Vision Freqformer for Vibration Monitoring using existing surveillance cameras

Fukuta, Tomonori

Kawaguchi, Hiroshi

Signal, Image and Video Processing (2025) 19:1375
https://doi.org/10.1007/s11760-025-04953-4

ORIGINAL PAPER

Vision Freqformer for Vibration Monitoring using existing surveillance cameras

Tomonori Fukuta[1,2] · **Hiroshi Kawaguchi**[1]

Received: 15 July 2025 / Revised: 20 September 2025 / Accepted: 1 November 2025

Abstract
Social infrastructure, such as road bridges and tunnels, is used in the long term; therefore, their structural integrity must be maintained during this period. Currently, the soundness of social infrastructure is confirmed through visual and sound inspections. However, these inspections are sensitive and difficult to perform and inexperienced inspectors may overlook them. Although camera-based inspection can examine wide areas simultaneously, they only examine the surface structures and not bridge components. Finite element method has been used to investigate the structural components by applying known vibrations and observing the frequency responses. Road bridges vibrate owing to traffic. The internal structure of road bridges can be investigated by measuring these vibrations. In this study, we propose a novel machine learning method that does not use a Fourier transform. Our method directly estimates vibration information from structural images by improving a transformer. We call this Vision Freqformer. Our method uses surveillance cameras to monitor road bridges. We assess the vibration estimation accuracy and robustness of the bit rate. Consequently, our method achieved an estimation accuracy exceeding 71.6 % in tests using vibration data from the damper equations and Z24 dataset simulations.

Keywords Frequency · Neural Network · Vibration Monitoring · Vision Transformer · Surveillance camera

1 Introduction

Maintaining social infrastructure such as road bridges and tunnels is extremely important for their long-term safe use. In recent years, infrared [1] and laser light [2] have been used for image analysis and visual inspection. Methods for analyzing structural vibrations have been studied as alternatives to hammer tests. Modal analysis [3] is a technique used to investigate the state of a structure and the dynamic properties of systems in the frequency domain. The structural health is monitored by determining the natural frequencies and modal damping ratios. Usually, in an operational modal analysis, it is necessary to install an acceleration sensor [4, 5] connected wired or wirelessly to measure the vibrations. Although these sensors are highly reliable for vibration measurements, they require considerable weight, costs, and installation time. Batteries and communication relays are required even when wireless units are used, resulting in considerable weight. Sensors such as accelerometers provide pointwise information. They have a low spatial resolution for measuring the entire structure and may be insufficient for the analysis of large structures if insufficient/large numbers of sensors are utilized. Contactless sensors have also been investigated for this purpose. A laser Doppler vibrometer [6] does not require the installation of a large number of sensors, as is the case with accelerometers. Laser Doppler vibrometers have a high spatial resolution; however, they are extremely expensive and time-consuming for large structures because of the time-sequential measurement. Digital video cameras are inexpensive, have an excellent spatial resolution, and can simultaneously measure large areas. The method of employing digital video cameras combines image correlation or optical flow [7, 8] to measure structural vibrations and perform modal analysis. However, many of these methods require speckle patterns [9] or additional markers [10, 11] to strengthen the correlations and simplify feature-point tracking, which is a serious limitation. Instead of using additional markers, some methods employ video magnification techniques [12, 13]. These methods use

✉ Tomonori Fukuta
214p604p@stu.kobe-u.ac.jp

1 Graduate School of Science, Technology and Innovation, Kobe University, 1-1, Rokkodai-cho, Nada-ku, 6578501 Kobe, Hyogo, Japan

2 Advanced R&D Center, Mitsubishi Electric Corporation, 8-1-1, TsukaguchiHonmachi, 6618661 Amagasaki, Hyogo, Japan

Published online: 17 November 2025

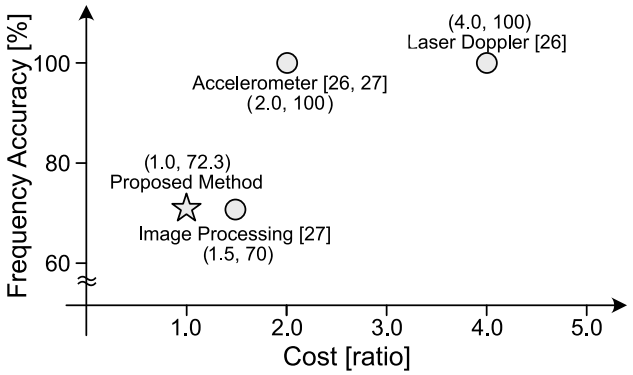

Fig. 1 Comparison of selected methods vs proposed method

video magnification technology to convert microvibrations into large vibrations and subsequently use image processing to perform modal analyses [14]. Methods using a multiscale pyramid [15] to perform modal analysis and machine learning [16, 17] improve the accuracy of vibration detection. A method using machine learning for image processing has also been proposed. The method using machine learning replaces the object selection of measurement areas [18, 19], frequency analysis using Fourier analysis [20], and vibration extraction using image processing [21].

