Kalpa Publications in Computing

Volume 22, 2025, Pages 578–588

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


# Bridge Inspection Using A Multi-Modal Vision Language Model

Zhengxing Chen[1], Yang Zou[1], Vicente A. González[2], Jason Ingham[1] and Liam M. Wotherspoon[1]

[1] University of Auckland, Auckland, New Zealand

[2] University of Alberta, Edmonton, Canada

Zche570@aucklanduni.ac.nz

**Abstract**

Using an Unmanned Aerial Vehicle (UAV) in bridge inspections can reduce human involvement in complex and hazardous inspection environments and automate the inspection process. Current practices require human operators to define task objectives, oversee safe flight operations, and evaluate bridge conditions. There is a growing demand for improving the seamless collaboration between UAVs and human inspectors to complete the inspection task efficiently and more safely, especially in post-disaster scenarios where critical bridges and other infrastructure facilities need to be inspected within hours or days. A significant gap exists in enabling UAVs to intelligently perceive and understand the bridge inspection scene according to human instructions. An intuitive human-UAV collaboration system using a multi-modal Vision Language Model (VLM) was proposed to partially fill this gap. This system leverages a few-shot Contrastive Language–Image Pretraining (CLIP)-based model to enable UAVs to visually and semantically understand the bridge inspection environment based on human commands. By incorporating text prompt learning with a cache adapter, the proposed model enhances the ability of CLIP to interpret both textual and visual inputs in the context of bridge inspection. The model was trained and evaluated in a bridge inspection image dataset and achieved an accuracy of 83.33%, outperforming other few-shot image classification methods, demonstrating its effectiveness in the bridge inspection domain. This approach is expected to improve collaboration between AI-empowered UAVs, inspectors, and bridge environments, thereby enhancing the overall efficiency of bridge inspections.

## 1 Introduction

Bridges constitute essential elements of the transportation infrastructure, and the maintenance of bridges is imperative for safeguarding public safety. Assessing the structural health of bridges is

I. Hkust, J. Cheng and Y. Yantao (eds.), ICCBEI 2025 (Kalpa Publications in Computing, vol. 22), pp. 578–588

particularly problematic because the inspection processes are mainly conducted manually (Zhang et al., 2022). The traditional inspection of bridges is particularly challenging because many bridge structures are either too costly or too dangerous for human inspectors to access directly (Dorafshan & Maguire, 2018). Unmanned aerial vehicles (UAVs) have been introduced in recent years for bridge inspection, aiming to enhance efficiency, reduce labour demand, and minimize human exposure to hazardous environments (Bolourian & Hammad, 2020). Current practices still require significant manual effort because UAVs cannot perform inspections independently without human input and skilled pilots. There is a growing demand for a collaborative environment where human operators and UAV robotic systems work side by side during bridge inspections. When faced with emergencies such as post-disaster bridges, emergency inspection and repair are essential ways to restore bridge transportation capacity quickly. There is a growing need of the collaboration between inspectors and UAVs to enhance the efficiency, safety, and effectiveness of bridge inspections in complex and challenging conditions.

Human-robot collaboration (HRC) is defined as the integration of human adaptability and decision-making capabilities with the physical precision, strength, and repeatability of robotic assistants to achieve common goals efficiently within shared workspaces (Ajoudani et al., 2018; Michalos et al., 2014). Current methods for controlling and interacting with UAVs in the physical world have been dominated by complex teleoperation controllers (Seo et al., 2018), hand gestures (Naseer et al., 2022), and rigid command protocols (Contreras et al., 2020), where the robots execute predefined tasks based on specialised programming languages. Among these methods, natural language-based HRC stands out for its intuitive and accessible nature. Natural language-based HRC allows non-experts in robot programming to intuitively communicate with robot assistants, making the interaction efficient and accessible (Park et al., 2024). Traditional natural language processing (NLP) models are usually trained on a limited dataset that shows limited adaptability in diverse working environments and with different inspectors. The application of HRC to bridge inspections comes with unique challenges and specifications. The inherent uncertainties and complexities in structural deterioration and failure contribute to varying probabilities and consequences of failure across different bridges. The advent of Large Language Models (LLMs), such as ChatGPT (OpenAI, 2020), offers the potential to develop an interactive and communicative approach to HRC. LLMs trained on extensive and diverse datasets bring a deep understanding of natural language and human intentions, especially in the context of HRC tasks.

