Response to Comments of Referee #1

Thank you for the instructive and constructive comments for our paper. Those comments are very helpful for and serve as significant guidance for our research. We have studied the comments carefully and revised our manuscript accordingly. The changes in our manuscript are highlighted in **red**. The point-to-point responses to your questions/comments are listed as follows.

Comments to the Author:

This paper introduces BRIGHT, a novel and timely benchmark dataset for building damage assessment using multimodal high-resolution optical and SAR imagery. Covering 14 globally distributed disaster events, BRIGHT provides pixel-level damage annotations for over 384,000 buildings. The dataset is designed to facilitate AI-based disaster response research, particularly in challenging all-weather conditions. The authors also benchmark a suite of machine learning and deep learning models on multiple tasks. The authors provided detailed documents and descriptions, making the data, related source code, and pretrained weights of models easy to understand and use.

In summary, this is quite interesting and solid work. I'd like to recommend the acceptance of this work since it represents an important contribution to Earth observation and disaster response communities. Yet before acceptance, several clarifications and refinements are suggested.

Response: We really appreciate your spot-on summary of our manuscript and such a positive endorsement of our work. Our responses to your valuable comments and suggestions are itemized below.

Q1: The manuscript would benefit from deeper exploration of what the models learn from multimodal fusion. Specifically, what roles do optical images play in multimodal building damage assessment? Is it beyond just building footprint localization? On the other words, are the features extracted from optical imagery actively compared with SAR representations? Some discussion (e.g., based on CAMs in Fig. 7) is provided but can be more explicitly elaborated.

R1: Thank you so much for this very insightful comment. To investigate the role of optical imagery in multimodal building damage assessment, we conducted additional experiments as suggested. Specifically, we evaluated UNet and DeepLabV3+ under two input settings: optical + SAR and SAR only. We chose these two models because they are a single-branch architecture, making it straightforward to adjust the number of input channels by modifying the first convolutional layer. In contrast, the other five methods adopt Siamese networks, where structural changes for different input modalities would require extensive reconfiguration. For UNet and DeepLabV3+, the modification introduces negligible

changes in the parameter count. To isolate the contribution of optical imagery beyond building footprint localization, we provided all models with perfect building masks as post-processing steps prior to evaluation.

The results, presented in **Table 7** of the revised manuscript, demonstrate that optical imagery contributes significantly to distinguishing different damage levels. When provided with optical + SAR inputs, both models show notable improvements in the IoU scores for the "Damaged" and "Destroyed" classes compared to SAR-only inputs. For example, UNet's IoU for "Damaged" improved from 35.83% to 44.83%. DeepLabV3+ also benefits from optical imagery, with IoU for "Damaged" changing from 39.63% (SAR only) to 40.45% (optical + SAR), and for "Destroyed" increasing substantially from 59.54% to 64.94%. These findings indicate that optical imagery provides critical complementary information that supports damage classification, rather than merely improving building localization.

Accompanying Table 7, we have added a new **Section 5.3** to the revised manuscript to provide a more detailed discussion of these findings. We show the revised part below for your convenience.

Table 7. Comparison of UNet and DeepLabV3+ performance using only SAR and optical-SAR inputs for damage classification. Here, accurate building masks are provided as the post-processing step to all models to isolate the effect of building localization on the damage classification task.

Method	Modality	F_1^{clf} (%)	Final mIoU (%)	IoU per class (%)			
				Background	Intact	Damaged	Destroyed
UNet	SAR	68.71	69.84	100.0	88.19	35.83	55.35
	Optical-SAR	73.59	72.41	100.0	89.38	44.83	55.42
DeepLabV3+	SAR	72.12	72.19	100.0	89.59	39.63	59.54
	Optical-SAR	73.90	73.93	100.0	90.32	40.45	64.94

5.3 The role of optical pre-event data in multimodal building damage mapping

In the last section, CAM visualizations revealed that DL models also exhibit responses to disaster-specific patterns in pre-event optical imagery. This observation suggests that optical data may play a more complex role in multimodal building damage mapping than simply supporting building localization. In other words, in a multimodal bi-temporal setup, does pre-event optical imagery act solely as a localization aid, or does it provide additional semantic cues that networks can exploit for more accurate damage classification?

