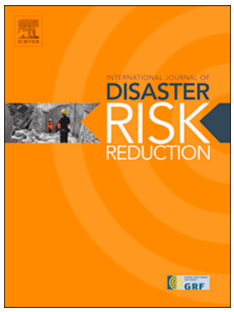

Towards Reliable Deep Learning For Post-Disaster Damage Assessment: An XAI-based Evaluation

Umut Lagap, Saman Ghaffarian, Sophie Gelinas-Gagne, Jasmin Jilma, Zhiyu Liu, Zhiyuan Luo

PII: S2212-4209(25)00663-6

DOI: https://doi.org/10.1016/j.ijdrr.2025.105839

Reference: IJDRR 105839

To appear in: *International Journal of Disaster Risk Reduction*

Received Date: 4 April 2025

Revised Date: 22 September 2025

Accepted Date: 23 September 2025

Please cite this article as: U. Lagap, S. Ghaffarian, S. Gelinas-Gagne, J. Jilma, Z. Liu, Z. Luo, Towards Reliable Deep Learning For Post-Disaster Damage Assessment: An XAI-based Evaluation, *International Journal of Disaster Risk Reduction*, https://doi.org/10.1016/j.ijdrr.2025.105839.

Towards Reliable Deep Learning For Post-Disaster Damage Assessment: An XAI-based Evaluation

Umut Lagap[a]*, Saman Ghaffarian[a]**, Sophie Gelinas-Gagne[a], Jasmin Jilma[a], Zhiyu Liu[a], and Zhiyuan Luo[a]

[a] *Department of Risk and Disaster Reduction, University College London (UCL), London, United Kingdom*

*Corresponding author: umutlagap@gmail.com, ucfbula@ucl.ac.uk

** Corresponding author: s.ghaffarian@ucl.ac.uk

Towards Reliable Deep Learning For Post-Disaster Damage Assessment: An XAI-based Evaluation

Abstract

The increasing frequency and severity of natural hazard-induced disasters necessitate rapid and reliable post-disaster damage detection (PDD) to inform disaster response and recovery. Deep learning (DL) models, when paired with remote sensing (RS) data, have shown potential in this domain, but challenges persist due to limited interpretability and inconsistent reliability, particularly for high-severity damage classes. This study investigates the use of attention mechanisms—Channel Attention (CA), Spatial Attention (SA), and Multihead Attention (MA)—to enhance the accuracy and interpretability of state-of-the-art DL models. Utilizing the xBD dataset, we evaluated eight DL architectures and their attention-augmented configurations, in total 32 model, using explainable AI (XAI) models, i.e., Grad-CAM and Saliency Maps to visualize decision-making processes. Results indicate that models enhanced with MA achieve the highest reliability, with MA_ShallowNetV2 and MA_InceptionV3 achieving accuracies of 81.9% and 80.0%, respectively. Grad-CAM analysis demonstrated precise localization of damaged areas, while Saliency Maps revealed well-concentrated pixel-level focus. In contrast, models with CA or certain SA configurations struggled with misplaced or diffused attention. These findings underscore the importance of incorporating explainable and interpretable AI approaches in disaster risk management. Specifically, MA generally improved interpretability and reliability in our evaluation, particularly for identifying high-severity damage levels in post-disaster scenarios.

Keywords: deep learning; attention mechanisms; Explainable AI; Grad-CAM; Saliency Maps; post-disaster damage detection, remote sensing

1. Introduction

The increasing frequency and intensity of certain natural hazard-induced disasters highlight the urgent need for rapid and reliable PDD to inform disaster response and recovery efforts. The International Disaster Database, EM-DAT, reveals that these

disasters impacted around 3.4 billion people and created an estimated economic loss of $4.2 trillion worldwide between 2004 and 2023 (EM-DAT, 2024). Storm related disasters, such as tornadoes, produce highly localized and varied patterns of destruction that can be difficult to discern (Crawford et al., 2020; Gokaraju et al., 2015; Mansour et al., 2021). While timely PDD is critical for identifying affected areas and guiding resource allocation (Abdullah M Braik & Maria Koliou, 2024; Wu et al., 2021), existing approaches often struggle to deliver both accurate and reliability in high-pressure disaster contexts. This creates a pressing need for approaches that are not only accurate but also operationally reliable and interpretable for informed decision-making, ultimately speeding up the recovery process and reducing long-term impacts on affected communities (Ghaffarian et al., 2023; Matin & Pradhan, 2021).

RS technologies, including satellite imagery and aerial photography, provide comprehensive and near real-time data covering vast and often inaccessible regions, facilitating quicker and more accurate assessments that make them indispensable for disaster monitoring (Da et al., 2022; Ghaffarian et al., 2018; Zhan et al., 2022; Zou et al., 2023). The introduction of benchmark studies such as xBD and a following version of it named Xview2 have driven major advances by providing standardized data for training and evaluation specifically designed for training DL models. These datasets provide a large-scale collection of annotated very-high resolution satellite imagery showing pre-disaster and post-disaster conditions with labelled building damages, and include data for various disaster types such as hurricanes, earthquakes, floods, and wildfires (Ritwik Gupta et al., 2019). They have enabled DL models to outperform traditional machine learning relying on predefined features and require extensive domain knowledge (Ahmadi et al., 2024; Song et al., 2019; X. Zhang et al., 2022). In this study, we specifically focus on tornado events, using three representative cases from the xBD

dataset. They present unique challenges due to their highly localized and heterogeneous damage patterns, which are difficult to detect and classify reliably. Thus, a key gap lies in evaluating whether DL models can provide consistent and interpretable assessments when applied to tornado-related disasters across diverse contexts.

Significant efforts have been made to improve the accuracy of DL models for PDD using various model architectures combined with attention mechanisms, that highlight the most informative features and filter out irrelevant inputs (Ghaffarian et al., 2021). Zhang et al. (2023) presented LRBNet combining a Siamese network and UNet++ architecture with components like lightweight compression module and efficient channel attention, achieved an accuracy of 85.0% for building localization and 70.7% for damage classification with xBD dataset. Rohit Gupta and Shah (2021) presented RescueNet integrating a novel localization-aware loss function and a multi-headed architecture, significantly improving accuracy to 84.0% for building localization and 74.0% for damage classification with xBD dataset. L. Deng and Wang (2022) designed an improved U-Net model architecture for PDD, leveraging the xBD dataset, and achieved 87.41% accuracy for building localization and 75.36% accuracy for damage classification by utilizing extra skip connections, asymmetric convolution blocks, and a shuffle attention module. Abdullah M Braik and Maria Koliou (2024) leveraged xBD dataset and Convolutional Neural Networks (CNN) by combining with Geographic Information Systems (GIS) for large-scale building damage assessment. After a fine-tuning process, the model achieved over 90% accuracy with a case study on Hurricane Ike in Galveston County. Leveraging the xBD dataset, Xia et al. (2023) presented BDANet integrating multi-scale feature fusion and cross-directional attention modules and achieved an accuracy of 80.43% for building localization and 84.17% for damage classification in the real-world scenario of the 2023 Turkey earthquake. Oludare et al. (2022) proposed ATS-

HRNet combining high-resolution network blocks with criss-cross attention modules to extract contextual information and achieved an accuracy of 89.2% for building localization and 87.5% for damage classification, outperforming other state-of-the-art methods in post-disaster damage assessment. However, the evaluation of attention mechanisms has remained almost entirely performance-centred and mainly neglected the explanation of the decision-making process

es, leaving these models as a "black box." This creates an important knowledge gap about the practical value of attention mechanism for enhancing trust and transparency in disaster contexts.

In parallel, XAI plays a crucial role in applications such as disaster response and recovery where reasoning behind model predictions is essential for decision-makers (Cheng et al., 2022; Ghaffarian et al., 2023; Wenli Yang et al., 2023). Even models with high accuracy may not be fully trusted or effectively utilized by stakeholders seeking a comprehensive basis for their decisions when these models lack explainability and reliability (Ghaffarian et al., 2023; Tursunalieva et al., 2024; Wenli Yang et al., 2023). Explainable models not only enhance trust and reliability but also enable users to identify potential errors, biases, or areas for improvement, facilitating more informed and effective disaster management (Tursunalieva et al., 2024). Seydi et al. (2023) presented BDD-Net+ combining convolution layers, transformer blocks, and self-attention layers to enhance damage detection accuracy on the Haiti Earthquake and Bata Explosion datasets and interpreted their model's decisions with Grad-CAM, one of the highly used XAI methods. Grad-CAM revealed that BDD-Net+ successfully focused key features, such as rough surfaces and collapsed areas, while detecting damaged areas, ensuring trust in the model's outputs (Seydi et al., 2023). Cheng et al. (2022) employed DoriaNET (Cheng et al., 2021a), xBD datasets and XAI techniques such as Grad-CAM to visualize

the model's focus areas during damage classification. Combining Grad-CAM and uncertainty metrics such as Bayesian inference and Monte Carlo dropout revealed that their model makes reliable predictions, aligning well with human-labeled assessments (Cheng et al., 2022). Despite its promise, XAI applications in PDD are fragmented and rarely scaled across multiple architectures especially with attention mechanisms. Moreover, most studies use a single XAI method, missing the opportunity to evaluate interpretability at multiple levels of granularity neglecting the potential of systematic and multi-level evaluation of model reliability in PDD.

Overall, most studies evaluate DL models for PDD primarily on accuracy, overlooking whether attention mechanisms actually improve interpretability and reliability. Despite their extensive usage, there is a lack of evidence that attention modules make models more trustworthy in practice (Islam et al., 2023; Victor et al., 2022; H. Zhang et al., 2022). Besides, XAI methods like Grad-CAM are frequently used to visualize model predictions, there is a notable gap in understanding how these tools can compare the effectiveness of various attention mechanisms specifically for post-disaster scenarios (Cheng et al., 2022; Islam et al., 2023; Seydi et al., 2023). The combination of XAI methods and attention mechanisms in PDD is crucial to know whether attention-driven performance gains are accompanied by improvements in transparency, or whether they come at the cost of interpretability. Addressing these gaps is necessary if AI systems are to move beyond benchmarks and be adopted in real-world disaster response, where decision-makers require both performance and transparency to act with confidence.

To address these shortcomings and bridge the gap between high performance and practical applicability, our contributions are threefold:

1. We incorporate three attention modules - channel, spatial, and multihead – into eight widely used DL architectures and perform a systematic evaluation of their

performance in terms of accuracy and reliability for PDD with 32 model combinations in total.

2. By coupling attention mechanisms with complementary XAI techniques, using Grad-CAM to initially filter and identify spatial focus, followed by Saliency Maps for a finer pixel-level analysis, this study introduces a complementary approach to evaluating model decision-making in the context of PDD.
3. We emphasize the role of both interpretability and reliability in distinguishing high-severity from lower-severity categories, demonstrating how reliable models can provide actionable insights that are directly relevant to decision-makers in disaster response and recovery efforts.

By bridging the gap between accuracy-driven DL research and interpretability-driven practical needs, this study advances both methodological and applied knowledge. It contributes a systematic, comparative framework for evaluating the role of attention in PDD and offers actionable insights into building transparent and trustworthy AI tools for disaster risk management.

2. Materials and Methods

2.1. Study Area and Dataset

This study utilizes the xBD dataset that contains a large-scale annotated collection of satellite imagery designed for PDD (Ritwik Gupta et al., 2019). A subset of the xBD dataset focused on tornado events has been selected for this study. Specifically, the Joplin, Moore, and Tuscaloosa tornadoes are used as the study area due to the availability of high-quality annotated imagery for these events. These tornadoes were significant disasters that resulted in extensive damage and loss of life across the central United States,

with each event exhibiting unique damage patterns and levels of destruction (Atkins et al., 2014; Prevatt et al., 2012).

The dataset includes pre- and post-disaster images with labeled damage classifications, making it suitable for training DL models aimed at identifying damage severity. It classifies damages into four levels: no damage, minor damage, major damage, and destroyed (Table 1). However, the dataset exhibits a substantial class imbalance, with "no damage" samples significantly outnumbering other categories. To mitigate this imbalance issue and improve generalization, we applied targeted data augmentation, particularly to the minority classes. The augmentation techniques included random rotations, horizontal and vertical flips, color space, noise addition, brightness adjustment, and contrast adjustment (S. Yang et al., 2022). These transformations increased the diversity of training samples, balanced class distributions, and enhanced the robustness of the model to variations in image orientation, illumination, and sensor noise (Mumuni & Mumuni, 2022; Shorten & Khoshgoftaar, 2019). Figure 1 shows the key information of selected case studies: their occurrence date, intensity on the Enhanced Fujita scale, and economic loss (Atkins et al., 2014; Crawford et al., 2017; Paul & Stimers, 2012). It also shows the number of buildings based on their damage levels for each event.

Table 1: Damage level description and the distribution of training and testing samples.

Damage Level	**Structure Description**	**Training samples**	**Testing samples**
No Damage	Undisturbed. No sign of structural or shingle damage or burn marks.	37677	500
Minor Damage	Building partially burnt, roof elements missing, or visible cracks.	4614	500
Major Damage	Partial wall or roof collapse.	1420	500

Destroyed	Scorched, completely collapsed, or otherwise no longer present.	5455	500

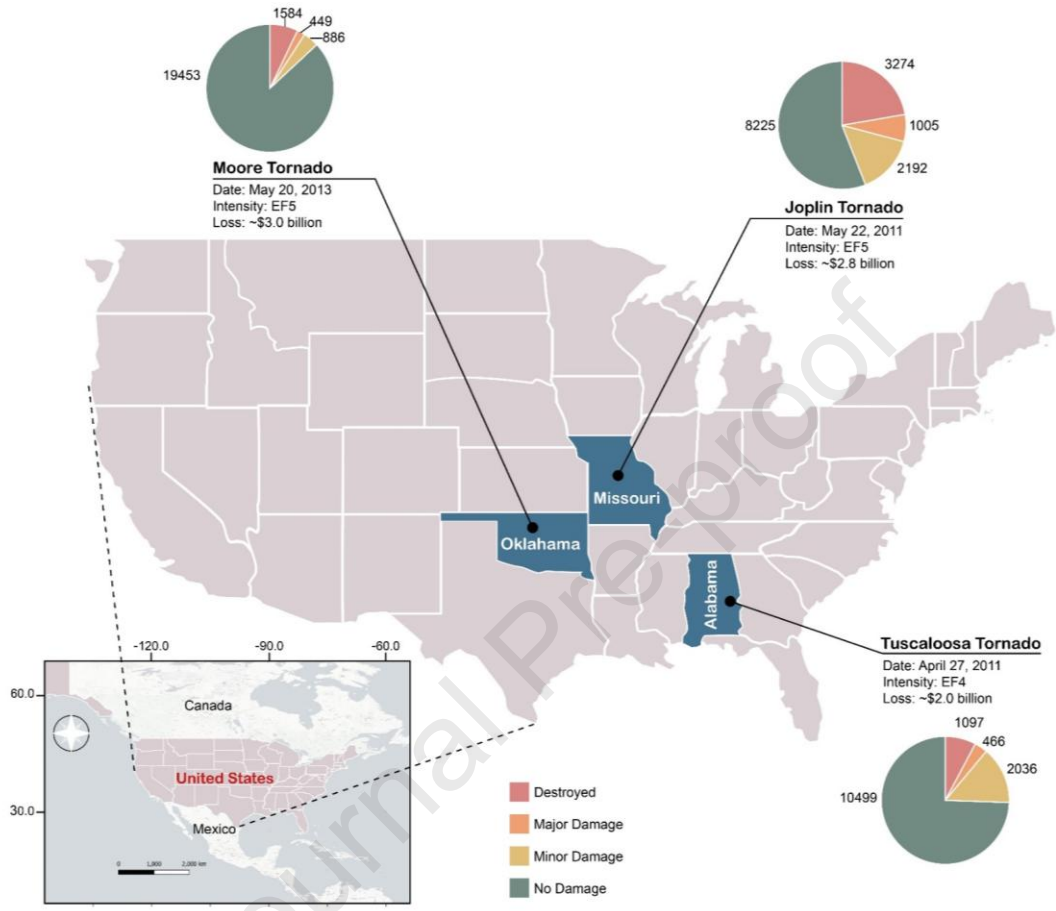

Figure 1: Overview of the selected tornado events, including key details such as damage intensities, economic losses, and the number of images per damage category.

Figure 2 shows that pre-disaster and post-disaster high resolution satellite images (1024x1024 pixel tiles with a spatial resolution of 0.5 m) of tornado affected areas were utilized to prepare the data for this study, consisting of three spectral bands: red, green, blue (RGB). The dataset includes building annotations for both pre-disaster locations and post-disaster damage levels, facilitating the identification and categorization of damage for each structure (Ritwik Gupta et al., 2019). The data preparation process begins with segmenting each annotated building area into smaller 128x128 pixel images. This allows the extraction of focused views for each individual building, isolating them from the

larger satellite image. The post-disaster images were then cropped to match the pre-disaster building locations, ensuring alignment and consistency for comparative analysis. This alignment enables the model to directly compare structural changes across time and ensures that explainable AI visualizations can meaningfully trace these differences. Doubling the paired pre- and post-disaster patches into a square format also standardized the input dimensions for stable processing in the deep learning models. Each cropped building image from the post-disaster layer is subsequently labeled based on its damage classification.

