# Satellite monitoring of war urban damage with a temporal knowledge-guided deep learning scheme

Zhengyang Hou[1†], Ying Qu[1†], Liqiang Zhang[1*], Qian Shi[2], Qiwei Yu[1*], Yuanyuan Zhao[3], Hong Tang[1], Yuebin Wang[1], Xingang Li[1], Yang Li[1], Shuwen Peng[1], Xin Yao[1] and Chenghu Zhou[4]

[1]State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, 100875, China.

[2]Guangdong Provincial Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou, 510006, China.

[3]College of Land Science and Technology, China Agricultural University, Beijing, 100193, China.

[4]State Key Laboratory of Resources and Environment Information System, Institute of Geographical Science and Natural Resources, Chinese Academy of Sciences, Beijing, 100101, China.

*Corresponding author(s). E-mail(s): zhanglq@bnu.edu.cn;

†These authors contributed equally to this work.

# The File Includes:

1 Technical Terms

2 Supplementary Note 1

3 Supplementary Figures

4 Supplementary Tables

# 1 Technical Terms

**Precision:** Precision measures the accuracy of the positive predictions, which is the proportion of the true positive predictions among all the positive predictions made by the model. Precision is given by $TP/(TP + FP)$, where $TP$ is the number of the true positive predictions and $FP$ is the number of the false positive predictions.

**Recall:** Recall measures the completeness of the positive predictions, which is the proportion of the true positive predictions among all actual positive samples in the dataset. Recall is given by $TP/(TP + FN)$, where $FN$ is the number of the false negative predictions.

**F1 score:** A evaluation metric for classification tasks that takes into account both precision and recall. It is the harmonic mean of precision and recall, and ranges from 0 to 1, with 1 indicating perfect precision and recall. A high F1 score indicates that a classifier has both high precision and high recall. It is expressed by $2 \times (precision \times recall)/(precision + recall)$.

**False positive rate (FPR):** The ratio of the number of the false positives to the total number of negatives in a binary classification problem. It measures the proportion of the negative instances that are incorrectly classified as positive, and is expressed by $FP/(TN + FP)$, where $TN$ is the number of the true negative predictions.

**True positive rate (TPR):** The ratio of the number of the true positives to the total number of the positives in a binary classification problem. It measures the proportion of the positive instances that are correctly classified as positive, and is expressed by $TP/(TP + FN)$.

**Receiver operating characteristic curve (ROC curve):** The ROC curve plots the TPR against the FPR for different threshold values. A higher ROC curve implies better classification performance, as it indicates a higher TPR for a given FPR.

**Area Under the Curve (AUC):** AUC measures the area under the ROC curve, and is a single scalar value that represents the overall performance of a classifier across all possible classification thresholds. AUC ranges from 0 to 1, with a value of 0.5 indicating a random classifier and a value of 1 indicating a perfect classifier.

**Precision-recall curve (PR curve):** The PR curve evaluates the PR trade-off of a model. The PR curve plots precision on the y-axis and recall on the x-axis, with each point on the curve representing a different threshold for the model classification decision. A higher PR curve implies better classification performance, as it indicates a higher precision for a given recall. Because both

precision and recall are calculated based on positive instances. Compared to the ROC curve, the PR curve can capture the classifier's performance on positive instances better, and is more suitable in situations when the positive class is rare and the goal is to maximize precision.

**Average precision (AP):** AP is a measure of the average value of the precision-recall curve over all possible recall values. It summarizes the precision-recall curve as the weighted mean of precision achieved at each threshold, with the recall as the weight. AP ranges from 0 to 1, with a value of 0 indicating a random classifier and a value of 1 indicating a perfect classifier.

