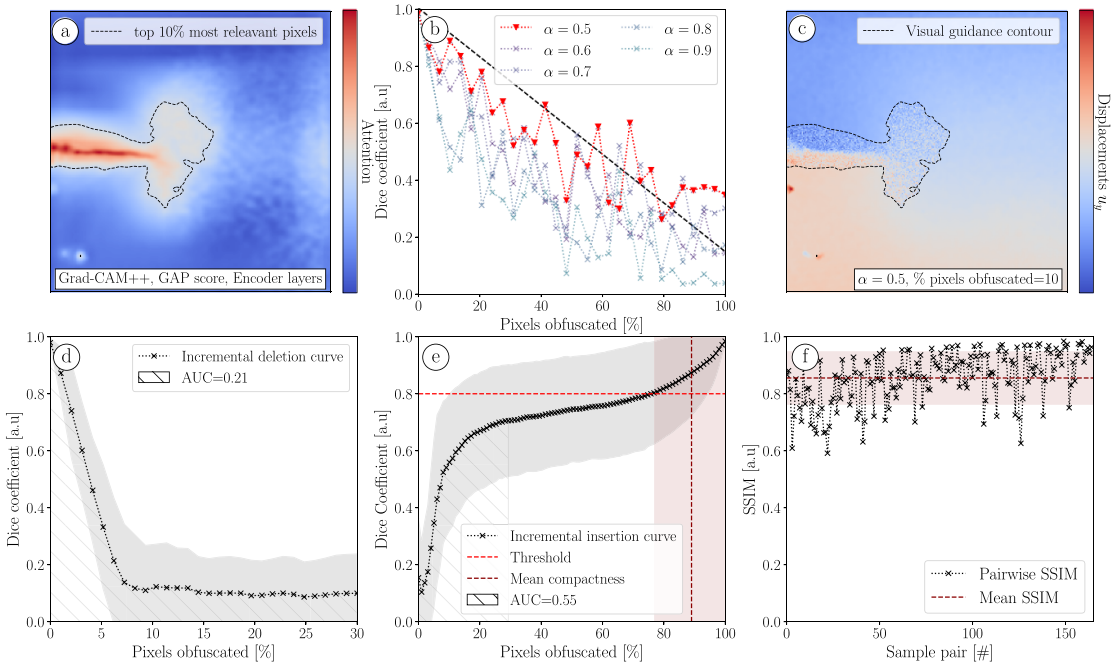

Figure 9. Quantitative evaluation of explanation quality using selected Co12 criteria. Mean and standard deviation are reported across all validation samples. (a) Example Grad-CAM++ attention heatmap computed from encoder layers using the GAP score. Pixels with the top 10% of attention highlighted. (b) Calibration curves showing Dice degradation under random pixel obfuscation with varying noise scales α, cf equation (11). (c) Input displacement u_y after Gaussian noise with $\alpha = 0.5$ is applied to the top 10% most relevant pixels. (d) Incremental deletion (correctness, Co_1) curve; steeper decline—as measured by AUC—indicates better correctness. (e) Incremental insertion (completeness, Co_2) and Compactness (Co_7) curve; AUC reflects completeness, while the threshold indicates how quickly the original performance is recovered. (f) continuity (Co_4) measured via pairwise SSIM across consecutive samples; higher values indicate explanation stability.

Completeness (Co_2) Complementary to correctness, completeness measures how much relevant information is preserved in the most important regions of the input [29]. It is typically assessed through incremental insertion, a process in which the most relevant pixels are gradually reintroduced into a fully obfuscated input to evaluate how well the model's predictions recover. The AUC of the resulting Dice curve (from $p = 0$ to $p = P$), similar to (11), defines the completeness score. Higher AUC indicates better completeness of the explanation.

Continuity (Co_4)
Continuity assesses whether small changes in the input lead to corresponding changes in the explanations. In fatigue crack growth experiments, the digital image correlation displacement data evolves smoothly over time. To evaluate continuity, predictions, and explanations are compared between consecutive samples in the sequence. Similarity of explanations is measured using the structural similarity index measure (SSIM) with the standard parameters introduced in [53]. The continuity score is computed as the average SSIM over all consecutive samples in a dataset:

$$Co_4 = \frac{1}{N-1}\sum_{i=2}^{N} \mathrm{SSIM}\left(\Phi(\mathbf{X}_{i-1})), \Phi(\mathbf{X}_i)\right) \tag{12}$$

Compactness (Co_7)
Compactness is evaluated by determining how quickly predictive accuracy recovers during the incremental insertion (completeness) curve. Specifically, the minimal percentage p^* of top-ranked pixels required to recover a Dice score $\geqslant 0.8$ is recorded. Methods with smaller p^* values are considered more compact.

