Kalpa Publications in Computing

Volume 22, 2025, Pages 222–232

Proceedings of The Sixth International Conference on Civil and Building Engineering Informatics

Semantic-Enriched Image Retrieval for Bridge Damage Assessment

Chengzhang Chai[1], Jiucai Liu[2], Yan Gao[3], Guanyu Xiong[4] and Haijiang Li[5*]

1) Ph.D. Candidate, School of Engineering, Cardiff University, Cardiff, UK. Email: chaic1@cardiff.ac.uk
2) Ph.D. Candidate, School of Engineering, Cardiff University, Cardiff, UK. Email: liuj151@cardiff.ac.uk
3) Ph.D., School of Engineering, Cardiff University, Cardiff, UK. Email: gaoy74@cardiff.ac.uk
4) Ph.D. Candidate, School of Engineering, Cardiff University, Cardiff, UK. Email: xiongg@cardiff.ac.uk
5) Prof., School of Engineering, Cardiff University, Cardiff, UK. Email: lih@cardiff.ac.uk
* Corresponding author

Abstract: Periodic bridge damage inspections result in a vast number of image records stored in a database, which can be used as a reference for the subsequent damage assessment. However, the currently used content-based image retrieval (CBIR) techniques are limited by the 'semantic gap'. They tend to consider only the low-level visual features in an image and ignore the high-level semantic information. This study proposes a semantic-enriched image retrieval framework (SEIR-Net) for bridge damage assessment. The framework enables the image encoder to extract low-level visual features and high-level semantic information by fine-tuning the multi-modal image captioning model (CNN-LSTM). The high-dimensional vectors extracted using the fine-tuned encoder are stored in the FAISS vector database, and efficient retrieval is achieved based on L2 Euclidean distance. Retrieval evaluation was performed on a damage dataset constructed on the real-world bridge inspection report, and our proposed method outperforms the commonly used VGG-16 and ResNet-50 models on the mAP and Recall@K (K=1, 2, 4) metrics. These results suggest that incorporating the semantic content of damage in image retrieval would be more beneficial for assessment references. In summary, this study effectively enhances the utility of historical image records in bridge damage assessment through semantic-enriched image retrieval techniques.

Keywords: Deep learning, Semantic enriched, Image retrieval, Bridge damage

1. INTRODUCTION

Bridges are one of the key components in the global transport network. The aging of bridges is now a growing problem. The American Society of Civil Engineers (ASCE) released the America's Infrastructure 2021 Report Card, which states that there are more than 617,000 bridges across the United States (ASCE, 2021). 42% of these bridges are at least 50 years old, and as many as 46,154 (7.5% of the nation's bridges) are considered to be at risk for structural deficiencies (Ali et al., 2022; Cha et al., 2024; Spencer et al., 2019). These figures highlight the urgency of carrying out inspections and raise a widespread concern.

Regular damage inspections generate many image records stored in a database, which can be a

valuable reference. As the saying goes, 'A picture is worth a thousand words.' Engineers can retrieve images of the same type or similar scenarios from the historical records as a reference to assist in on-site diagnosis during subsequent damage assessment. Therefore, the effective use of such image information has become a vital issue in bridge maintenance assessment.

Image retrieval techniques are generally classified into two categories: tag-based methods and content-based methods. Tag-based image retrieval (TBIR) relies on information provided by metadata or text for retrieval (Lee et al., 2017). It usually includes pre-defined template information such as name, time of capture, geographic location, device information, and other relevant details. For example, Sun et al. proposed a generic framework for tag-driven social image retrieval, starting from five orthogonal dimensions (tag relatedness, tag discrimination, tag length normalization, tag query matching model, and query model) to improve the accuracy of image retrieval (Sun et al., 2011). Li et al. developed a retrieval system called 'BIMSeek', which calculates the similarity between components based on the attribute information (including geometric features and semantic attributes) of BIM components, thus enabling efficient retrieval of components (Li et al., 2020). Ma et al. used photos' metadata (including information such as photographed location, camera perspective, and image semantic content) to build an image retrieval system that efficiently retrieves large amounts of facility management data (Ma et al., 2021).