We propose an improved implementation of the transformer [22] that achieves state-of-the-art language processing and vision transformer [23], which serves as a reference for image input to the transformer. It is necessary to use multiple consecutive camera images when measuring and estimating the vibration of a structure. This time-series analysis involves analyzing the relationship between images in successive time periods. Therefore, we focused on the attention structure proposed by Transformer [22]. We also extended the transformer [22] to handle images [24, 25]. Methods that use image processing are more complicated, and the number of operations varies significantly depending on the image. However, the proposed method can be easily configured using machine learning libraries, and they can achieve scalability in cloud services. This is because our method does not have conditional branching, unlike image processing techniques.

A comparison of selected methods vs proposed method introduced above is shown in Fig. 1. The accuracy of each method was based on the results of previous research, and the cost was estimated based on our own market research. First, regarding frequency accuracy, laser Doppler vibrometers are equivalent to or slightly more accurate than accelerometers. This is because minute vibrations caused by light vehicles cannot be measured using accelerometers because they are buried in the noise [26]. Methods using image processing have been reported to be less accurate than accelerometers [27]. The proposed method can infer frequencies with more than 71.6 % accuracy, which is comparable to that of the method using image processing. Second, the laser Doppler vibrometer is the most expensive. Installation costs are extremely high for the methods using accelerometers. Some image processing methods require high-resolution images of structures to measure vibrations, making it impossible to use existing surveillance cameras. However, the proposed method uses images obtained from existing surveillance cameras. Therefore, the proposed method is slightly cheaper than the image processing method.

2 Vision Freqformer

The model structure of the proposed Vision Freqformer is illustrated in Fig. 2. Instead of dividing a single image into multiple patches as a vision transformer [23], the Vision Freqformer takes as input a patch from which a specific area of a temporally continuous image is extracted. Focusing on the oscillation periodicity, t and $t+i$ are similar, which represents a half-period, and $t+2i$ is one period. Inferring the position of i, that is, time-series analysis, allows the frequency to be inferred.

2.1 Input Layer of the Vision Freqformer

The bottom-right side of Fig. 2 shows how the input image is handled in the input layer of the Vision Freqformer. $\mathbf{x}^t$ where (H, W) is the resolution of the original image, and C is the number of channels in the input image at time t in the video, which includes the object for the frequency inference. The small patch is defined as $\tilde{\mathbf{x}}_p^t$ where (P, P) is the resolution of the image patch at position (x_c, y_c) in $\mathbf{x}^t$. The small patch $\tilde{\mathbf{x}}_p^t$ is equivalent to the image patch in the vision transformer [23]. $\tilde{\mathbf{x}}_p^{t+1}$ denotes a patch image at time $t+1$ in the video. Similarly, $\tilde{\mathbf{x}}_p^{t+N_p}$ denotes the patch image at time $t+N_p$ in the video. These are combined as $\tilde{\mathbf{x}}_p \in \mathbb{R}^{N_p \times (P^2 \cdot C)}$. Then, if the oscillation is periodic and at period T, it returns to the same state at time $t+T$ with respect to time t. We define difference image $\mathbf{x}_p^{t+i} \in \mathbb{R}^{N_p \times (P^2 \cdot C)}$, where $0 \leq i \leq N_p$, as follows:

$$\mathbf{x}_p^{t+i} = \tilde{\mathbf{x}}_p^{t+i} - \tilde{\mathbf{x}}_p^t \tag{1}$$

"-" in Fig. 2 represents the operation of Eq. (1). The difference image, $\mathbf{x}_p^{t+i}$, becomes nearly zero in phases, for example, T and $T/2$, where T is the period of the vibration. Therefore, the vibration frequency can be determined from time T using the frame interval of the video. The difference image improves the ability to infer frequency. Next, $\mathbf{x}_p^{t+i}$ is flattened and embedded. The proposed Vision Freqformer uses the same process as the vision transformer [23] until the position is embedded, where the length of the embedded vector is

