Large Vision Model for Structural Damage Diagnosis

Yang XU[1,2][0000-0002-8394-9224]

[1] Key Lab of Smart Prevention and Mitigation of Civil Engineering Disasters of the Ministry of Industry and Information Technology, Harbin Institute of Technology, Harbin 150090, China
[2] School of Civil Engineering, Harbin Institute of Technology, Harbin, 150090, China
xyce@hit.edu.cn

Abstract. Conventional deep-learning-based structural damage recognition using images usually requires well-designed network structures for various damage categories and complicated procedures of hyperparameter tuning and retraining. Recently, pre-trained large foundation models in unimodal of vision and multimodal of vision-language have been established to perceive fundamental knowledge of large-scale visual and linguistic datasets, which provides great potential for structural damage diagnosis with full use of structural inspection images and texts. This study introduced a large model pipeline for structural damage diagnosis by establishing a large vision model for visual damage segmentation and a vision-language model for linguistic damage description towards intelligent structural inspection. First, a large vision model based on DINO was proposed via cross-level feature alignment and contrastive learning for universal structural damage segmentation. The recognition accuracy and model robustness were validated by various types of structural components and surface damage for actual bridges and buildings. Then, a large vision-language model based on miniGPT-v2 was developed via domain knowledge embedding and cross modal learning for multi-round dialogue of human-agent interaction to describe details of structural damage. The model capacity and generalization ability were further demonstrated on several downstream tasks in vision-language cross modality of image captioning, visual question answering, and visual grounding. The results preliminarily indicated the feasibility and effectiveness of the proposed large model paradigm for structural damage diagnosis.

Keywords: Unsupervised Structural Damage Segmentation, Cross-level Feature Alignment, Contrastive Learning

1 Introduction

For bridges, it is of great significance to maintain operational durability, maintenance safety, and structural reliability during the entire service period. Timely detection and accurate recognition of multi-type surface damage is essential for maintenance decision making to mitigate the risk of potential structural failures. For past several decades, time-consuming and labor-intensive manual inspection has been adopted as the basic solution to detect structural damage, which faces severe limitations due to high

dependences on subjective experiences and lacking stability, reliability, and efficiency for real-world applications [1].

Computer vision has demonstrated its efficacy and efficiency in image-based structural health monitoring and damage recognition. Conventional digital image processing algorithms have been widely utilized to detect surface damage. However, the model performance heavily relies on manually selected parameters, including the type of edge detector, the sequence and number of image open and close operations, and the size of selected structural elements. Subsequently, machine-learning-based approaches are introduced as intelligent solutions for extracting image features of structural damage and training the corresponding recognition models. However, hand-crafted features are required as well-chosen inputs, and the robustness under complex real-world background and generalization ability could not be guaranteed [2].

Deep learning has achieved significant progress in establishing end-to-end mapping between input image and object annotation as automatic multi-level feature extractors, in which convolutional neural networks (CNNs) are the most widely investigated [3-5]. Generally, the CNN-based methods for structural damage recognition are often confined to specific datasets that contain only one or few damage types and application scenarios. This limitation may compromise the generalization ability to new damage categories and disaster scenarios [6]. Additionally, they often require a substantial number of annotated images to obtain satisfactory model performance, which is inconsistent with the occasionality and sparsity of structural damage for certain real-world applications. Therefore, the recognition accuracy is significantly influenced by the quantity of labeled samples, inter-class balance, and the comprehensiveness of collected damage categories [7]. However, practical engineering applications always necessitate that the trained structural damage recognition model possesses a reasonable generalization capability across various scenarios while maintaining high accuracy on unlabeled images. To avoid training various deep networks on diverse datasets in a fragmented manner, it is essential to develop a universal, accurate, and stable vision recognition model for multi-type structural damage, which should be capable of effectively functioning under multi-scale real-world scenarios with complex background interferences [8].