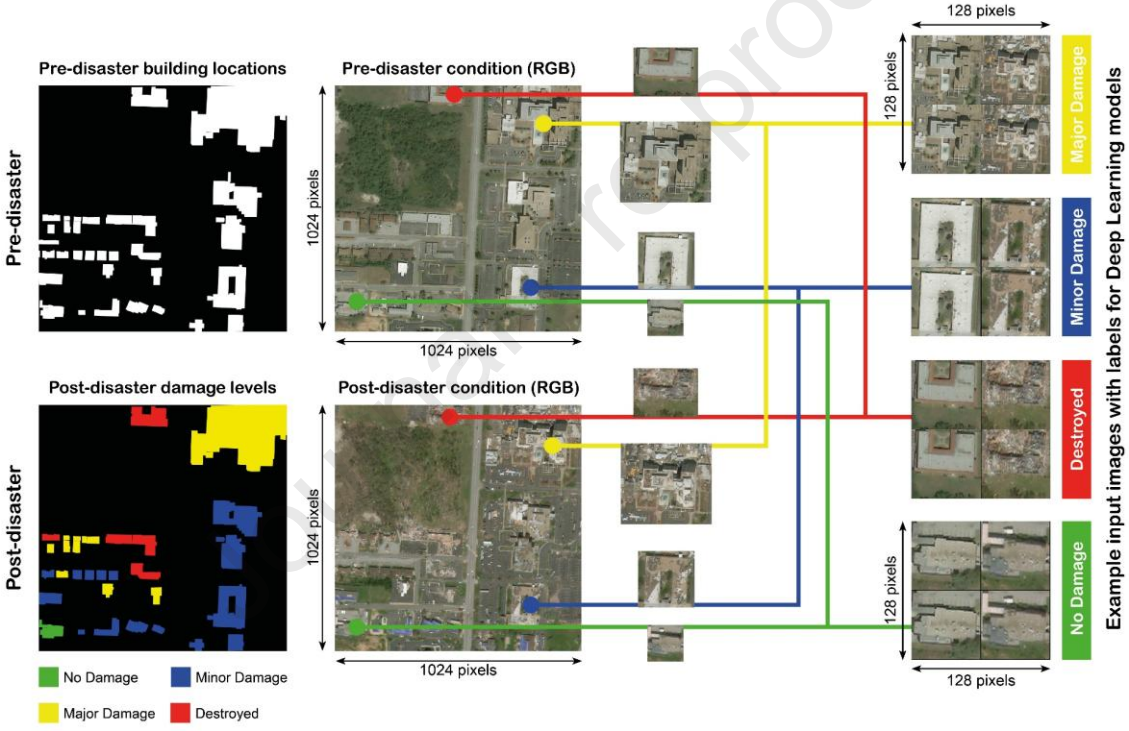

Figure 2: Data preparation workflow illustrating the process of pre- and post-disaster satellite images into 128×128 pixel patches, aligning building locations, and labeling each patch based on its damage classification.

This data preparation approach reduces computational complexity by shifting from pixel-level segmentation to image-level classification, allowing for faster processing of high-resolution images (Csillik, 2017; Wang et al., 2011). It enhances

generalizability by focusing on broader structural features rather than pixel details, and it improves model accuracy by isolating critical damage patterns in each building image. For XAI, this method enables clear and focused visual explanations, making interpretations accessible to non-technical stakeholders.

2.2. Our Proposed DL Models

In this study, our approach leverages eight widely used CNN architectures, that reflect diverse design strategies and have been commonly applied in remote sensing tasks, to evaluate the impact of incorporating attention mechanisms into DL models for PDD. These eight models are InceptionV3 (C. Szegedy et al., 2016), Xception (Chollet, 2017), InceptionResNetV2 (Christian Szegedy et al., 2017), ResNet152 (He et al., 2016), DenseNet169 (Huang et al., 2017), ShallowNetV2, MobileNetV2 (Howard, 2017), and NasNetMobile (Zoph et al., 2018), each of which serves as a base model for feature extraction. These architectures capture different design aims including deep residual and dense networks (ResNet152, DenseNet169), multi-scale inception variants (InceptionV3, InceptionResNetV2), lightweight architectures suitable for rapid deployment (MobileNetV2, NASNetMobile), a separable convolution model (Xception), and a baseline shallow network (ShallowNet). This diversity allows us to examine whether attention and XAI improve model interpretability consistently across networks with distinct strengths and limitations, ensuring that our findings are broadly applicable to remote sensing–based damage detection.

To enhance the focus of these models on critical aspects of the input data, we integrate three types of attention mechanisms: Channel Attention (CA), Spatial Attention (SA), and Multihead Attention (MA). These mechanisms are designed to dynamically prioritize different elements in the feature maps, improving the models' performance,

robustness, and interpretability. In all eight base models, the attention modules were consistently inserted at the end of the feature extraction stage and immediately before the flattening and classification layers. This placement was chosen to ensure that attention operates on the highest-level semantic features, providing a uniform integration point across diverse architectures, reducing computational overhead by acting on compact feature maps, and directly influencing the decision-making process of the classifier. Figure 3 illustrate the architecture of each attention module and Figure 4 demonstrate how these modules are integrated with the base models.

CA is designed to emphasize the importance of certain feature channels over others. It employs both global average pooling and max pooling, followed by a shared multilayer perceptron with two dense layers ($C \rightarrow C/8$ with ReLU activation, then C with sigmoid activation), generating channel weights that rescale the input feature maps. This mechanism dynamically adjusts the weights of each feature channel based on their relevance to the task, allowing the model to focus on the most significant attributes of the input data (Woo et al., 2018). Besides, SA emphasizes the specific regions within input data by concatenating average-pooled and max-pooled feature descriptors across channels and passing them through a 7×7 convolution, which is crucial for accurate classification. By creating attention maps, this mechanism directs the model's attention to spatially relevant areas (Woo et al., 2018). MA applies scaled dot-product attention across multiple heads, implemented with four heads and a key dimension of 64. Queries, keys, and values are projected from the feature maps, attention outputs are concatenated, and residual connections with layer normalization (ε = 1e-6) are applied, ensuring stable training and contextual refinement (Vaswani et al., 2017).

For each of the eight base models, we create three variants by applying one of the attention mechanisms, resulting in a total of 32 unique model configurations. Each

configuration processes input images of size 128x128x3, with a series of layers following the attention mechanism: a flattening layer, dropout layers (with a rate of 0.5) to reduce overfitting, and a final softmax layer for multi-class classification.

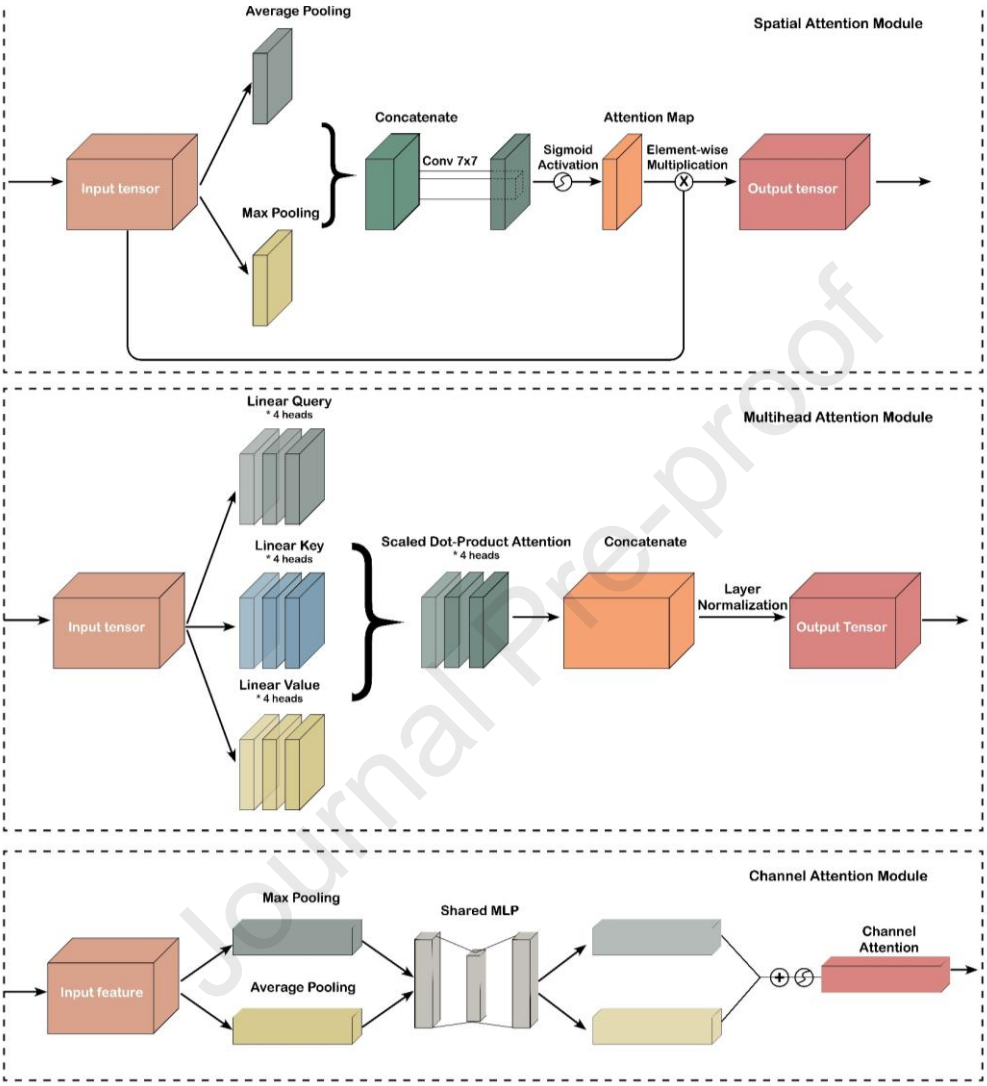

Figure 3: Schematics of the three attention mechanisms in the study: SA, MA, and CA.

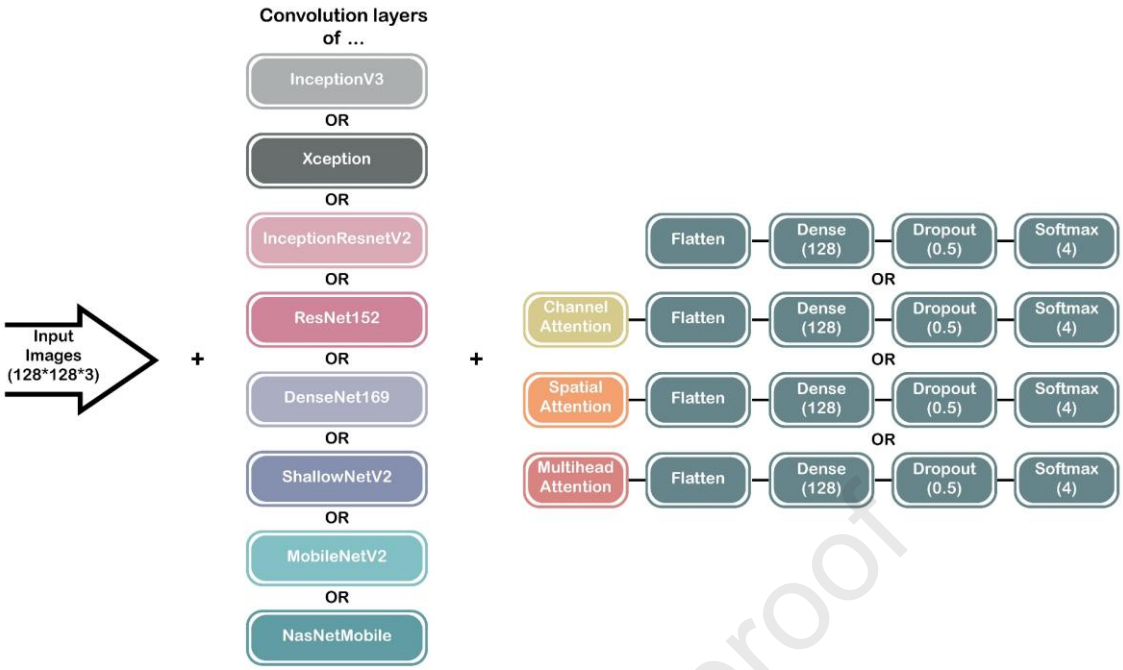

Figure 4: Overview of the 32 model configurations, showing how attention modules are incorporated into eight baseline CNN architectures for PDD.

This study employs transfer learning for the majority of DL models, except ShallowNetV2, leveraging pretrained weights on the ImageNet dataset to enhance the models' ability to extract meaningful features from post-disaster imagery (J. Deng et al., 2009). Using pretrained models on large-scale datasets like ImageNet allows these architectures to start with a solid foundation of feature recognition, which is particularly beneficial in domains with limited labeled data, such as post-disaster damage detection (Zhuang et al., 2021). ImageNet is a large-scale dataset containing over 14 million labeled images across thousands of categories, widely used to pretrain deep learning models for image recognition tasks (Arbulu & Ballard, 2004; Krizhevsky et al., 2012; Christian Szegedy et al., 2017).

In addition to these pretrained architectures, we developed a custom model, ShallowNetV2, specifically designed and trained from scratch for this study. ShallowNetV2's architecture, as illustrated in Figure 5. ShallowNetV2 is a simple

convolutional neural network with four convolutional blocks, each consisting of a 3×3 convolutional layer (filters = 32, 64, 128, and 256, respectively) with ReLU activation and same padding, followed by 2×2 max-pooling. After feature extraction, the network includes a flattening operation, a fully connected dense layer with 128 units (ReLU activation), a dropout layer (rate = 0.5) to reduce overfitting, and a final softmax output layer with four units corresponding to the damage classes. Several configurations were initially explored (e.g., varying the number of convolutional blocks, filter sizes, and dense units), and the final design of ShallowNetV2 was selected based on training and validation performance. By including ShallowNetV2 alongside pretrained models, we aim to assess the comparative performance of a simpler, customized architecture against more complex, pretrained models.

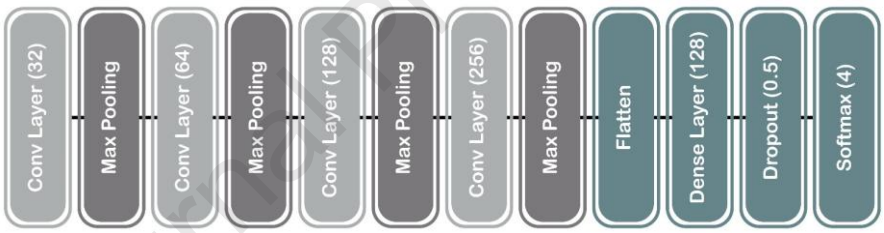

Figure 5: Architecture of base ShallowNetV2 model.

2.3. Evaluation Metrics and Training Procedure

To assess model performance, we utilized the following metrics: accuracy, precision, recall, F1 score and the confusion matrix. These metrics are well-suited for DL-based classification tasks and provide a thorough understanding of model effectiveness in the context of PDD (A. M. Braik & M. Koliou, 2024; Lu et al., 2024; Wiguna et al., 2024).

Accuracy represents the overall proportion of correct predictions across all classes, calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (E.1)$$

where *TP* and *TN* indicate true positives and true negatives, respectively, while *FP* and *FN* denote false positives and false negatives.

Precision indicates the ratio of correctly identified positive predictions to the total positive predictions, reflecting the model's reliability in identifying actual damage cases.

$$Precision = \frac{TP}{TP + FP} * 100 \quad (E.2)$$

Recall measures the proportion of actual positive cases that the model correctly identifies, capturing the model's ability to detect damage when it truly exists.

$$Recall = \frac{TP}{TP + FN} * 100 \quad (E.3)$$

F1-score provides a balanced measure that combines precision and recall into a single metric by calculating their harmonic mean.

$$F1\ Score = \frac{Precision * Recall}{Precision + Recall} * 100 \quad (E.4)$$

Additionally, we utilized a confusion matrix to provide a detailed breakdown of the model's classification across each damage category, offering insights into its performance in distinguishing between levels of damage severity.

We trained the models using the Adam optimizer with a cross-entropy loss function, a common choice for multi-class classification (Ahmadi et al., 2024; Wanting Yang et al., 2021). Hyperparameters, including learning rate, batch size, and number of epochs, were fine-tuned through a grid search to maximize model performance on the validation set. Specifically, we experimented with the following values:

- Learning Rates: 0.01, 0.005, 0.001, 0.0005, and 0.0001.
- Batch Sizes: 16, 32, 64, and 128.
- Number of Epochs: 50, 100, and 200.

The optimal configuration determined was a learning rate of 0.001, a batch size of 64, and 200 epochs, balancing accuracy with computational efficiency. To further

enhance model robustness, we applied 5-fold cross-validation, reporting the average performance across folds to ensure reliable and unbiased results. To prevent overfitting, early stopping was applied based on validation loss, stopping the training once improvements plateaued. After training and validation, the models were evaluated on the test set, with accuracy, precision, recall, and the confusion matrix used as final performance metrics, ensuring robust evaluation.

2.4. Explainable AI

Model transparency is crucial for fostering trust and enabling informed, accountable decisions among stakeholders. To improve the interpretability of our models, we implement two widely recognized Explainable AI (XAI) methods: Saliency Maps and Grad-CAM. They are particularly effective for image-based tasks, as they visually indicate which parts of an image most significantly impact the model's predictions, enhancing our understanding of the decision-making process (Weber et al., 2023). These XAI methods provide insights into whether the model's attention aligns with human judgment regarding post-disaster images. In high-stakes disaster management, this interpretability is essential, as it enables first responders, decision-makers, and other stakeholders to make more transparent, confident decisions based on a clear understanding of the model's reasoning (Ghaffarian et al., 2023; Kakogeorgiou & Karantzalos, 2021).

Saliency Maps utilize gradients to identify the most critical pixels influencing the model's output for a given input image. This technique computes the gradient of the class score concerning each input pixel, thereby revealing the areas that contribute most to the model's classification decisions. This is invaluable in disaster scenarios, where it is

essential to know precisely which image regions, such as damaged buildings or collapsed infrastructure, are driving the model's predictions.

For a particular class c, the Saliency Map S is derived as:

$$S = \left|\frac{\partial y^c}{\partial I}\right| \qquad \text{(E. 3)}$$

where $\frac{\partial y^c}{\partial I}$ represents the gradient of the class score y^c in relation to the input image I. The absolute values of these gradients highlight key pixels, allowing us to visualize which regions significantly influence the outcome (Simonyan, 2013).

Grad-CAM is another technique used to visualize model decision-making but focuses on high-level feature maps within the model's final convolutional layers, producing a heatmap that emphasizes the areas relevant to a specific class prediction. Unlike Saliency Maps, which operate at the pixel level, Grad-CAM provides a broader perspective, highlighting entire regions within the image that the model finds important for a given class.