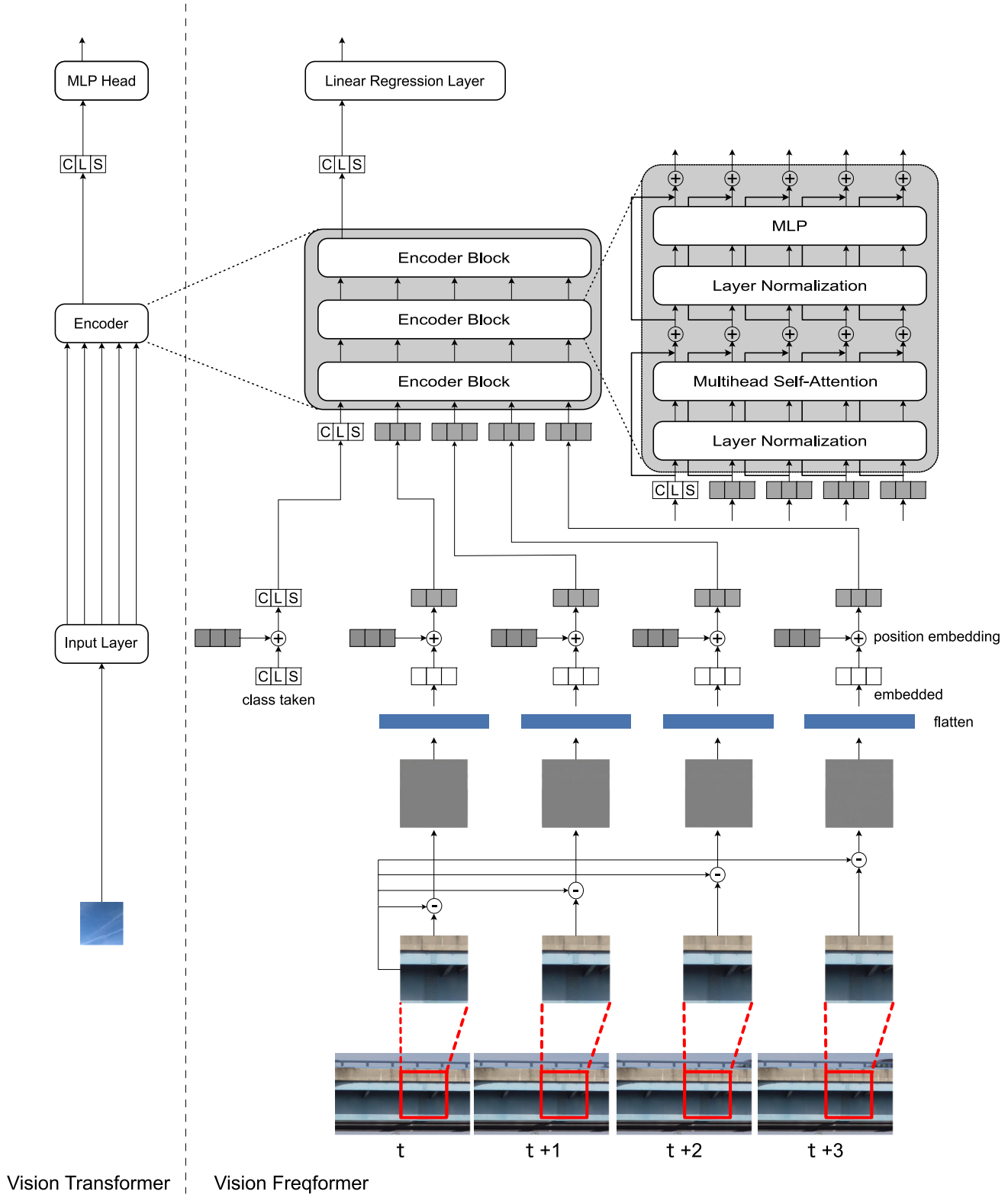

Fig. 2 Left figure shows the vision transformer model [23] and the right figure shows the proposed method

D, and the weight of the embedding layer is $\mathbb{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$. $\mathbf{x}_p^{t+i} \in \mathbb{R}^{(P^2 \cdot C)}$ is the $t+i$th patch vector, and ";" denotes the coupling in the patch direction. $\mathbf{x}_p^{t+i}\mathbb{E} \in \mathbb{R}^D$ is the $t+i$th embedded vector whose length is D. Here, $\mathbf{x_{class}} \in \mathbb{R}^D$ is the class token indicated by the CLS in Fig. 2 and the position embedding is $\mathbf{E_{pos}} \in \mathbb{R}^{(N_p+1) \times D}$, and the output $\mathbf{z}_0$ of the input layer is as follows:

$$\mathbf{z}_0 = [\mathbf{x_{class}}; \mathbf{x}_p^{t+1}\mathbb{E}; \mathbf{x}_p^{t+2}\mathbb{E}; \cdots \mathbf{x}_p^{t+N_p}\mathbb{E}] + \mathbf{E_{pos}}$$
$$\mathbf{z}_0 \in \mathbb{R}^{(N_p+1) \times D} \quad (2)$$

"+" in Fig. 2 represents position embedding. As described above, the proposed method determines the frequency by observing the periodicity of the difference images, rather than by tracking the object.