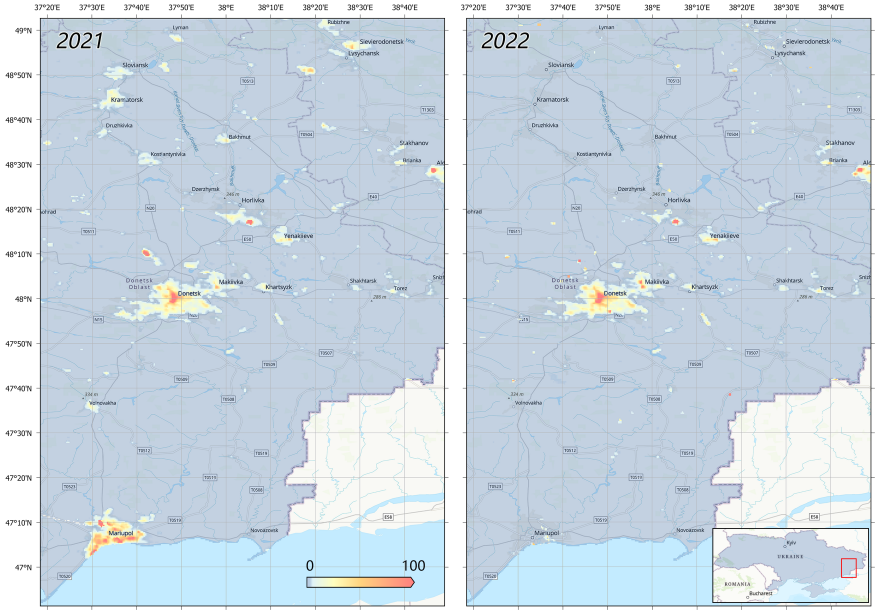
**Gradient-weighted class activation mapping (Grad-CAM):** Grad-CAM is a visual interpretation technique used to visualize important feature regions in a network model to aid in understanding the mechanism behind model decisions. It determines which regions in the feature map contribute more to the model classification decision by calculating the gradient weights of the model output with respect to the input feature map. Specifically, the Grad-CAM can compute the importance of each feature map channel to the predicted class label by backpropagating the predicted output label and calculating the gradients. The importance of these channels is then summed with weights to produce a heat map, where the colors represent the contribution of each region to the classification result.

**10-fold cross validation:** It involves dividing the available data into 10 equal parts, or "folds", and using 9 of them for training and the remaining ones for testing. This process is repeated 10 times, with each fold used as the test set, while the other 9 folds are used for training. The final performance metric is the average of the 10 metrics. Cross-validation can evaluate the model's ability to adapt to different data distributions by randomly shuffling the data in each fold, which helps assess the model performance on unknown data.

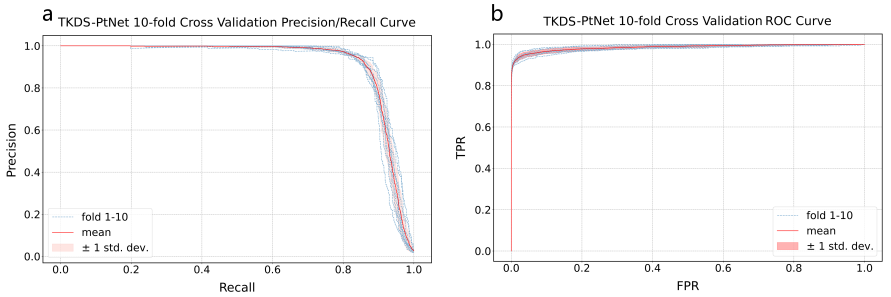
# 2 Supplementary Note 1

**Performance of the TKDS-PtNet-SCL** On the 10m-resolution Syrian dataset, the improvement in accuracy of the TKDS-PtNet-SCL come from two aspects: utilization of temporal pattern knowledge in the TKDS and extraction of inter-class difference features learned by SCL. First, the TKDS improved the F1 score of PtNet by 27.4 (66.0 vs 38.6). Then, we integrated PtNet with SCL, and obtained the weights of the upstream task after 1000 epochs of training as the initial weights of PtNet in the TKDS-PtNet model. The TKDS-PtNet model with SCL weights (i.e. TKDS-PtNet-SCL) achieved the highest F1 score (72.4), higher than that of the TKDS-PtNet model (66.0). Moreover, the F1 score of PtNet increased to 40.26 from 38.6 after the SCL weights were added, which was the same detection accuracy level as ResNet-50 on 0.5m-resolution remote sensing images (F1=40.5). Since our method fully utilized the temporal and spectral information of pre- and post-damage features, it achieved higher detection accuracy for building damage from both high- and medium-resolution remote sensing images, even with extremely imbalanced samples.