While LLMs excel in understanding human interaction commands, they lack the visual and semantic comprehension needed to interpret the environment around robots. Previous research has primarily focused on using deep learning-based object detection algorithms, such as You Only Look Once (YOLO) (Redmon et al., 2016), Faster R-CNN (Ren et al., 2015), and Single Shot MultiBox Detector (SSD) (Liu et al., 2016), to enhance the scene understanding of robots through image processing. Applying these algorithms in bridge inspections presents challenges, including the high costs of data annotation, significant computational demands, and the scarcity of large-scale training datasets (Liang et al., 2024). There is a need for computer vision algorithms that can operate effectively with minimal training samples in bridge inspection tasks. With the further development of LLMs and the increasing demand for integrating multiple modalities such as language and vision, Vision-Language Models (VLMs) have emerged (Zhou et al., 2022b). VLMs facilitate open-vocabulary visual recognition and make complex inferences about interactions between objects and agents within images (Kirillov et al., 2023). VLMs provide “eyes” to explore and find arbitrary objects described by humans, understand the environments, and provide context for decision-making. VLMs such as Contrastive Language–Image Pretraining (CLIP) (Pan et al., 2022), DALL-E (Ramesh et al., 2021), and Vision-and-Language BERT (ViLBERT) (Lu et al., 2019) have demonstrated strong zero-shot image classification performance on public datasets as a result of extensive training on large-scale image-text pairs. VLMs allow robots to navigate to and identify objects they have not been explicitly trained on, enhancing their adaptability and performance in unstructured and unforeseen

scenarios (Gadre et al., 2022). Although the robotic research community has integrated VLMs for robotic control, these systems are primarily used for industrial applications (Shibata et al., 2024) or household services (Brohan et al., 2023). There is a need to develop a low-cost, highly efficient human-UAV collaboration method that enables UAVs to perform semantic scene understanding based on human instructions for UAV-assisted bridge inspections.

In early 2021, OpenAI released a large-scale multi-modal model for aligning images and texts called CLIP (Pan et al., 2022), which was trained on over 400 million image-text pairs. CLIP exhibits powerful zero-shot inference capabilities that can recognise and understand unseen images without being explicitly trained on specific tasks or datasets. The core concept of the CLIP model is to embed both text and images into a shared semantic space, where related text descriptions and image representations are positioned closely together. In contrast unrelated image-text pairs are placed farther apart. The CLIP model comprises two main components: an image encoder and a text encoder. The image encoder converts images into feature vectors, while the text encoder, typically a Transformer model, converts text into feature vectors. These two encoders operate within the same vector space, enabling cross-modal information interaction and fusion. CLIP has demonstrated excellent zero-shot performance on public datasets and performs effectively in many everyday tasks. Due to regulatory constraints, bridge inspection datasets are highly domain-specific and often unavailable on the Internet. The zero-shot capabilities of CLIP may be limited because it has not been trained in the context of bridge inspection. This highlights the need to adapt CLIP for this domain through transfer learning. Completely retraining CLIP for bridge inspection poses several challenges: 1) Due to the large number of parameters in CLIP, fine-tuning the entire network requires substantial computing resources. 2) UAV-assisted bridge inspection is an emerging technology that has not yet been widely adopted, resulting in a limited dataset that cannot cover all inspection object categories. Therefore the data distribution in bridge inspection often diverges from the pre-training data of CLIP. 3) Labelling bridge inspection data is time-consuming and labour-intensive, requiring domain expertise, further complicating the process. To address these challenges, it is essential to leverage few-shot learning techniques. By fine-tuning the CLIP model with a limited number of labelled samples, domain-specific knowledge can be transferred, enabling CLIP to perform well in bridge inspection tasks despite the scarcity of training data.

The authors propose a few-shot CLIP-based model to enable UAVs to visually and semantically understand the bridge inspection environment based on human commands. First a few-shot CLIP model integrated with text prompt learning and a cache adapter was proposed to enable UAVs to visually and semantically interpret the bridge inspection environment based on human commands. Second a bridge inspection dataset, collected by UAVs, was developed to test the proposed model. This dataset includes four bridge components: pier, girder, railing, and pavement; four structural details: bearing, cover plate termination, gusset plate connection, and out-of-plane stiffener; and two types of damage: cracks and corrosion. To evaluate the performance of the proposed method in bridge inspection tasks, the accuracy of the proposed model was compared with baseline CLIP and other fine-tuned CLIP-based models.