2.2 Encoder Layer

Similar to the vision transformer [23], the encoder layer consists of alternating layers of multiheaded self-attention and multilayer perceptron (MLP) blocks, as shown in Fig. 2. Layer normalization is applied before each block, and residual functions are applied after each block.

2.3 Linear Regression Layer

The output of the encoder layer of the Vision Freqformer is $\mathbf{z}_L^0 \in \mathbb{R}^D$, the weight of the linear layer is $\mathbf{W}^y \in \mathbb{R}^D$ and the output of the Vision Freqformer is as follows:

$$\mathbf{y} = \mathbf{LN}(\mathbf{z}_L^0)\mathbf{W}^y \quad (3)$$

The output $\mathbf{y}$ is the inference of the vibration frequency.

3 Simulation

The Vision Freqformer utilizes surveillance cameras to monitor the infrastructure and bridges are used as infrastructure subjects. We assessed the robustness of the vibration estimation accuracy and bit rate. To confirm the effectiveness of the proposed method, we verified the vibration inference performance using images of the simulated vibrations. For this simulation, we used the International Organization for Standardization (ISO) resolution chart [28] and pictures of the Hirakata Bridge in Japan shown in Fig. 3.

3.1 Dataset construction

The dataset was developed using vibration data derived from the general equation of an underdamped damper system as follows:

Fig. 3 The upper one is Hirakata Bridge, and the lower one is Tennozan Bridge

$$x = Ae^{-\gamma t}\cos(\omega t + \alpha)$$
$$\forall A, \alpha \in \mathbb{R} \quad (4)$$

The frequency of this equation is $f = \omega/2\pi$ and the amplitude is A. The damping ratio was controlled using γ. The values of f, A, and γ were randomly generated between 0.5 Hz and 15.0 Hz, 0.5 pix and 3 pix, and 0.5 and 2.0, respectively, resulting in multiple samples of damped oscillations, each with a duration of 3 s. Based on the vibration data, we used numerical simulations to create an ISO chart and bridge image to simulate bridge vibration. The ground truth is f in Eq. (4). For training, validation, and testing, 7395156, 1811813, and 36976 images, respectively were generated. In this simulation, camera frame rate was 30 frames per second (fps).

3.2 Parameters of the Vision Freqformer

The parameters of the proposed Vision Freqformer are listed in Table 1. Batch learning and dropout were introduced to prevent overfitting. High-definition videos were used because current surveillance cameras are high-definition and can provide a wide view of the structure. $N_p = 36$ is equal to 1.2 s under 30 fps. Therefore, the proposed method could predict values ranging from 0.833 to 15.0 Hz. The camera was installed 70 m away from the Hirakata Bridge. The bridge was located approximately 14 m above the camera position. The spatial resolution of the bridge image was approximately 0.02 m/pixel. Therefore, a 24 pixels patch corresponded to an area of 0.48 m.

3.3 Training Parameters

The following parameters used for the learning phase are described in Table 2.

Table 1 Network parameters

Parameter	Symbol	value
image height	H	1080 pix
image width	W	1920 pix
camera frame rate	-	30 fps
number of color channels	C	3
patch size	P	24 pix
number of patches	N_p	36
dimensions of embedded vector	D	576
number of encoders	L	3
number of SA heads	k	16
dimensions of output vector from SA	D_h	1152