To calculate the Grad-CAM heatmap for a class c, we begin by obtaining the gradients of the class score y^c with respect to the feature maps A^k in a chosen convolutional layer. These gradients are averaged over spatial dimensions to produce weights α_k^c:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \qquad \text{(E. 4)}$$

where Z is the total number of pixels in the feature map. We then compute a weighted combination of the feature maps, followed by a ReLU activation to create the Grad-CAM heatmap:

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha_k^c A^k\right) \qquad (\text{E. } 5)$$

Applying ReLU ensures that only positive influences on the class score are visualized, resulting in a heatmap that highlights the image regions most pertinent to the model's prediction (Selvaraju et al., 2017).

3. Results and Discussion

3.1. Performance Comparison of the models

This section presents the performance of the proposed base models and their configurations using different attention mechanisms. To streamline the discussion, models are prefixed based on their configuration: Base_ for the original model without attention mechanisms, CA_ for models with CA, SA_ for those with SA, and MA_ for those with MA. For example, Base_DenseNet169 represents the original DenseNet169 model, while MA_ResNet152 indicates the ResNet152 model configured with MA. In

Table 2, blue text highlights the configurations with the highest accuracy among variations of the same base model, while red text indicates the configurations achieving the highest precision.

The results shows that Base_InceptionV3 and Base_Xception exhibited consistently high accuracy, achieving 81% and 80.5%, respectively, and high precision, reaching 80.8% and 80.6%, respectively, among the eight base models. These findings highlight the robustness of these architectures even without additional enhancements from attention mechanisms. However, the integration of attention mechanisms significantly improved both accuracy and precision for most models, emphasizing their potential to refine the performance of deep learning models for PDD.

Table 2: Performance comparison of deep learning models with different attention mechanisms for PDD, (Pre: Precision, Acc: Accuracy, Rec: Recall, F1-S: F1 Score). The highest accuracy achieved within each base model configuration is highlighted in blue, the highest precision is highlighted in red, the highest recall is highlighted in green, the highest f1 score is highlighted in purple, and the overall best performance across all metrics is highlighted in yellow.

Model Name	Transfer Learning	Base Model				Channel Attention				Spatial Attention				Multihead Attention			
		Pre	Acc	Rec	F1-S	Pre	Acc	Rec	F1-S	Pre	Acc	Rec	F1-S	Pre	Acc	Rec	F1-S
DenseNet169	✓	79.5	79.2	79.2	79.3	81.3	**80.9**	**80.9**	**81.1**	79.5	79.4	79.4	79.4	**81.6**	80.5	80.5	81.0
InceptionV3	✓	80.8	81.0	81.0	80.9	81.1	81.0	81.0	81.1	80.8	81.0	81.0	80.9	**82.9**	**81.9**	**81.9**	**82.4**
InceptionResNetV2	✓	80.5	**80.2**	**80.2**	**80.3**	79.9	78.3	78.3	79.1	79.2	79.2	79.2	79.2	**80.9**	79.7	79.7	**80.3**
MobileNetV2	✓	74.8	75.0	75.0	74.9	75.6	73.8	73.8	74.7	77.5	77.7	77.7	77.6	**83.1**	**82.5**	**82.5**	**82.8**
NasNetMobile	✓	69.1	69.1	69.1	69.1	73.3	72.3	72.3	72.8	68.5	67.7	67.7	68.1	**78.7**	**77.7**	**77.7**	**78.2**
ResNet152	✓	78.7	78.3	78.3	78.5	78.9	78.8	78.8	78.9	**79.6**	79.0	79.0	**79.3**	78.6	**79.2**	**79.2**	78.9
ShallowNetV2	-	73.0	69.9	69.9	71.5	79.7	79.7	79.7	79.7	**81.0**	**80.8**	**80.8**	**80.9**	80.6	80.0	80.0	80.3
Xception	✓	80.6	80.5	80.5	80.6	80.9	79.7	79.7	80.3	**81.9**	**81.2**	**81.2**	**81.6**	81.1	79.7	79.7	80.4

Among the attention mechanisms, MA consistently produced the best results, particularly with models like MobileNetV2 and NasNetMobile. For example, MA_MobileNetv2 achieved an accuracy of 82.5% and a precision of 83.1%, surpassing its base model by substantial margins of 7.5% in accuracy and 8.3% in precision. While SA did not consistently outperform the other mechanisms, it was particularly beneficial for models like ShallowNetV2, improving its accuracy from 69.9% (Base_ ShallowNetV2) to 80.8% (SA_ ShallowNetV2), demonstrating the role of SA in emphasizing critical regions within the images. CA also demonstrated its effectiveness, especially for feature-rich models like DenseNet169. The CA_DenseNet169 configuration improved both accuracy and precision compared to its base variant, achieving 80.9% accuracy and 81.3% precision.

Overall, MA_MobileNetV2 achieved the best accuracy (82.5%) and precision (83.1%) performances, among 32 model configurations, followed by MA_InceptionV3 with an accuracy of 81.9% and precision of 82.9%. Despite a transformative potential of attention mechanisms in improving model performances, configurations of InceptionResNetV2 demonstrated only marginal improvements, suggesting that the choice of base model plays a crucial role in determining the effectiveness of attention mechanisms. Another interesting finding is that ShallowNetV2 performed competitively despite being a custom-designed model without pretrained weights.

A critical aspect of PDD is accurately identifying buildings with major damage and those that are destroyed, as these classes represent the highest severity levels requiring urgent response (Shen et al., 2022; Wu et al., 2021). Misclassification of these categories could delay relief efforts and resource allocation, emphasizing the need for models capable of precise predictions in these critical classes (Kaur et al., 2023). Figure

6, the detailed confusion matrix, reveals that the integration of MA significantly improved the prediction accuracy for major damage and destroyed classes across several models. For example, in MA_InceptionV3, the destroyed class was predicted with an impressive accuracy of 87.6%, while the major damage class improved to 80.5%.

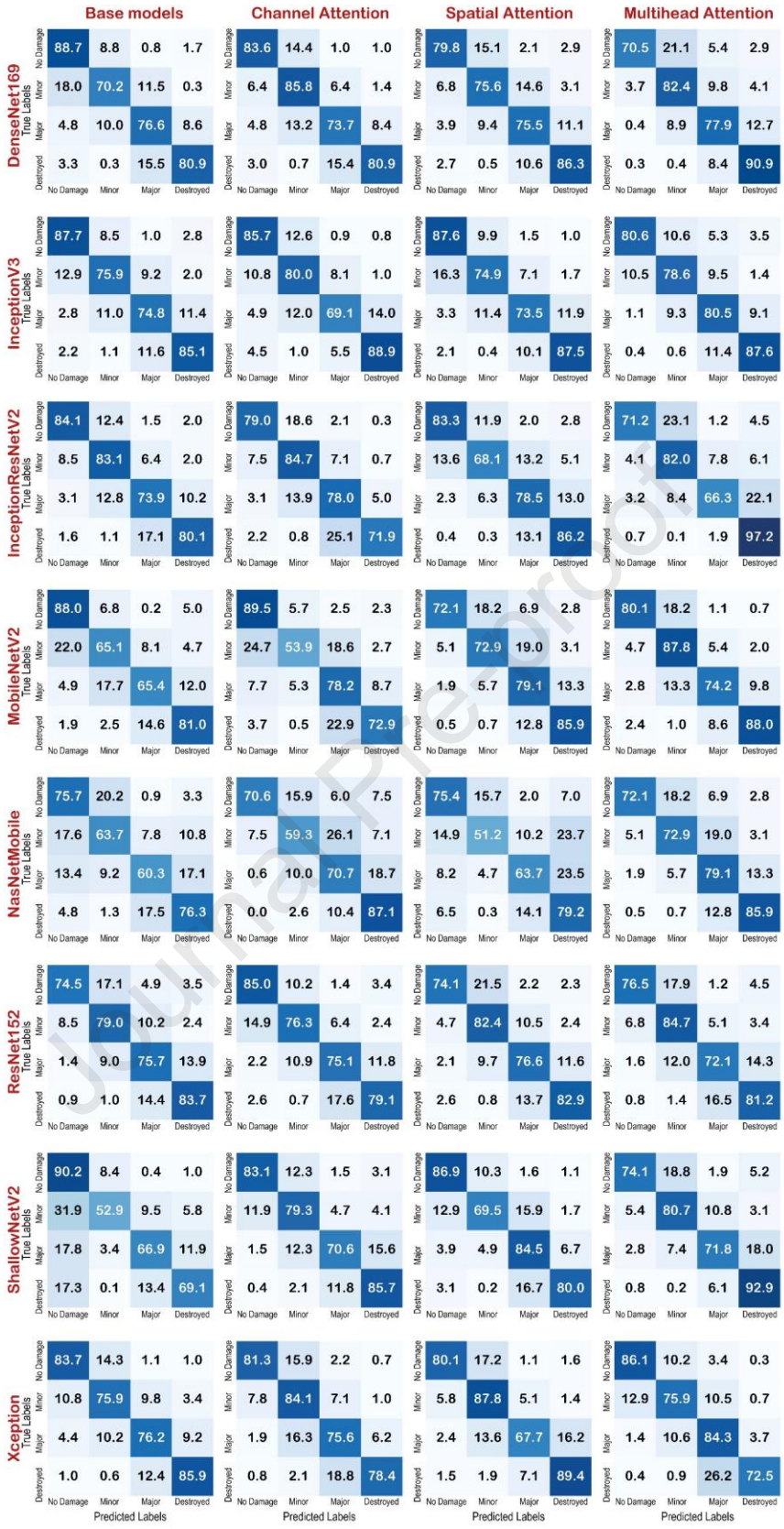

Figure 6: Confusion matrices showing the classification performance of the base models and their variants with attention mechanisms for the four damage classes.

Notably, SA_ShallowNetV2 and MA_InceptionV3 emerged as the only models capable of predicting both the major damage and destroyed classes with over 80% accuracy, achieving a balance between high accuracy for severe damage categories and overall robustness. Figure 7 reveals the predictions made by each base model and its attention-augmented variants (CA, SA, and MA) on a sample post-disaster image Across all base models, the most common trend is a systematic underestimation of severity, with frequent confusions between adjacent categories—most notably between major damage and destroyed, or between minor damage and no damage. For example, Base_MobileNetV2 and Base_NasNetMobile misclassify large proportions of destroyed structures as major damage, and minor damage as no damage, reflecting a bias toward underestimating severity. The introduction of attention mechanisms reduces these systematic errors, particularly Spatial and Multihead Attention, which sharpen class boundaries and improve the separation of high-severity categories. For instance, MA_InceptionV3 and MA_ShallowNetV2 exhibit precise identification of destroyed structures, closely aligning with the ground truth. The figure further demonstrates performance variations across architectures, revealing that base models such as Base_NasNetMobile and Base_MobileNetV2 struggle to differentiate between high-severity damage levels without attention mechanisms. However, their accuracy improves considerably when attention is incorporated, as seen in MA_NasNetMobile and MA_MobileNetV2. Consistency in predictions is another critical observation, with attention-enhanced models such as MA_InceptionV3 and SA_ShallowNetV2 consistently predicting major damage and destroyed classes with higher accuracy. These findings align with earlier results showing that these configurations achieve over 80% accuracy for these critical categories. Finally, the visualization underscores the operational relevance of attention mechanisms, as stakeholders can easily verify model

predictions by cross-referencing the visual outputs with actual damage patterns. This aspect highlights the importance of transparency and interpretability in post-disaster scenarios, ensuring the reliability and usability of the models in real-world applications.

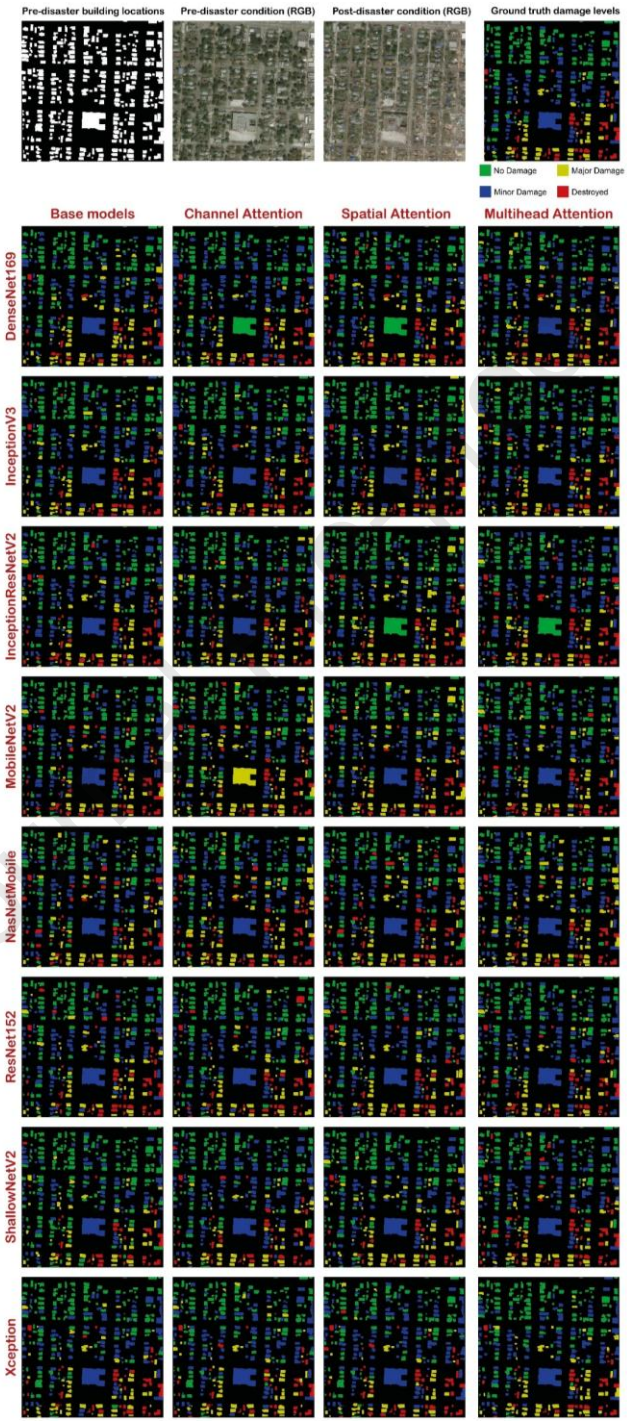

Figure 7: Visual comparison of predictions made by base models and their variants with attention mechanisms on a sample post-disaster image.

These results highlight the crucial potential of attention mechanisms, particularly MA, in improving model focus on high-severity damage classes. By enhancing the models' ability to correctly identify major damage and destroyed structures, these configurations align well with the priorities of disaster response efforts, ensuring more efficient resource allocation and faster decision-making during critical phases. This underscores the importance of designing models not just for overall accuracy but with a targeted emphasis on the classes that hold the greatest operational significance in post-disaster contexts.

3.2. Reliability assessment of the models with XAI

The reliability of the proposed DL models for PDD were evaluated using Grad-CAM and Saliency Map techniques. These XAI methods provided visual insights into how each model, including their attention-augmented variants, focused on critical features while predicting specific damage level classes. To ensure a fair and comprehensive analysis, the example images from each damage class presented in this study are those that 1) all 32 models predicted correctly and 2) no model predicted correctly (Figure 8). These representative examples were selected to illustrate consistent patterns observed across the broader dataset, rather than isolated cases, and therefore provide overall insights into model behavior and limitations. The analysis demonstrated the impact of attention mechanisms on enhancing the explainability and reliability of the models, particularly for high-severity damage classes.

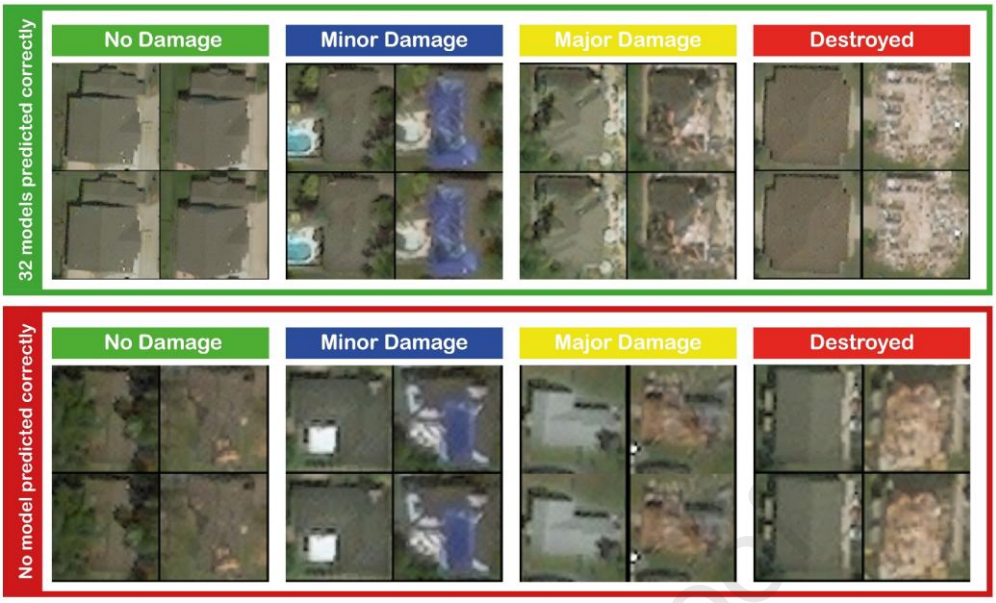

Figure 8: Example images used in XAI visualizations, illustrating the four damage categories for 1) all 32 models predicted correctly, and 2) no model predicted correctly.

In this study, interpretability was assessed at two complementary levels. Global reliability reflects whether the models focus on the correct areas of interest across the dataset as a whole, indicating their overall decision-making behaviour. For example, for "Major Damage" or "Destroyed" categories, a globally reliable model is expected to focus pre- and post-disaster images and compare them side by side by especially concentrating on buildings where structural changes are visible, rather than unrelated regions. Local reliability relates to how well individual predictions are explained at the pixel or region level, namely whether the model highlights the specific damaged parts of a building within a single image. To capture these two perspectives, Grad-CAM was mainly employed to evaluate global behaviour, as it illustrates where the model attends when forming its predictions, while Saliency Maps were used to examine local reliability, as they indicate which pixels are most influential for a decision. Using these complementary methods allowed us to assess both the overall logic of the models and the detailed evidence behind specific predictions.