# 2 Method

The core objective of this research was to develop an algorithm that enables UAVs to navigate toward specific target objects within an unknown bridge environment. Figure 1 shows the framework of the proposed human-UAV collaboration method. Successful navigation requires UAVs to possess semantic scene understanding and natural language processing capabilities. The capabilities allow UAVs to identify objects in the environment based on task goals defined by human inspectors and translate these goals into a semantic context in textual form. To achieve this goal the authors leverage

the CLIP model, which is a VLM that facilitates cross-modal understanding by learning to compare text and images. CLIP captures the semantic relationships between text and images through contrastive learning without supervision labels.

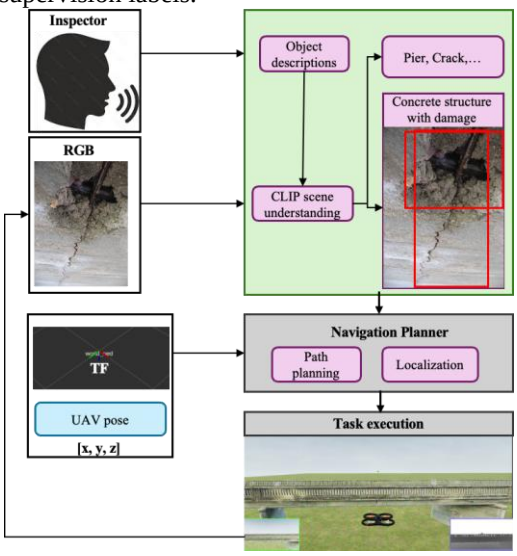


**Figure 1:** The framework of human-UAV collaboration in bridge inspections

The task is formulated as follows: A UAV is randomly placed within an unseen environment $E$, with a sequence of predefined navigation goals $G = \{g_1, g_2, \ldots, g_n\}$, decoded from natural language inputs (e.g., "pier," "girder," or "crack"). The objective of the UAV is to navigate to any specified goal object. At time $t$, the UAV receives an observation in the form of an RGB image $I_t$ from its onboard camera and must select an action from the action space $A$. Navigation is considered successful if the UAV stops within a safe distance of the object, and the object is visible without further movement. In this study, the authors focused on scene object recognition using CLIP. Given a set of navigation goals $G$ and the collected image $I_t$ at time $t$, the objective is to find if the image $I_t$ contains the target object $g_m \in G$. If the target is present, a semantic scene understanding prompt $P_{g_m,t}$ is generated to describe the goal object. The task can be formulated as the following equation:

$$P_{g_m,t} = CLIP(I_t, G) \quad (1)$$

## 2.1 Few-shot CLIP for UAV-assisted Bridge Inspections

Few-shot image classification offers a solution for fine-tuning CLIP with fewer training datasets, reduced computational resources, and shorter training time. Authors have explored few-shot adaptation techniques, leading to two primary strategies: prompt-based and adapter-based approaches (Liu et al., 2024). Prompt-based fine-tuning methods, such as CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a), transform the fixed textual prompts of CLIP's text encoder into learnable vectors. These vectors are then fine-tuned using a small dataset to improve the performance of CLIP on domain-specific tasks. Adapter-based methods, such as CLIP-Adapter (Gao et al., 2021) and Tip-Adapter (Zhang et al., 2022), introduce lightweight adapter modules into the pre-trained model. These adapter parameters are fine-tuned with a small dataset, allowing the foundation model to address domain-specific tasks with minimal resource requirements effectively.

CoOp and Tip-Adapter are representative works in prompt-based and adapter-based fine-tuning methods, respectively. CoOp is designed to optimise the prompt's context to enhance the image recognition performance of CLIP. The goal is to iteratively refine the prompt based on the

performance of CLIP on domain-specific tasks, ultimately finding the optimal prompt for classification. In the conventional zero-shot method, CLIP uses prompt templates such as "A photo of a {label}," where class labels are inserted into predefined text prompts, and image-text similarity is calculated for classification. CoOp improves this process by introducing a set of learnable vectors $\{[V]_1, [V]_2, \dots, [V]_M\}$ that model contextual text alongside class labels within the prompt. Both text and image features are computed during forward propagation, and cross-entropy loss is calculated concerning the labels. The learnable vectors are updated during backpropagation to minimise the loss, while the weights of the pre-trained CLIP model remain fixed, with only the learnable tokens being fine-tuned. CoOp focuses solely on prompt optimisation of the text. CoOp does not incorporate domain knowledge transfer on the image side, which may limit its effectiveness in domain-specific tasks.