Currently, unsupervised and self-supervised learning are cutting-edge techniques as potential approaches for autonomous structural damage detection, particularly in the context of small labeled datasets [9]. Although unsupervised semantic segmentation methods based on contrastive learning have begun investigations in the computer vision field, it is still challenging to establish a universal vision recognition model for pixel-level structural damage segmentation tasks, especially for dealing with a large volume of unlabeled images under various real-world scenarios [10].

To address the above challenges, this study proposes a large vision model for universal structural damage segmentation. Section 2 introduces the network architecture of the proposed universal unsupervised damage segmentation model. Section 3 describes the investigated imageset of multi-scale multi-type structural components and surface damage. Section 4 presents a series of test results to demonstrate the effectiveness, robustness, and generalization capability of the established model under real-world inspection scenarios with complex background disturbances for cable-supported bridges and concrete bridges. Finally, Section 5 concludes this paper.

2 Methodology

Figure 1 shows the overall schematic of the proposed model architecture for universal structural damage segmentation. Different from conventional supervised-learning-based semantic segmentation models, this study establishes a knowledge distillation pipeline of teacher-student networks in an end-to-end unsupervised learning manner for the model training process. Each teacher and student branch comprises a data augmentation module, a frozen visual backbone subnetwork based on transformer, a fine-tuned segmentation head based on CNN in sequence. The self-supervised model updating strategy is designed based on a synthetical loss function of correlation loss and contrastive loss. Upon completion of the training process, new images are directely fed into the frozen visual backbone, the well-trained segmentation head, and a post-processing module of semantic clustering to perform pixel-level segmentation of structural damage as the prediction phase.

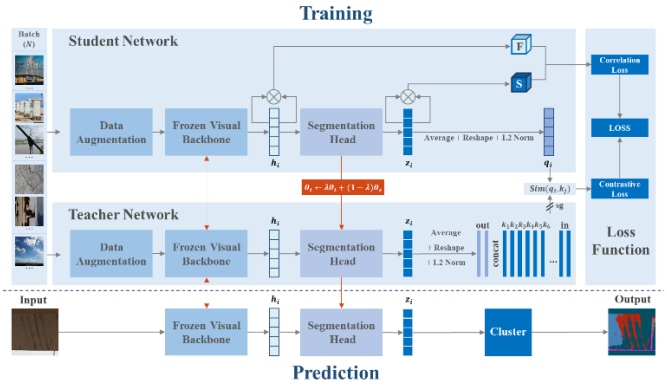


Fig. 1. Model architecture for universal unsupervised segmentation of structural damage.

For each individual instance of input image, feature maps with the same dimensions of channel, height, and width are obtained before and after the segmentation head. For each branch of student and teacher networks, spatial points on feature maps before the segmentation head are noted as f_{chw} and $g_{ch'w'}$, while spatial points on feature maps after the segmentation head are noted as s_{chw} and $t_{ch'w'}$, respectively. The feature correspondence $F_{hwh'w'}$ between f_{chw} and $g_{ch'w'}$ and segmentation correspondence $S_{hwh'w'}$ between s_{chw} and $t_{ch'w'}$ are obtained by calculating the point-wise cosine similarity as

$$F_{hwh'w'} = \frac{\sum_{c=1}^{C} f_{chw} \times g_{ch'w'}}{\|\boldsymbol{f}_{hw}\|_2 \times \|\boldsymbol{g}_{h'w'}\|_2}, S_{hwh'w'} = \frac{\sum_{c=1}^{C} s_{chw} \times t_{ch'w'}}{\|\boldsymbol{s}_{hw}\|_2 \times \|\boldsymbol{t}_{h'w'}\|_2} \quad (1)$$