Figure 9 illustrates the Grad-CAM results for all 32 model configurations, showing the focus of base models and their attention-augmented variants on the "Destroyed" class. The figure reveals that base models such as Base_InceptionV3, Base_ShallowNetV2, and Base_MobileNetV2 struggle to localize severely damaged areas, as their heatmaps display scattered or diffused attention. These models fail to consistently highlight regions of structural damage, indicating limitations in extracting relevant features for high-severity classes. The addition of attention mechanisms improved model focus in general, but their effectiveness varied.

In some cases, CA-augmented models, such as CA_DenseNet169 and CA_ResNet152, produced heatmaps that primarily emphasized pre-disaster image regions rather than post-disaster damage. This misplaced focus suggests that CA, while effective at prioritizing feature channels, struggles to integrate spatial or temporal context effectively. For example, in CA_ShallowNetV2, the model highlighted areas in the pre-disaster image with little connection to the visible damage in the post-disaster image, reducing its reliability for interpreting structural changes. In contrast, SA and MA consistently enhanced model focus on critical areas in models such as SA and MA variations of InceptionV3, ShallowNetV2, and NasNetMobile, effectively combined spatial and temporal information from pre- and post-disaster images. Their Grad-CAM heatmaps demonstrated precise and well-localized attention on regions of significant change, ensuring contextually accurate predictions. These models excelled at correlating structural changes observed in the post-disaster image with their pre-disaster state, enhancing interpretability and reliability. However, as shown in Figure 9, even high-performing configurations like MA_MobileNetV2 occasionally misplaced focus on visually distinct regions in the post-disaster image, leading to potential unreliability.

Figure 10 illustrates the Grad-CAM results for all 32 model configurations, showing the focus of base models and their attention-augmented variants on the "Major Damage" class. The base models display weak and diffuse activation patterns, often spreading attention across large portions of the image or focusing on irrelevant regions such as Base_InceptionV3, Base_ShallowNetV2, and Base_MobileNetV2. This inconsistent focus makes it difficult to reliably identify the areas of partial roof or wall collapse that define major damage, underscoring their limited ability to isolate features critical for this class. In contrast, attention-augmented models generally display stronger and more localized activation on damaged building regions, and more effectively compare pre- and post-disaster images. This was particularly evident in SA and MA variations of InceptionV3 and ShallowNetV2, where attention maps concentrated on structurally relevant regions. Similar to the "Destroyed" class, these models appear more reliable across other variations by combining spatial and contextual focus, producing more stable and interpretable activations.

Figure 11and Figure 12 present Grad-CAM visualizations for the "Destroyed" and "Major Damage" classes where none of the models produced correct predictions, respectively. These heatmaps reveal that both base and attention-augmented models fail to concentrate on the critical damaged areas. Instead, their attention is either scattered across large portions of the image or misplaced on undamaged regions and pre-disaster structures. In contrast, correctly predicted samples of the same classes showed sharper and more localized focus on collapsed roofs, missing walls, and other key structural changes, particularly in SA- and MA-augmented models. This pattern highlights a recurring source of error that misclassifications might be often associated with diffuse or misplaced attention maps, underscoring the limits of current models in consistently capturing ambiguous or complex damage cues, especially in lower resolution images.

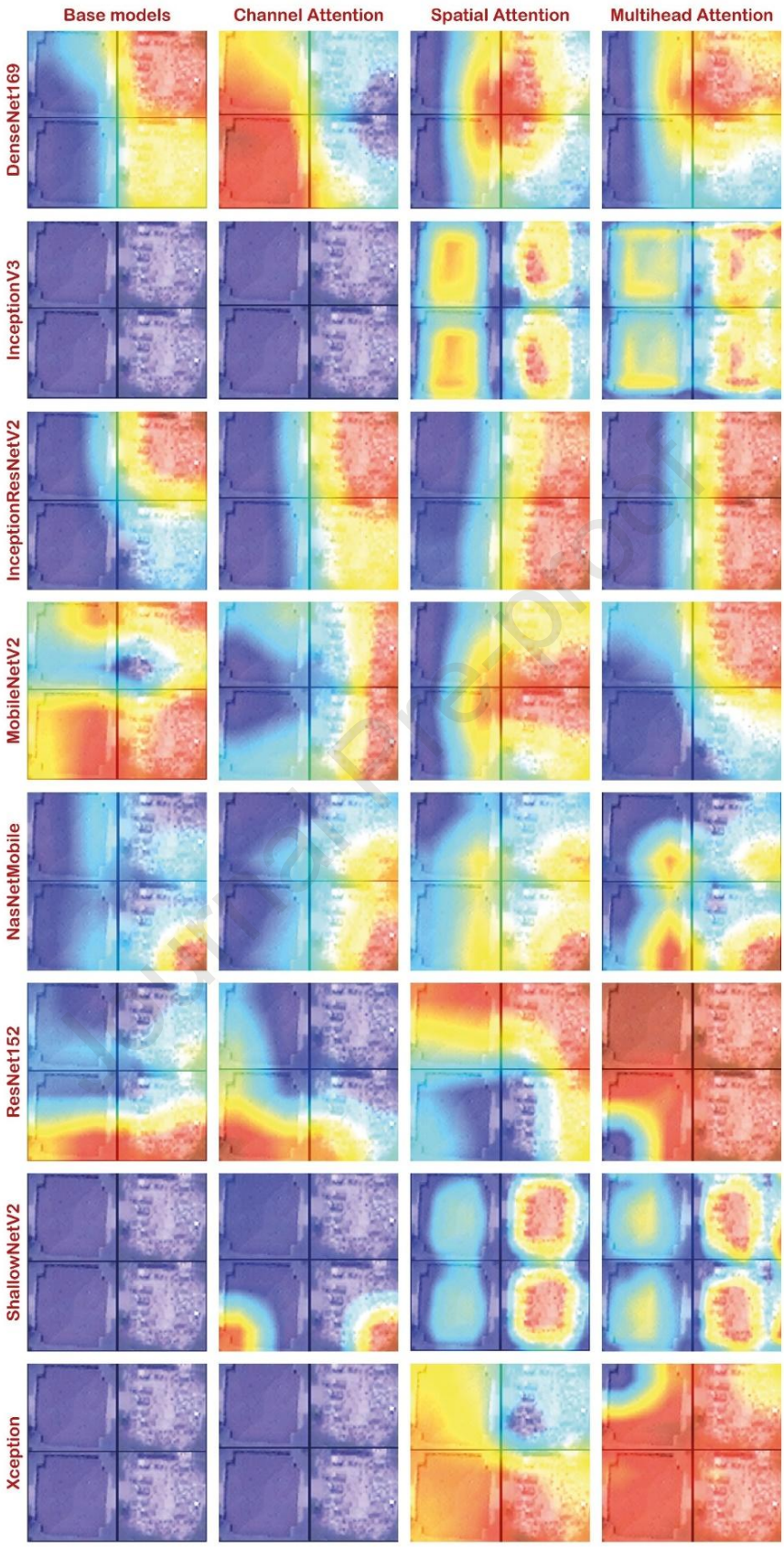

Figure 9: Grad-CAM visualizations for the "Destroyed" class, comparing base models and their attention-augmented variants (all 32 models predicted correctly).

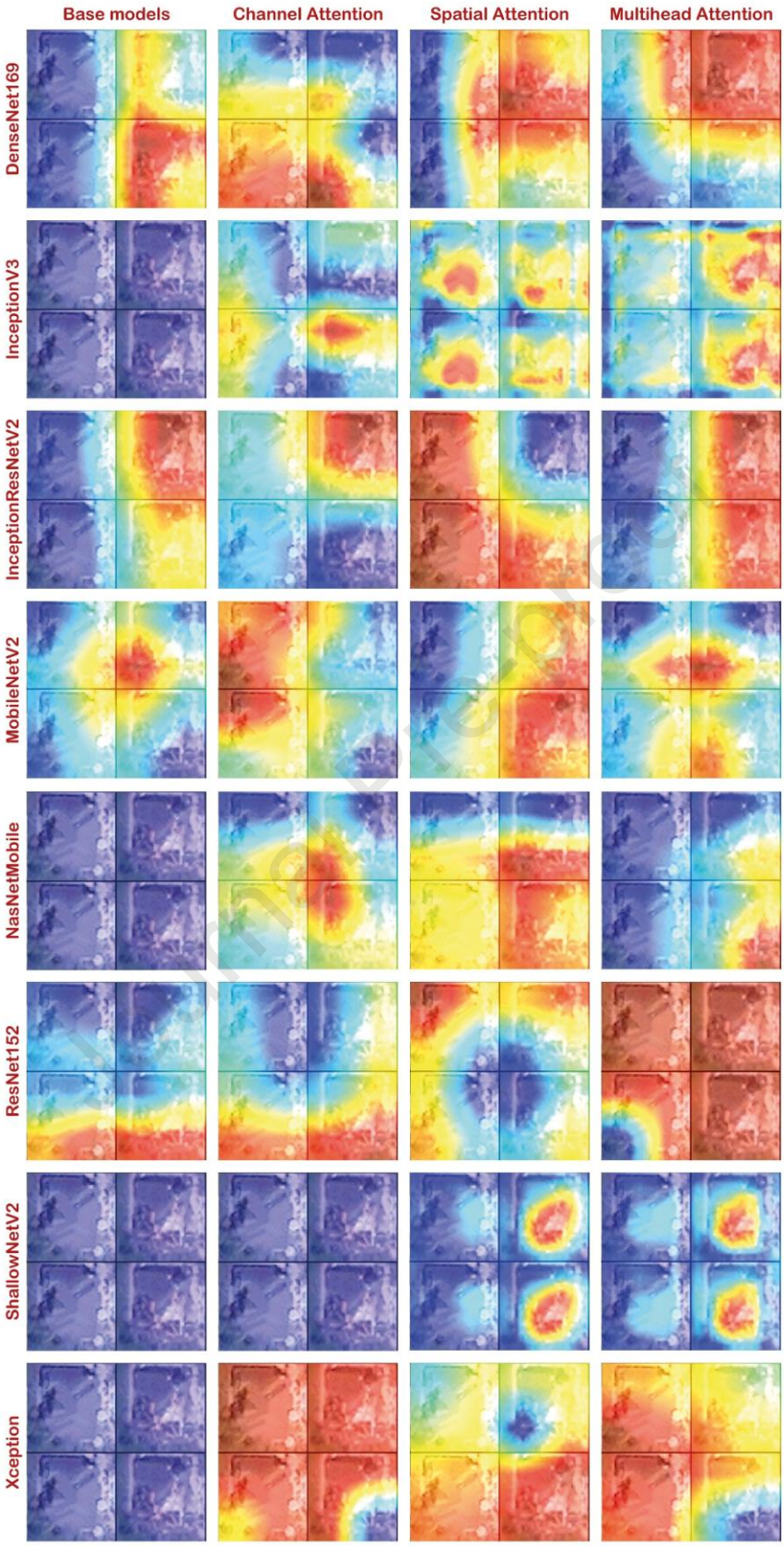

Figure 10: Grad-CAM visualizations for the "Major damage" class, comparing base models and their attention-augmented variants (all 32 models predicted correctly).

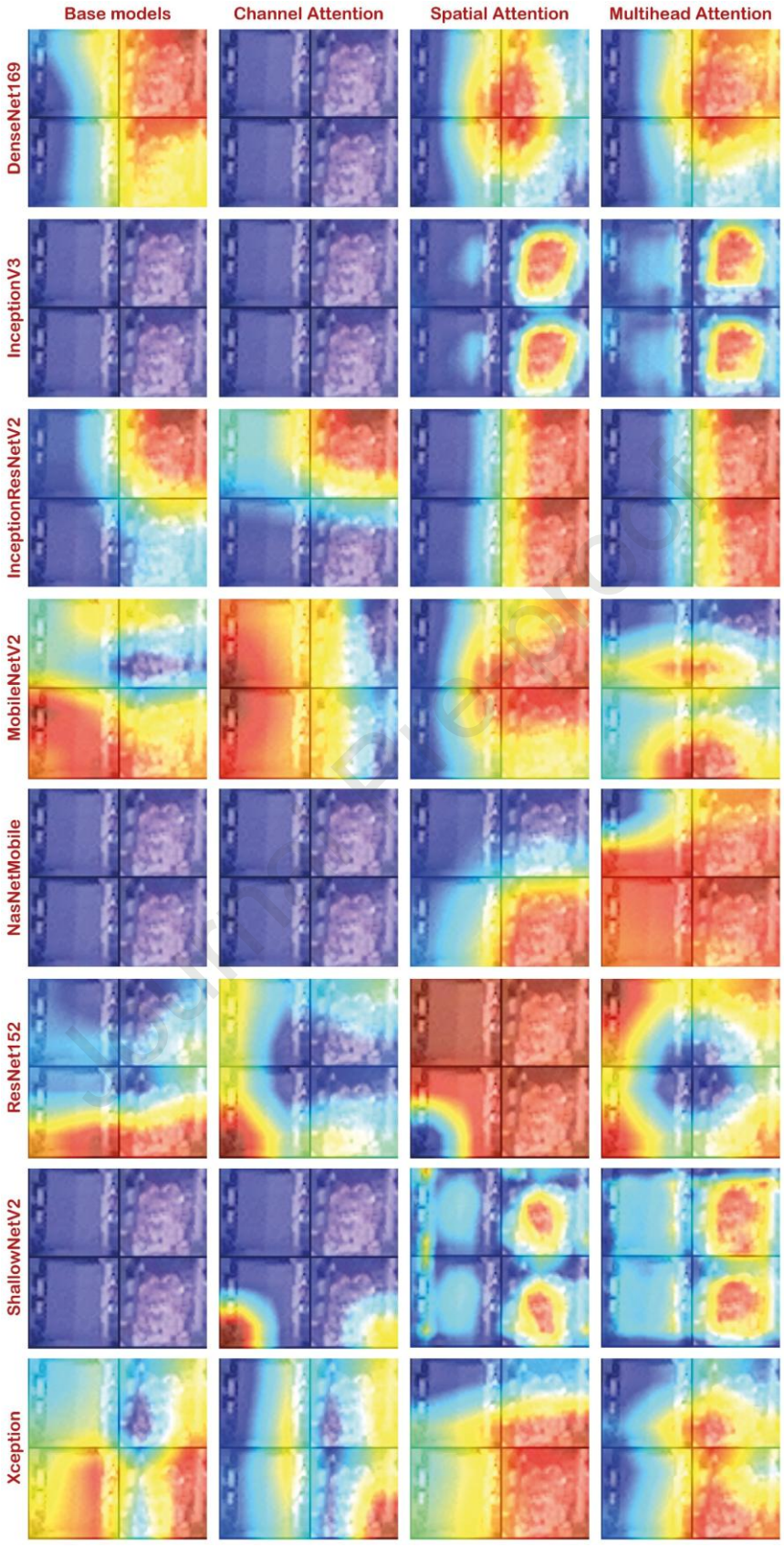

Figure 11: Grad-CAM visualizations for the "Destroyed" class, comparing base models and their attention-augmented variants (no model predicted correctly).

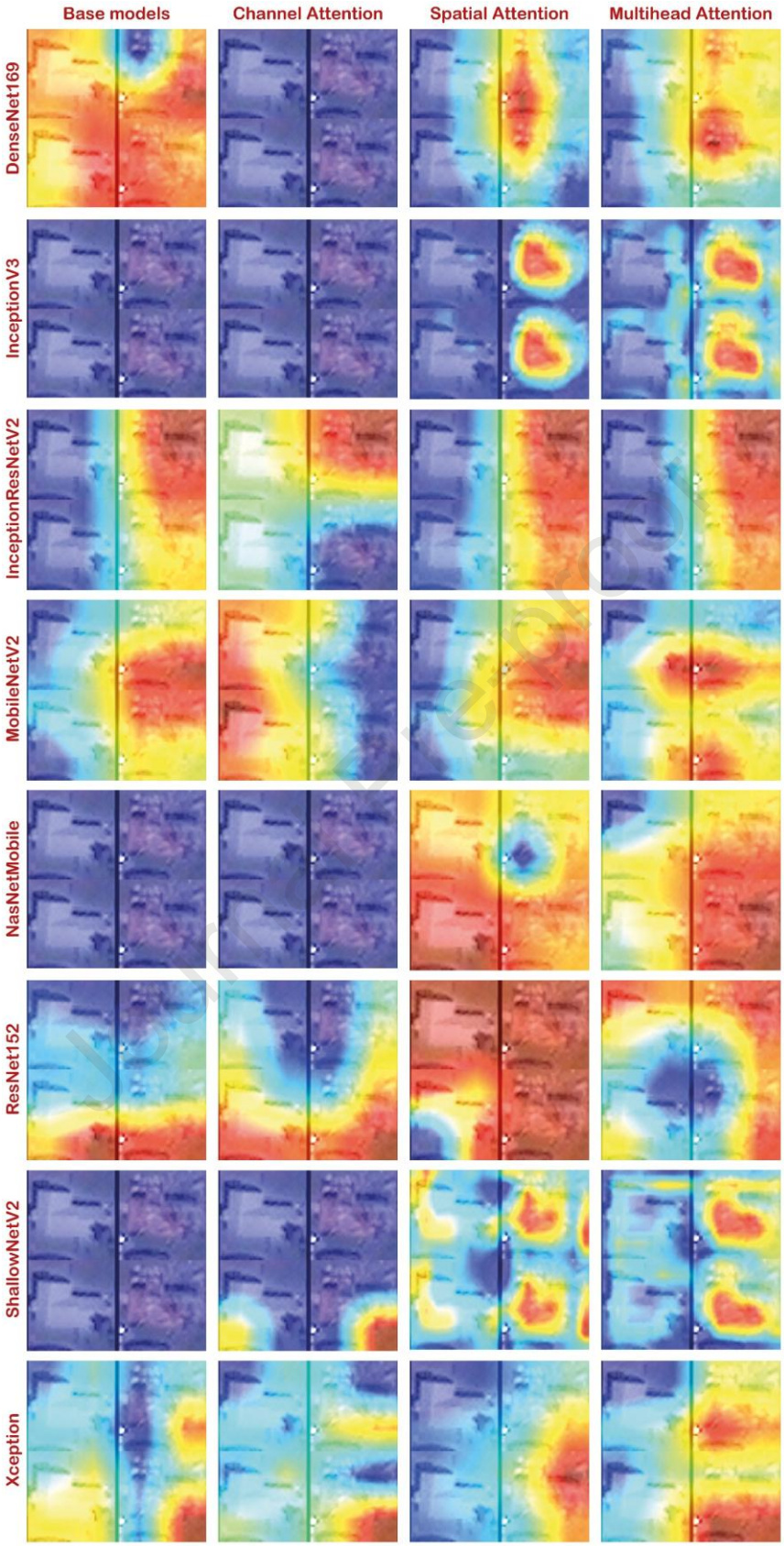

Figure 12: Grad-CAM visualizations for the "Major damage" class, comparing base models and their attention-augmented variants (no model predicted correctly).