To explore this, we conducted controlled experiments using UNet and DeepLabV3+. Both networks were trained under two configurations: (1) using post-event SAR imagery only, and (2) using multimodal pre- and post-event inputs (optical-SAR). To isolate the contribution of pre-event optical data beyond building localization, we provided perfect building masks for postprocessing in both settings. This design ensures that any observed differences in performance are attributable to the additional information from pre-event optical imagery, rather than differences in network architecture or localization accuracy.

The results, summarized in Table 7, show that incorporating pre-event optical imagery leads to notable improvements in distinguishing building damage levels. For UNet, the IoU for the "Damaged" class increased from 35.83% (SAR only) to 44.83% (Optical-SAR), and for the "Destroyed" class from 55.35% to 55.42%. DeepLabV3+ exhibited significant gains also, with IoU improvements from 39.63% to 40.45% for "Damaged" category, and from 59.54% to 64.94% for "Destroyed" category. These results suggest that pre-event optical imagery contributes beyond mere building localization, enriching the feature space for more effective semantic comparison for different building damage levels across modalities.

Q2: The manuscript makes extensive evaluations of supervised and unsupervised change detection models, but the conceptual and methodological relationship between building damage assessment and generic change detection remains unclear, which is largely implied rather than discussed. An explicit and clearer explanation would be great for readers who lack of related background.

R2: Thank you for this insightful comment. We agree that clarifying the conceptual and methodological relationship between building damage assessment (BDA) and generic change detection (CD) will help readers unfamiliar with the field.

Specifically, a common view is to treat BDA as a special case of "one-to-many" semantic change detection tasks [1]-[4], where the goal is to assess not just whether a change has occurred but also to characterize the type and severity of the change (*i.e.*, levels of damage). In this sense, BDA extends beyond binary change detection by requiring finer-grained semantic interpretation of pre- and post-event imagery. Many existing methods for BDA are thus derived from or adapted versions of generic change detection models. Furthermore, in some unified change detection frameworks [3]-[5], BDA is explicitly included as one of the downstream tasks, highlighting their methodological overlap.

It is important to note that this discussion focuses on the formulation of BDA tasks that take bi-temporal inputs (i.e., both pre- and post-disaster images). Alternative approaches that rely solely on post-disaster imagery exist but are outside the scope of our evaluation and discussion.

We have added the above description in Section 4.1 of the revised manuscript to clarify this problem. We show the revised part below for your convenience.

[1] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sensing of Environment*, vol. 265, p. 112636, 2021.

[2] W. Lu, L. Wei and M. Nguyen, "Bitemporal Attention Transformer for Building Change Detection and Building Damage Assessment," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 4917-4935, 2024.

[3] H. Chen, J. Song, C. Han, J. Xia and N. Yokoya, "ChangeMamba: Remote Sensing Change Detection with Spatiotemporal State Space Model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-20, 2024.

[4] Z. Zheng, Y. Zhong, J. Zhao, A. Ma, and L. Zhang, "Unifying remote sensing change detection via deep probabilistic change models: From principles, models to applications," *ISPRS Journal of Photogrammetry Remote Sensing*, vol. 215, pp. 239–255, 2024.

[5] H. Guo, X. Su, C. Wu, B. Du and L. Zhang, "SAAN: Similarity-Aware Attention Flow Network for Change Detection with VHR Remote Sensing Images," *IEEE Transactions on Image Processing*, vol. 33, pp. 2599-2613, 2024.

are employed in the experiments to compare their results.