In comparison, content-based image retrieval (CBIR), which typically relies on visual features of images such as texture, shape, colour, and spatial information, has made more significant progress in recent years (Dubey, 2021; Li et al., 2021). In an early exploratory phase, Brilakis and Soibelman completed an attempt at content-based construction image retrieval using the technique of blind relevance feedback (Brilakis & Soibelman, 2005). The developed system makes better use of the image's texture features and shape features to quickly find relevant images compared to traditional tag-based retrieval methods. In addition, Brilakis et al. used a clustering approach to assign the material feature information extracted by Fourier analysis and edge detection to the nearest centroid clusters (Brilakis et al., 2006). The presence of construction materials in the image is detected by matching with pre-defined material samples, which in turn completes the image retrieval. The above explorations mainly started from a single feature, i.e., only the visual features of the image were considered for retrieval. Based on this, Brilakis and Soibelman made further improvements by introducing meta-information, including date and location, to achieve more accurate multi-feature construction material image retrieval (Brilakis & Soibelman, 2008). The win of the AlexNet model in the 2012 ImageNet image classification competition marked an essential breakthrough in deep learning and convolutional neural networks (CNN). Feature extraction methods in image retrieval tasks have thus gradually shifted from traditional manual feature-based to deep learning-based approaches. Ha et al. used a rendered BIM model to construct the dataset and a pre-trained VGG network for image feature extraction (Ha et al., 2018). Firstly, the similarity between the rendered image and the real captured indoor image is evaluated, after which the positional and directional information contained in the rendered image is combined to jointly complete the estimation of indoor positioning. Wang et al. argued that using existing content-based methods to retrieve construction site scenes is limited to utilising simple visual features of the entire image, which makes it difficult to distinguish similar details in related scenes (Wang et al., 2023). Therefore, they used the object detection model to acquire the critical sub-regions in the construction image and employed CNN to extract these fine features, thus effectively improving the performance of the retrieval system.

It is worth noting that existing applications of image retrieval tasks in civil engineering fall into two main categories. One category is the safety management and identification of workers, materials and machinery in construction site scenarios, and the other category is the identification of components in BIM models, thus enabling the alignment of images with BIM data. However, there are few applications in the field of bridge damage detection. The closest to our study is the retrieval of bridge deck images by Wogen et al. (Wogen et al., 2024). This study used a Siamese convolutional neural network to extract image features. Based on a composite similarity measure that combines image

features with deck geolocation information, effective retrieval of bridge inspection images was achieved. However, this study's shortcomings are firstly that fine-tuning the Siamese convolutional neural network relies on manually labelled 14,959 positive and negative samples. Such labelling is time-consuming and challenging. Secondly, this study focuses on single-deck component retrieval and does not explore different damage types in real-world bridges. Finally, GPS data may be incomplete or inaccurate in real-world applications, especially in the defective areas underneath the bridge structure. This situation can directly affect the retrieval validity of the model. In the scenario of bridge damage assessment, relying solely on the simple visual features of the images extracted by CNN is often insufficient to retrieve the corresponding damaged images accurately and comprehensively. Therefore, additional information should be provided to assist the retrieval as much as possible.

In this study, a novel semantic-enriched image retrieval framework (SEIR-Net) is developed to address these challenges. This framework is different from existing methods that rely solely on CNN encoders to extract image low-level visual features. Our core idea is to make this encoder capable of spanning from low-level visual features, such as edges, textures, etc. to high-level semantic features by fine-tuning the multi-modal image captioning model (CNN-LSTM). This feature extraction method can capture rich semantic information in the image, such as the type of bridge damage, the location of occurrence, etc., which enables the retrieval process to match the image accurately based on visual and semantic cues. The main contributions of this research are as follows:

1. A semantic-enriched image retrieval framework is proposed. By fine-tuning the CNN-LSTM multi-modal model, the cross-entropy loss function is used to adjust the positional distribution of the encoder vectors in the feature space to provide it with the ability to capture the high-level semantic information implicit in the images.

2. A Faiss-based vector database is adopted for efficient storage and fast retrieval. By storing the extracted high-dimensional image features in the Faiss vector database, the L2 Euclidean distance is used as the retrieval algorithm to calculate the similarity of the feature vectors in the space and ultimately achieve accurate retrieval.

3. Experiments on a real-world constructed bridge damage dataset verify its effectiveness as an assessment tool. This method can assist engineers in efficiently retrieving historical damage images, which provides practical support for bridge damage assessment.