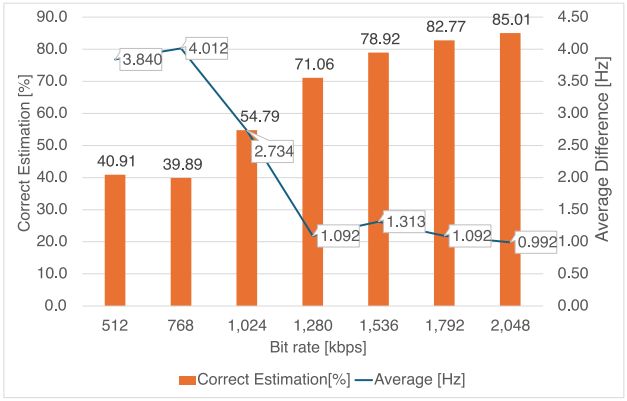

Fig. 4 Robustness of bit rate

3.4 Results

During the inference phase, we used 36976 images that were not used for training. The absolute error (AE) was used to evaluate the inference accuracy. Consequently the mean, maximum, and minimum AEs were 0.161 Hz, 9.36 Hz, and 8.77×10^{-6} Hz, respectively. The frequency resolution of the fast Fourier transform (FFT) was approximately 1.67 Hz, where the camera frame rate was 30 fps and the number of patches was 36. The absolute difference between the inference value and the ground truth was less than 1.67 Hz, which was considered correct. Therefore, the accuracy was 98.5 %.

4 Experiment

4.1 Bit rate Robustness

Here, the images of the Tennozan Bridge in Fig. 3 were used because they were not used for training. The length of the Tennozan Bridge is approximately the same as that of the Hirakata Bridge. The spatial resolution of the Tennozan Bridge image was the same as that of the Hirakata Bridge. We investigated the performance at the expected bit rate for surveillance cameras. The encoding format was set to H.264, and the bit rates were set to 512, 768, 1024, 1280, 1536, 1792, and 2048 kilobits per second (kbps). To generate damped oscillations, the vibrations were simulated using the underdamped damper equation shown in Eq. (4). f, A, and γ were randomly generated to create 3 s data. The results are presented in Fig. 4. Given a camera frame rate of 30 fps, corresponding to intervals of approximately 33 ms and a frequency resolution of 1.67 Hz in the short-time FFT over 36 frames, which constitutes the analysis unit of the Vision Freqformer. The accuracies at 1,280 kbps and 2,048 kbps were 71.06 % and 85.01 %, respectively. These bit rates are higher than those used for the long-term storage of surveillance cameras because they are a practical rate for live streaming. If it is necessary to reduce the amount of data, a sufficiently small data size can be achieved by streaming only a selected portion of the bridge image.

4.2 Z24 dataset validation

In the second validation, we utilized the dataset accumulated at the Z24 bridge in the canton of Bern and spanned the A1 highway between Bern and Zürich [30–33]. The Z24 dataset consists of the long-term continuous monitoring test during the year and the progressive damage test over a month. We used one-week data from "Week 01 day 3 14 h, channel 10" to "Week 02 day 2 14 h, channel 10" in the Z24 dataset. The Z24 dataset was sampled at a frequency of 100 Hz. With

Table 2 Training parameters

parameter	value
loss function	Mean Squared Error
batch size	128
epochs	50
optimization method	Adam: A Method for Stochastic Optimization [29]
learning rate	$1e^{-6}$
γ(adam parameter)	0.7
dropout rate	0.01

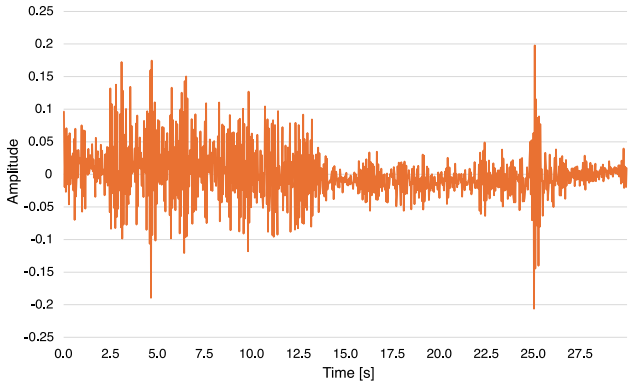

Fig. 5 Resampled Z24 Data "Week 01 day 3."

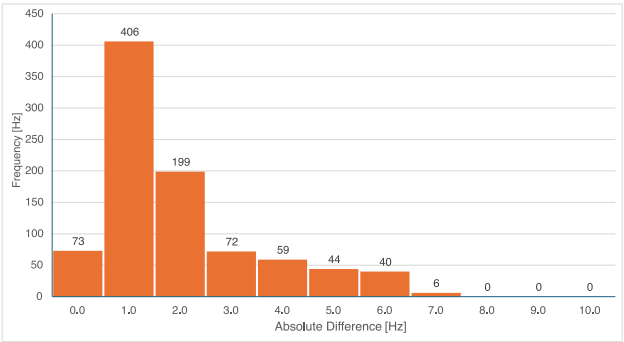

Fig. 6 Distribution of absolute differences between the ground truth and inference on "Week 01 day 3."