It is worth noting that in this work, we focus on the formulation of building damage assessment as a bi-temporal task, where both pre- and post-event images are used as inputs. This formulation aligns closely with generic change detection tasks, which aim to identify changes between two time points. Conceptually, building damage assessment can be viewed as a specialized "one-to-many" semantic change detection problem (Zheng et al., 2021, 2024; Lu et al., 2024), where the objective is not only to detect whether a change has occurred but also to categorize the type and severity of changes (damages) to buildings. Many existing methods are thus derived from or adapted versions of generic change detection frameworks (Chen et al., 2024; Zheng et al., 2024; Guo et al., 2024).

4.2 Benchmark suites

Q3: Since UMCD methods underperform, consider including a random guessing baseline for reference. This would contextualize the difficulty of BRIGHT and help readers understand the performance floor under UMCD setup.

R3: Thank you for your insightful suggestion. We have added the results of a random guessing baseline to **Table 12** for reference. As shown below, the different methods achieve improvements over random guessing; however, the gains are not very significant. This highlights the challenging nature of applying UMCD methods to the BRIGHT dataset. We show the revised part below for your convenience.

Table 12. Results of representative unsupervised multimodal change detection methods. KC is the acronym of kappa coefficient. The highest values are highlighted in **bold**, and the second-highest results are highlighted in underline. The accuracies on the UMCD benchmark dataset are the accuracies on the four datasets presented in Figure G1, obtained from their literature. Details of methods and benchmark datasets are presented in Appendix G. The random guessing baseline is included to indicate the performance floor under the UMCD setup. The "-" symbol indicates that the corresponding method did not report results on that dataset in their original publications.

Method	UMCD benchmark datasets			BRIGHT			
	OA	F1	KC	OA	F1	IoU	KC
Random guessing	50.0	8.4 / 6.0 / 11.0 / 11.4	0.0	50.00	7.83	4.08	0.00
IRG-McS (Sun et al., 2021)	98.3 / - / 97.1 / 97.2	80.4 / - / 75.4 / 73.7	79.4 / - / 73.9 / 75.1	**90.03**	12.65	6.75	7.65
SR-GCAE (Chen et al., 2022b)	98.6 / 98.5 / - / -	82.9 / 77.6 / - / -	82.1 / 76.9 / - / -	77.83	14.35	7.73	5.64
FD-MCD (Chen et al., 2023)	98.2 / 97.8 / - / 96.7	81.4 / 72.2 / - / 73.2	82.3 / 71.1 / - / 71.4	80.96	**15.84**	**8.60**	7.94
AOSG (Han et al., 2024)	- / - / - / 96.4	- / - / - / 77.7	- / - / - / 75.9	77.93	10.75	5.68	3.98
AGSCC (Sun et al., 2024a)	98.3 / - / 95.9 / 97.6	78.2 / - / 68.0 / 77.9	77.3 / - / 65.8 / 76.6	88.49	14.82	8.00	**9.54**
AEKAN (Liu et al., 2025)	98.7 / - / - / 98.3	83.8 / - / - / 84.7	83.1/ - / - / 83.9	81.60	13.09	7.00	3.56

Q4: While Table 1 offers a comprehensive comparison of datasets, several datasets seem relevant and should be included to enhance its completeness, like CRASAR-U-DROIDs [arXiv:2407.17673] and Noto-Earthquake building damage dataset [10.5194/essd-2024-363].

R4: Thank you for your valuable suggestion. We have reviewed the CRASAR-U-DROIDs [arXiv:2407.17673] and the Noto-Earthquake Building Damage Dataset [10.5194/essd-2024-363] and

have updated **Table 1** to include them for a more comprehensive comparison.

We show the corresponding revised part below for your convenience.

Table 1. Comparison of BRIGHT with the existing building damage assessment datasets. The OA indicates whether the dataset is open access (OA) or not, and GSD is an acronym for ground sampling distance (GSD). Note that since some datasets integrate other datasets, we summarize only the largest one to avoid duplication here. For example, the BDD dataset (Adriano et al., 2021) includes the Tohoku-Earthquake-2011 dataset (Bai et al., 2018) and Palu-Tsunami-2018 dataset (Adriano et al., 2019).