2. METHODOLOGY

2.1 Problem Statement

As can be seen from the research related to content-based image retrieval, using convolutional neural network to extract visual features from image data for similarity matching has become one of the most effective methods for image retrieval. This success can be attributed to CNN's automatic feature learning capability and the fact that the extracted high-dimensional feature vectors can be efficiently metricised by the distance function for similar image matching.

However, the core challenge in image retrieval is the semantic gap between its high-level semantic information and low-level visual features (Wan et al., 2014). The CNN network used in existing retrieval methods is mainly responsible for extracting low-level visual features, but has limited ability to capture high-level semantic information contained in the image, thus affecting the accuracy and relevance of the retrieval effect. This problem is even more critical in the scenario of bridge damage assessment. When engineers conduct the bridge damage assessment, they need to retrieve historical images with similar damage information to assist in diagnosis. Therefore, high-level semantic information in images (such as damage type and location) plays a vital role in retrieval. Effectively bridging the semantic gap between low-level visual features and high-level semantic information has become a critical challenge in applying the image retrieval technique for bridge damage assessment.

2.2 Overall Framework

This study proposes a semantic-enriched image retrieval framework (shown in Figure. 1) to

solve the above conflict, which consists of four parts: dataset construction, semantic enrichment, feature indexing, and image retrieval. Firstly, the dataset is constructed by preparing the damaged images with corresponding descriptions based on the bridge inspection report, which will be described in detail in the Section 3.1. Following that, in the semantic enrichment stage, the encoder is trained through supervised learning to capture high-level semantic information and low-level visual features from the images. Next, in the feature indexing stage, the feature vectors of the test image set are extracted using the fine-tuned encoder and saved to the FAISS vector database for efficient storage. Finally, the query image is input in the image retrieval stage, and the fine-tuned encoder extracts the query features. Similarity matching is then performed with the stored feature vectors in the FAISS vector database to retrieve the most similar historical images.

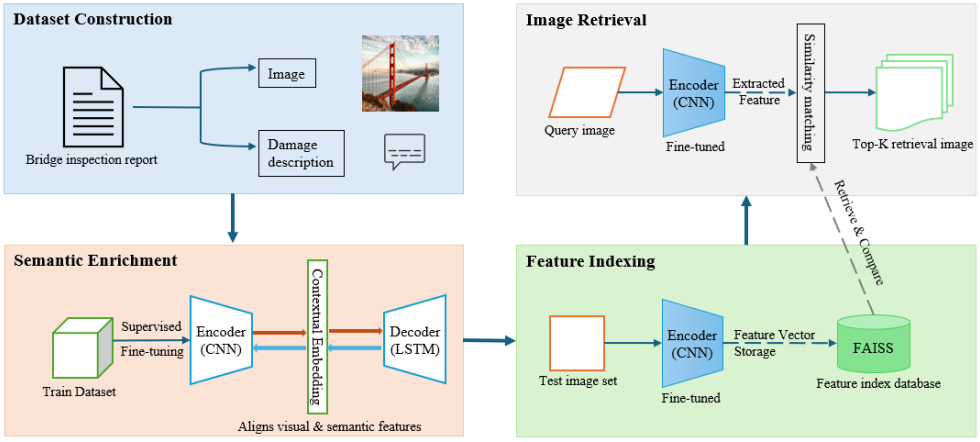


Figure 1. Semantic-enriched image retrieval framework

2.3 Semantic Enrichment

The semantic enrichment component is the core part of this SEIR-Net framework, as shown in Figure 2. Its purpose is to fine-tune the multi-modal image captioning model (CNN-LSTM) so that the fine-tuned encoder not only extracts the image's low-level visual features but also considers the high-level semantic information simultaneously. In this framework, the CNN acts as an encoder, which is mainly responsible for extracting low-level visual features (edges, textures, and colours) within an image, and the LSTM acts as a decoder, which can convert the visual features extracted by the CNN into semantic text sequences through sequence modelling. The model performs supervised learning on image-text pairs through a cross-entropy loss function. The position of the encoder in the feature space is continuously optimised through the backpropagation of parameters. As training proceeds, the CNN encoder gradually learns to extract visual features that are highly correlated with the semantic sequence of text. This process allows the encoder to capture both low-level visual and high-level semantic information from images in the end.