the Vision Freqformer, the camera rate had a sampling frequency of 30 fps. Therefore, with the Vision Freqformer, it is possible to estimate up to approximately 15 Hz based on the sampling theorem. The natural frequency of the bridge is between 0.1 and 11.0 [Hz] [34, 35]. Therefore, a low-pass filter with a cutoff frequency of 15 Hz was applied to the Z24 dataset to remove the high-frequency components. Subsequently, normalization was performed using Eq. (5), where x is the original data point, μ is the mean value of the data, and $max(x)$ is the maximum value of the data. Therefore, $z \leqq 1$.

$$z = \frac{x - \mu}{max(x) - \mu} \tag{5}$$

z is resampled to match the camera frame rate using linear interpolation. The normalized and resampled data are shown in Fig. 5. (See Fig. 6)

The resampled dataset, as shown in Fig. 5 has a maximum amplitude of 0.2, which was scaled such that the maximum amplitude would be 2.0 for the simulation. Similar to the bit rate robustness, it was applied to the image, creating approximately 30 s of vibration video and allowing for 900 inference cycles. The frequency spectrum was obtained from the resampled dataset using a short-time FFT over a

Table 3 Results for one week

Day	correct inference	Accuracy
Week 01 day 3	641	71.2 %
Week 01 day 4	676	75.7 %
Week 01 day 5	609	62.3 %
Week 01 day 6	740	77.9 %
Week 01 day 7	587	65.2 %
Week 02 day 1	735	86.0 %
Week 02 day 2	563	62.6 %
Average Accuracy	650	71.6 %

1.2 s interval. The Ground truth was the frequency with the strongest power obtained by the short-time FFT. Similar to the bit rate robustness, the accuracy was calculated by considering the predictions correctly if the absolute difference between the predicted and ground truth was within 1.67 Hz, based on the short-time FFT frequency step. The results for one week are summarized in Table. 3. The accuracy of the Z24 data was 71.6 %. A spectrogram for approximately 30 s is shown in Fig. 7. The inference result from the Vision Freqformer was superimposed with "·" markers. The areas where the Vision Freqformer inferred values coincided with the areas of high spectral power were correctly inferred. Real bridges do not vibrate at one frequency but at multiple frequencies. The Vision Freqformer estimates the frequency with the highest power. Therefore the absolute difference between the frequency calculated from the accelerometer and that calculated by the Vision Freqformer was 1.67 Hz or less was considered the correct inference.

$$\mathbf{AE_t} = |\max(F_t) - I_t| \tag{6}$$

where t is time and $\mathbf{AE_t}$ is the absolute difference between frequency F_t calculated from the accelerometer and frequency I_t inferred by the Vision Freqformer. Therefore, the accuracy of the inferences is as follows:

$$C = \frac{\text{count}(AE_t < 1.67)}{N} \tag{7}$$

where count(x) is the frequency under condition x and N is the number of inferences.

5 Future Work

Actual bridges are constantly vibrating owing to traffic, such as cars and trucks, as well as vibrations resulting from natural phenomena, such as wind. Furthermore, these vibrations are not limited to one dimension; they also occur in three dimensions. These vibrations can also be considered com-

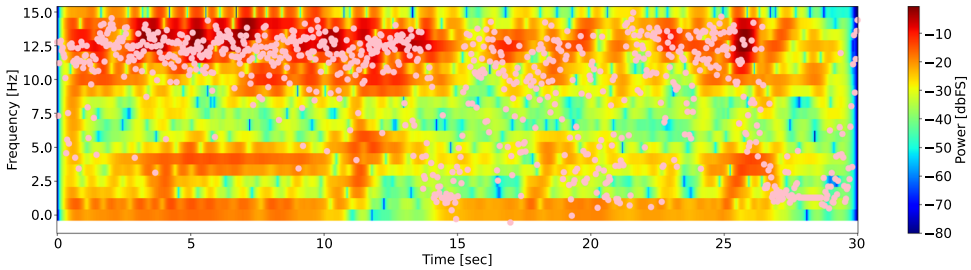

Fig. 7 Spectrum of the Z24 dataset and of the Vision Freqformer on "Week 01 day 3."

binations of multiple damped oscillations. Therefore, it is necessary to verify the use of more realistic vibrations. The spatial resolution of the bridge images strongly correlates with the magnitude of the vibrations. Therefore, it is necessary to verify the spatial resolution. The verification of the proposed method was performed under the condition that the camera was completely fixed. In reality, cameras vibrate because of factors such as traffic and wind. It is much easier to attach an accelerometer to a camera to measure vibrations. Therefore, the separation of these vibrations must be considered. Finally, while benchmarking against other methods such as those using image processing, long short-term memory (LSTM)[36], or a convolutional neural network-LSTM (CNN-LSTM)[37] is required, these methods do not directly determine frequencies. Instead, they derive a frequency analysis by obtaining temporal variations in the amplitude and then performing an FFT. However, a direct comparison is difficult. Therefore, it is necessary to consider methods for comparing similar methods. As discussed in this study, vibrations in bridges caused by traffic cannot be analyzed through long-term frequency analysis owing to damping, making the analysis using a short-time FFT meaningful in this context.