The Tip-Adapter method enhances few-shot classification by utilising a pre-trained CLIP model and constructing a key-value cache model. In a typical few-shot setting, there are $N$ classes, each with $K$ samples (K-shot), resulting in $NK$ images in the training set. Visual features are extracted from these $NK$ images using the visual encoder of CLIP, which serves as the keys, while the corresponding one-hot encoded labels act as the values in the cache model. This cache model is integrated with the pre-trained CLIP classifier without requiring additional parameter tuning. During inference, the affinity between the test image features and the cache keys is calculated, and the corresponding values are aggregated to form the prediction of the adapter. Tip-Adapter effectively combines the zero-shot prediction capabilities of CLIP with domain-specific visual knowledge from few-shot tasks. Tip-Adapter still relies on manually designed prompts for image classification, limiting its ability to fully harness the extensive knowledge embedded in the text encoder of CLIP. This dependency on manual prompts reduces its potential to leverage the complete semantic understanding of the model across modalities.

This study incorporated text prompt learning by CoOP with the cache model of Tip-Adapter, as shown in Figure 2. Given the new bridge inspection dataset (K samples and N classes), image dataset $I_K$), text labels LN, the process of text prompt learning with Tip-Adapter is as follows. A context vector $\mathrm{L_{initial}}$ was randomly initialised by drawing from a zero-mean Gaussian distribution. $L_N$ was converted into N-dimensional one-hot vectors $OneHot(L_N)$. The learnable vector $L_{learn}$ is trained through a tunable *TextEncoder*, with the input as $OneHot(L_N)$ and $\mathrm{L_{initial}}$ to get the optimised prompt $L_{learn}$ in the training phrase:

$$L_{learn} = TextEncoder(L_N, \mathrm{L_{initial}}) \tag{2}$$

According to CoOp, the prompt can be designed to put class labels at the end of $L_{learn}$: $[V]_1[V]_2 \dots [V]_M[Class\}$ or in the middle of $L_{learn}$: $[V]_1 \dots V_{\frac{M}{2}}[Class][V]_{\frac{M}{2}+1} \dots [V]_M$.

Then the parameters of the text prompt were frozen, and the pre-trained CLIP *VisualEncoder* was used to extract the L2-normalized C-dimensional visual features of each image $I_K$ in the training set:

$$F_{train}^T \in \mathbb{R}^{NK \times C} = VisualEncoder(I_K) \tag{3}$$

During inference the visual features $f_{test}$ from the test image $I_{test}$ are extracted using *VisualEncoder*:

$$f_{test} \in \mathbb{R}^{1 \times C} = VisualEncoder(I_{test}) \tag{4}$$

Then the affinities A between $F_{train}^T$ and $f_{test}$ are computed as follows:

$$A = \exp\left(-\beta(1 - f_{test}F_{train}^T)\right) \tag{5}$$

Where $\beta$ is used to control the sharpness of the similarity distribution, which ensures better classification performance, the ache model $AL_{learn}$ is built as the dot product of the affinities $A$ and

the optimised prompt $L_{learn}$. The final logits are computed as the summation of cache model $AL_{learn}$ and prior knowledge $f_{test}W_c^T$ from the pre-trained CLIP:

$$logits = \alpha \cdot \mathrm{A}L_{train} + f_{test}W_c^T \tag{6}$$

where $\alpha$ is a weighting factor, $W_c^T$ is the weight *TextEncoder* generates.

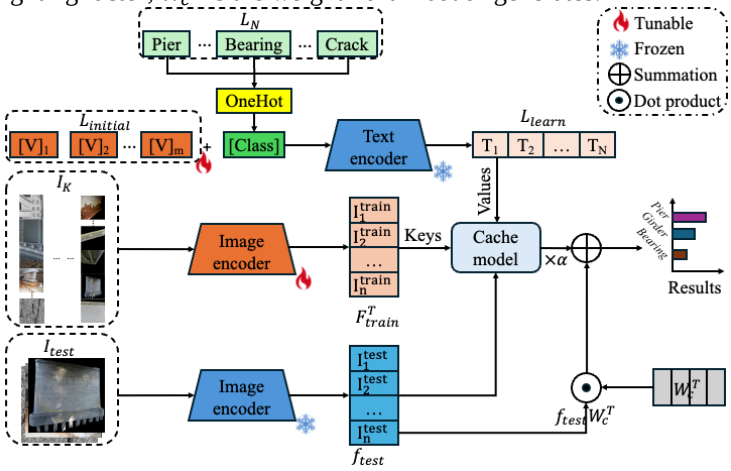


**Figure 2:** Few-shot CLIP incorporated text prompt learning with a cache model

# 3 Experiment and Results

To evaluate the performance of the proposed few-shot CLIP algorithm for object recognition in bridge inspections, experiments were conducted using both public datasets and a newly developed domain-specific dataset. The evaluation involved 1-shot, 2-shot, 4-shot, 8-shot, and 16-shot image classification tasks, where a limited number of labelled examples were available for each class. The accuracy of the proposed method was compared with that of the baseline CLIP and other fine-tuned CLIP-based models.