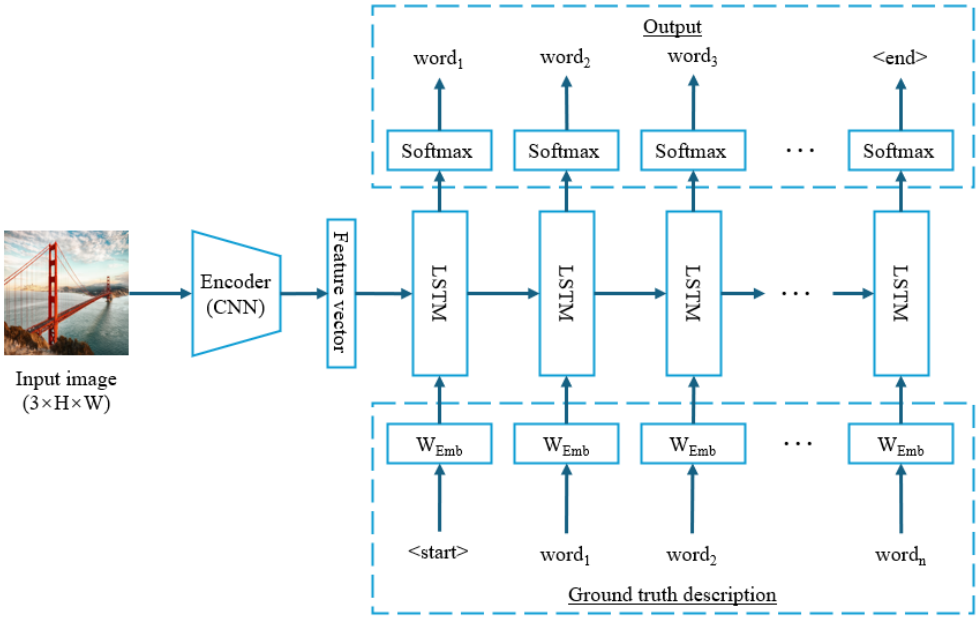


Figure 2. Semantic enrichment using CNN-LSTM

2.4 Feature Indexing and Image Retrieval

In the feature indexing stage, the features of the image set are first extracted using the encoder part that has been fine-tuned by supervised learning in the semantic enrichment stage. After high-dimensional vectorisation, these extracted features are stored in the Facebook AI Similarity Search (FAISS) vector database. FAISS is an open-source library for efficient similarity search and clustering of dense vectors (Douze et al., 2024). In the framework of SEIR-Net, the vector storage uses the IndexFlatL2 index structure. This index structure implements vector retrieval based on L2 Euclidean distance, which can process large-scale data quickly and supports similarity search of vectors in high-dimensional spaces. At the same time, the index will be preliminarily optimised to obtain the distribution characteristics of the vectors before storage to better accelerate the subsequent similarity matching.

In the retrieval stage, the engineer will provide an image of the bridge damage to be queried, and the system will perform feature extraction using a fine-tuned encoder (CNN part) to obtain high-dimensional vectors. Then, the distance metric function (L2 Euclidean distance) is used to match the similarity with the feature vectors in the FAISS database, as shown in Equation 1. The Euclidean distance (L2 paradigm) is a metric formula between two vectors in Euclidean space. It is calculated by summing the squares of the differences between two vectors in each dimension and then taking the square root. The smaller the distance value, the more similar the two vectors are in the feature space. The system will sort by similarity distance and retrieve the K historical images from the FAISS database that are most similar to the query image, i.e., Top-K retrieval results. These retrieval results will be returned to the engineer as a matched result set of the query image to provide reference support for bridge damage assessment.

$$d(I_1, I_2) = \sqrt{\sum_{i=1}^{k} (I_{1i} - I_{2i})^2} \quad (1)$$

2.5 Loss Function

The loss function is mainly used for supervised training of multi-modal image captioning models in the semantic enrichment stage of the SEIR-Net framework. We use the cross-entropy loss function shown in Equation 2, and the parameters of the encoder and decoder will be continuously updated by backpropagation during the training process. This update allows the encoder to gradually acquire high-level semantic information based on its ability to extract low-level visual features. The function continuously compares the difference between the text sequence output by the LSTM decoder and the reference description to measure the difference between the predicted probability distribution and the actual target distribution. The minimisation of this loss can contribute to a better match between the image features extracted by the model and the semantic description of the text sequence. Eventually, by alignment, the model encoder can represent both low-level visual features and high-level semantic information within the feature space.

$$L_{ce}(\theta) = -\sum_{t=1}^{T} log\left(p_\theta(w_t^* | w_{1:t-1}^*)\right) \quad (2)$$

In the encoder-decoder framework of the image captioning model, the attention mechanism is sometimes prone to focus on local regions and ignore the global context (Zohourianshahzadi and Kalita, 2022). This phenomenon can lead to insufficiently accurately generated descriptions, and the encoder cannot extract enough semantic information in the feature space. Therefore, this study additionally uses the bi-stochastic attention regularisation to address this potential problem. This regularisation term enables the model to generate more comprehensive descriptions using penalty weights to ensure the attention is more evenly distributed.