Coherence (Co_{12})
Coherence measures agreement between model explanations and (expert-derived) attention targets, such as the physical crack tip field. Explanations Φ and attention targets $\hat{\Phi}$ are compared using SSIM and cosine similarity (CSI). This metric is not used to judge explanation quality in isolation but serves as the key component of our attention-guided training (AGT) loss (see appendix C).

These five metrics provide a multi-faceted, quantitative basis for comparing explanation methods beyond visual inspection or subjective plausibility. All evaluations are conducted on the validation dataset.

Appendix C. Attention-guided training

To perform AGT, domain-specific target attentions are added to the training and validation datasets as supervision signals. For fatigue crack growth in samples with a small plastic zone w.r.t. the crack length, the Williams series was chosen as a well-established and robust analytical representation of the stress field near the crack tip. A per-sample parameterization of equation (3) was omitted for simplicity, using instead a fixed, representative configuration.

The parameters were obtained using a least-squares fit method, available in the CrackPy library, yielding:

- $A_1 = \frac{K_I}{\sqrt{2*\pi}} = \frac{23.71*\sqrt{1000}}{\sqrt{2*\pi}}\ \mathrm{MPa}\sqrt{\mathrm{mm}}$
- $A_2 = \frac{T}{4} = \frac{-7.22}{4}\ \mathrm{MPa}$
- $A_3 = -2.87\ \frac{\mathrm{MPa}}{\sqrt{\mathrm{mm}}}$
- $A_4 = 0.03\ \frac{\mathrm{MPa}}{\mathrm{mm}^{-1/4}}$
- $A_n = 0\ \forall\ n > 4$
- $B_n = 0\ \forall\ n$

The proposed AGT framework integrates the components described above into a unified approach that explicitly steers model attention toward domain-specific priors during training. Model explanations, generated using CAM techniques, are evaluated using four objective criteria—correctness, compactness, continuity, completeness—and are aligned with target explanations derived from fracture mechanics. This alignment is enforced through a coherence criterion incorporated into the overall loss function.

The approach follows a two-stage training procedure:

1. An initial pretraining phase using only prediction loss.
2. A joint training phase using both prediction and explanation loss.

Model initialization and explanation selection

The U-Net model [34] is initialized using random weights. Initial training is required to reach a state of salient explanations; this phase is empirically set to 30 epochs. Explanations are then generated using the selected CAM methods mentioned in appendix A. Variations of targeted layers, methods, and score functions were considered. Preliminary qualitative evaluations and discussions led to a focused exploration of the encoder branch (Blocks `Down1` - `Down4` and `Base`) in combination with the GAP score function. This choice is supported by several arguments: early encoder layers preserve diverse, high-fidelity features; deeper layers (e.g. the bottleneck) capture more abstract representations; and GAP considers the entire input signal when computing explanation scores. Finally, Grad-CAM++ was chosen for AGT using the metrics found in figure 4.

Target attentions

Target attention maps $\hat{\Phi}$ were computed from the von Mises stress field using the Williams series expansion (equation (3)) and manual crack tip annotations. For comparison, two physical (domain-informed) and two unphysical attention target strategies were considered. The variations originating from the physical crack tip field are:

- *BW*: The continuous field is binarized using an empirical threshold of 162 MPa. Regions above this threshold were set to 1, others to 0.
- *GW*: The continuous field is obtained by clipping the Williams stress field to 162 MPa and below 75 MPa. Values below the threshold were set to 0, and values above to 1. The intermediate region was re-scaled, resulting values in the range $[0, 1]$, yielding a continuous, non-binary attention map with gradually fading intensity.

For comparison, additional unphysical attention targets were considered:

- *Binary misleading (BM)*: The target attention is set to 1 within a 76-pixel square located at the bottom-right corner of the domain and 0 elsewhere.

- *Multi-gradual misleading (MGM)*: The target attention is set to 1 at the top- and bottom-right corners and gradually decays within a radius of 70 pixels, using a radius-dependent exponential decay $e^{-0.013 \cdot r}$.

The empirical parameters chosen here were determined by balancing total relevant pixels and attention with (ir)relevant areas. All attention maps were clipped and normalized to $[0, 1]$ for comparability.