## 3.1 Dataset

To train and evaluate the performance of few-shot image classification algorithms for bridge inspection tasks, experiments were conducted on three public datasets, including Common Objects in Context Dataset for Structural Detail Detection of Bridges (COCO-Bridge) (Bianchi et al., 2021), Labelled Cracks in the Wild (LCW) (Bianchi & Hebdon, 2022), and Corrosion Condition State Semantic Segmentation Dataset (CCSSSD) (Bianchi & Hebdon, 2022), along with a newly developed dataset, the Bridge Member Dataset (BMD). COCO-Bridge comprises 774 images and over 2,500 object instances collected by UAV, targeting the detection of four key structural bridge details: bearing, cover plate termination, gusset plate connection, and out-of-plane stiffener. The dataset provides a broad range of structural features essential for bridge inspection. LCW contains 3,817 finely annotated images of segmented cracks gathered from structural inspection reports provided by the Virginia Department of Transportation (VDOT). This dataset is focused on identifying and segmenting cracks, which are critical for assessing structural integrity.

Similarly CCSSSD includes 440 finely annotated images of segmented corrosion sourced from VDOT Bridge Inspection Reports. Corrosion is another primary concern in bridge inspections, and this dataset enables detailed analysis and classification of such defects. In addition to these public

datasets, a new small-scale dataset, BMD, was developed specifically for this study. BMD comprises 150 images and includes four segmented bridge components: pier, girder, railing, and pavement. Including this dataset allows for further testing on distinct bridge elements often inspected during UAV-assisted bridge inspection tasks. Together, these datasets provide a comprehensive testing environment for assessing the effectiveness of few-shot image classification models in real-world bridge inspection scenarios.

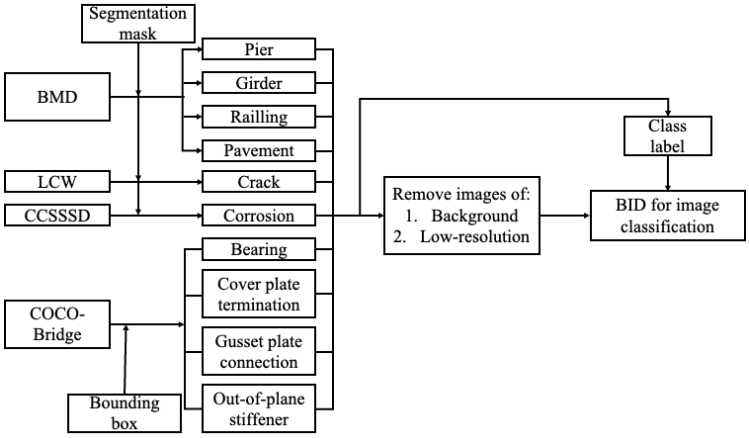


**Figure 3:** Preprocessing of datasets

The authors implemented a customisation process to adapt the datasets for classification, as shown in Figure 3. First the segmentation areas for each object in the BMD, LCW, and CCSSSD datasets were extracted using the annotated segmentation masks and class labels. In contrast the remaining areas were filled with black. Regions of Interest (ROIs) were cropped for the COCO-Bridge dataset based on the annotated bounding boxes and class labels. Next task-unrelated background areas in the cropped images were removed, and any segmented or cropped areas with low image quality were discarded to ensure the overall quality of the dataset. Finally the pre-processed datasets were split into 80% for training, 10% for testing, and 10% for evaluation. Sample images and their corresponding descriptions are shown in Figure 4.