3. EXPERIMENTS AND RESULTS

3.1 Experimental Design

Dataset: The core of the SEIR-Net framework is fine-tuning the CNN-LSTM multi-modal model in the semantic enrichment stage. This training process requires the preparation of bridge damage image-description datasets for supervised learning. Considering that no damage image-description dataset is publicly available, we chose the bridge inspection report provided by Centregreat Rail (CGR) to construct the dataset. This bridge inspection report details the investigation of bridge damage by engineers on the site. From it, we extracted images related to various damage types, each containing a textual description of the damage type and the location of the damage. In addition, we cleaned the data on the above images with corresponding descriptions to ensure that this constructed dataset was of high quality. The cleaned dataset has 429 sets of samples, divided into training set, validation set, and testing set in the ratio of 70%, 10%, and 20%.

Evaluation metrics: To comprehensively evaluate the performance of the SEIR-Net framework in the semantic enrichment and image retrieval stages, we used different evaluation metrics for measurement. To evaluate the effectiveness of the multi-modal model in generating text, we used the Bilingual Evaluation Understudy (BLEU) series of metrics, as shown in Equation 3, which has a score between 0 and 1. BLEU is a widely used evaluation criterion in natural language processing (Papineni et al., 2002). It is classified into BLEU-1 to BLEU-4 depending on the n-gram, where n-gram refers to the number of consecutive words. Thus, BLEU measures the model's ability to generate text at different levels, and a higher BLEU score also ensures that the encoder has captured relevant high-level semantic information in the feature space. For the evaluation of the effectiveness of the retrieval

stage, we use the mean average precision (mAP) and recall, which are commonly used in image retrieval tasks, as evaluation metrics. The precision (P) is the proportion of relevant images in the returned retrieved images during a single retrieval, as shown in Equation 4. The recall (Q) is the ratio of the number of relevant images in the returned retrieved images to the number of actual relevant images in a single retrieval process, as shown in Equation 5. mAP evaluates the overall performance of the retrieval system by calculating the precision rate and averaging it using all the data with different recall. A higher mAP indicates that the model can retrieve more accurate images across various queries. recall@K, on the other hand, measures the model's ability to find relevant images in the first K retrieval results.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{3}$$

$$\text{P}=\frac{\text{Number of retrieved images related to the query image}}{\text{Number of retrieval images}} \tag{4}$$

$$\text{R}=\frac{\text{Number of retrieved images related to the query image}}{\text{Number of relevant images}} \tag{5}$$

Implementation detail: The implementation of the SEIR-Net framework is based on the environment of Python 3.8 and Pytorch 2.0.1. For hardware, the CPU processor is Intel(R) Core (TM) i7-10700KF CPU@ 3.80 GHz, and the GPU processor is NVIDIA GeForce RTX 3060 Ti. Fine-tuning the multi-modal model is at the core of the whole framework, and we set the number of training epochs to 50 and the batch size to 8 to better balance the training efficiency and the convergence effects. The learning rate is set to 1e-4 for the encoder (CNN), and 4e-4 for the decoder (LSTM), and this differential learning strategy will better balance the optimisation process of the two components. In addition, to effectively improve the generalisation ability of the model, we also set up a BLEU-4-based early stopping mechanism. If the effectiveness of the validation set does not improve in 10 consecutive epochs, the training stop will be triggered. This early stopping strategy can prevent the model from overfitting during training.

3.2 Results and Performance Evaluation

(1) Semantic-Enriched Multi-modal Model Fine-tuning Result

We first trained the multi-modal model and evaluated the fine-tuning results on the test set using the BLEU family of metrics mentioned in 3.1. The evaluation results include comparing the generated text with the reference text from BLEU-1 to BLEU-4, i.e., at the 1-gram level to the 4-gram level, as shown in Table 1.