Training procedure
Subsequently, attention supervision was activated via a joint loss:

$$L_{\text{total}} = (1 - \text{Dice}(\mathbf{y}, \hat{\mathbf{y}})) + \lambda \cdot \text{CSI}\left(\mathbf{\Phi}, \hat{\mathbf{\Phi}}\right),$$

where y is the model output, $\hat{\mathbf{y}}$ the binary segmentation label, $\mathbf{\Phi}$ the Grad-CAM++ attention heatmaps, and $\hat{\mathbf{\Phi}}$ the target attention heatmaps. The CSI was chosen for its scale invariance and stable gradients. A weighting factor $\lambda = 2$ was used to balance the two loss terms, selected empirically to ensure that domain supervision did not dominate training.

Appendix D. Statistical tests

To assess statistically significant differences between attention guidance strategies, we employed the Mann–Whitney-U (MWU) test. The MWU test is a non-parametric alternative to the two-sample t-test and does not assume normality, making it suitable for the small sample size of $n = 10$, independent runs per strategy, and potentially skewed performance metrics considered in this study. All tests were conducted using the implementation contained in the Python package *SciPy*.

Each training run constitutes one independent sample. Runs were grouped according to their attention targets: *Binary Williams(BW)*, *GW*, *BM*, *MGM*, and a *Reference* configuration without attention guidance. For high-level analysis, strategies were further pooled into *physical* (BW + GW) and *misleading* (BM + MGM) groups.

Two families of directional hypotheses were evaluated at a significance level of $\alpha = 0.05$. First, three pre-specified overview comparisons were performed to assess the general effect of attention guidance: (i) Physical vs. Reference, (ii) Misleading vs. Reference, and (iii) Physical vs. Misleading. Second, four follow-up comparisons were conducted to assess the performance of the best-performing attention strategy BW relative to all remaining strategies. Directional hypotheses reflect the expected improvement direction (lower is better for validation loss and correctness AUC, higher is better for reliability).

Overall, physical attention guidance yields statistically significant improvements in validation loss across both grouped and individual strategy comparisons. Reliability differences are largely non-significant for in-distribution and mildly out-of-distribution datasets, which exhibit near-saturated performance across all strategies. Statistically meaningful reliability differences emerge only for the far out-of-distribution dataset $S_{950,1.6}$, where the BW target consistently outperforms all baselines. These findings indicate that physical attention guidance improves optimization behavior and robustness under distribution shift, with the BW attention targets achieving the strongest and most consistent performance gains.

Table 2. Directional Mann–Whitney-U test results comparing attention guidance strategies. Each row reports the p-value for the stated directional hypothesis. Statistically significant results ($p < 0.05$) are highlighted in bold.

Metric	Hypothesis	p-value
Validation loss	**Physical < Reference**	**0.0001**
	Physical < Misleading	**0.0**
	Misleading < Reference	0.105
	BW < R	**0.0018**
	BW < BM	**0.027**
	BW < MGM	**0.0009**
	BW < GW	0.7146
Reliability $S_{160,4.7}$	Physical > Reference	0.1586
	Physical > Misleading	0.124
	Misleading > Reference	0.4821
	BW > R	0.121
	BW > BM	**0.0373**
	BW > MGM	0.3758
	BW > GW	0.3909
Reliability $S_{160,2.0}$	Physical > Reference	0.1778
	Physical > Misleading	0.3149
	Misleading > Reference	0.3038
	BW > R	0.5616
	BW > BM	0.7473
	BW > MGM	0.8004
	BW > GW	0.9836
Reliability $S_{950,1.6}$	Physical > Reference	0.3789
	Physical > Misleading	0.2161
	Misleading > Reference	0.7015
	BW > R	**0.0154**
	BW > BM	**0.0069**
	BW > MGM	**0.0045**
	BW > GW	**0.002**
AGT correctness (smaller is better)	**Physical < Reference**	**0.0003**
	Physical < Misleading	**0.0**
	Misleading < Reference	0.2476
	BW < R	**0.0036**
	BW < BM	**0.0036**
	BW < MGM	**0.0129**
	BW < GW	0.8276

ORCID iDs

Jesco Talies 0009-0000-0786-7908
Eric Breitbarth 0000-0002-3479-9143
David Melching 0000-0001-5111-6511