6 Conclusions

We proposed the Vision Freqformer, an improved vision transformer [23] that can infer the periodic motion frequency of structures such as bridges efficiently. The proposed method directly determines the frequencies of the structures. Our proposed method can infer the frequency with an accuracy exceeding 71.6 %. Using surveillance cameras, we confirmed that the method is robust to bit rates and achieves an accuracy of over 71 % at 1280 kbps or higher. As described in the Future Work section, further validation is required; however, we demonstrated the potential of using surveillance cameras to monitor vibrations.

Acknowledgements The authors would like to thank T. Nagano for the useful discussions. This study was supported by Mitsubishi Electric Corporation.

Author Contributions T.F. wrote the manuscript and prepared the figures. All authors reviewed the manuscript.

Funding This study was supported by Mitsubishi Electric Corporation.

Data Availability Z24 datasets is available in the homepage, https://bwk.kuleuven.be/bwm/z24. The other datasets generated and analyzed during the study are not publicly available due to specific reasons, but they may be accessible upon reasonable request from the corresponding author.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Materials availability Not applicable.

Code availability Not applicable.

Competing interests Tomonori Fukuta is an employee of Mitsubishi Electric Corp.

References

1. Bagavathiappan, S., Lahiri, B.B., Saravanan, T., Philip, J., Jayakumar, T.: Infrared thermography for condition monitoring - a review. Infrared Phys. & Technol. **60**, 35–55 (2013)
2. Park, H.S., Lee, H.M., Adeli, H., Lee, I.: A new approach for health monitoring of structures: terrestrial laser scanning. Comput.-Aided Civil Infrastruct. Eng. **22**(1), 19–30 (2007)
3. Amezquita-Sanchez, J.P., Adeli, H.: Signal processing techniques for vibration-based health monitoring of smart structures. Archives Comput. Meth. Eng. **23**(1), 1–15 (2016)
4. Pan, N.: A sensor data fusion algorithm based on suboptimal network powered deep learning. Alex. Eng. J. **61**(9), 7129–7139 (2022)