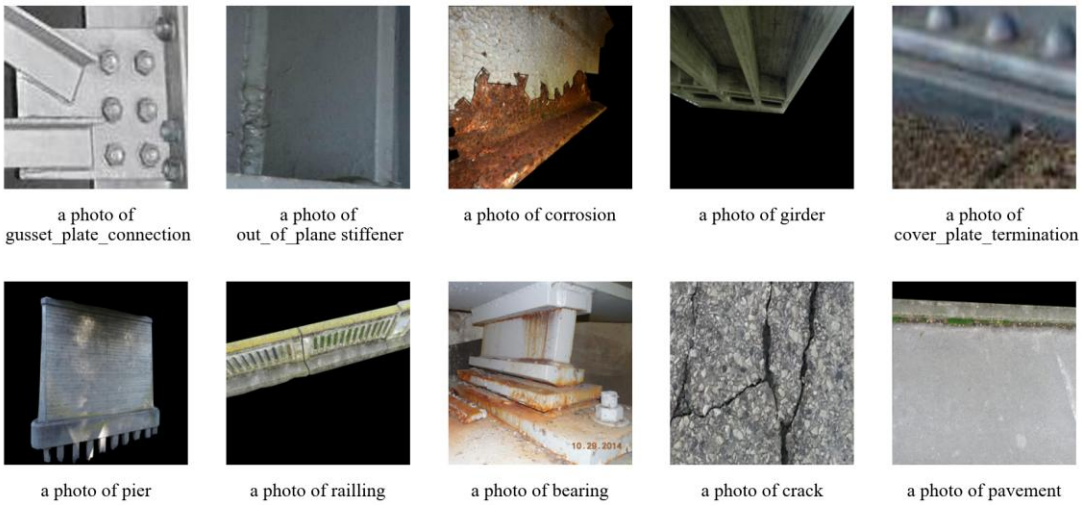


**Figure 4:** Sample images and corresponding descriptions

## 3.2 Evaluation of the Proposed Few-shot CLIP model

The authors conducted experiments using the proposed few-shot CLIP model incorporating text prompt learning with a cache model on the bridge inspection dataset. To evaluate few-shot learning, performances were compared across 1, 2, 4, 8, and 16-shot training sets, with testing conducted on the whole test set. ViT-B/16 (Dosovitskiy et al., 2021) was used as the visual encoder, and the training epoch for prompt tuning via CoOp was set to 50. Once the text prompt was trained, the optimised text weights were frozen and extracted as pre-trained text features, which were then used as input for the text encoder in Tip-Adapter. During Tip-Adapter training, the parameters were configured with 50 training epochs, a batch size 256, and a learning rate of 0.001.

Performance comparison was conducted between the model proposed by authors and Zero-shot CLIP (Pan et al., 2022), CoOp (Zhou et al., 2022b), COCoOp (Zhou et al., 2022a), MaPLe (Khattak et al., 2023), and Tip-Adapter ( Zhang et al., 2022). All experiments were performed on 1, 2, 4, 8, and 16-shot training sets and evaluated on the complete test sets. For a fair comparison, the visual encoder backbone for all models was standardised to ViT-B/16. The training epoch was all set as 50. Following the framework of CoOp, four variants were tested: class token placed at the end or middle of the prompt, unified context (UC) versus class-specific context (CSC). The number of context tokens for CoOp was set to 16.

Figure 5 illustrates the performance of various models across different shot settings, ranging from 1 to 16 shots per class. At 16 shots, CoOp variants exhibit strong performance, with accuracy ranging from 79.43% to 80.33%, while Tip-Adapter, COCOOP, and MaPLe achieve 76.58%, 27.90%, and 30.50%, respectively. The proposed model surpasses all these methods, achieving the highest accuracy of 83.33%, highlighting its effectiveness in improving accuracy for bridge inspection tasks. The proposed model consistently outperforms the other methods in lower-shot settings (1, 2, 4, and 8 shots). For instance, in the 8-shot setting, it achieves 75.89%, outperforming CoOp (68.40% to 71.67%) and Tip-Adapter (71.64%). This trend continues in the 1, 2, and 4-shot settings, where the proposed model remains competitive or superior, underscoring its robustness and adaptability in few-shot bridge inspection image classification tasks.