Figure 13 represents the Saliency Map analysis providing critical pixel-level insights into the decisions made by the models for the "Destroyed" class. The reliability hinges on two key factors: the ability to focus on damaged areas and avoiding overly diffuse saliency patterns that attribute importance to irrelevant or excessive pixels. Building on the Grad-CAM analysis, which identified SA and MA variations of InceptionV3, ShallowNetV2, and NasNetMobile as the most reliable models for predicting the "Destroyed" class, we conducted a further evaluation using Saliency Maps. While Grad-CAM provided insights into where the models focused spatially, Saliency Maps offered a pixel-level perspective, highlighting the most influential pixels driving the models' decisions. This complementary analysis aimed to validate the reliability of these models and uncover any limitations.

For SA_NasNetMobile and MA_NasNetMobile, the Saliency Map revealed a significant limitation: a diffused focus across the image, with nearly all pixels marked as influential. Instead of isolating the most critical pixels corresponding to damaged areas, the model attributed importance to large portions of the post-disaster image, including undamaged regions. This behavior undermines the interpretability of the model's predictions, as the diffused saliency pattern does not provide clear evidence of a specific decision-making process. Besides, both SA_ShallowNetV2 and SA_InceptionV3 exhibited problems in localizing logically damaged areas. Their Saliency Maps frequently highlighted regions that did not correspond to visibly damaged structures in the post-disaster image. Instead of concentrating on critical damaged areas, these models produced saliency patterns that were either diffused or misplaced, emphasizing irrelevant regions with no connection to actual destruction. However, MA_ShallowNetV2 and MA_InceptionV3 emerged as the only consistently reliable models based on Saliency

Map analysis. Both configurations demonstrated sharp and well-localized saliency patterns concentrated on pixels corresponding to areas of severe structural damage in the post-disaster image. These models not only avoided diffused focus but also aligned their influential pixels closely with visible damage, making their decision-making process both interpretable and trustworthy. The integration of MA allowed these models to effectively balance spatial and temporal context, ensuring accurate predictions.

Figure 14 shows the Saliency Map analysis for the "Major Damage" class in cases where models produced correct predictions. Compared to the "Destroyed" class, reliable identification of major damage requires finer discrimination, as the damage is often partial and less visually obvious. The image showed that many base and attention-augmented models produced diffuse or misplaced saliency patterns, failing to isolate the damaged pixels; however, the MA_ShallowNetV2 model stood out by showing a consistent alignment between Grad-CAM and Saliency Maps. In this case, the global attention (Grad-CAM) highlighting damaged roof regions coincided with the pixel-level attribution (Saliency), indicating that the model not only looked at the right area but also relied on the correct pixels when making its decision. This consistency across two interpretability methods was not observed in other architectures, where Grad-CAM and Saliency Maps often emphasized different or overly broad regions. These results reinforce the finding that reliable predictions depend on both global and local focus, with MA_ShallowNetV2 emerging as the most trustworthy configuration for major damage detection.

Figure 15 and Figure 16 show Saliency Map analyses for "Destroyed" and "Major Damage" cases where all models produced incorrect predictions, respectively. Unlike the correctly classified samples, these visualizations reveal diffuse, misplaced, or in some cases almost absent pixel attributions. Many models scattered their focus across irrelevant

regions of the image, while others (notably MA_ShallowNetV2) showed very little activation at all, indicating that the decision was made without reliance on meaningful visual cues. This lack of concentrated or interpretable saliency contrasts sharply with the successful cases, where reliable models aligned their pixel-level focus with damaged roofs, collapsed walls, or other structural features. These failed examples therefore highlight two recurring sources of error: diffuse or misplaced focus, and in some cases, the near-complete absence of focus, both of which might undermine interpretability and point to the difficulty of detecting ambiguous or subtle damage patterns in post-disaster imagery.

These differences highlight the importance of analyzing models not only through Grad-CAM but also at a more granular level using Saliency Maps. While MA_ShallowNetV2 and MA_InceptionV3 remain reliable in both Grad-CAM and Saliency Map evaluations, SA_ShallowNetV2 and SA_InceptionV3 struggle with misplaced pixel focus, and SA_NasNetMobile and MA_NasNetMobile are undermined by diffused saliency patterns. These findings emphasize the need for refined attention mechanisms that balance spatial coherence with pixel-level precision, ensuring reliable and interpretable predictions in post-disaster damage detection tasks.

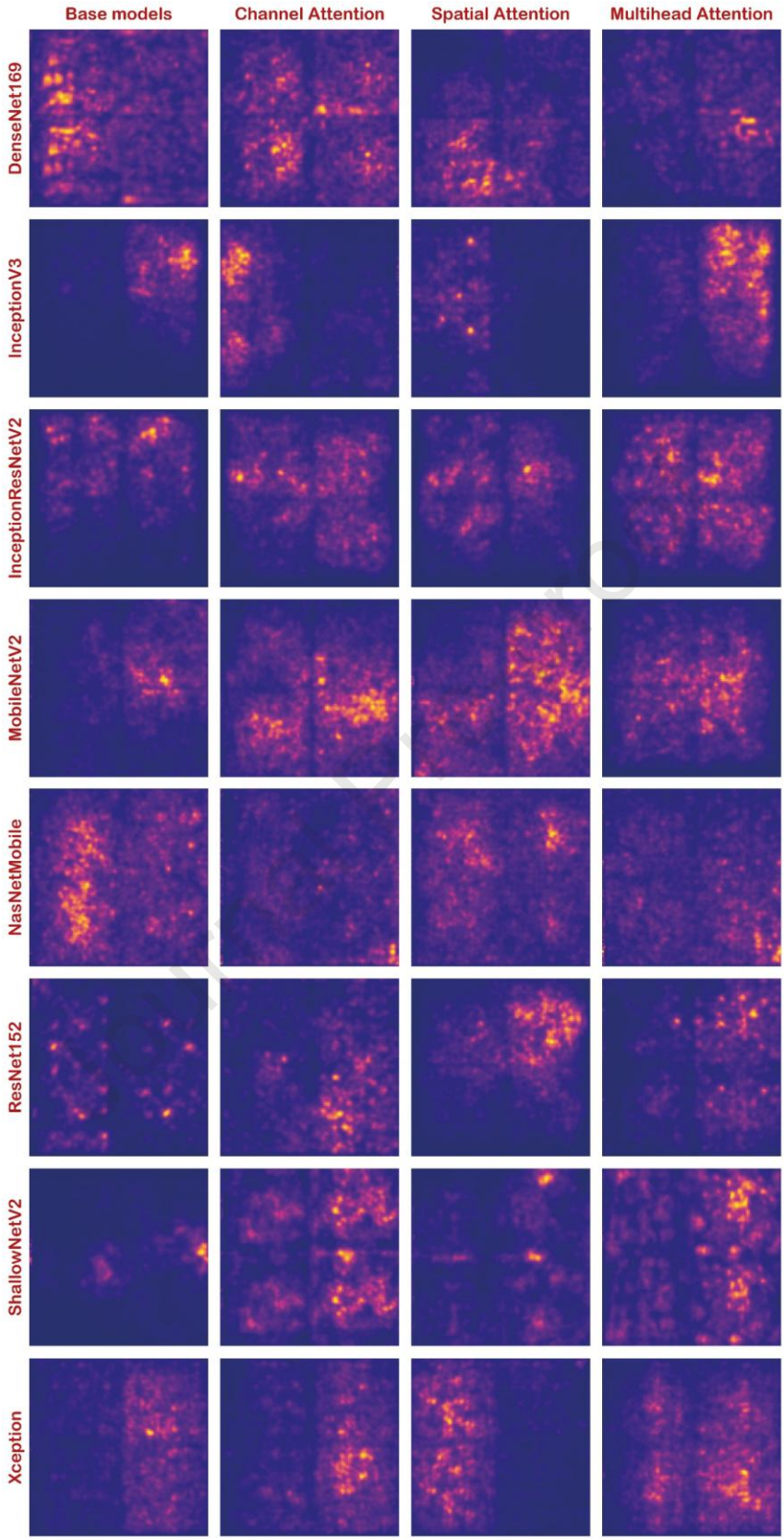

Figure 13: Saliency Map visualizations for the "Destroyed" class, comparing base models and their attention-augmented variants (all 32 models predicted correctly).

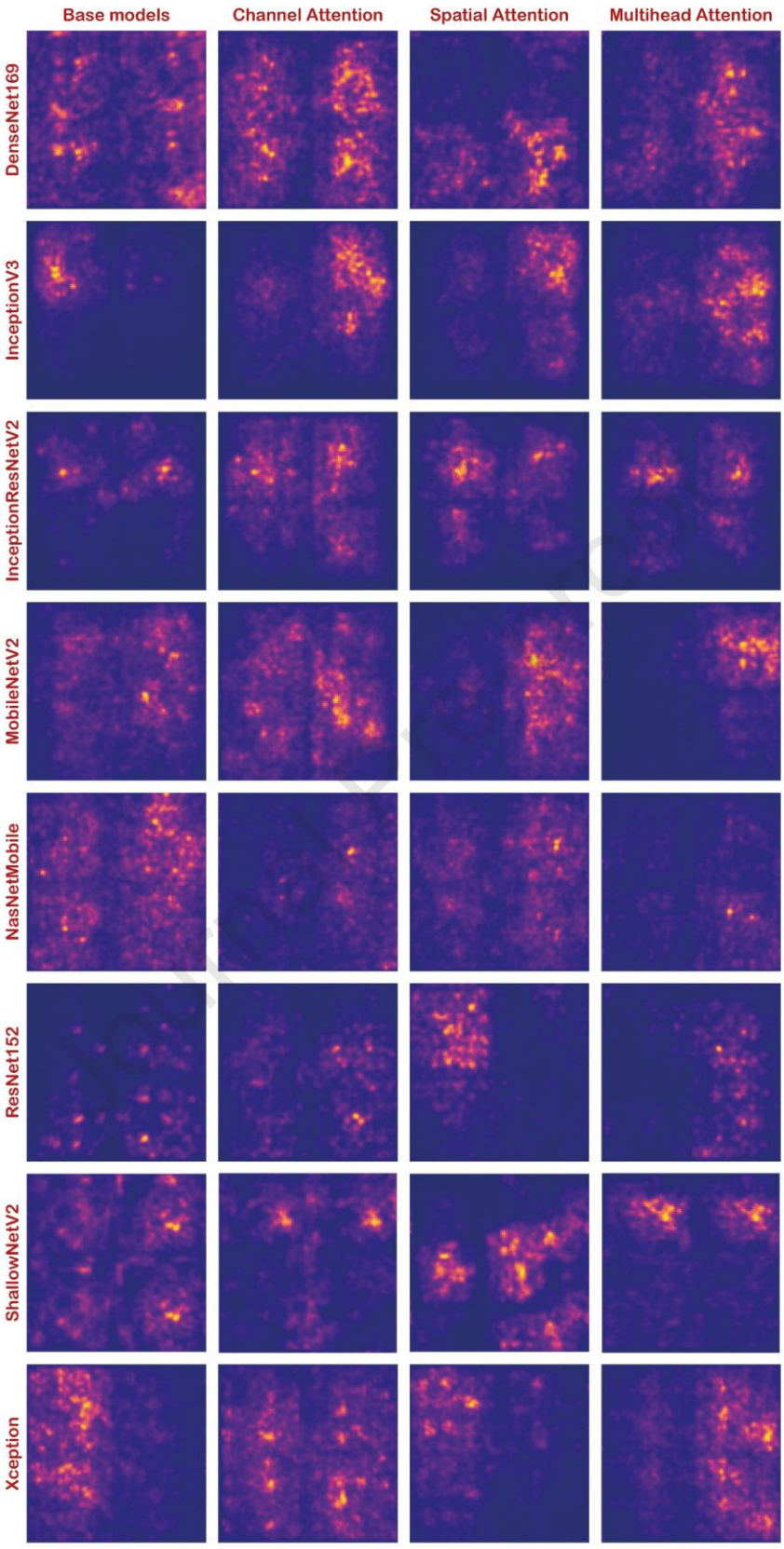

Figure 14: Saliency Map visualizations for the "Major damage" class, comparing base models and their attention-augmented variants (all 32 models predicted correctly).

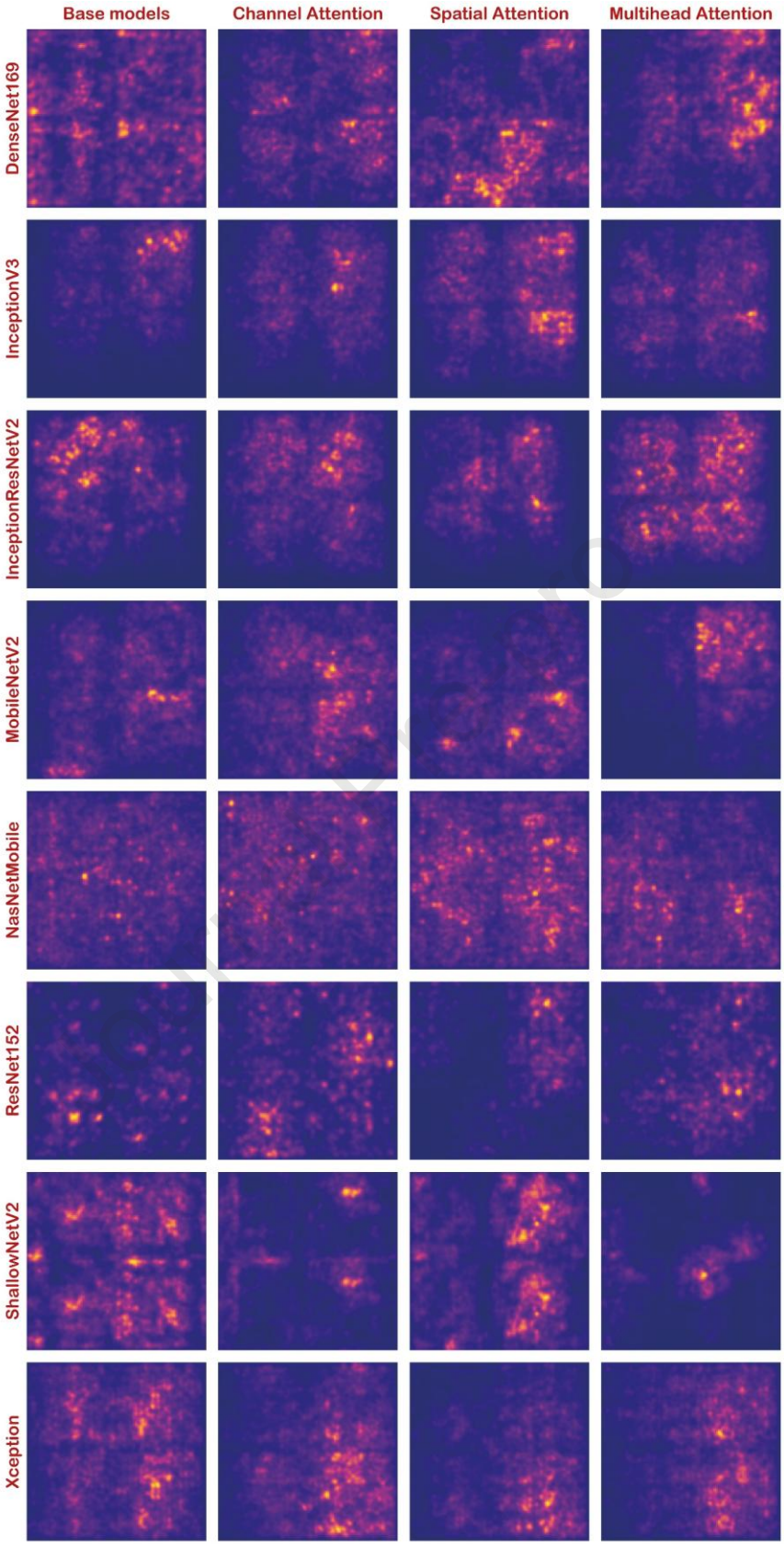

Figure 15: Saliency Map visualizations for the "Destroyed" class, comparing base models and their attention-augmented variants (no model predicted correctly).