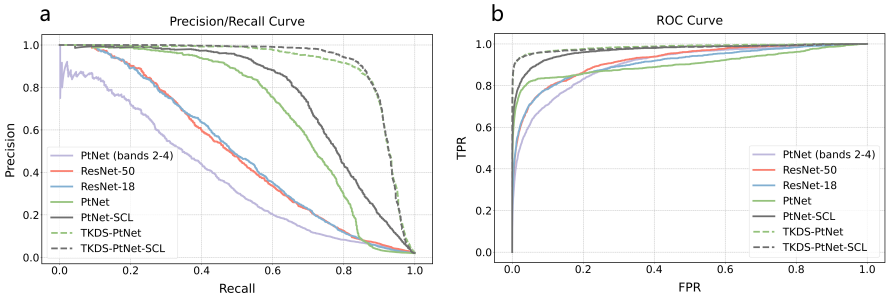
# 3 Supplementary Figures



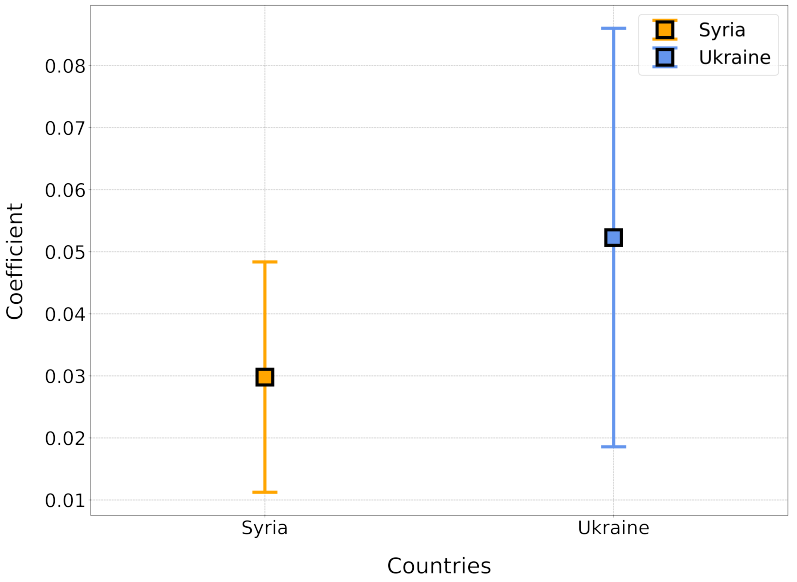

**Supplementary Fig. 1 | Average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band in 2021 (left) and March to June 2022 (right) in southeastern Ukraine.** The average radiance of the four Ukrainian cities (i.e. Sievierodonetsk, Rubizhne, Volnovakha and Mariupol cities) in March to June 2022 was only 8.1% in 2021, and Mariupol city was only 6.5% in 2021 (those in Sievierodonetsk, Rubizhne and Volnovakha areas were 18.85%, 32.67% and 10.62%, respectively).



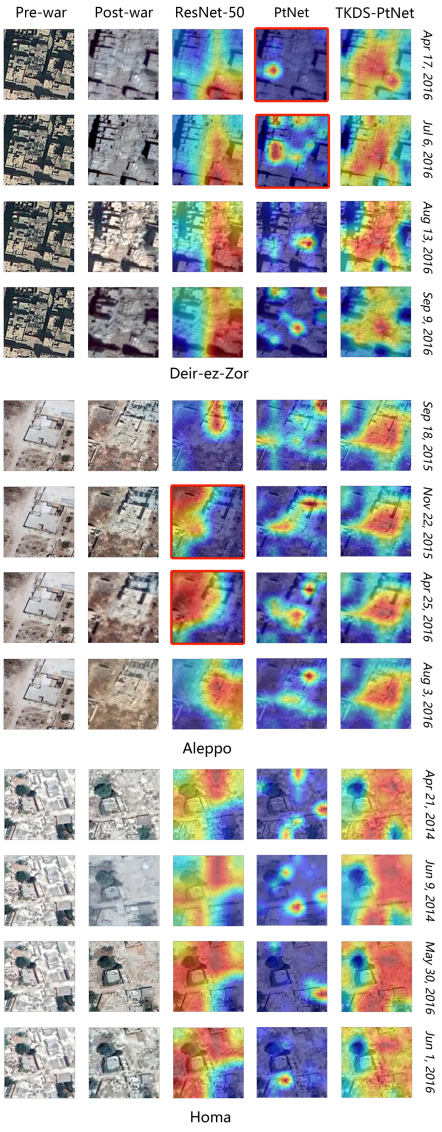

**Supplementary Fig. 2 | The performance of the TKDS-PtNet on the Ukrainian urban damage with 10-fold cross validations.** The blue dashed lines represent the PR curves in **a**, and the AUC/ROC curves in **b** for the folds 1-10. The red solid lines represent the mean for the folds 1-10. The pink region represents the positive and negative one standard deviation.