5. Bai, C.C., Guo, J.F., Zheng, H.X.: Three dimensional vibration-based terrain classification for mobile robots. IEEE ACCESS **7**, 63485–63492 (2019)
6. Teter, A., Gawryluk, J.: Experimental modal analysis of a rotor with active composite blades. Compos. Struct. **153**, 451–467 (2016)
7. Dong, C.-Z., Celik, O., Catbas, F.N., O'Brien, E.J., Taylor, S.: Structural displacement monitoring using deep learning-based full field optical flow methods. Struct. Infrastruct. Eng. **16**(1), 51–71 (2020)
8. Jana, D., Nagarajaiah, S.: Computer vision-based real-time cable tension estimation in dubrovnik cable-stayed bridge using moving handheld video camera. Struct. Control. Health Monit. **28**(5), 2713 (2021)
9. Torbol, M., Park, K.T.: (2018) Machine learning and digital image processing for non-contact modal parameters identification of structures. In: Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2018, vol. 10598, pp. 139–146. SPIE
10. Xu, Y., Brownjohn, J., Kong, D.: A non-contact vision-based system for multipoint displacement monitoring in a cable-stayed footbridge. Struct. Control. Health Monit. **25**(5), 2155 (2018)
11. Sysyn, M., Przybylowicz, M., Nabochenko, O., Kou, L.: Identification of sleeper support conditions using mechanical model supported data-driven approach. Sensors **21**(11), 3609 (2021)
12. Wadhwa, N., Rubinstein, M., Durand, F., Freeman, W.T.: Phase-based video motion processing. ACM Transactions on Graphics **32**(4), 1–10 (2013)
13. Oh, T.-H., Jaroensri, R., Kim, C., Elgharib, M., Durand, F., Freeman, W.T., Matusik, W.: (2018) Learning-Based Video Motion Magnification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018 vol. 11208, pp. 663–679. Springer, Cham
14. Chen, J.G., Wadhwa, N., Cha, Y.-J., Durand, F., Freeman, W.T., Buyukozturk, O.: Modal identification of simple structures with high-speed video using motion magnification. J. Sound Vib. **345**, 58–71 (2015)
15. Yang, Y., Dorn, C., Mancini, T., Talken, Z., Kenyon, G., Farrar, C., Mascareñas, D.: Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification. Mech. Syst. Signal Process. **85**, 567–590 (2017)
16. Peng, C., Zeng, C., Wang, Y.: Phase-based noncontact vibration measurement of high-speed magnetically suspended rotor. IEEE Trans. Instrum. Meas. **69**(7), 4807–4817 (2020)
17. Bao, Y., Li, H.: Machine learning paradigm for structural health monitoring. Struct. Health Monit. **20**(4), 1353–1372 (2021)
18. Pan, X., Yang, TY., Xiao, YF., Yao, HC., Adeli, H.: Vision-based real-time structural vibration measurement through deep-learning-based detection and tracking methods. ENGINEERING STRUCTURES **281** (2023)
19. Cai, Z., Peng, C., Yang, B., Liu, X.: An Intelligent Area Localization Framework for Rotating Machine Vision Vibration Measurement. In: 2022 IEEE 20th International Conference on Industrial Informatics (INDIN), pp. 347–352 (2022)
20. Yang, R., Singh, S.K., Tavakkoli, M., Amiri, N., Yang, Y., Karami, M.A., Rai, R.: CNN-LSTM deep learning architecture for computer vision-based modal frequency detection. Mech. Syst. Signal Process. **144**, 106885 (2020)
21. Yang, R., Singh, S.K., Tavakkoli, M., Amiri, N., Karami, M.A., Rai, R.: Continuous video stream pixel sensor: a CNN-LSTM based deep learning approach for mode shape prediction. Struct. Control. Health Monit. **29**(3), 2892 (2022)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv (2017)
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021)
24. Wang, J., Xia, S., Zou, C., Wu, G., He, Z.: FreqFormer: frequency-enhanced face super-resolution via dual-synergy learning. IEEE Signal Process. Lett. **32**, 2169–2173 (2025)
25. Dai, T., Wang, J., Guo, H., Li, J., Wang, J., Zhu, Z.: FreqFormer: Frequency-aware Transformer for Lightweight Image Super-resolution. In: Thirty-Third International Joint Conference on Artificial Intelligence, vol. 2, pp. 731–739 (2024)
26. Rossi, G., Marsili, R., Gusella, V., Gioffrè, M.: Comparison between accelerometer and laser vibrometer to measure traffic excited vibrations on bridges. Shock. Vib. **9**(1–2), 11–18 (2002)
27. Kalybek, M., Bocian, M., Nikitas, N.: Performance of optical structural vibration monitoring systems in experimental modal analysis. Sensors (Basel, Switzerland) **21**(4), 1239 (2021)
28. 14:00-17:00: ISO 12233:2023. https://www.iso.org/standard/79169.html (2023)
29. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
30. Maeck, J., De roeck, G.: Description of z24 benchmark. Mechanical Systems and Signal Processing **17**(1), 127–131 (2003)
31. Reynders, E., De Roeck, G.: Vibration-Based Damage Identification: The Z24 Bridge Benchmark. In: Encyclopedia of Earthquake Engineering, pp. 3871–3879. Springer (2015)
32. Reynders, E., Roeck, G.D.: Continuous Vibration Monitoring and Progressive Damage Testing on the Z24 Bridge. In: Encyclopedia of Structural Health Monitoring. John Wiley & Sons, Ltd (2009)
33. Yousefpour, H., Amiri, S.A., Mohammadpoory, Z.: Bridge anomaly detection via structure health monitoring with ResNet-152 and scalogram techniques on vibration data. SIViP **19**(4), 310 (2025)
34. Yajima, Y., Petladwala, M., Kumura, T., Kim, C.-W.: Natural Frequency and Displacement Ratio Based Probabilistic Damage Identification for Bridges Using FE Model Update. STRUCTURAL HEALTH MONITORING 2023 (0) (2023)
35. Zhang, M., Yu, H., Zhang, Z., Xu, F.: Vortex-induced vibration control of bridge decks using energy dissipative devices: a review of recent developments. Advances Wind Eng. **2**(3), 100069 (2025)
36. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
37. Wang, J., Yu, L.-C., Lai, K.R., Zhang, X.: Dimensional sentiment analysis using a regional CNN-LSTM model. In: Erk, K., Smith, N.A. (eds.) Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 225–230. Association for Computational Linguistics, Berlin, Germany (2016)