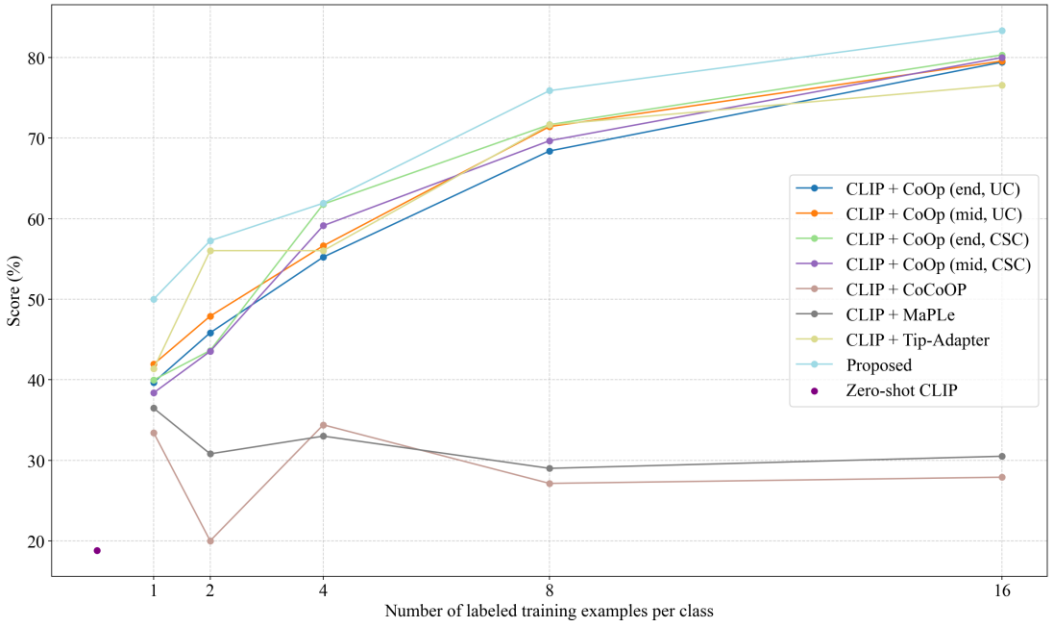


**Figure 5:** Few-shot classification accuracy of different models using 1, 2, 4, 8, 16 shots

# 4 Discussion and Conclusions

In this study the authors proposed a few-shot CLIP model incorporating text prompt learning with a cache model to enhance HRC in UAV-assisted bridge inspection and scene understanding tasks. The proposed model achieved the highest accuracy of 83.33%, outperforming other few-shot image classification methods, demonstrating its effectiveness in leveraging few-shot learning for bridge inspection tasks. By enabling UAVs to interpret the bridge environment based on human instructions, the model shows promise in improving the accuracy and efficiency of visual bridge inspections.

The contributions of this paper are twofold. First the proposed approach combines a text prompt learning method introduced by CoOP with a cache model developed by Tip-Adapter within a unified CLIP-based framework to capture both language and visual knowledge for bridge inspection tasks. Second the newly developed BMD dataset is introduced along with three modified datasets: COCO Bridge, LCW, and CCSSSD, to train and evaluate the proposed algorithm. This work demonstrates the potential of vision language models to enhance HRC in UAV-assisted bridge inspections.

Although the proposed model with pre-trained prompts outperforms manually designed prompts, the text prompts trained by CoOp are a string of vectors that are relatively difficult to interpret. Recent research has begun exploring external knowledge, such as knowledge graphs (e.g., CuPL by Pratt et al. (2022)), to help models better handle unseen samples, enhance semantic comprehension and robustness, improve interpretability, and specialise in specific domains. Incorporating domain knowledge from bridge inspection into both prompt engineering and visual adapter design presents a promising direction for improving performance in future bridge inspection scene understanding tasks. Future work will also focus on refining and integrating the model into the Robotic Operating System (ROS) for deployment on real UAVs to perform bridge inspection tasks. This study underscores the potential for AI-empowered UAVs to revolutionise bridge inspection processes, making UAVs more efficient and reliable in inspection tasks.

# Acknowledgements

This research is supported financially by the University of Auckland and the China Scholarship Council (CSC) (Grant number: 202207000017).

# References

Ajoudani, A., Zanchettin, A. M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., & Khatib, O. (2018). Progress and prospects of the human–robot collaboration. *Autonomous Robots*, *42*(5), 957–975. https://doi.org/10.1007/s10514-017-9677-2

Bianchi, E., Abbott, A. L., Tokekar, P., & Hebdon, M. (2021). COCO-Bridge: Structural Detail Data Set for Bridge Inspections. *Journal of Computing in Civil Engineering*, *35*(3), 04021003. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000949

Bianchi, E., & Hebdon, M. (2022). Development of Extendable Open-Source Structural Inspection Datasets. *Journal of Computing in Civil Engineering*, *36*(6), 04022039. https://doi.org/10.1061/(ASCE)CP.1943-5487.0001045

Bolourian, N., & Hammad, A. (2020). LiDAR-equipped UAV path planning considering potential locations of defects for bridge inspection. *Automation in Construction*, *117*, 103250. https://doi.org/10.1016/j.autcon.2020.103250

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman,

K., Herzog, A., Hsu, J., Ichter, B., … Zitkovich, B. (2023). *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control* (arXiv:2307.15818). arXiv. https://doi.org/10.48550/arXiv.2307.15818