For N input images within an input batch, the feature correspondence tensors and segmentation correspondence tensors could be denoted as $F_1, \ldots, F_N \in \mathcal{R}^{H\times W\times H\times W}; S_1, \ldots, S_N \in \mathcal{R}^{H\times W\times H\times W}$ with four-dimensional elements of $F_{hwh'w'}$ and $S_{hwh'w'}$. The dense semantic correlation loss is calculated based on feature correspondence tensors and segmentation correspondence tensors by

$$F^{SC}_{hwh'w'} = F_{hwh'w'} - \frac{1}{HW}\sum_{h',w'} F_{hwh'w'}, \tag{2}$$

$$L_{corr} = -\sum_{h,w,h',w'} (F^{SC}_{hwh'w'} - b) max\,(S_{hwh'w'}, 0)$$

where $F^{SC}_{hwh'w'}$ denotes the feature correspondence tensor after spatial centralization, b is a hyperparameter to avoid model collapse and ensure a positive correlation loss value. The contrastive loss between the teacher-student networks is defined as

$$L_{cont} = -\sum_{i=1}^{N} log\left\{\frac{exp[Sim(q_i,k_+)/\tau]}{\sum_{j=1}^{K} exp[Sim(q_i,k_j)/\tau]}\right\} \tag{3}$$

where Sim denotes the cosine similarity between two vectors, q_i denotes the *i*th query feature vector obtained from the student branch for the *i*th image in a batch, k_+ denotes the feature vector obtained from the teacher branch as the positive sample of the corresponding query image, N denotes the batch size, k_j denotes the *j*th referenced feature vector in the feateure dictionary, K denotes the queue length of the preset feature dictionary, τ denotes a temperature hyperparameter.

The synthetic loss function is defined by a weighted sum of correlation loss and contrastive loss as

$$Loss = \alpha L_{corr} + (1-\alpha) L_{cont} \tag{4}$$

where α denotes the weight coefficient of the correlation loss.

3 Implementation Details

This study focuses on the unsupervised semantic segmentation of universal multi-type structural damage images under real-world applications with complex background interferences. A hierarchical structural damage imageset is constructed to include multi-scale information of environment, structure, component, and damage. Some representative samples in the hierarchical structural damage imageset are shown in Figure 2.

A total of 20K images with distinct resolutions are uniformly resized into a consistent resolution of 1,024 × 1,024. Each resized image is cropped into 224 × 224 with a sliding window of 100 pixels to generate sufficient image patches instead of directly downsampling, avoiding possible feature leakage of minor damage. 128 image patches are randomly selected to form an input batch of the proposed method.

The selection of training hyperparameters plays a vital role in ensuring the optimal model performance for deep learning. After several trials, the training hyperparameters are determined. Note that although the reported configurations might not be globally optimal, the trained large vision model for universal structural damage segmentation could obtain a satisfactory segmentation accuracy, good robustness to complex background, and generalization capacity under new scenes. Therefore, this study does not focus on the determination of the best hyperparameter setups and instead seeks to validate the feasibility and effectiveness of the proposed large vision model for universal structural damage segmentation under real-world inspection scenarios.

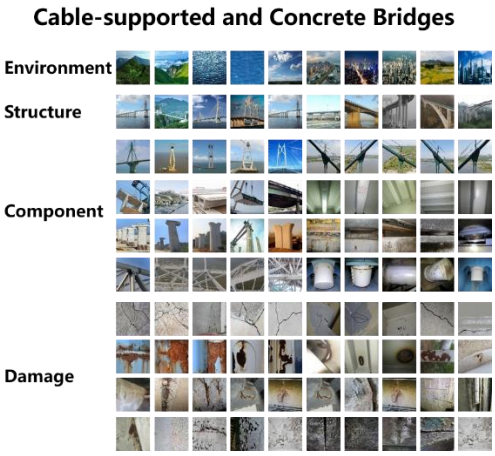
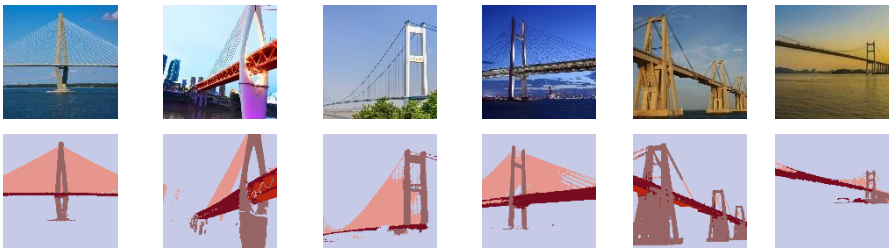


Fig. 2. Representative images of hierarchical structural damage with multi-scale information.