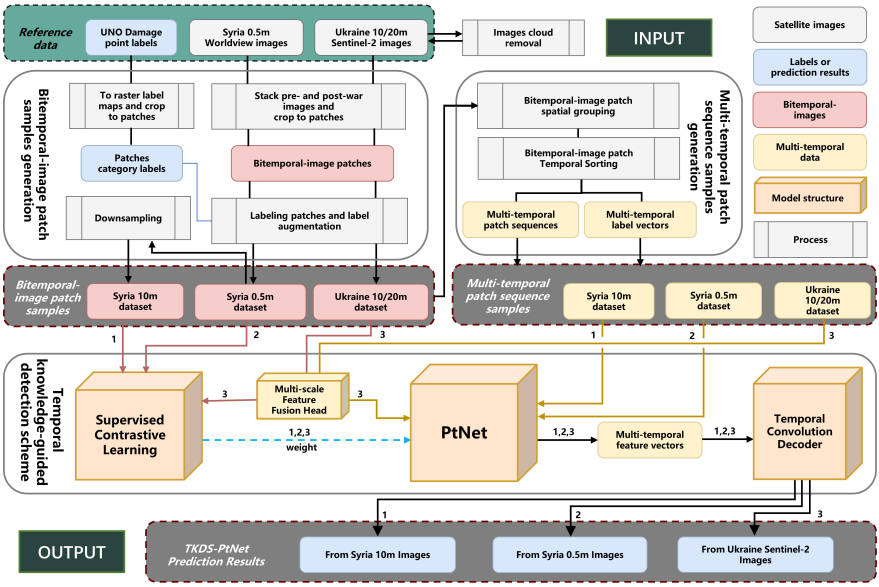

**Supplementary Fig. 3 | Precision performance of the models using Sentinel-2 images in Ukrainian cities. a.** The precision-recall curve. **b.** The ROC curve.



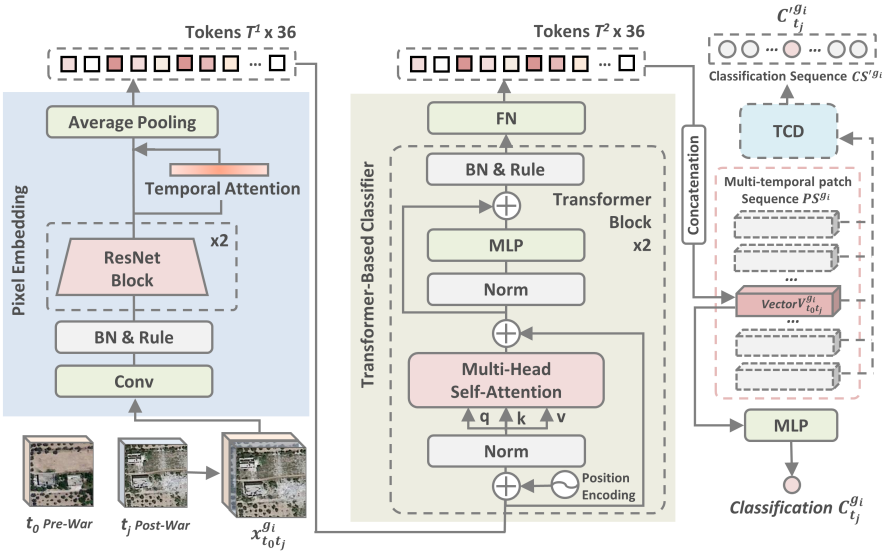

**Supplementary Fig. 4 | The associations of the predicted results of building damage with bombing events from LiveUAmap in Syria and Ukraine.** Error bars represent 95% confidence intervals. The robust standard error is clustered at the location of patch level.