Contreras, R., Ayala, A., & Cruz, F. (2020). Unmanned Aerial Vehicle Control through Domain-Based Automatic Speech Recognition. *Computers*, *9*(3), Article 3. https://doi.org/10.3390/computers9030075

Dorafshan, S., & Maguire, M. (2018). Bridge inspection: Human performance, unmanned aerial systems and automation. *Journal of Civil Structural Health Monitoring*, *8*(3), 443–476. https://doi.org/10.1007/s13349-018-0285-4

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Gadre, S. Y., Wortsman, M., Ilharco, G., Schmidt, L., & Song, S. (2022). *CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object Navigation* (arXiv:2203.10421). arXiv. https://doi.org/10.48550/arXiv.2203.10421

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., & Qiao, Y. (2021). *CLIP-Adapter: Better Vision-Language Models with Feature Adapters* (arXiv:2110.04544). arXiv. https://doi.org/10.48550/arXiv.2110.04544

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2023). *MaPLe: Multi-modal Prompt Learning* (arXiv:2210.03117). arXiv. https://doi.org/10.48550/arXiv.2210.03117

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything* (arXiv:2304.02643). arXiv. https://doi.org/10.48550/arXiv.2304.02643

Liu, F., Zhang, T., Dai, W., Zhang, C., Cai, W., Zhou, X., & Chen, D. (2024). Few-shot adaptation of multi-modal foundation models: A survey. *Artificial Intelligence Review*, *57*(10), 268. https://doi.org/10.1007/s10462-024-10915-y

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 21–37). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 13–23). Curran Associates Inc.

Michalos, G., Makris, S., Spiliotopoulos, J., Misios, I., Tsarouchi, P., & Chryssolouris, G. (2014). ROBO-PARTNER: Seamless Human-Robot Cooperation for Intelligent, Flexible and Safe Operations in the Assembly Factories of the Future. *Procedia CIRP*, *23*, 71–76. https://doi.org/10.1016/j.procir.2014.10.079

Naseer, F., Ullah, G., Siddiqui, M. A., Jawad Khan, M., Hong, K.-S., & Naseer, N. (2022). Deep Learning-Based Unmanned Aerial Vehicle Control with Hand Gesture and Computer Vision. *2022 13th Asian Control Conference (ASCC)*, 1–6. https://doi.org/10.23919/ASCC56756.2022.9828347

OpenAI. (2020). *ChatGPT* [Computer software]. https://chat.openai.com

Pan, X., Ye, T., Han, D., Song, S., & Huang, G. (2022). *Contrastive Language-Image Pre-Training with Knowledge Graphs* (arXiv:2210.08901). arXiv. https://doi.org/10.48550/arXiv.2210.08901

Park, S., Wang, X., Menassa, C. C., Kamat, V. R., & Chai, J. Y. (2024). Natural language instructions for intuitive human interaction with robotic assistants in field construction work. *Automation in Construction*, *161*, 105345. https://doi.org/10.1016/j.autcon.2024.105345

Pratt, S., Covert, I., Liu, R., & Farhadi, A. (2022, September 7). *What does a platypus look like? Generating customized prompts for zero-shot image classification*. arXiv.Org. https://arxiv.org/abs/2209.03320v3

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). *Zero-Shot Text-to-Image Generation* (arXiv:2102.12092). arXiv. https://doi.org/10.48550/arXiv.2102.12092

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, 91–99.

Seo, J., Duque, L., & Wacker, J. (2018). Drone-enabled bridge inspection methodology and application. *Automation in Construction*, *94*, 112–126. https://doi.org/10.1016/j.autcon.2018.06.006

Shibata, K., Deguchi, H., & Taguchi, S. (2024). *CLIP feature-based randomized control using images and text for multiple tasks and robots* (arXiv:2401.10085). arXiv. https://doi.org/10.48550/arXiv.2401.10085

Zhang, C., Zou, Y., Wang, F., del Rey Castillo, E., Dimyadi, J., & Chen, L. (2022). Towards fully automated unmanned aerial vehicle-enabled bridge inspection: Where are we at? *Construction and Building Materials*, *347*, 128543. https://doi.org/10.1016/j.conbuildmat.2022.128543

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., & Li, H. (2022). Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (pp. 493–510). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19833-5_29

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a). *Conditional Prompt Learning for Vision-Language Models* (arXiv:2203.05557). arXiv. https://doi.org/10.48550/arXiv.2203.05557

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022b). Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, *130*(9), 2337–2348. https://doi.org/10.1007/s11263-022-01653-1