Dataset	OA	Modality	GSD (m/pixel)	No. of events	Disaster type	No. of building	Granularity
ABCD (Fujita et al., 2017)	✓	Optical EO	0.4	1	Tsunami	N/A	Image-level
(Nguyen et al., 2017)	✓	Images on social media	N/A	4	3 natural disasters	N/A	Image-level
(Cheng et al., 2021)	✓	Optical EO	N/A	1	Hurricane	1,802	Image-level
(Xue et al., 2024)	✓	Street-view image	N/A	1	Hurricane	2,468	Image-level
FloodNet (Rahnemoonfar et al., 2021)	✓	Optical EO	N/A	1	Flood	6,675	Pixel-level
RescueNet (Rahnemoonfar et al., 2023)	✓	Optical EO	N/A	1	Hurricane	10,903	Pixel-level
Ida-BD (Kaur et al., 2023)	×	Optical EO	0.5	1	Hurricane	18,083	Pixel-level
CRASAR-U-DROIDs (Manzini et al., 2024)	✓	Optical EO	0.02-0.12	10	4 natural disasters 1 man-made disaster	21,716	Pixel-level
Noto-BDA-MV (Vescovo et al., 2025)	✓	Optical EO	N/A	1	Earthquake	140,208	Pixel-level
xBD (Gupta et al., 2019)	✓	Optical EO	<0.8	15	6 natural disasters	>700,000	Pixel-level
QQB (Sun et al., 2024b)	✓	Optical and SAR EO	<1	1	Earthquake	4,029	Pixel-level
BDD (Adriano et al., 2021)	×	Optical and SAR EO	1.2-3.3	9	3 natural disasters	123,453	Pixel-level
BRIGHT	✓	Optical and SAR EO	0.3-1	14	5 natural disasters 2 man-made disasters	384,596	Pixel-level

Q5: The paper describes careful multimodal alignment but omits the software used, e.g., ENVI, ArcGIS, or QGIS. Please provide related details.

R5: Thank you for mentioning this detail. We have added information about the multimodal registration process in the **Appendix B** of the revised manuscript. Specifically, we used QGIS as the registration software, employing the [Georeferencer] plugin to align SAR images to the optical imagery as the reference. The transformation type was set to [Thin Plate Spline], and [Lanczos resampling (6×6 kernels)] was applied to ensure high-quality interpolation.

We show the corresponding revised part below for your convenience.

Appendix B: Manual registration and estimating registration errors

We performed the manual registration process using QGIS, with the "Georeferencer" plugin to align SAR images to the optical imagery as the reference. The transformation type was set to "Thin Plate Spline", and "Lanczos resampling (6×6 kernels)" was applied to achieve high-quality interpolation. The manually selected control points by EO experts on some disaster scenes are shown in Figure B1.

Q6: Appendix G includes important new experimental setups and evaluation methods for UMCD. However, too much content is composed together now. It is not easy for people to grasp information. Adding section subtitles could improve readability.

R6: Thank you for this helpful suggestion. According to your suggestion, we have revised Appendix G by dividing it into two parts for improved clarity. The first part introduces the unsupervised multimodal change detection methods, while the second part describes the proposed more practical evaluation protocol. This restructuring makes it easier for readers to grasp the key information.

Q7: 8: Please specify in the figure or caption that the values represent average ± standard deviation across models.

R7: Thank you for your thoughtful comment. We believe you were referring to **Figure 8**. We have clarified in the caption that each bar represents the mean IoU of seven deep learning models for a specific class under each disaster type, and the error bars indicate the standard deviation of IoU scores across the seven models.

We show the corresponding revised part below for your convenience.