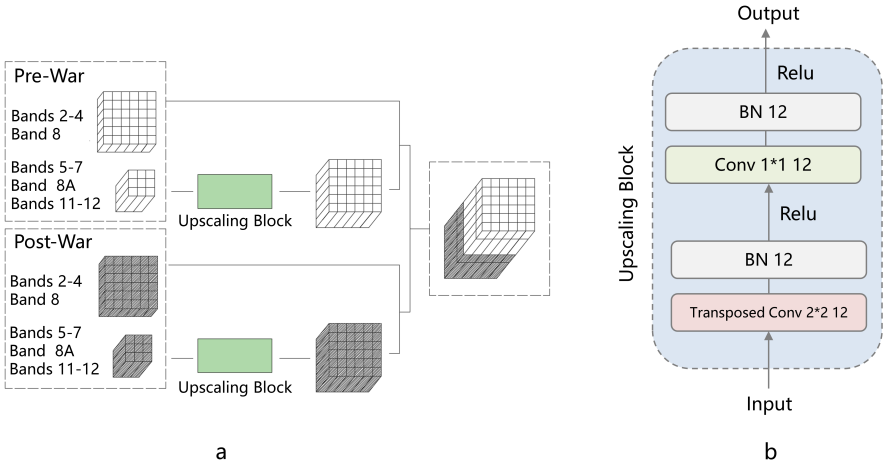

**Supplementary Fig. 5 | Model interpretability based on the Grad-CAM, showing four groups of 0.5m-resolution multi-temporal images from different damaged areas in Syrian cities.** The sub-figures in the first and second columns from the left to the right show the pre-war images from the same date and post-war damage images from four different dates. Those in the third, fourth, and fifth columns show the activation maps of ResNet-50, PtNet, and TKDS-PtNet, respectively, with red borders representing the incorrect class output.



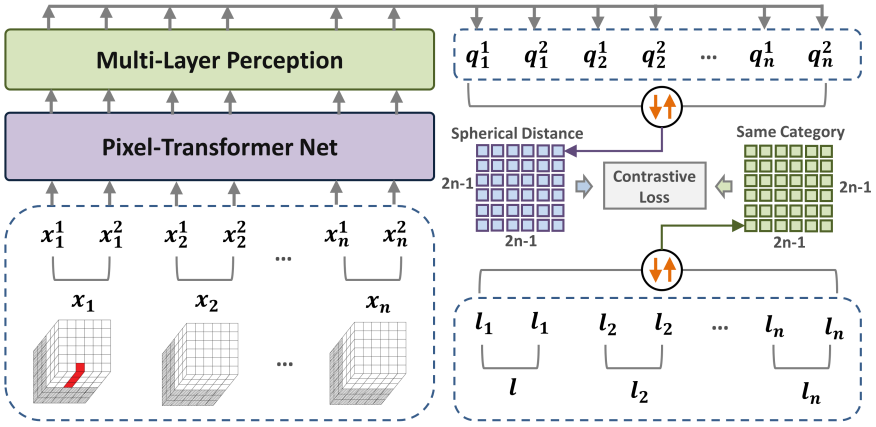

**Supplementary Fig. 6 | The urban destruction monitoring workflow.**



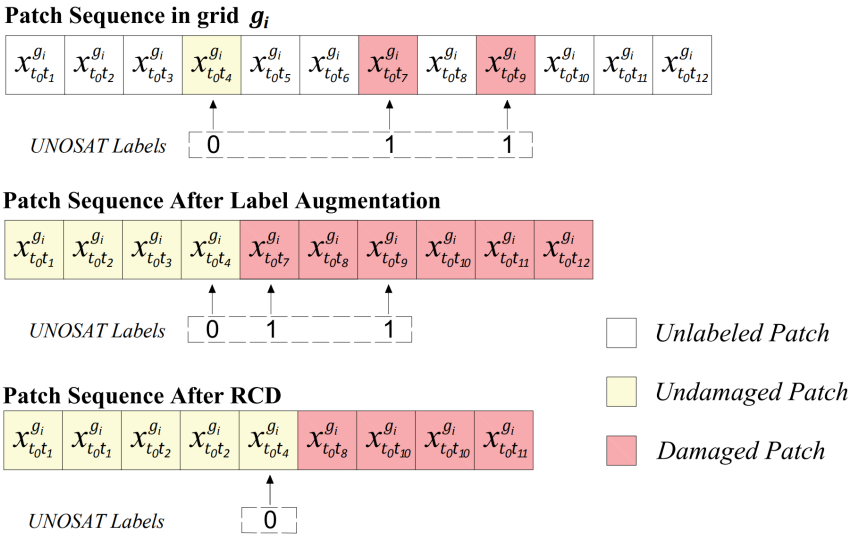

**Supplementary Fig. 7 | The TKDS-PtNet architecture.**




**Supplementary Fig. 8 | The architecture of the multi-scale feature fusion head.**




**Supplementary Fig. 9** | **The supervised contrastive learning architecture.**




154

155 **Supplementary Fig. 10 | Random Copy and Delete (RCD) used to eliminate**
156 **pseudo temporal regularity.** $t_j$ **is the capture time of the images.**