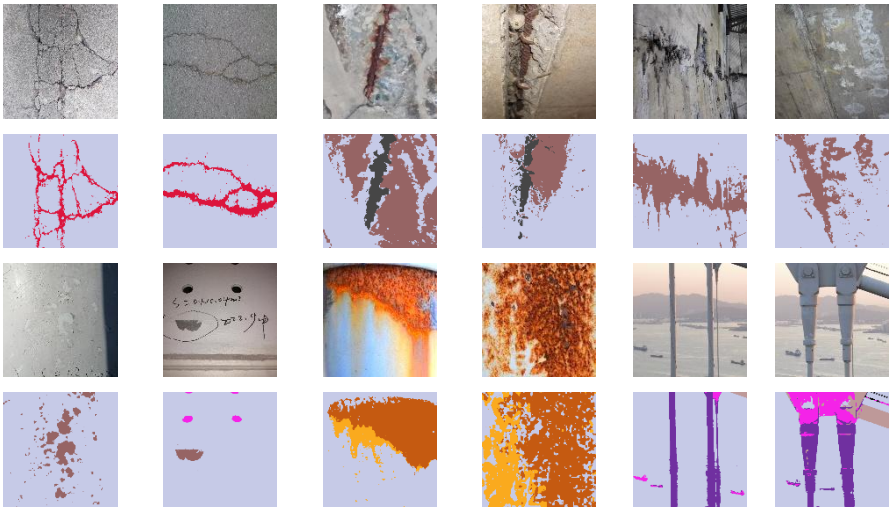
The proposed large vision model for universal structural damage segmentation is trained and tested under the software environment of PyTorch 1.8 and Python 3.7 on a 48G GPU of NVIDIA RTX A6000, and the average training time with the reported hyperparameter configurations is about 48 hours to obtain a well-trained model.

4 Results and Discussion

Figure 3(a) shows some representative prediction results on coarse-grained segmentation of main bridge structures: (a) for cable-supported bridges and (b) for concrete bridges. The test PA, mIoU, and FWIoU are 97.17%, 91.47%, 94.64% for cable-support bridges and 92.72%, 82.46%, 86.98% for concrete bridges. The results show that main components of pylon, cable, girder, deck, and pier can be generally identified from entire images of bridge structures. Figure 3(b) shows some representative prediction results on fine-grained segmentation of multi-type structural damage for bridges, including concrete crack, concrete spalling, rebar exposure, water seepage, saltpetering, steel fatigue crack, coating spalling, steel corrosion, and fire burning. Table 1 shows the evaluation metrics for multi-type structural damage segmentation of bridges. Additionally, it is observed that the proposed method can make clear distinctions between coupled damage of concrete spalling and rebar exposure, and it also successfully separates severe corrosion regions from slight corrosion regions.



(a) coarse-grained segmentation of main bridge structures



(b) fine-grained segmentation of multi-type structural damage

Fig. 3. Representative segmentation results of multi-scale bridge structures and surface damage.

Table 1. Evaluation metrics for multi-type structural damage segmentation of bridges.