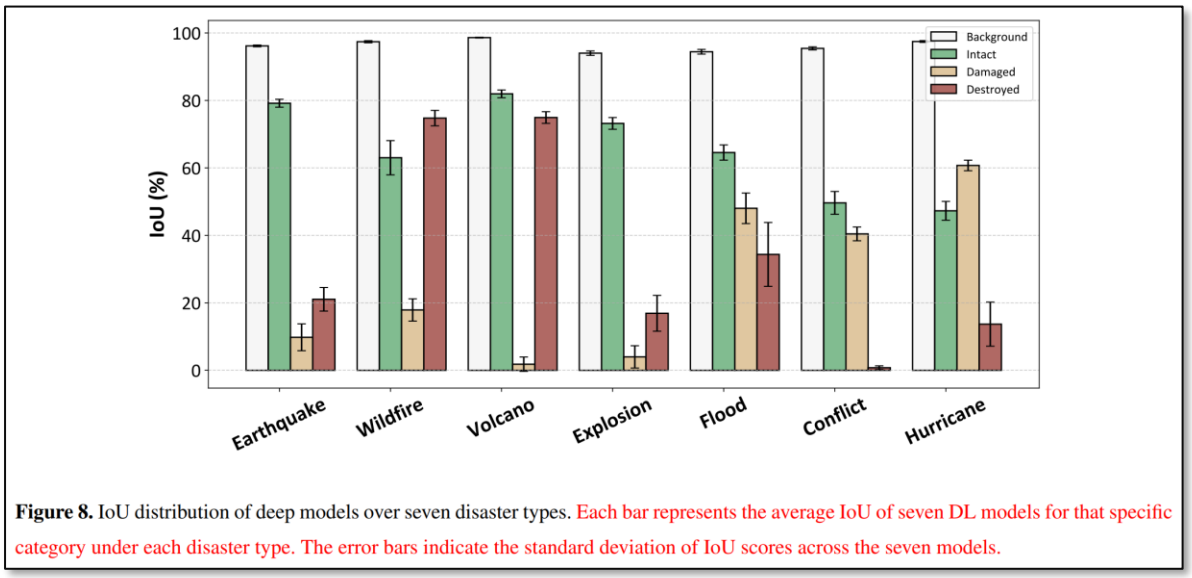

Figure 8. IoU distribution of deep models over seven disaster types. Each bar represents the average IoU of seven DL models for that specific category under each disaster type. The error bars indicate the standard deviation of IoU scores across the seven models.

Q8: 10: Add a note in the caption to clarify that each dot corresponds to performance on a single test event under cross-event transfer.

R8: Thank you for your thoughtful comment. We believe you were referring to **Figure 10**. We have added a note in the caption to clarify that each dot represents the performance on a single test event under cross-event transfer.

We show the corresponding revised part below for your convenience.

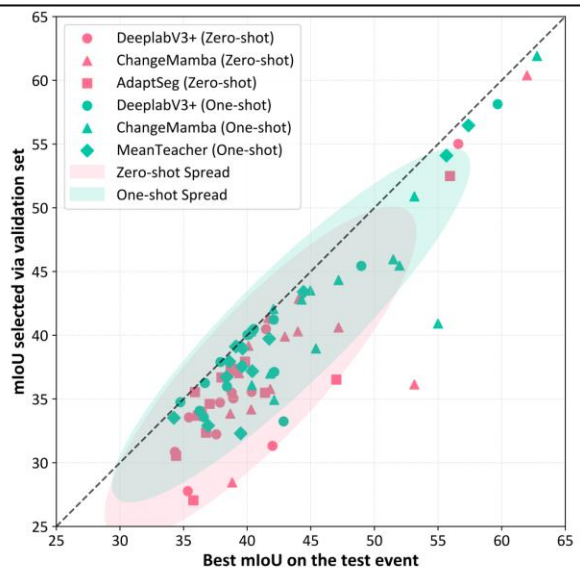

Figure 10. Comparison of models' best performance (mIoU) on test events versus the best checkpoints selected on validation sets under the cross-event transfer setting. Each point represents the performance on a single test event under cross-event transfer. The farther a point lies from the diagonal line, the larger the gap between the model's selected performance and its true upper bound.