References

[1] Voulodimos A, Doulamis N, Doulamis A and Protopapadakis E 2018 Deep learning for computer vision: a brief review *Comput. Intell. Neurosci.* **2018** 1–13
[2] Jumper J *et al* 2021 Highly accurate protein structure prediction with alphafold *Nature* **596** 583–9
[3] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 Recent advances and applications of machine learning in solid-state materials science *npj Comput. Mater.* **5** 83
[4] Merchant A, Batzner S, Schoenholz S S, Aykol M, Cheon G and Cubuk E D 2023 Scaling deep learning for materials discovery *Nature* **624** 80–85
[5] Zeni C *et al* 2025 A generative model for inorganic materials design *Nature* **639** 624–32
[6] Herrmann L and Kollmannsberger S 2024 Deep learning in computational mechanics: a review *Comput. Mech.* **74** 281–331
[7] Aldakheel F, Satari R and Wriggers P 2021 Feed-forward neural networks for failure mechanics problems *Appl. Sci.* **11** 6483
[8] Strohmann T, Starostin-Penner D, Breitbarth E and Requena G 2021 Automatic detection of fatigue crack paths using digital image correlation and convolutional neural networks *Fatigue Fract. Eng. Mater. Struct.* **44** 1336–48
[9] European Commission 2021 AI Act: the first regulatory framework for artificial intelligence in the european union *Proposed by the European Commission, Accessible on the EU* (Law and Publications website) (available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX,

[10] European Union Aviation Safety Agency 2023 Artificial intelligence roadmap 2.0: a human-centric approach to AI in aviation *Technical Report* (European Union Aviation Safety Agency) (available at: www.easa.europa.eu/en/domains/research-innovation/ai (Accessed 08 January 2025)
[11] National Aeronautics and Space Administration 2024 Guidance-990: artificial intelligence *Technical Report* (National Aeronautics and Space Administration) (available at: https://eirb.jsc.nasa.gov/EIRB/sd/resource/doc/DOC8DC90EE48B6684A/ORA-990-GUIDANCE_Artificial (Accessed 08 January 2025)
[12] Arrieta A B *et al* 2020 Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible AI *Inf. Fusion* **58** 82–115
[13] Zeng J, Ustun B and Rudin C 2016 Interpretable classification models for recidivism prediction *J. R. Stat. Soc. A* **180** 689–722
[14] Simonyan K, Vedaldi A, and Zisserman A 2014 Deep inside convolutional networks: visualising image classification models and saliency maps
[15] Lundberg S M and Lee S-I 2017 *A Unified Approach to Interpreting Model Predictions* (Curran Associates, Inc.) (available at: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
[16] Ribeiro M T, Singh S and Guestrin C 2016 *why Should i Trust you?: Explaining the Predictions of any Classifier* pp 1135–44
[17] Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A 2016 Learning deep features for discriminative localization *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 2921–9
[18] Vinogradova K, Dibrov A and Myers G 2020 Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract) *Proc. AAAI Conf. Artif. Intell.* **34** 13943–4
[19] Melching D, Strohmann T, Requena G and Breitbarth E 2022 Explainable machine learning for precise fatigue crack tip detection *Sci. Rep.* **12** 9513
[20] Rudin C 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead *Nat. Mach. Intell.* **1** 206–15
[21] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2016 Grad-cam: Visual explanations from deep networks via gradient-based localization *Int. J. Comput. Vis.* **128** 336–59
[22] Chattopadhay A, Sarkar A, Howlader P and Balasubramanian V N 2018 Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks *2018 IEEE Winter Conf. on Applications of Computer Vision (WACV)* pp 839–47
[23] Lerma M and Lucas M 2022 Grad-cam++ is equivalent to grad-cam with positive gradients (arXiv:2205.10838)
[24] Pillai V and Pirsiavash H 2021 Explainable models with consistent interpretations *Proc. AAAI Conf. Artif. Intell.* **35** 2431–9
[25] Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, Mardziel P and Hu X 2020 Score-cam: Score-weighted visual explanations for convolutional neural networks *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)* pp 111–9
[26] Muhammad M B and Yeasin M 2020 Eigen-cam: class activation map using principal components *2020 Int. Joint Conf. on Neural Networks (IJCNN)* (IEEE) pp 1–7
[27] Desai S and Ramaswamy H G 2020 Ablation-cam: visual explanations for deep convolutional network via gradient-free localization *2020 IEEE Winter Conf. on Applications of Computer Vision (WACV)* (IEEE) (https://doi.org/10.1109/WACV45572.2020.9093360)
[28] Vilone G and Longo L 2021 Notions of explainability and evaluation approaches for explainable artificial intelligence *Information Fusion* **76** 89–106
[29] Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, Schlötterer J, van Keulen M and Seifert C 2023 From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI *ACM Comput. Surv.* **55** 1–42
[30] Selvaraju R R, Lee S, Shen Y, Jin H, Ghosh S, Heck L, Batra D and Parikh D 2019 Taking a hint: leveraging explanations to make vision and language models more grounded *2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV)* pp 2591–600
[31] Brack M, Schramowski P, Deiseroth B and Kersting K 2023 ILLUME: Rationalizing vision-language models through human interactions *Proc. 40th Int. Conf. on Machine Learning (Proc. of Machine Learning Research)* vol 202, ed A Krause, E Brunskill, K Cho, B Engelhardt, S Sabato and J Scarlett (PMLR) (available at: https://proceedings.mlr.press/v202/brack23a.html), pp 3021–37
[32] Schramowski P, Stammer W, Teso S, Brugger A, Herbert F, Shao X, Luigs H-G, Mahlein A-K and Kersting K 2020 Making deep neural networks right for the right scientific reasons by interacting with their explanations *Nat. Mach. Intell.* **2** 476–86
[33] Stammer W, Friedrich F, Steinmann D, Brack M, Shindo H and Kersting K 2024 Learning by self-explaining *Trans. Mach. Learn. Res.* **2024**
[34] Ronneberger O, Fischer P and Brox T 2015 *U-Net: Convolutional Networks for Biomedical Image Segmentation* (Springer) pp 234–41
[35] Williams M L 1957 On the stress distribution at the base of a stationary crack *J. Appl. Mech.* **24** 109–14
[36] Tavares S M O and de Castro P M S T 2017 An overview of fatigue in aircraft structures *Fatigue Fract. Eng. Mater. Struct.* **40** 1510–29
[37] Roux S, Réthoré J and Hild F 2009 Digital image correlation and fracture: an advanced technique for estimating stress intensity factors of 2d and 3d cracks *J. Phys. D: Appl. Phys.* **42** 214004
[38] Zhao J, Sang Y and Duan F 2019 The state of the art of two-dimensional digital image correlation computational method *Eng. Rep.* **1** e12038
[39] Jiang P-T, Zhang C-B, Hou Q, Cheng M-M and Wei Y 2021 Layercam: exploring hierarchical class activation maps for localization *IEEE Trans. Image Process.* **30** 5875–88
[40] Jung H and Oh Y 2021 Towards better explanations of class activation mapping *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)* (IEEE) pp 1316–24 (http://dx.doi.org/10.1109/ICCV48922.2021.00137)
[41] Veit A, Wilber M J and Belongie S 2016 Residual networks behave like ensembles of relatively shallow networks *Proc. 30th Int. Conf. on Neural Information Processing Systems* (Curran Associates Inc.) pp 550–8 (available at: https://dl.acm.org/doi/10.5555/3157096.3157158)
[42] Sudre C H, Li W, Vercauteren T, Ourselin S and Cardoso M J 2017 *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations* (Springer) pp 240–8
[43] Kuna M 2013 *Finite Elements in Fracture Mechanics: Theory - Numerics - Applications* (Springer) (http://dx.doi.org/10.1007/978-94-007-6680-8)
[44] Strohmann T, Melching D, Paysan F, Klein A, Dietrich E, Requena G, and Breitbarth E 2022 Crack analysis tool in python - crackpy version 1.0.0. *Zenodo* (https://doi.org/10.5281/zenodo.7319653)
[45] Sapoval N *et al* 2022 Current progress and open challenges for applying deep learning across the biosciences *Nat. Commun.* **13** 1728
[46] Scorzato L 2024 Reliability and interpretability in science and deep learning *Minds Mach.* **34** 27