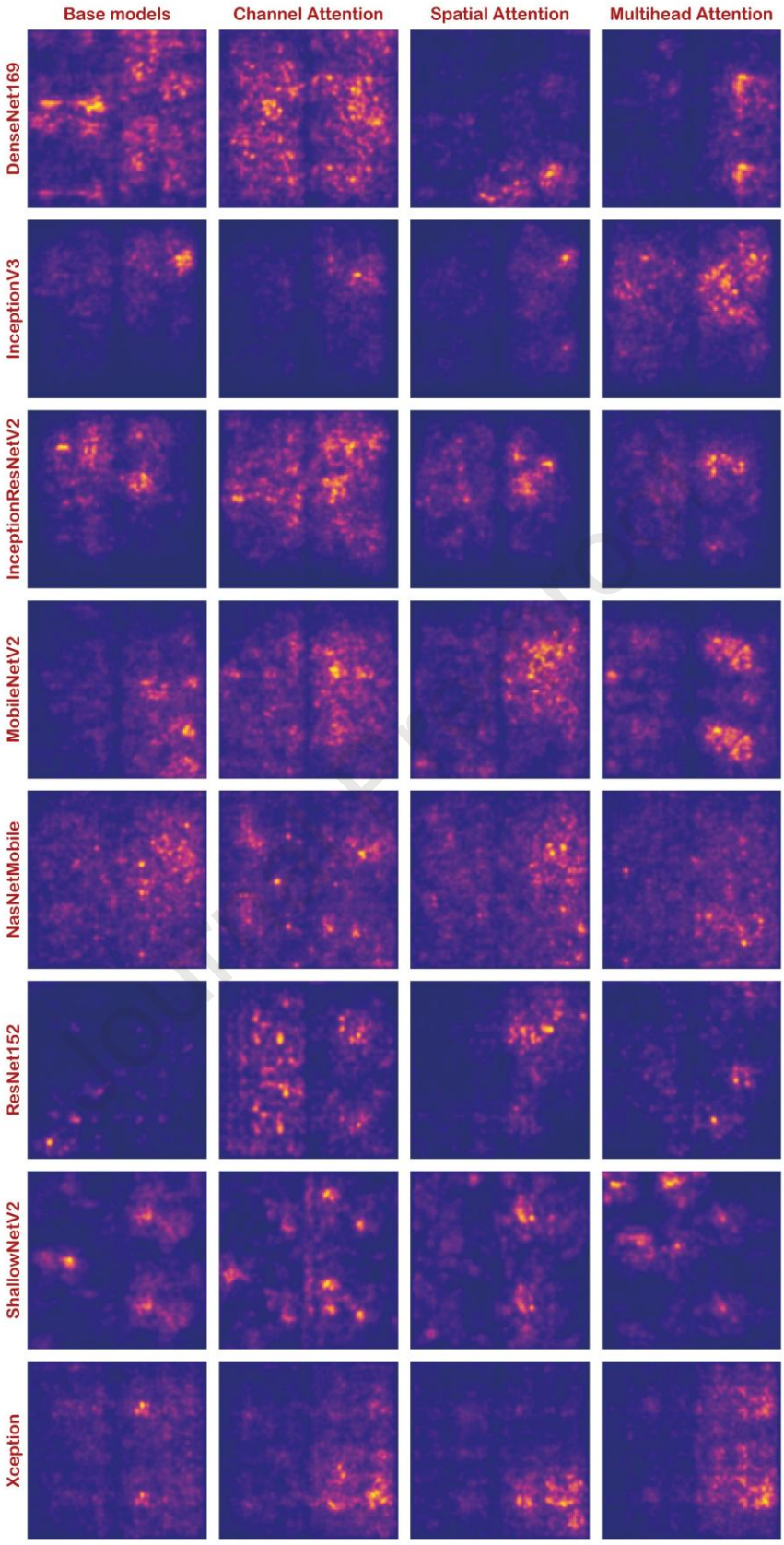

Figure 16: Saliency Map visualizations for the "Major damage" class, comparing base models and their attention-augmented variants (no model predicted correctly).

4. Conclusions

This study highlights the critical role of XAI in evaluating the reliability of deep learning models for PDD. Although these models achieve high accuracy, our XAI-based analysis uncovers reliability limitations, emphasizing the necessity for enhanced interpretability to ensure trustworthy deployment in real-world scenarios. Furthermore, we show the potential of attention mechanisms in DL models for PDD. Through systematic evaluation using a tornado subset of the xBD dataset, we demonstrated that integrating attention mechanisms, particularly MA and SA improve model reliability, accuracy, and interpretability. These mechanisms were shown to enhance the focus of models on critical damage areas, especially in high-severity categories like "Destroyed" and "Major Damage," which are vital for prioritizing disaster response efforts.

XAI techniques, the complementary analysis of Grad-CAM and Saliency Maps together, played a pivotal role in evaluating model reliability. Grad-CAM visualizations identified MA variations of ShallowNetV2 and InceptionV3 as the most reliable configurations, with heatmaps consistently highlighting regions of structural damage while maintaining spatial coherence. Conversely, models with CA often struggled to focus on post-disaster features, misdirecting attention to pre-disaster regions and diminishing their interpretability. Saliency Map analysis further validated these findings at a pixel level, revealing that MA_ShallowNetV2 and MA_InceptionV3 provided sharp, focused saliency patterns aligning with actual damage, providing interpretable and trustworthy decision paths.

The comparative analysis underscores the importance of selecting appropriate attention mechanisms and explainability methods in PDD tasks. While MA emerged as the most effective mechanism, its success depends on the underlying architecture's ability to integrate spatial and temporal information effectively. These findings emphasize the

need for robust and interpretable models that balance predictive accuracy with explainability, ensuring their applicability in real-world disaster management scenarios.

The novelty of this study is primarily comparative and application-focused. Its contribution lies in systematically integrating and evaluating attention mechanisms across multiple architectures, rather than in proposing new algorithms or interpretability frameworks. This applied focus nevertheless provides practical value by highlighting consistent patterns of reliability and limitations, supporting researchers and practitioners in selecting suitable approaches for operational disaster management.

At the same time, several limitations must be acknowledged. Although this study employed qualitative XAI analyses to assess model reliability, the inclusion of quantitative metrics such as the Intersection over Union between attention heatmaps and annotated damage masks, or the percentage of salient pixels falling within damaged regions, would provide a more rigorous evaluation of model interpretability. Incorporating such quantitative measures could complement the qualitative insights presented here and further strengthen the assessment of model reliability in PDD. Another limitation is the use of the standard cross-entropy loss function, which treats all misclassifications equally and does not capture the ordinal relationship between damage classes. This limitation is particularly important in real-world applications, where confusing “destroyed” with “no damage” carries far greater consequences than misclassifying it as “major damage.” Recent studies have begun addressing this issue through ordinal-aware loss functions. For example, Tsai and Lin (2024) employed the Ordinal Class Distance Penalty Loss (OCDPL), introducing penalties that increase with the class distance between the prediction and ground truth to reduce the impact of severe misclassifications. Cheng et al. (2021b) adopted a more advanced approach by using the squared Earth Mover’s Distance (EMD), which evaluates the dissimilarity between the

entire predicted probability distribution and the true label distribution. Unlike OCDPL, which directly applies distance-based penalties at the class level, EMD considers the underlying geometry of the probability space, offering a principled way to enforce ordinal consistency across predictions. Moreover, this study is the use of fixed 128×128 patches centered on individual buildings, which, while simplifying training and ensuring consistency across models, may exclude valuable contextual information such as nearby damage or infrastructure. This approach can also distort building scales and introduce artificial uniformity, limiting the generalizability of the models to more complex post-disaster scenes. To mitigate such issues, previous studies have proposed strategies such as multi-scale feature fusion to capture both fine detail and broader contextual cues (Shen et al., 2022), patch-based discriminative learning with multi-level and pyramid representations to enhance discriminative power across scales (Muhammad et al., 2022), and scale-adaptive approaches that dynamically adjust patch size or receptive fields according to scene complexity (Liu et al., 2023). These strategies underscore the importance of integrating both local and contextual information when designing damage assessment models.

Future research should address these limitations by developing context-aware loss functions, incorporating quantitative interpretability metrics, and exploring variable patch sizes or scene-level models. Future studies should also explore ordinal-aware approaches, such as ordinal regression techniques or cost-sensitive loss functions, which can penalize severe misclassifications more heavily than adjacent ones and thereby better align model behavior with the practical needs of disaster response. Moreover, future studies could investigate variable patch sizes, multi-scale architectures, or scene-level models that capture both local and contextual information more effectively to address the constraints of fixed 128×128 patches. Advances in generative AI also hold promises for addressing

the scarcity of high-resolution disaster datasets, enabling synthetic augmentation to improve model robustness across diverse hazard types. Finally, extending this approach to multi-hazard scenarios, such as earthquakes, floods, or cascading hazard sequences, represents an important next step, as operational models must ultimately account for the compound and interacting risks that characterize real-world disaster environments.

In summary, this work demonstrates that while attention mechanisms can improve both performance and interpretability in PDD, their effectiveness varies across architectures. The systematic evaluation presented here provides actionable insights for developing more reliable and interpretable AI tools, helping bridge the gap between technical advances and operational needs in disaster risk management.

Acknowledgments

We gratefully acknowledge the support and contributions of Yulu Li, Daniel Richards, Jiayi Xiong, and Amanda Yang, whose collaboration and insights were invaluable in the completion of this project. This study was mainly conducted as part of a master's course IRDR0047: Geospatial Data Science (2023/2024) at the Department of Risk and Disaster Reduction, University College London (UCL).

Disclosure statement

The authors report there are no competing interests to declare.

Data availability statement

The data that support the findings of this study are available from the xBD dataset, which is publicly accessible at https://xview2.org/dataset. The dataset includes pre- and post-disaster satellite images labeled with damage levels, enabling reproducibility and further research.

Appendix

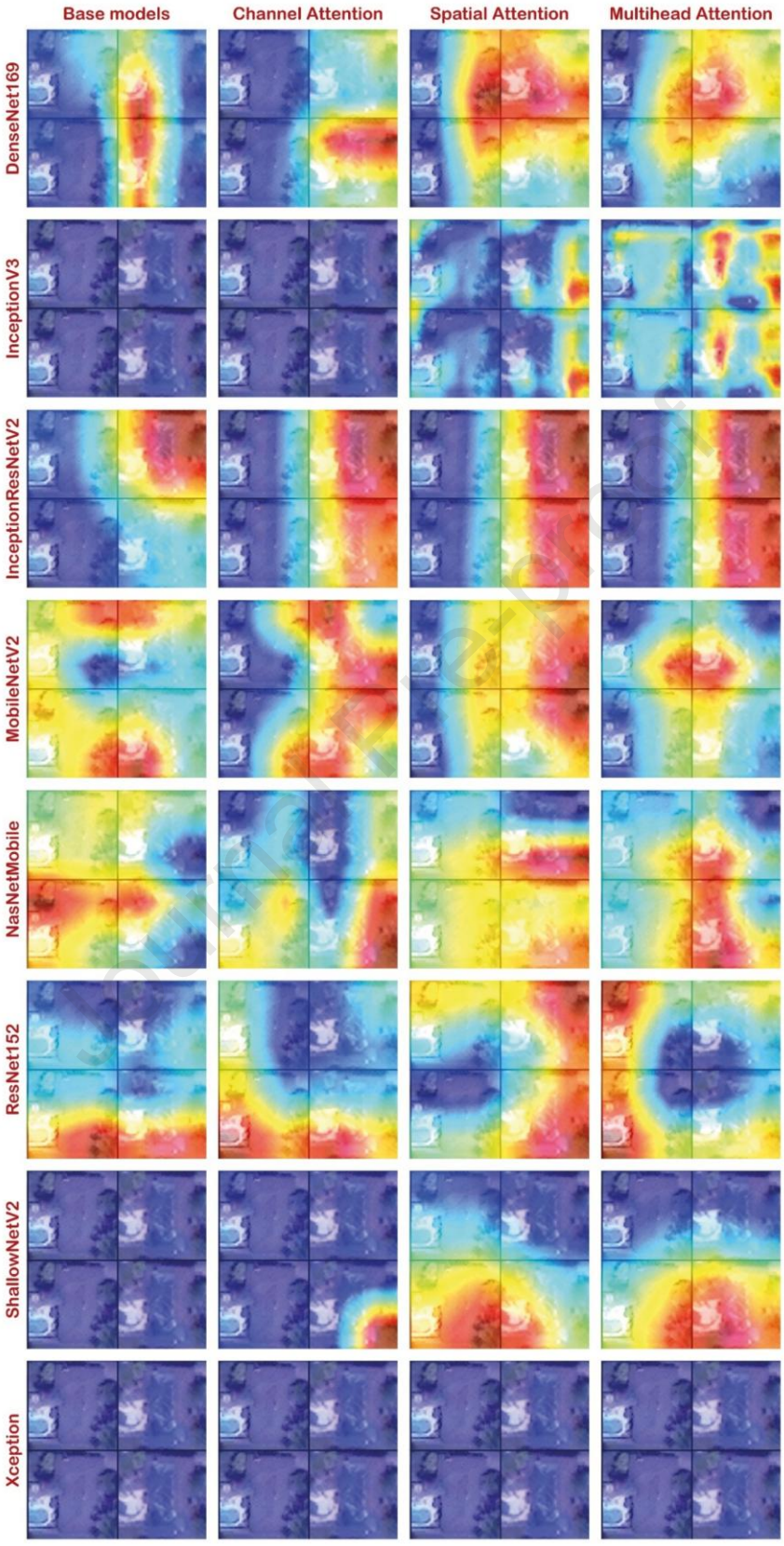

Figure 17: Grad-CAM visualizations for the "Minor damage" class, comparing base models and their attention-augmented variants (all 32 models predicted correctly).

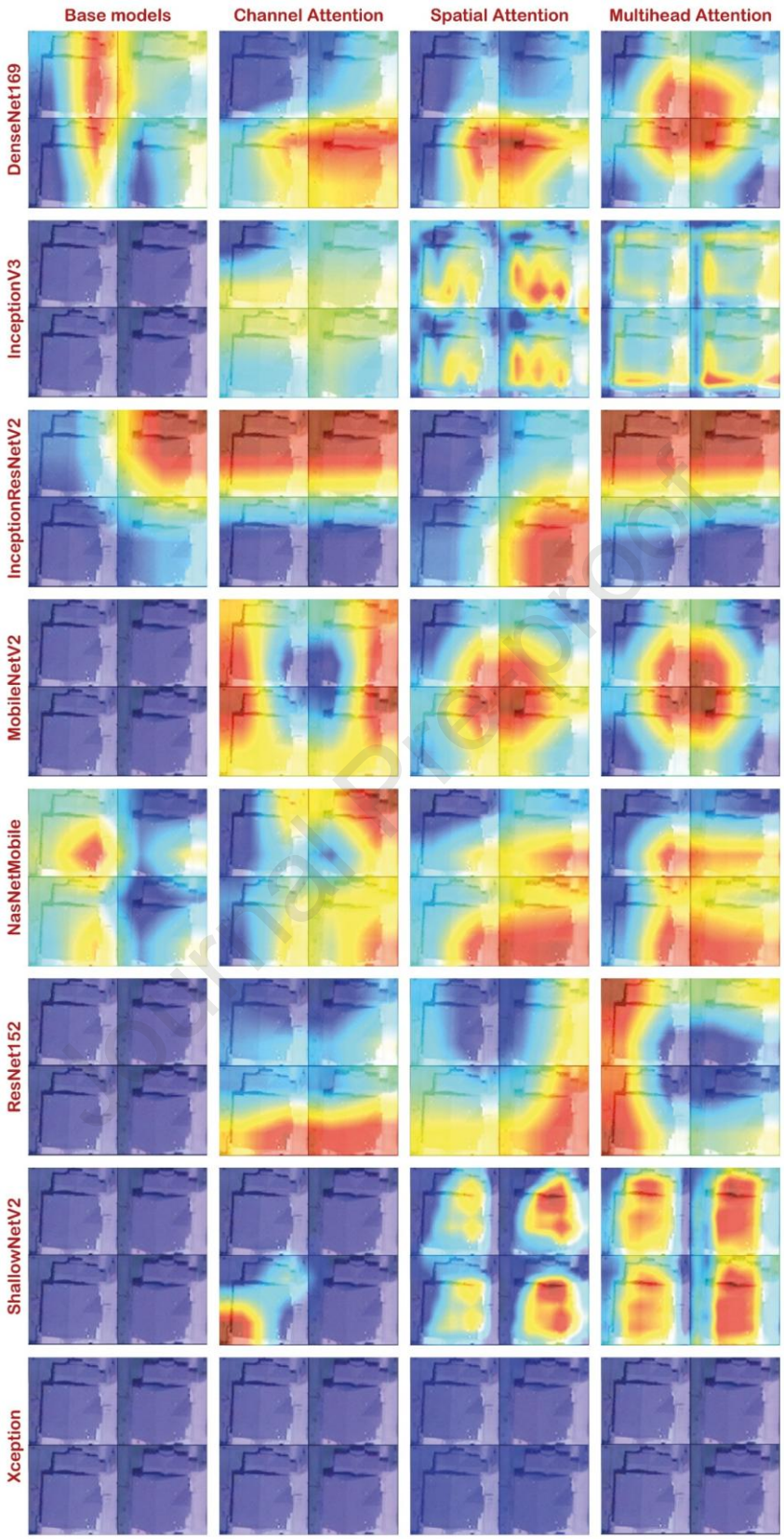

Figure 18: Grad-CAM visualizations for the "No damage" class, comparing base models and their attention-augmented variants (all 32 models predicted correctly).