Q9: Typo in Table 7: "Object-based major voting" should be corrected to "Object-based majority voting".

R9: Thank you for your careful review. We have corrected "Object-based major voting" to "Object-based majority voting" in Table 7 (now **Table 8** in the revised manuscript).

We show the corresponding revised part below for your convenience.

Table 8. Further contributions to mIoU from post-processing algorithms. ChangeMamba (Chen et al., 2024) is used here as the baseline. Details on these algorithms are provided in Appendix D.

Method	mIoU (set)	mIoU (event)
Baseline	67.63	51.39
Test-time augmentation	68.50	51.95
Object-based majority voting	67.22	52.08
Ensembling multiple models	68.45	52.14
All	68.86	52.31

Q10: Clarify the meaning of "–" symbols in Table 11. Do they indicate missing data or inapplicability? This should be stated explicitly.

R10: Thank you for pointing this out. The "–" symbols in Table 11 indicate that the corresponding methods did not report results on that dataset in their original publications. We have clarified this in the caption of Table 11 (now **Table 12** in the revised manuscript). We show the corresponding revised

part below for your convenience.

Table 12. Results of representative unsupervised multimodal change detection methods. KC is the acronym of kappa coefficient. The highest values are highlighted in **bold**, and the second-highest results are highlighted in underline. The accuracies on the UMCD benchmark dataset are the accuracies on the four datasets presented in Figure G1, obtained from their literature. Details of methods and benchmark datasets are presented in Appendix G. The random guessing baseline is included to indicate the performance floor under the UMCD setup. The "-" symbol indicates that the corresponding method did not report results on that dataset in their original publications.

Q11: "ML" should be defined on its first use and consistently used thereafter instead of alternating with [machine learning].

R11: Thank you for carefully checking this detail. We have defined "ML" (machine learning) at its first occurrence in the manuscript and have revised the text to ensure consistent use of the abbreviation thereafter.

Q12: Standardize currency formatting (e.g., USD vs. US$).

R12: Thank you for carefully noting this. We have standardized the currency formatting throughout the manuscript and now consistently use "USD" to avoid ambiguity.

Q13: Define abbreviations such as IGN and GSI when first mentioned as data providers.

R13: Thank you for kindly reminding us of this. We have defined the abbreviations "GSI" and "IGN" in the caption of **Table 2** in the revised manuscript. Specifically, GSI refers to The Geospatial Information Authority of Japan, and IGN refers to The Instituto Geográfico Nacional (National Geographic Institute) of Spain.

We show the corresponding revised part below for your convenience.

Table 2. Summary of basic information of the BRIGHT dataset with disaster events listed in chronological order. GSI refers to the Geospatial Information Authority of Japan, and IGN refers to the Instituto Geográfico Nacional (National Geographic Institute) of Spain.

Disaster area	Type of disaster	Date	GSD (m/pixel)	Data provider / source	No. of tiles	No. of building
Beirut, Lebanon	Explosion (EP)	04 Aug. 2020	1	Maxar & Capella	133	25,496
Bata, Equatorial Guinea	Explosion (EP)	07 Mar. 2021	0.5	Maxar & Capella	107	8,893
Goma, DR Congo	Volcano eruption (VE)	22 May 2021	0.33	Maxar & Capella	123	18,741
Les Cayes, Haiti	Earthquake (EQ)	14 Aug. 2021	0.48	Maxar & Capella	73	18,918
La Palma, Spain	Volcano Eruption (VE)	19 Sept. 2021 - 13 Dec. 2021	0.3-0.35	IGN (Spain) & Capella	933	30,239

Q14: The format of references should be standardized. Some of these entries use abbreviations for journals, while others have full titles.

R14: Thank you for your kind reminder. We have standardized all references in the revised manuscript. Journal names are now uniformly abbreviated according to the Journal Title Abbreviations by Caltech Library, as required by ESSD.