# 4 Supplementary Tables

**Supplementary Table 1 | Performance and generalization of the models in Ukrainian cities**

| Country | Region | Train/ validation patches | Destroyed patches, % | Train/ validation MPS* | MPS* with destroyed patches, % |
|---|---|---|---|---|---|
| Syria | Aleppo | 469,558/201,24 | 2.84% | 29,365/12,586 | 6.21% |
| | Hama | 98,896/42,456 | 5.90% | 8,325/3,569 | 9.98% |
| | Homs | 47,328/20,282 | 9.05% | 6,756/2,896 | 10.16% |
| | Raqqa | 46,369/19,889 | 4.73% | 5,115/2,193 | 7.38% |
| | Deir-Ez-Zor | 50,570/21,706 | 4.78% | 4,571/1,961 | 6.34% |
| | Idlib | 21,518/9,217 | 2.18% | 1,790/768 | 4.14% |
| | All | 734,239/314,794 | 3.88% | 55,922/23,973 | 7.30% |
| Ukrain | Mariupol | 53,530/22,878 | 2.41% | 12,877/5,524 | 1.66% |
| | Rubizhne | 57,269/24,573 | 2.50% | 5,429/2,328 | 4.19% |
| | Siwvierodoetsk | 66,426/28,463 | 0.39% | 6,438/2,761 | 0.85% |
| | Volnovskha | 23,960/10,277 | 3.29% | 3,928/1,685 | 3.60% |
| | All | 201,185/86,191 | 1.87% | 28,672/12,298 | 2.22% |

MPS*: multi-temporal patch sequence

**Supplementary Table 2 | Performance and generalization of the models in Ukrainian cities**

| Resolution | City | ResNet-50 | | PtNet | | TKDS-PtNet | |
|---|---|---|---|---|---|---|---|
| | | F1 | AUC | F1 | AUC | F1 | AUC |
| 10m | Aleppo | 24.14 | 82.85 | 33.32 | 86.77 | 57.61 | 95.59 |
| | Hama | 35.40 | 75.63 | 48.46 | 83.96 | 70.84 | 95.21 |
| | Homs | 35.32 | 77.47 | 46.35 | 84.85 | 92.71 | 98.83 |
| | Raqqa | 36.01 | 78.91 | 46.87 | 82.86 | 86.16 | 98.23 |
| | Deir-Ez-Zor | 3.09 | 73.25 | 31.13 | 79.66 | 54.35 | 91.67 |
| | Idlib | 5.69 | 65.35 | 22.22 | 72.08 | 34.80 | 80.17 |
| | All | 28.52 | 80.74 | 38.56 | 85.57 | 65.99 | 95.86 |
| 0.5m | Aleppo | 39.29 | 87.28 | 52.89 | 88.83 | 82.97 | 98.66 |
| | Hama | 47.98 | 79.88 | 63.67 | 85.45 | 87.60 | 83.20 |
| | Homs | 46.07 | 84.38 | 58.26 | 88.58 | 98.26 | 99.94 |
| | Raqqa | 40.85 | 82.76 | 59.06 | 88.59 | 95.41 | 99.41 |
| | Deir-Ez-Zor | 19.37 | 70.92 | 38.13 | 79.79 | 72.06 | 95.14 |
| | Idlib | 16.90 | 72.36 | 22.77 | 73.81 | 70.23 | 96.12 |
| | All | 40.50 | 84.22 | 54,68 | 88.15 | 86.05 | 98.51 |

**Supplementary Table 3 | The associations of the predicted building damage results with bombing events from LiveUAmap in Syrian and Ukrainian cities**

| | (1)<br>Syria<br>Predicted value | (2)<br>Ukraine<br>Predicted value |
|---|---|---|
| After bombing event or not | 0.029799***<br>(0.009479) | 0.052290***<br>(0.017204) |
| Grid Fixed Effects | Yes | Yes |
| City-Time Fixed Effects | Yes | Yes |
| Obs. | 1,310,123 | 2,012,674 |
| Mean | 0.026675 | 0.011871 |

**Notes.** The first row lists the coefficients from the two-way fixed effects model, and robust standard errors in parentheses. $*P<0.1$, $**P<0.05$, $***P<0.01$. Dependent variables in Columns (1) and (2) are the prediction scores of building damage in the six Syrian cities and four in Ukrainian cities using the TKDS-PtNet