Bridge structural damage type	PA	mIoU	FWIoU
Concrete crack	96.19%	69.85%	93.35%
Concrete spalling/rebar exposure	98.97%	74.12%	98.30%
Water seepage/saltpetering	88.28%	75.15%	79.21%
Steel fatigue crack	96.21%	68.07%	94.69%
Coating spalling/steel corrosion	91.07%	75.39%	84.83%
Fire burning	95.85%	76.22%	93.02%

5 Conclusions

This study proposes a universal unsupervised structural damage segmentation model in a self-supervised paradigm based on correlation learning and contrast learning to address challenges of high dependences on sufficient, complete, and high-quality image-annotation pairs of fragmented recognition models by conventional supervised learning. The main conclusions are obtained as follows:

(1) A unified semantic segmentation architecture for multi-type structural components and surface damage is established following a knowledge distillation pipeline of teacher-student networks.

(2) Unlabelled image pairs after random data augmentation are utilized as inputs. correlation learning strategy between high-level feature maps of frozen backbone

network and dense segmentation maps of fine-tuned segmentation head is introduced to ensure the cross-level feature alignment of various component and damage regions inside each image. A contrastive learning module between the normalized aggregated feature vectors across student and teacher branches is employed to quantify the intra-instance similarity and inter-instance separability among different images.

(3) A synthetic loss function comprising a correlation loss and a contrastive loss is designed. A multi-scale image dataset of multi-type components and damage for various bridge structures is constructed. Comparative studies validate the segmentation accuracy, generalization ability, and robustness under complex background disturbances.

Acknowledgments

Financial support for this study was provided by the National Key R&D Program of China [2023YFC3805900], National Natural Science Foundation of China [52192661], Heilongjiang Provincial Natural Science Foundation [LH2022E070], and Fundamental Research Funds for the Central Universities [HIT.NSRIF202334].

References

1. Xu, Y., Qian, W., Li, N., Li, H.: Typical advances of artificial intelligence in civil engineering. Advances in Structural Engineering, 25(16): 3405-3424 (2022).
2. Xu, Y., Fan, Y., Li, H.: Lightweight semantic segmentation of complex structural damage recognition for actual bridges. Structural Health Monitoring, 22(5), 3250-3269 (2023).
3. Wang, Y., Jing, X., Cui, L., Zhang, C., Xu, Y., Yuan, J., Zhang, Q.: Geometric consistency enhanced deep convolutional encoder-decoder for urban seismic damage assessment by UAV images. Engineering Structures, 286, 116132 (2023).
4. Wang, Y., Jing, X., Xu, Y., Cui, L., Zhang, Q., & Li, H.: Geometry‐guided semantic segmentation for post‐earthquake buildings using optical remote sensing images. Earthquake Engineering & Structural Dynamics, 52(11), 3392-3413 (2023).
5. Xu, Y., Qiao, W., Zhao, J., Zhang, Q., & Li, H.: Vision-based multi-level synthetical evaluation of seismic damage for RC structural components: a multi-task learning approach. Earthquake Engineering and Engineering Vibration, 22(1), 69-85 (2023).
6. Xu, Y., Bao, Y., Zhang, Y., Li, H.: Attribute-based structural damage identification by few-shot meta learning with inter-class knowledge transfer. Structural Health Monitoring, 20(4), 1494-1517 (2021).
7. Xu, Y., Fan, Y., Bao, Y., Li, H.: Task-aware meta-learning paradigm for universal structural damage segmentation using limited images. Engineering Structures, 284, 115917 (2023).
8. Zhong, J., Fan, Y., Zhao, X., Zhou, Q., Xu, Y.: Multi-type structural damage image segmentation via dual-stage optimization-based few-shot learning. Smart Cities, 7(4), 1888-1906 (2024).
9. Xu, Y., Fan, Y., Bao, Y., Li, H.: Few-shot learning for structural health diagnosis of civil infrastructure. Advanced Engineering Informatics, 2024, 62, A, 102650 (2024).
10. Fan, Y., Li, H., Bao, Y., Xu, Y.: Cycle-consistency-constrained few-shot learning framework for universal multi-type structural damage segmentation. Structural Health Monitoring (2024).