Table 1. The assessment metrics for multimodal model

BLEU-1	BLEU-2	BLEU-3	BLEU-4
0.822	0.793	0.770	0.749

As shown in Table 1, the BLEU scores decrease as the n-gram in the assessment increases. These scores indicate that the model is most accurate at the word-level assessment dimension and can predict individual words accurately. As the assessment dimension becomes longer, moving to longer dimensions of phrases and sentence structures, the match between the generated text and the reference text decreases. This decline is because longer dimensions are evaluated with more demands on sentence fluency and semantic integrity. BLEU-4 achieves a score of 0.749, indicating that the model can

generate text more consistent with the reference descriptions, meaning it could effectively capture semantic information about the context of the image in the feature space. Therefore, the encoder part of the model can be effectively used in the subsequent feature indexing and image retrieval stages. In addition, we provide some examples of generated descriptions in Figure. 3 to facilitate a more intuitive demonstration of the effectiveness of the multi-modal model.

Ref: section loss at the lower crank angle of stiffener
Gen: section loss at the lower crank angle of stiffener

Ref: hole at the deck plate
Gen: hole at the deck plate

Figure 3. Generation description of multi-modal model (Ref: Reference text, Gen: Generated text)

(2) Evaluation of Retrieval Performance

Before performing image retrieval, the fine-tuned model encoder is first used to extract the image features on the test set. The extracted high-dimensional vectors need to be saved into the FAISS vector database for feature indexing. Then, mAP and Recall, as mentioned in 3.1 are used as the primary evaluation metrics in the image retrieval stage to assess the similarity retrieval performance of the images. Meanwhile, to better demonstrate the advancement of the SEIR-Net framework in retrieval performance, we compare it with the benchmark models of two CNN encoders (VGG-16 and ResNet-50) that are widely used in the current image retrieval field (Fang et al., 2022; Wang et al., 2023). The specific results are shown in Table 2.

Table 2. The comparative experiments on retrieval performance

Method	mAP	Recall@1	Recall@2	Recall@4
VGG-16	0.270	0.256	0.279	0.302
ResNet-50	0.628	0.581	0.663	0.724
SEIR-Net	**0.642**	**0.628**	**0.686**	**0.744**

As can be seen from Table 2, SEIR-Net outperforms VGG-16 and ResNet-50 in both mAP and Recall@K (K=1, 2, 4). This comparison fully demonstrates the advantages of the encoder fine-tuned with the multi-modal model. It can better capture high-level semantic information while extracting the low-level visual feature of the image, thus making the retrieved image more compatible with the query image in terms of visual and semantic content. In contrast, VGG-16 and ResNet-50 rely only on the extracted low-level visual features for image matching. They cannot effectively utilise the semantic information present within the image, resulting in a lower retrieval effect. In addition, to demonstrate the effectiveness of the SEIR-Net framework more intuitively, we provide some retrieval examples in Figures 4 and 5.

Figure 4. Visualisation examples of image retrieval (Green: Relevant, Red: Irrelevant)

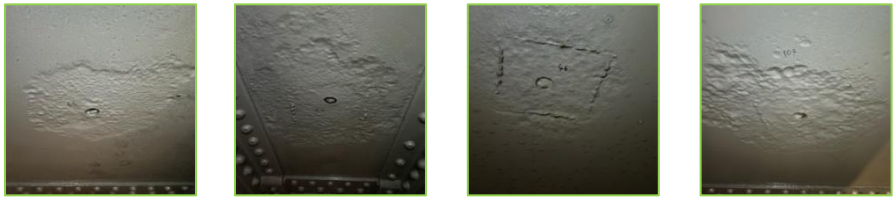
SEIR-Net 

Figure 5. Visualisation examples of image retrieval (Green: Relevant, Red: Irrelevant)

These examples show that the retrieved images returned by the SEIR-Net framework have a higher correlation with the query image. It can effectively retrieve images from the historical damage database that are more consistent with the query image in terms of the semantic content of the damage. Thus, it has a higher utility in bridge damage assessment.

4. CONCLUSION

In this study, we propose a novel semantic-enriched image retrieval framework for bridge damage assessment. By incorporating semantic information into the retrieval model, we significantly enhance the relevance and accuracy of retrieval. Notably, the fine-tuning of the multi-modal image captioning model enables the image encoder to effectively capture the high-level semantic information embedded in the images while extracting visual low-level features. Comparative experiments on a damage dataset constructed based on the real-world bridge inspection report demonstrate that our proposed method outperforms the commonly used VGG-16 and ResNet-50 models on the mAP and Recall@K (K=1, 2, 4) metrics. It also shows that incorporating the semantic content of damage in image retrieval will be more beneficial for the damage assessment.