[47] Breitbarth E, Strohmann T and Requena G 2020 High-stress fatigue crack propagation in thin aa2024-t3 sheet material *Fatigue Fract. Eng. Mater. Struct.* **43** 2683–93
[48] Strohmann T, Melching D, Paysan F, Dietrich E, Requena G and Breitbarth E 2024 Next generation fatigue crack growth experiments of aerospace materials *Sci. Rep.* **14** 14075
[49] Salton G and McGill M J 1983 *Introduction to Modern Information Retrieval* (McGraw-Hill, Inc.)
[50] Rong Y, Leemann T, Borisov V, Kasneci G and Kasneci E 2022 A consistent and efficient evaluation strategy for attribution methods *Proc. 39th Int. Conf. on Machine Learning* vol 162 (PMLR) p 18770–95
[51] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R and Samek W 2015 On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation *PLoS One* **10** e0130140
[52] Yeh C-K, Hsieh C-Y, Suggala A S, Inouye D I and Ravikumar P 2019 *On the (in)Fidelity and Sensitivity of Explanations* (Curran Associates Inc.)
[53] Wang Z, Bovik A, Sheikh H and Simoncelli E 2004 Image quality assessment: from error visibility to structural similarity *IEEE Trans. Image Process.* **13** 600–12
[54] Melching D, Strohmann T, Requena G and Breitbarth E 2022 Full-field displacements and strains obtained by digital image correlation during fatigue crack growth experiments *Zenodo* (https://doi.org/10.5281/zenodo.5740216)