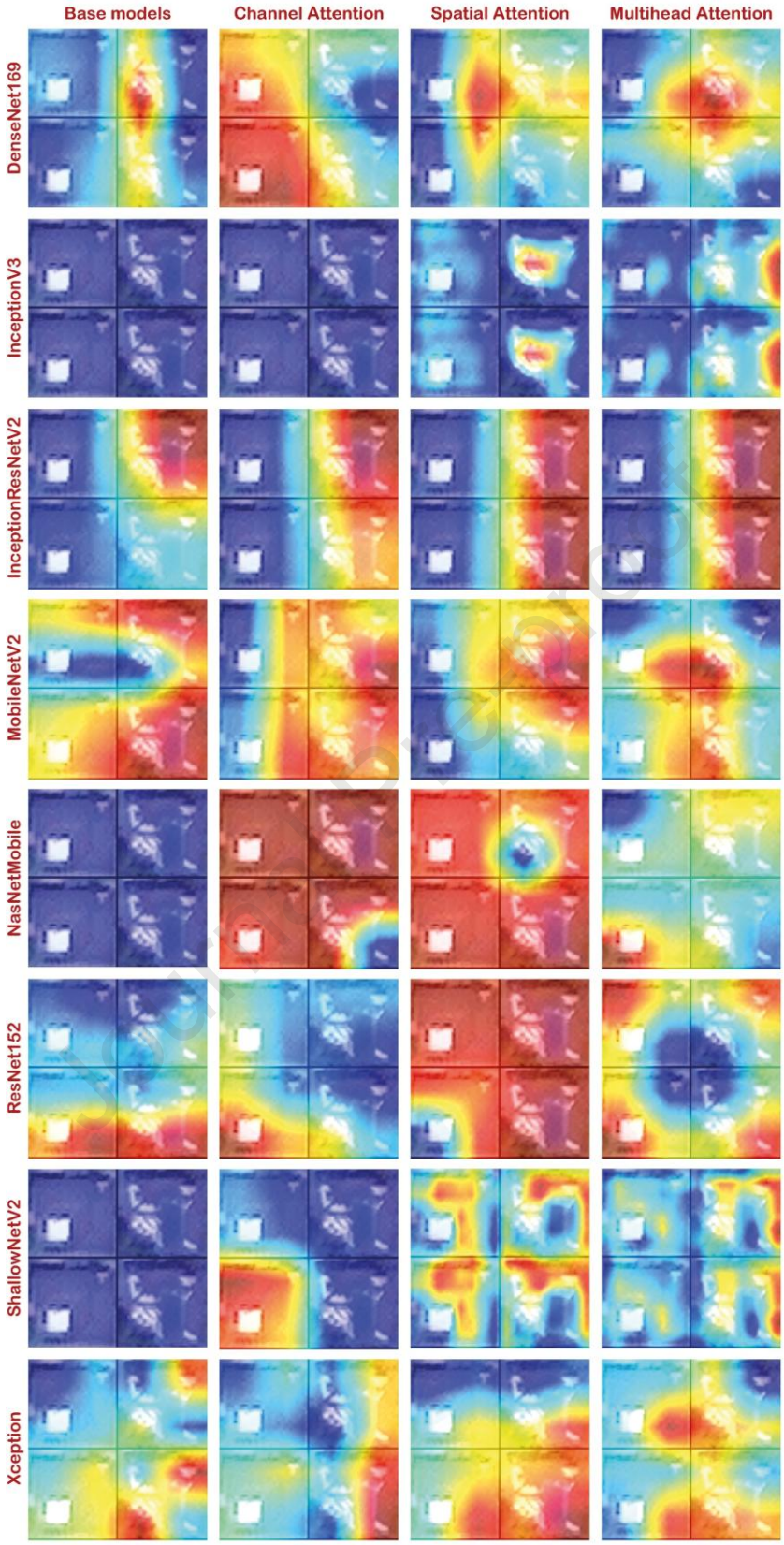

Figure 19: Grad-CAM visualizations for the "Minor damage" class, comparing base models and their attention-augmented variants (no model predicted correctly).

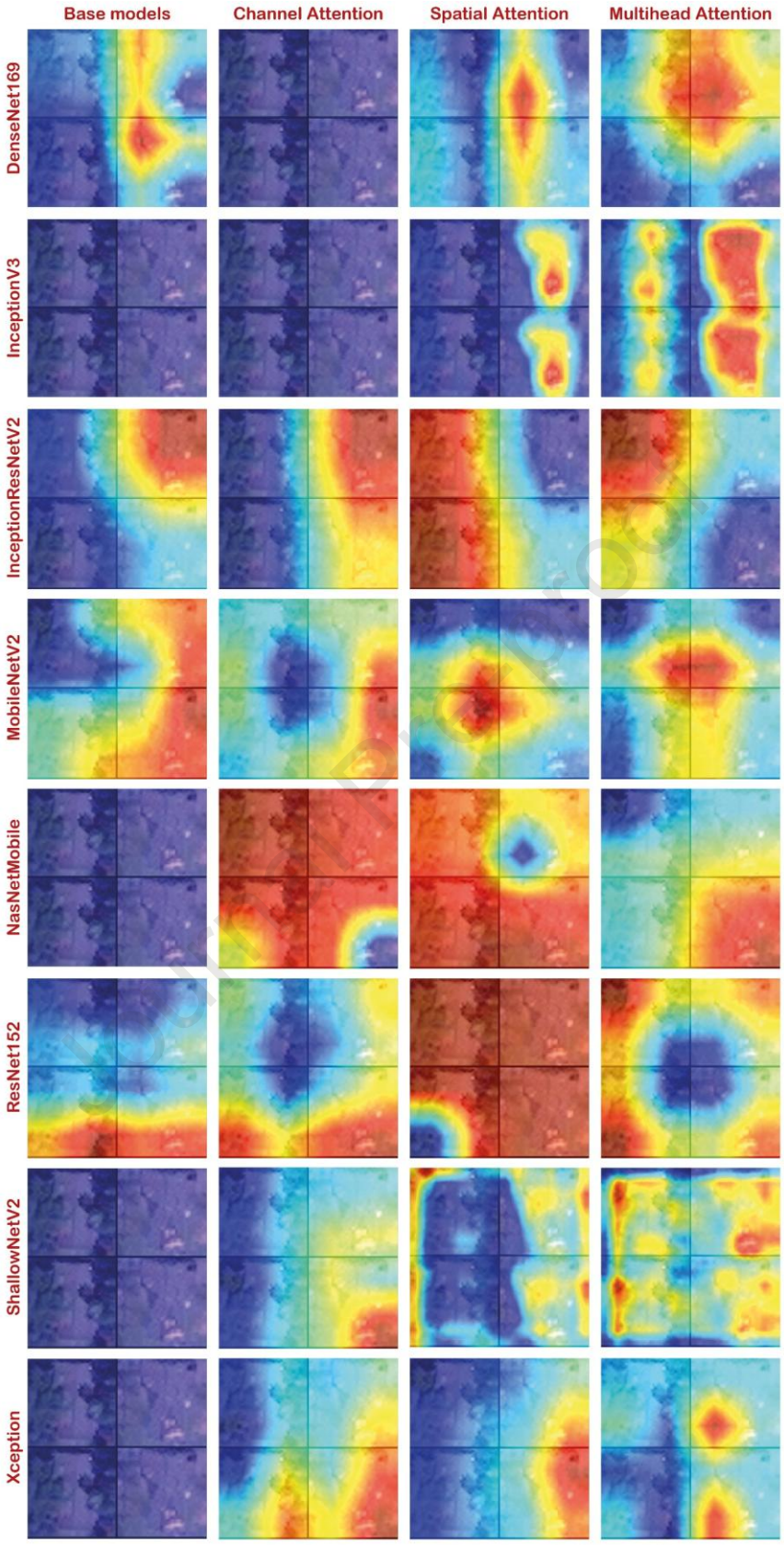

Figure 20: Grad-CAM visualizations for the "No damage" class, comparing base models and their attention-augmented variants (no model predicted correctly).

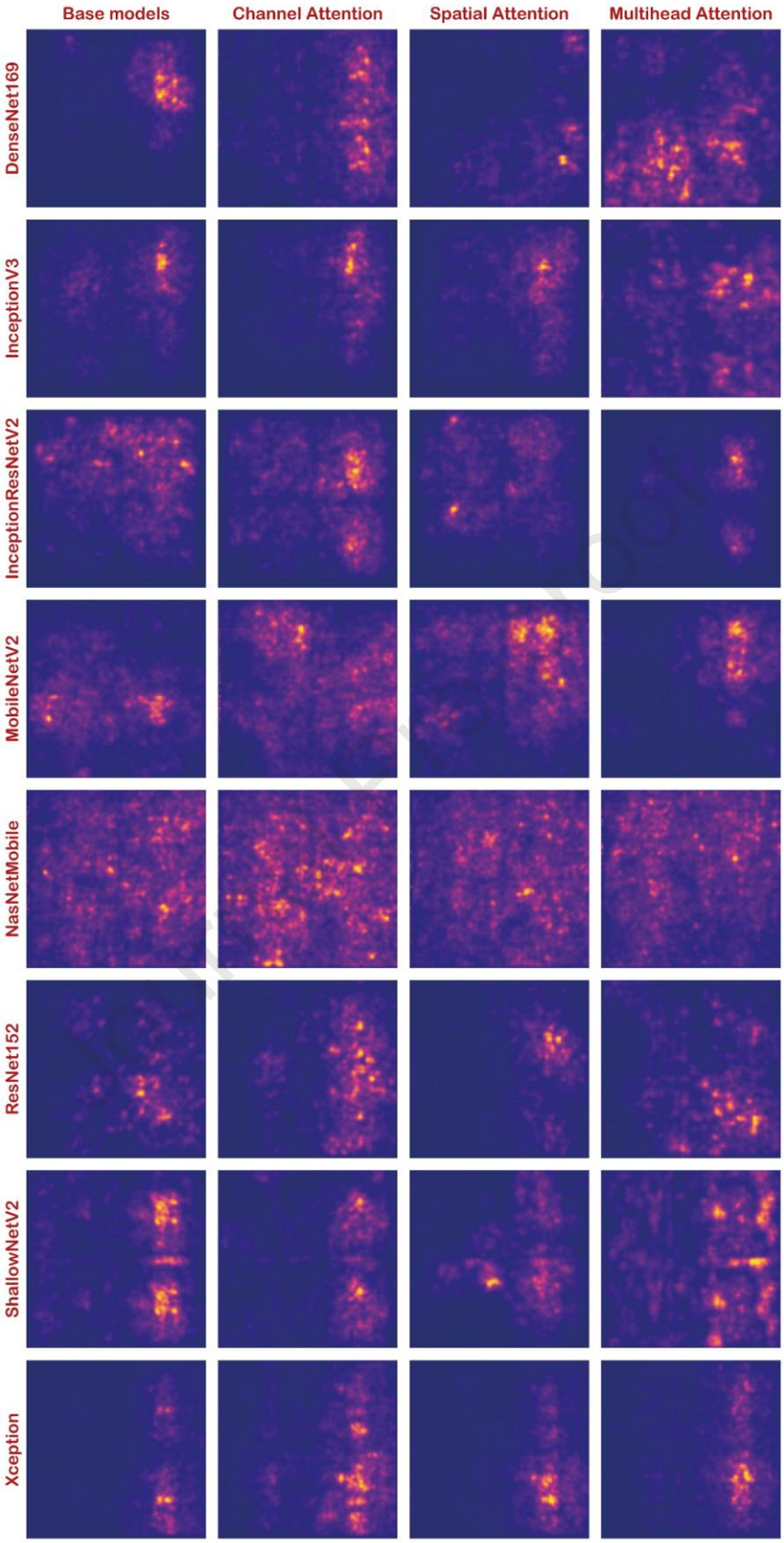

Figure 21: Saliency Map visualizations for the "Minor damage" class, comparing base models and their attention-augmented variants (all 32 models predicted correctly).

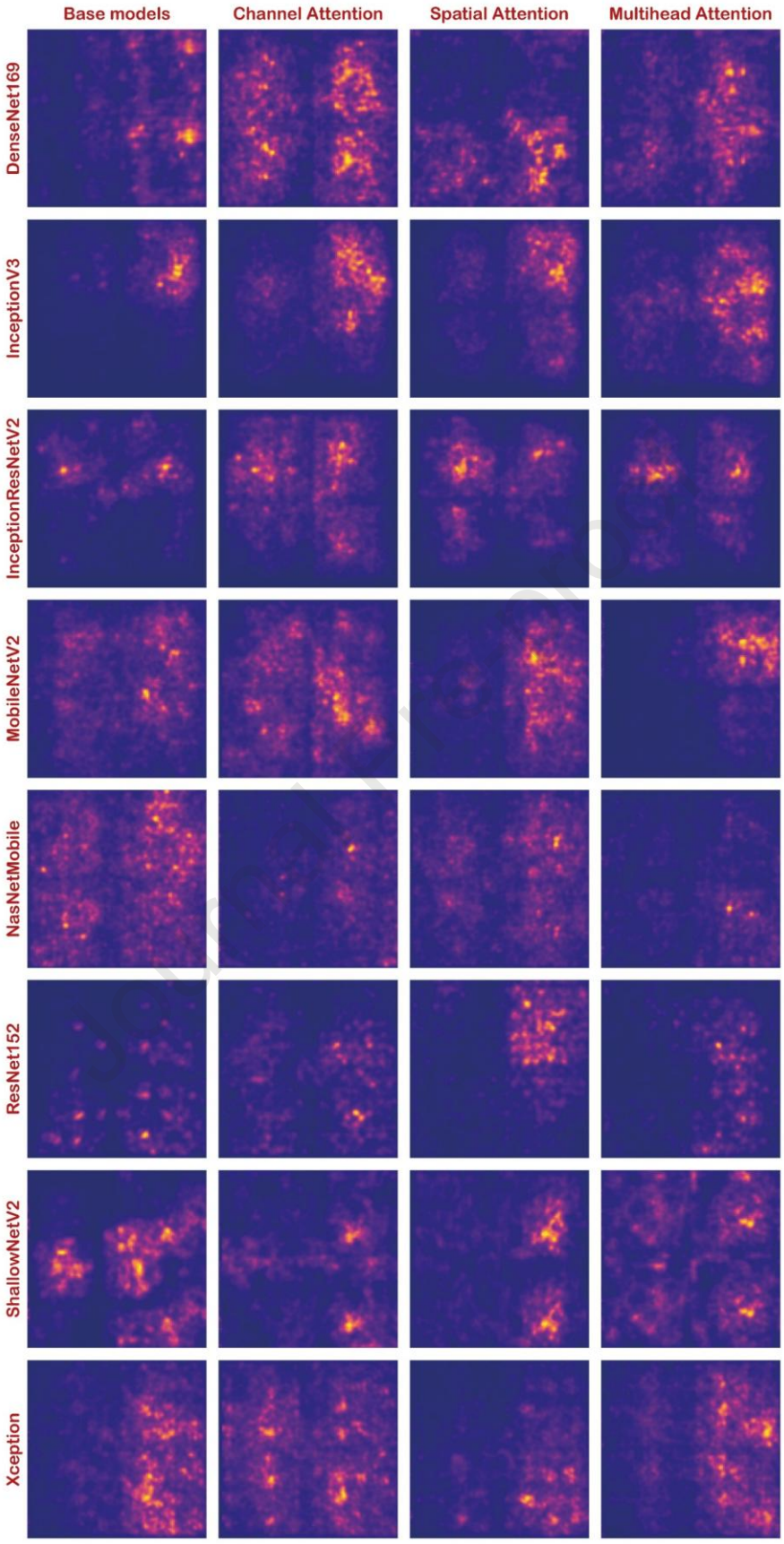

Figure 22:Saliency Map visualizations for the "No damage" class, comparing base models and their attention-augmented variants (all 32 models predicted correctly).

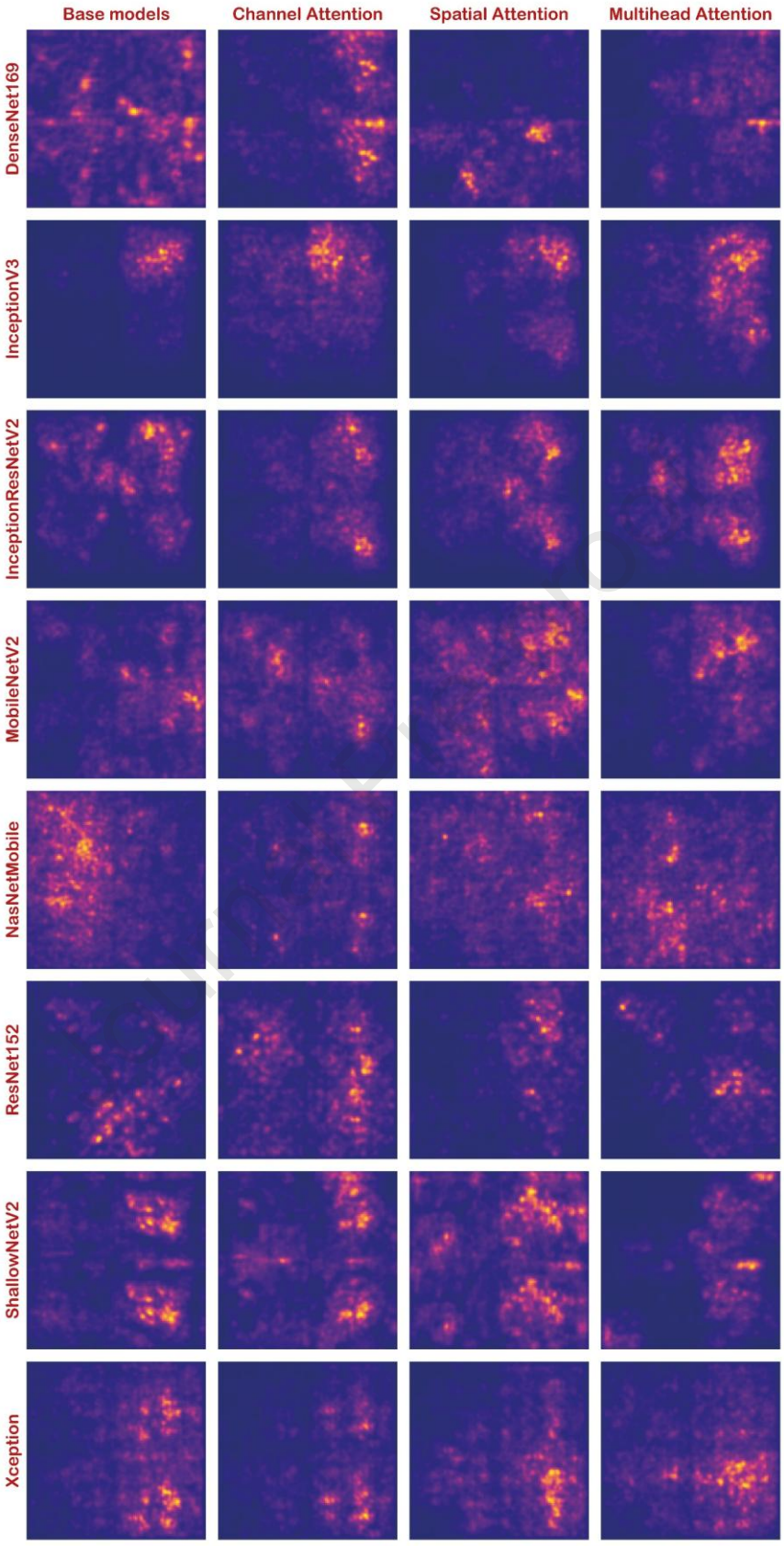

Figure 23: Saliency Map visualizations for the "Minor damage" class, comparing base models and their attention-augmented variants (no model predicted correctly).