At the same time, this study has some limitations that need to be further addressed. Firstly, due to the lack of publicly available bridge damage datasets, the dataset we constructed is small and needs to be further expanded. Secondly, we need to try other deep learning models to explore better retrieval methods combined with multi-modal information to make the model more useful.

ACKNOWLEDGMENTS

This work is part of the Knowledge Transfer Partnerships (KTP) project, DIGIBRIDGE: BIM and Digital Twins in support of Smart Bridge Structural Surveying. The project receives funding from Innovate UK with reference number 10003208.

REFERENCES

Ali, R., Chuah, J.H., Talip, M.S.A., Mokhtar, N. and Shoaib, M.A., 2022. Structural crack detection using deep convolutional neural networks. *Automation in Construction*, 133, pp.103989.

ASCE, (2021). U.S. bridge report card. Retrieved from: https://infrastructurereportcard.org/cat-item/bridges-infrastructure.

Brilakis, I. and Soibelman, L., 2005. Content-based search engines for construction image databases. *Automation in Construction*, 14(4), pp.537-550.

Brilakis, I. and Soibelman, L., 2006. Multimodal image retrieval from construction databases and model-based systems. *Journal of Construction Engineering and Management*, 132(7), pp.777-785.

Brilakis, I. and Soibelman, L., 2008. Shape-based retrieval of construction site photographs. *Journal of Computing in Civil Engineering*, 22(1), pp.14-20.

Cha, Y.J., Ali, R., Lewis, J. and Büyüköztürk, O., 2024. Deep learning-based structural health monitoring. *Automation in Construction*, 161, p.105328.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L.

and Jégou, H., 2024. The faiss library. arXiv preprint arXiv:2401.08281.
Dubey, S.R., 2021. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5), pp.2687-2704.
Fang, W., Love, P.E., Luo, H. and Xu, S., 2022. A deep learning fusion approach to retrieve images of People's unsafe behavior from construction sites. *Developments in the Built Environment*, *12*, p.100085.
Ha, I., Kim, H., Park, S. and Kim, H., 2018. Image retrieval using BIM and features from pretrained VGG network for indoor localization. *Building and Environment*, *140*, pp.23-31.
Lee, S., Masoud, M., Balaji, J., Belkasim, S., Sunderraman, R. and Moon, S.J., 2017. A survey of tag-based information retrieval. *International Journal of Multimedia Information Retrieval*, 6, pp.99-113.
Li, N., Li, Q., Liu, Y.S., Lu, W. and Wang, W., 2020. BIMSeek++: Retrieving BIM components using similarity measurement of attributes. *Computers in Industry*, *116*, p.103186.
Li, X., Yang, J. and Ma, J., 2021. Recent developments of content-based image retrieval (CBIR). *Neurocomputing*, 452, pp.675-689.
Ma, J.W., Czerniawski, T. and Leite, F., 2021. An application of metadata-based image retrieval system for facility management. *Advanced Engineering Informatics*, 50, p.101417.
Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
Spencer Jr, B.F., Hoskere, V. and Narazaki, Y., 2019. Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering*, *5*(2), pp.199-222.
Sun, A., Bhowmick, S.S., Nam Nguyen, K.T. and Bai, G., 2011. Tag‐based social image retrieval: An empirical evaluation. *Journal of the American Society for Information Science and Technology*, 62(12), pp.2364-2381.
Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y. and Li, J., 2014, November. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 157-166).
Wang, Y., Xiao, B., Bouferguene, A., Al-Hussein, M. and Li, H., 2023. Content-based image retrieval for construction site images: leveraging deep learning–based object detection. *Journal of Computing in Civil Engineering*, 37(6), p.04023035.
Wogen, B.E., Choi, J., Zhang, X., Liu, X., Iturburu, L. and Dyke, S.J., 2024. Automated Bridge Inspection Image Retrieval Based on Deep Similarity Learning and GPS. *Journal of Structural Engineering*, 150(3), p.04023238.
Zohourianshahzadi, Z. and Kalita, J.K., 2022. Neural attention for image captioning: review of outstanding methods. *Artificial Intelligence Review*, *55*(5), pp.3833-3862.