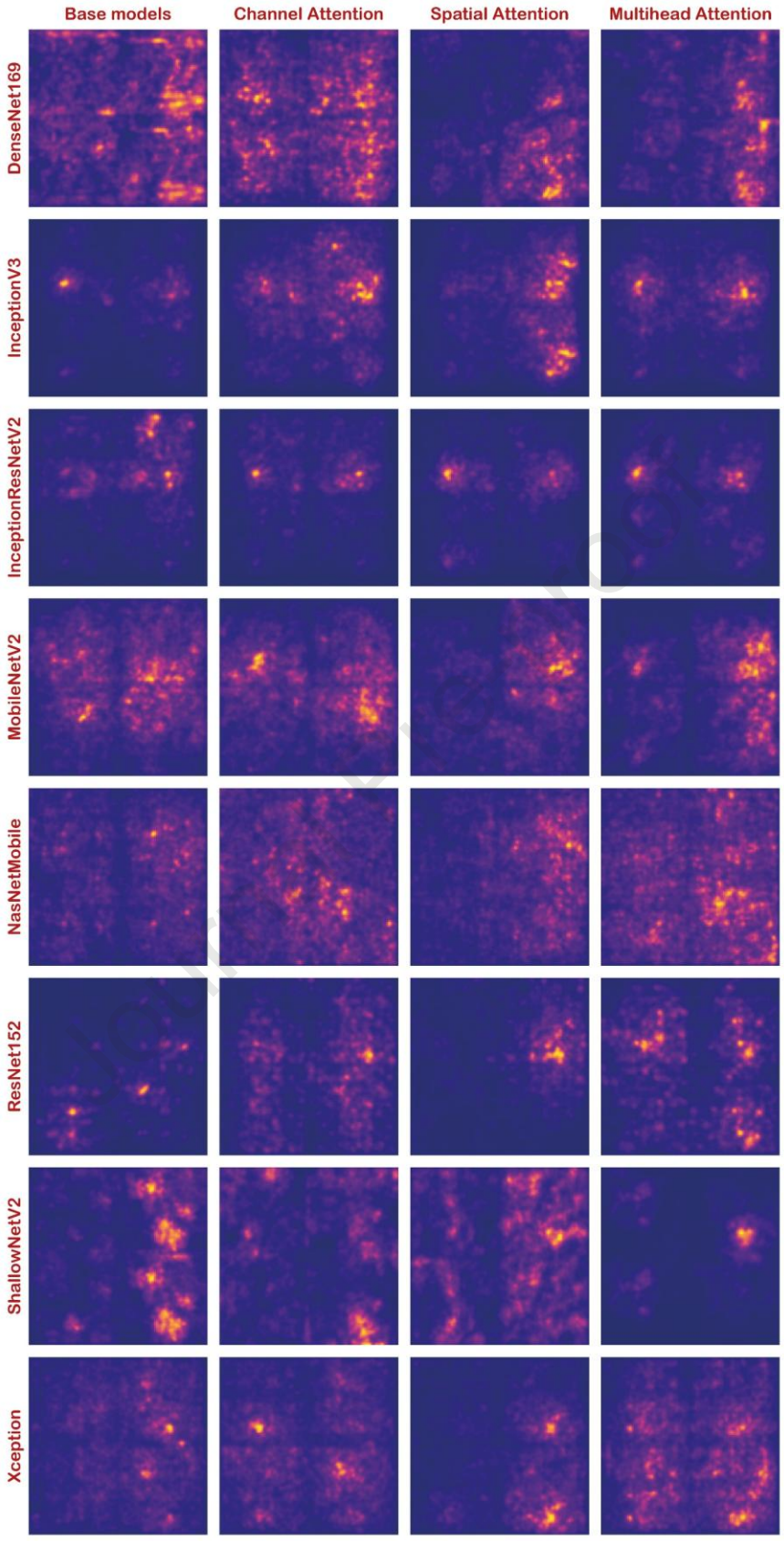

Figure 24: Saliency Map visualizations for the "No damage" class, comparing base models and their attention-augmented variants (no model predicted correctly).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT for language improvement. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication

References

Ahmadi, S. A., Mohammadzadeh, A., Yokoya, N., & Ghorbanian, A. (2024). BD-SKUNet: Selective-Kernel UNets for Building Damage Assessment in High-Resolution Satellite Images. *Remote Sensing*, *16*(1), 182. https://www.mdpi.com/2072-4292/16/1/182

Arbulu, R., & Ballard, G. (2004). Lean supply systems in construction. *Proceedings of the 12th Annual Conference for International Group for Lean Construction, Carlos Formoso and Marton Marosszeky, 2004*.

Atkins, N. T., Butler, K. M., Flynn, K. R., & Wakimoto, R. M. (2014). An Integrated Damage, Visual, and Radar Analysis of the 2013 Moore, Oklahoma, EF5 Tornado. *Bulletin of the American Meteorological Society*, *95*(10), 1549-1561. https://doi.org/https://doi.org/10.1175/BAMS-D-14-00033.1

Braik, A. M., & Koliou, M. (2024). Automated building damage assessment and large-scale mapping by integrating satellite imagery, GIS, and deep learning [Article]. *Computer-Aided Civil and Infrastructure Engineering*, *39*(15), 2389-2404. https://doi.org/10.1111/mice.13197

Braik, A. M., & Koliou, M. (2024). Automated building damage assessment and large-scale mapping by integrating satellite imagery, GIS, and deep learning. *Computer-Aided Civil and Infrastructure Engineering*.

Cheng, C.-S., Behzadan, A., & Noshadravan, A. (2021a). *DoriaNET: A visual dataset from Hurricane Dorian for post-disaster building damage assessment*.Designsafe-CI. https://doi.org/10.17603/DS2-GQVG-QX37

Cheng, C.-S., Behzadan, A. H., & Noshadravan, A. (2021b). Deep learning for post-hurricane aerial damage assessment of buildings. *Computer-Aided Civil and Infrastructure Engineering*, *36*(6), 695-710. https://doi.org/https://doi.org/10.1111/mice.12658

Cheng, C.-S., Behzadan, A. H., & Noshadravan, A. (2022). Uncertainty-aware convolutional neural network for explainable artificial intelligence-assisted disaster damage assessment. *Structural Control and Health Monitoring*, *29*(10), e3019. https://doi.org/https://doi.org/10.1002/stc.3019

Chollet, F. (2017). *Xception: Deep Learning with Depthwise Separable Convolutions*. https://doi.org/10.1109/CVPR.2017.195

Crawford, S., Graettinger, A., Powell, L., Awondo, S., Back, E., & Spector, S. (2017). *Five Years after the April 27, 2011, Tuscaloosa Tornado: A Study in Community Resilience*. https://doi.org/10.1061/9780784480502.063

Crawford, S., Hainen, A., Graettinger, A., Lindt, J., & Powell, L. (2020). Discrete-Outcome Analysis of Tornado Damage Following the 2011 Tuscaloosa, Alabama, Tornado. *Natural Hazards Review*, *21*. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000396

Csillik, O. (2017). Fast Segmentation and Classification of Very High Resolution Remote Sensing Data Using SLIC Superpixels. *Remote Sensing*, *9*(3), 243. https://www.mdpi.com/2072-4292/9/3/243

Da, Y., Ji, Z., & Zhou, Y. (2022). Building Damage Assessment Based on Siamese Hierarchical Transformer Framework. *Mathematics*, *10*(11), 1898. https://www.mdpi.com/2227-7390/10/11/1898

Deng, J., Dong, W., Socher, R., Li, L. J., Kai, L., & Li, F.-F. (2009, 20-25 June 2009). ImageNet: A large-scale hierarchical image database. (Ed.),^(Eds.). 2009 IEEE Conference on Computer Vision and Pattern Recognition.

Deng, L., & Wang, Y. (2022). Post-disaster building damage assessment based on improved U-Net. *Scientific Reports*, *12*(1), 15862. https://doi.org/10.1038/s41598-022-20114-w

EM-DAT. (2024). *The Impact of Disasters between 2004 and 2023*.https://public.emdat.be/data

Ghaffarian, S., Kerle, N., & Filatova, T. (2018). Remote Sensing-Based Proxies for Urban Disaster Risk Management and Resilience: A Review. *Remote Sensing*, *10*(11), 1760. https://www.mdpi.com/2072-4292/10/11/1760

Ghaffarian, S., Taghikhah, F. R., & Maier, H. R. (2023). Explainable artificial intelligence in disaster risk management: Achievements and prospective futures. *International Journal of Disaster Risk Reduction*, *98*, 104123. https://doi.org/https://doi.org/10.1016/j.ijdrr.2023.104123

Ghaffarian, S., Valente, J., van der Voort, M., & Tekinerdogan, B. (2021). Effect of Attention Mechanism in Deep Learning-Based Remote Sensing Image Processing: A Systematic Literature Review. *Remote Sensing*, *13*(15), 2965. https://www.mdpi.com/2072-4292/13/15/2965

Gokaraju, B., Turlapaty, A. C., Doss, D. A., King, R. L., & Younan, N. H. (2015, 13-15 Oct. 2015). Change detection analysis of tornado disaster using conditional copulas and Data Fusion for cost-effective disaster management. (Ed.),^(Eds.). 2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR).

Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E., Choset, H., & Gaston, M. (2019). *xBD: A Dataset for Assessing Building Damage from Satellite Imagery*. https://doi.org/10.48550/arXiv.1911.09296

Gupta, R., & Shah, M. (2021). *RescueNet: Joint Building Segmentation and Damage Assessment from Satellite Imagery*. 2020 25th International Conference on Pattern Recognition (ICPR). https://doi.ieeecomputersociety.org/10.1109/ICPR48806.2021.9412295

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. (Ed.),^(Eds.). Proceedings of the IEEE conference on computer vision and pattern recognition.

Howard, A. G. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G., Liu, Z., Maaten, L. V. D., & Weinberger, K. Q. (2017, 21-26 July 2017). Densely Connected Convolutional Networks. (Ed.),^(Eds.). 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Islam, A. M., Masud, F. B., Ahmed, M. R., Jafar, A. I., Ullah, J. R., Islam, S., Shatabda, S., & Islam, A. K. M. M. (2023). An Attention-Guided Deep-Learning-Based Network with Bayesian Optimization for Forest Fire Classification and Localization. *Forests*, *14*(10), 2080. https://www.mdpi.com/1999-4907/14/10/2080

Kakogeorgiou, I., & Karantzalos, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *103*, 102520. https://doi.org/https://doi.org/10.1016/j.jag.2021.102520

Kaur, N., Lee, C.-C., Mostafavi, A., & Mahdavi-Amiri, A. (2023). Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images. *Computer-Aided Civil and Infrastructure Engineering*, *38*(15), 2072-2091. https://doi.org/https://doi.org/10.1111/mice.12981

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, *25*. https://doi.org/10.1145/3065386

Liu, Y., Shi, S., Wang, J., & Zhong, Y. (2023, 1-6 Oct. 2023). Seeing Beyond the Patch: Scale-Adaptive Semantic Segmentation of High-resolution Remote Sensing Imagery based on Reinforcement Learning. (Ed.),^(Eds.). 2023 IEEE/CVF International Conference on Computer Vision (ICCV).

Lu, W., Wei, L., & Nguyen, M. (2024). Bitemporal Attention Transformer for Building Change Detection and Building Damage Assessment [Article]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 4917-4935. https://doi.org/10.1109/JSTARS.2024.3354310

Mansour, M. A., Rhee, D. M., Newson, T., Peterson, C., & Lombardo, F. T. (2021). Estimating Wind Damage in Forested Areas Due to Tornadoes. *Forests*, *12*(1), 17. https://www.mdpi.com/1999-4907/12/1/17

Matin, S. S., & Pradhan, B. (2021). Earthquake-Induced Building-Damage Mapping Using Explainable AI (XAI). *Sensors (Basel)*, *21*(13). https://doi.org/10.3390/s21134489

Muhammad, U., Hoque, M. Z., Wang, W., & Oussalah, M. (2022). Patch-Based Discriminative Learning for Remote Sensing Scene Classification. *Remote Sensing*, *14*(23), 5913. https://www.mdpi.com/2072-4292/14/23/5913

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/https://doi.org/10.1016/j.array.2022.100258

Oludare, V., Kezebou, L., Jinadu, O., Panetta, K., & Agaian, S. (2022). *Attention-based two-stream high-resolution networks for building damage assessment from satellite imagery*. https://doi.org/10.1117/12.2618901

Paul, B. K., & Stimers, M. (2012). Exploring probable reasons for record fatalities: The case of 2011 Joplin, Missouri, Tornado [Article]. *Natural Hazards*, *64*(2), 1511-1526. https://doi.org/10.1007/s11069-012-0313-3

Prevatt, D. O., Lindt, J. W. v. d., Back, E. W., Graettinger, A. J., Pei, S., Coulbourne, W., Gupta, R., James, D., & Agdas, D. (2012). Making the Case for Improved Structural Design: Tornado Outbreaks of 2011. *Leadership and Management in Engineering*, *12*(4), 254-270. https://doi.org/doi:10.1061/(ASCE)LM.1943-5630.0000192

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. (Ed.),^(Eds.). Proceedings of the IEEE international conference on computer vision.

Seydi, S. T., Hasanlou, M., Chanussot, J., & Ghamisi, P. (2023). BDD-Net+: A Building Damage Detection Framework Based on Modified Coat-Net. *IEEE*

Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 4232-4247. https://doi.org/10.1109/JSTARS.2023.3267847

Shen, Y., Zhu, S., Yang, T., Chen, C., Pan, D., Chen, J., Xiao, L., & Du, Q. (2022). BDANet: Multiscale Convolutional Neural Network With Cross-Directional Attention for Building Damage Assessment From Satellite Images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-14. https://doi.org/10.1109/TGRS.2021.3080580

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Simonyan, K. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Song, J., Gao, S., Zhu, Y., & Ma, C. (2019). A survey of remote sensing image classification based on CNNs. *Big Earth Data*, *3*(3), 232-254. https://doi.org/10.1080/20964471.2019.1657720

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. (Ed.),^(Eds.). Proceedings of the AAAI conference on artificial intelligence.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016, 27-30 June 2016). Rethinking the Inception Architecture for Computer Vision. (Ed.),^(Eds.). 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Tsai, F. J., & Lin, S.-Y. (2024). A Class Distance Penalty Deep Learning Method for Post-disaster Building Damage Assessment. *KSCE Journal of Civil Engineering*, *28*(5), 2005-2019. https://doi.org/https://doi.org/10.1007/s12205-024-1587-1

Tursunalieva, A., Alexander, D. L. J., Dunne, R., Li, J., Riera, L., & Zhao, Y. (2024). Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications. *Applied Sciences*, *14*(2), 496. https://www.mdpi.com/2076-3417/14/2/496

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Victor, O., Landry, K., Obafemi, J., Karen, P., & Sos, A. (2022). Attention-based two-stream high-resolution networks for building damage assessment from satellite imagery. (Ed.),^(Eds.). Proc.SPIE.

Wang, X.-Y., Wang, T., & Bu, J. (2011). Color image segmentation using pixel wise support vector machine classification. *Pattern Recognition*, *44*(4), 777-787. https://doi.org/https://doi.org/10.1016/j.patcog.2010.08.008

Weber, L., Lapuschkin, S., Binder, A., & Samek, W. (2023). Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion*, *92*, 154-176. https://doi.org/https://doi.org/10.1016/j.inffus.2022.11.013

Wiguna, S., Adriano, B., Mas, E., & Koshimura, S. (2024). Evaluation of Deep Learning Models for Building Damage Mapping in Emergency Response Settings [Article]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 5651-5667. https://doi.org/10.1109/JSTARS.2024.3367853

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. (Ed.),^(Eds.). Proceedings of the European conference on computer vision (ECCV).

Wu, C., Zhang, F., Xia, J., Xu, Y., Li, G., Xie, J., Du, Z., & Liu, R. (2021). Building Damage Detection Using U-Net with Attention Mechanism from Pre- and Post-Disaster Remote Sensing Datasets. *Remote Sensing*, *13*(5), 905. https://www.mdpi.com/2072-4292/13/5/905

Xia, H., Wu, J., Yao, J., Zhu, H., Gong, A., Yang, J., Hu, L., & Mo, F. (2023). A Deep Learning Application for Building Damage Assessment Using Ultra-High-Resolution Remote Sensing Imagery in Turkey Earthquake. *International Journal of Disaster Risk Science*, *14*(6), 947-962. https://doi.org/10.1007/s13753-023-00526-6

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. (2022). Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.

Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M. B., & Kang, B. (2023). Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems*, *3*(3), 161-188. https://doi.org/10.1007/s44230-023-00038-y

Yang, W., Zhang, X., & Luo, P. (2021). Transferability of Convolutional Neural Network Models for Identifying Damaged Buildings Due to Earthquake. *Remote Sensing*, *13*(3), 504. https://www.mdpi.com/2072-4292/13/3/504

Zhan, Y., Liu, W., & Maruyama, Y. (2022). Damaged building extraction using modified mask R-CNN model using post-event aerial images of the 2016 Kumamoto earthquake. *Remote Sensing*, *14*(4), 1002.

Zhang, H., Wang, M., Zhang, Y., & Ma, G. (2022). TDA-Net: A Novel Transfer Deep Attention Network for Rapid Response to Building Damage Discovery. *Remote Sensing*, *14*(15), 3687. https://www.mdpi.com/2072-4292/14/15/3687

Zhang, X., Zhou, Y. n., & Luo, J. (2022). Deep learning for processing and analysis of remote sensing big data: a technical review. *Big Earth Data*, *6*(4), 527-560. https://doi.org/10.1080/20964471.2021.1964879

Zhang, Y., Yang, G., Gao, A., Lv, W., Xie, R., Huang, M., & Liu, S. (2023). An efficient change detection method for disaster-affected buildings based on a lightweight residual block in high-resolution remote sensing images. *International Journal of Remote Sensing*, *44*(9), 2959-2981. https://doi.org/10.1080/01431161.2023.2214274

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. (Ed.),^(Eds.). Proceedings of the IEEE conference on computer vision and pattern recognition.

Zou, L., Chen, F., Huang, X., & Kar, B. (2023). Big Earth data for disaster risk reduction. *Big Earth Data*, *7*(4), 931-936. https://doi.org/10.1080/20964471.2023.2291901

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: