Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AFSAF-Net: Adaptive Frequency-Spatial Attention Fusion Network with IMU Guidance for Robust Indoor Localization under Low-Light and Texture-Degraded Environments

Nan Yin
Guangzhou University of Software
Hua Zhou
Guangzhou University of Software
Wenjing Yu
Guangzhou University of Software
Luojun Yang
Guangzhou University of Software
Yuxiang Sun
sunyuxiang.jack@gmail.com

Canada Advanced Institute of Science and Technology

Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11188223/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

AFSAF-Net: Adaptive Frequency-Spatial Attention Fusion Network with IMU Guidance for Robust Indoor Localization under Low-Light and Texture-Degraded Environments

Nan Yin[1], Hua Zhou[1], Wenjing Yu[1], Luojun Yang[1], Yuxiang Sun[2*]

[1]School of Cyberspace Security, Guangzhou University of Software, Guangzhou, China.
[2]Canada Advanced Institute of Science and Technology, Ottawa, Canada.

*Corresponding author(s). E-mail(s): sunyuxiang.jack@gmail.com;
Contributing authors: yinnan.lisa@gmail.com;

Abstract

Indoor localization under low-light and texture-degraded environments remains challenging because insufficient visual information leads to unstable feature extraction, while inertial measurements are inevitably affected by accumulated drift. Previous MC4L-IMU-based studies demonstrated the effectiveness of handcrafted texture descriptors, frequency-domain enhancement, deep visual representation learning, and visual-inertial fusion for improving localization performance in dark indoor scenarios. However, these approaches mainly rely on manually designed frequency representations and fixed fusion strategies, limiting their adaptability to dynamically changing sensor reliability. To address these limitations, this paper proposes an Adaptive Frequency-Spatial Attention Fusion Network with IMU Guidance (AFSAF-Net) for robust indoor localization. The proposed framework introduces an adaptive frequency learning branch to capture spectral information beyond handcrafted transformations, an attention-driven frequency-spatial fusion module to dynamically integrate complementary visual representations, and an IMU reliability gate to adjust the contribution of inertial information according to sensor confidence. By jointly modeling frequency characteristics, spatial semantics, and inertial reliability, AFSAF-Net enables adaptive multimodal perception under challenging indoor conditions.

Experiments conducted on the MC4L-IMU indoor localization platform demonstrate that AFSAF-Net achieves a mean localization error of 0.85 cm and the highest coefficient of determination ($\boldsymbol{R^2 = 0.9986}$) among the evaluated methods. Furthermore, the proposed framework maintains more stable localization performance under low-light conditions compared with conventional visual and visual-inertial approaches. These results indicate that adaptive spectral representation learning and reliability-aware sensor fusion can effectively improve localization robustness without relying on fixed feature extraction or predefined fusion rules.

Keywords: Indoor localization, Multimodal sensor fusion, Inertial measurement unit, Frequency-spatial representation learning, Low-light perception

1 Introduction

With the rapid development of intelligent robots, autonomous navigation systems, and smart indoor environments, reliable indoor perception and localization have become increasingly important for applications such as mobile robots, human–machine interaction, and intelligent services. Unlike outdoor positioning systems that can utilize global navigation satellite systems (GNSS), indoor environments suffer from signal blockage, multipath propagation, and complex spatial structures, making accurate localization a challenging task. Therefore, various sensor-based indoor localization approaches have been investigated, including magnetic-field-based localization, inertial navigation, acoustic positioning, and vision-based positioning methods [1–4].

Among existing approaches, vision-based localization methods have attracted considerable attention due to their rich environmental information and low deployment cost. Camera sensors can capture spatial structures and semantic information, providing strong perception capabilities for indoor environments. However, the performance of visual localization systems is highly dependent on image quality. In practical scenarios, illumination variation, low-light conditions, insufficient textures, and appearance changes can significantly degrade visual feature extraction and reduce localization reliability [5, 6]. Therefore, improving the robustness of visual representations under degraded conditions remains a critical challenge for reliable indoor localization.

Traditional visual localization methods mainly rely on handcrafted feature descriptors to extract discriminative image information. Local Binary Pattern (LBP) has been widely adopted for texture representation because of its robustness to local illumination changes [7, 8]. However, handcrafted descriptors depend heavily on predefined feature structures and lack the ability to adapt to complex environmental variations. With the rapid development of deep learning, convolutional neural networks (CNNs) have demonstrated powerful representation learning capabilities and have been widely applied in image recognition and localization tasks [9, 10]. Nevertheless, CNN-based approaches still suffer from performance degradation when visual information is severely affected by illumination changes or texture loss.

Frequency-domain representation learning has recently been explored as an effective strategy for enhancing visual perception under challenging conditions. Fast Fourier

Transform (FFT) converts spatial information into frequency components, allowing models to capture global structural characteristics and complementary information beyond spatial features [11, 12]. In our previous research, frequency-enhanced visual representation based on LBP and FFT was investigated for indoor localization in dark environments. The experimental results demonstrated that frequency-domain information could effectively compensate for insufficient texture information and improve localization robustness [13]. However, the frequency features in these approaches were generated through predefined transformations, which limits their adaptability when different environments exhibit diverse visual degradation patterns.

To further improve robustness, visual-inertial fusion approaches have been extensively studied by integrating camera information with inertial measurement units (IMUs). IMUs provide continuous motion information and are less sensitive to illumination changes, making them an effective complementary sensor for visual perception systems. Existing visual-inertial odometry and fusion frameworks, such as optimization-based methods and Extended Kalman Filter (EKF)-based approaches, have demonstrated the effectiveness of combining visual and inertial measurements under relatively stable visual conditions [14, 15]. However, when low illumination and texture degradation occur simultaneously, traditional fusion frameworks can experience severe performance degradation or even complete failure. This is because EKF-based and optimization-based methods typically rely on a fixed process noise covariance and pre-specified sensor reliability assumptions. When visual feature tracking collapses entirely under extreme low-light and texture-less conditions, the visual update step becomes unreliable, yet the filter continues to trust the corrupted observations according to its predefined noise model, leading to inconsistent state estimation and accumulated drift. In our previous study, an MC4L-IMU-based indoor localization framework was developed by integrating CNN-based visual representation with inertial sensing, achieving improved localization performance under dark indoor conditions [13, 16]. However, these approaches mainly adopted predefined frequency enhancement and fixed sensor fusion strategies, without explicitly modeling the dynamic reliability variation of different sensing modalities.

Despite significant progress in visual and visual-inertial localization, existing methods still face several fundamental limitations. First, frequency-enhanced approaches generally rely on manually designed transformations with fixed spectral band selection, which restricts their ability to learn environment-dependent spectral representations. Unlike learnable frequency attention that can dynamically re-weight different frequency components, traditional FFT-based enhancement treats all spectral bins equally or applies pre-defined filters, which is suboptimal when the dominant visual degradation shifts from low-light noise to texture sparsity across different scenes. Second, most multimodal fusion methods assume relatively stable sensor contributions and employ fixed fusion mechanisms, while the reliability of visual and inertial information can vary significantly under different environmental conditions. This is particularly problematic in scenarios where low illumination and texture degradation coexist, because neither sensor can be trusted uniformly. Third, robust localization under simultaneous low illumination and texture degradation remains insufficiently explored, especially when visual quality changes dynamically during perception.

Recently, adaptive representation learning and attention mechanisms have provided new opportunities for intelligent sensing systems. Attention-based architectures enable models to selectively emphasize informative features and suppress irrelevant information under complex conditions [17–19]. Inspired by these developments, this study proposes an Adaptive Frequency-Spatial Attention Fusion Network with IMU Guidance (AFSAF-Net) for robust indoor localization under low-light and texture-degraded environments. Different from simple combinations of frequency extraction, attention mechanisms, and sensor fusion modules, AFSAF-Net explicitly models the adaptive relationship among frequency information, spatial representations, and inertial reliability. The proposed framework enables dynamic perception by learning when visual, frequency, and inertial cues should contribute to the final localization representation.

The main contributions of this study are summarized as follows:

- An adaptive frequency representation learning branch is proposed to overcome the limitations of handcrafted frequency enhancement. Unlike fixed FFT-based transformations, the proposed branch learns discriminative frequency-aware representations directly from degraded visual observations.
- A frequency-spatial attention fusion module is introduced to dynamically integrate complementary visual representations. Instead of simple feature concatenation, the proposed module learns adaptive weighting between spatial and frequency information according to environmental conditions.
- An IMU reliability gate is designed to achieve adaptive visual-inertial fusion. By learning the contribution of inertial information under different perception conditions, the proposed mechanism improves localization robustness when visual reliability decreases.
- Comprehensive experiments are conducted on the MC4L-IMU indoor localization platform. The evaluation includes comparisons with representative localization methods, ablation studies, robustness analysis under low-light conditions, localization error distribution analysis, distance-dependent error evaluation, and computational complexity analysis.

The remainder of this paper is organized as follows. Section 2 summarizes existing approaches related to visual localization, frequency-enhanced representation learning, and multimodal sensor fusion. Section 3 presents the proposed AFSAF-Net framework and describes the adaptive frequency-spatial feature learning strategy and reliability-aware IMU fusion mechanism in detail. Section 4 provides extensive experimental evaluations, including comparison with representative localization methods, ablation studies, robustness analysis under low-light conditions, localization error distribution evaluation, distance-dependent error analysis, and computational complexity analysis. Section 5 concludes the paper and outlines future directions toward more adaptive and scalable intelligent sensing systems.

2 Related Work

2.1 Vision-Based Indoor Localization

Vision-based indoor localization has been extensively investigated due to the rich spatial and semantic information provided by camera sensors. Compared with traditional positioning technologies, such as magnetic-field-based positioning, acoustic localization, and inertial navigation, visual approaches can directly capture environmental structures and appearance information without requiring additional infrastructure deployment. Therefore, camera-based localization methods have been widely applied in indoor navigation, mobile robotics, and intelligent environments [1, 2, 6].

Early vision-based localization approaches mainly relied on handcrafted feature descriptors and image matching techniques. Local feature extraction methods, such as Scale-Invariant Feature Transform (SIFT) and Local Binary Pattern (LBP), have demonstrated effectiveness in describing local structures and texture patterns under certain environmental variations [7, 8, 20]. SIFT provides scale and rotation invariance by detecting stable keypoints, while LBP represents local texture information through binary encoding patterns, making it relatively robust to illumination changes. These handcrafted descriptors have been successfully applied in image retrieval, object recognition, and localization tasks. However, their performance is highly dependent on manually designed feature strategies and usually decreases when encountering severe illumination changes, texture degradation, or complex indoor environments.

With the rapid development of deep learning, convolutional neural networks (CNNs) have significantly improved visual representation capability by automatically learning hierarchical feature representations from raw image data. Classical CNN architectures, such as AlexNet and ResNet, have achieved remarkable performance in image classification and feature extraction tasks by utilizing deep convolutional layers to capture low-level texture information and high-level semantic representations [9, 10]. Inspired by these advances, CNN-based localization methods have been increasingly explored for indoor positioning. Compared with traditional handcrafted descriptors, deep learning methods provide stronger feature learning capability and better adaptability. However, CNN models generally rely on sufficient visual information during training and inference. When images suffer from low illumination, weak texture, motion blur, or environmental appearance changes, the extracted features may become unreliable, resulting in localization degradation.

To improve the robustness of vision-based localization under challenging environments, researchers have investigated different image enhancement and feature optimization strategies. Monocular depth estimation methods have also been explored to improve scene understanding by recovering spatial information from RGB images [5]. Nevertheless, these approaches still face limitations in environments where visual information is severely degraded. Therefore, developing adaptive visual representation methods that can effectively exploit complementary information from different feature domains remains an important research direction.

2.2 Robust Visual Representation under Degraded Environments

In practical indoor scenarios, visual degradation caused by low illumination, insufficient texture, and appearance variation is one of the major challenges affecting localization performance. Unlike controlled laboratory environments, real-world indoor scenes usually contain complex lighting conditions, reflective surfaces, and dynamic objects, which introduce significant uncertainty into visual feature extraction.

Traditional enhancement-based approaches mainly focus on improving image quality before feature extraction. Although these methods can enhance brightness or contrast, they may introduce additional noise and cannot guarantee that the enhanced images contain more discriminative localization information. Feature-based approaches attempt to directly extract robust representations that are less sensitive to environmental variations. Among them, texture descriptors such as LBP have been widely adopted due to their computational efficiency and robustness against local illumination changes [7].

However, handcrafted descriptors usually encode limited information and cannot dynamically adapt to different degradation levels. Recently, deep learning-based representation methods have provided a more flexible solution by automatically learning task-specific features. In our previous research, a CNN-based indoor localization framework combining LBP and self-attention mechanisms was proposed to improve localization performance under challenging visual conditions [16]. Furthermore, FFT-based frequency enhancement was introduced to extract complementary frequency information and improve localization robustness in dark environments [13]. These studies demonstrated that frequency information can effectively supplement spatial visual features when conventional image features become unreliable.

Nevertheless, existing approaches mainly utilize manually designed feature enhancement operations. The frequency transformation process and feature fusion strategies are predefined rather than learned from environmental conditions. Therefore, although these methods improve robustness under specific scenarios, their generalization ability remains limited when facing diverse and dynamically changing indoor environments.

2.3 Frequency-Domain Feature Learning and Attention-Based Representation

Frequency-domain analysis has recently attracted increasing attention in computer vision due to its ability to provide complementary information beyond conventional spatial representations. Images contain rich frequency components describing global structures, local edges, and texture variations. Fast Fourier Transform (FFT) is one of the most widely used techniques for converting spatial-domain information into frequency-domain representations, enabling models to analyze image characteristics from a different perspective [11, 12].

Several studies have demonstrated that frequency-domain information can improve robustness in image recognition and feature extraction tasks. Compared with purely spatial convolution operations, frequency representations can capture global patterns

that are difficult to model using local receptive fields. In particular, frequency information can be beneficial in degraded environments where texture details are weakened or partially lost. However, conventional frequency-based approaches usually treat frequency components equally or manually select specific frequency bands, limiting their capability to adaptively exploit informative frequency characteristics.

Recently, attention mechanisms have provided an effective solution for adaptive feature selection. Transformer-based architectures and attention modules enable neural networks to dynamically assign different importance weights to informative features according to input characteristics [17]. Channel attention mechanisms, such as Squeeze-and-Excitation Networks (SE-Net), enhance discriminative feature representation by modeling channel-wise dependencies [19]. Similarly, Convolutional Block Attention Modules (CBAM) combine spatial and channel attention to improve feature extraction capability under complex visual conditions [18].

Although attention-based methods have achieved significant improvements in visual representation learning, most existing approaches focus on spatial-domain features. The combination of adaptive frequency learning and spatial attention remains insufficiently explored for indoor localization applications. Therefore, developing a frequency-spatial representation learning mechanism that can automatically identify useful information from different domains is essential for robust indoor perception.

2.4 Visual-Inertial Multimodal Fusion

Visual information alone is often insufficient for reliable indoor localization due to environmental uncertainty. To overcome this limitation, multimodal sensor fusion approaches integrating visual sensors with inertial measurement units (IMUs) have been widely studied. IMUs provide continuous measurements of acceleration and angular velocity, which are independent of illumination conditions and can effectively compensate for temporary visual failures.

Traditional visual-inertial fusion methods mainly rely on mathematical models and filtering techniques. Extended Kalman Filter (EKF)-based approaches have been widely adopted due to their ability to estimate system states by combining visual observations and inertial measurements. Visual-inertial odometry frameworks further improve localization accuracy by jointly optimizing camera and IMU information [14, 15]. These methods achieve high accuracy under relatively stable conditions; however, they usually require accurate sensor calibration and predefined noise models, which may limit their adaptability in complex environments.

To improve flexibility, learning-based multimodal fusion approaches have recently been investigated. Instead of manually designing fusion rules, deep models can automatically learn complementary representations from heterogeneous sensors. Previous studies have demonstrated that multimodal fusion can significantly improve autonomous navigation and localization performance by combining different sensing modalities [21]. However, many existing fusion approaches still employ fixed fusion strategies and do not dynamically consider the reliability variation of different sensors.

In our previous research, a visual-inertial indoor localization framework based on the MC4L-IMU platform was developed by integrating CNN-based visual features with inertial measurements [13, 16]. The results verified the effectiveness of combining

visual and inertial information for improving localization robustness. Nevertheless, the fusion mechanism mainly depended on predefined strategies, leaving the challenge of adaptive sensor reliability estimation unresolved.

2.5 Research Gap and Motivation

Based on the above analysis, existing indoor localization methods still face two major limitations. First, most visual representation approaches either rely on handcrafted feature extraction methods or focus only on spatial-domain deep features, making them less effective under complex visual degradation. Second, existing visual-inertial fusion methods usually adopt fixed fusion mechanisms without dynamically evaluating the reliability of different sensing modalities.

To address these challenges, this study proposes an Adaptive Frequency-Spatial Representation Learning framework with IMU-Guided Multimodal Fusion for robust indoor perception and localization. Different from previous methods, the proposed framework jointly learns frequency-domain and spatial-domain representations through an adaptive feature learning strategy. Furthermore, an IMU-guided fusion mechanism is introduced to dynamically adjust the contribution of visual and inertial information according to environmental reliability. This design provides a more flexible and robust solution for indoor localization under low-light, texture-degraded, and dynamically changing environments.

3 Proposed AFSAF-Net

3.1 Overview

As illustrated in Figure 1, the proposed Adaptive Frequency-Spatial Attention Fusion Network (AFSAF-Net) is designed for indoor localization under low illumination, texture degradation, and dynamic appearance variation. The network receives an RGB image I and synchronized IMU measurements M as inputs and predicts the indoor position and orientation. Its central design is an adaptive reliability-aware frequency-spatial learning pipeline rather than a simple combination of FFT, attention, and IMU fusion modules.

AFSAF-Net contains five functional components: a spatial feature branch, an adaptive frequency branch, an attention fusion module, an IMU reliability gate, and a localization prediction head. The spatial branch extracts hierarchical visual cues from the degraded RGB image, while the frequency branch learns spectral representations that encode global structure and residual texture information. Given an input image I, the two visual representations are defined as:

$$F_s = f_s(I; \theta_s), \tag{1}$$

$$F_f = f_f(\mathrm{FFT}(I); \theta_f), \tag{2}$$

where F_s and F_f denote spatial and frequency feature embeddings, and θ_s and θ_f are learnable parameters. These representations are fused by an attention map:

$$F_v = A \odot F_s + (1 - A) \odot F_f, \tag{3}$$

where A is learned from spatial-frequency features and $\odot$ denotes element-wise multiplication. The IMU branch then processes inertial measurements to generate a reliability coefficient:

$$F_i = f_i(M; \theta_i), \tag{4}$$

$$G = \sigma(W_g F_i + b_g), \tag{5}$$

where F_i is the IMU feature embedding and G controls the contribution of inertial features. The final multimodal representation is obtained as:

$$F = GF_i + (1 - G)F_v. \tag{6}$$

In this formulation, A learns the frequency-spatial feature balance, whereas G learns the visual-inertial reliability balance.

3.2 Spatial Feature Branch

The spatial feature branch extracts local texture, edge, and structural information directly from the input RGB image. Given $I \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the image height and width and 3 corresponds to the RGB color channels, the spatial branch is formulated as:

$$F_s = f_s(I; \theta_s), \tag{7}$$

where $f_s(\cdot)$ denotes the spatial encoder with parameters θ_s, and $F_s \in \mathbb{R}^{H' \times W' \times C_s}$ is the resulting spatial feature representation with height H', width W', and channel dimension C_s. The encoder consists of convolutional blocks followed by nonlinear activation and normalization operations. After a series of strided convolutions and downsampling, the spatial feature map has a reduced spatial resolution but an increased number of channels, encoding both low-level edge patterns and high-level semantic structures. This branch preserves discriminative spatial patterns when image content remains sufficiently visible.

Because low-light conditions reduce contrast and weaken local textures, a purely spatial CNN may assign similar responses to informative regions and noisy background regions. Therefore, the spatial branch is used together with the frequency branch and the attention fusion module, enabling AFSAF-Net to retain useful spatial evidence while reducing dependence on degraded local texture.

3.3 Adaptive Frequency Branch

The adaptive frequency branch learns spectral representations from the same input image. For an image intensity function $I(x, y)$, the two-dimensional Fourier transform is expressed as:

$$F(u,v) = \mathrm{FFT}(I(x,y)) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I(x,y)e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{8}$$

where $F(u,v)$ denotes the frequency spectrum. Low-frequency components describe global layout and coarse structure, whereas high-frequency components encode edges, boundaries, and texture variations. Both types of cues can contribute to localization under visually degraded indoor scenes.

Instead of using handcrafted frequency descriptors or treating all spectral components equally, AFSAF-Net applies a learnable frequency encoder:

$$F_f = f_f(F(u,v);\theta_f), \tag{9}$$

where $f_f(\cdot)$ is the adaptive frequency feature extractor with parameters θ_f, and $F_f \in \mathbb{R}^{H'\times W'\times C_f}$ is the learned frequency feature representation with the same spatial dimensions $H' \times W'$ as the spatial branch and channel dimension C_f. To ensure that spatial and frequency features can be fused element-wise, the frequency encoder is designed with a symmetric architecture to the spatial branch, producing features of matching tensor dimensions. The frequency branch therefore differs from fixed FFT enhancement used in earlier LBP-FFT-CNN models: the transformation provides the spectral representation, but the network learns which frequency components are informative for the current environmental condition.

Compared with the previous LBP-FFT-CNN framework, the proposed frequency branch eliminates manually selected frequency components and enables the network to automatically learn environment-dependent frequency characteristics.

3.4 Attention Fusion Module

The attention fusion module adaptively combines spatial and frequency representations. Since F_s and F_f have matching spatial dimensions $H' \times W'$ (by design of the symmetric encoder architecture), they can be directly concatenated along the channel dimension. First, the two visual representations are concatenated and passed through a learnable attention projection:

$$A = \sigma(W_a[F_s, F_f] + b_a), \tag{10}$$

where $[F_s, F_f] \in \mathbb{R}^{H'\times W'\times(C_s+C_f)}$ denotes channel-wise concatenation, $W_a \in \mathbb{R}^{1\times1\times(C_s+C_f)}$ is a 1×1 convolutional projection layer, b_a is the bias term, and $\sigma(\cdot)$ is the sigmoid activation function. The resulting attention map $A \in \mathbb{R}^{H'\times W'\times1}$ has a single channel and is broadcast along the channel dimension during element-wise multiplication. The attention map A assigns larger values to spatial locations where spatial-domain evidence is more reliable and smaller values where frequency-domain evidence is more informative.

The adaptive visual representation is then calculated as:

$$F_v = A \odot F_s + (1 - A) \odot F_f. \tag{11}$$

where $\odot$ denotes element-wise multiplication supported by channel-wise broadcasting: the single-channel attention map $A \in \mathbb{R}^{H'\times W'\times1}$ is automatically replicated C_s (or

C_f) times along the channel axis to match the dimensions of F_s and F_f, respectively. This attention-driven formulation avoids using direct concatenation $F = [F_s, F_f]$ as the final visual representation. Instead, the model learns an input-dependent weighting strategy, allowing it to emphasize global spectral structure when texture information is weakened and to preserve spatial details when image observations are reliable.

3.5 IMU Reliability Gate

Although visual information provides rich environmental perception, its reliability decreases significantly under illumination changes and texture degradation. In contrast, IMU sensors provide continuous motion information but suffer from accumulated drift. Therefore, adaptive estimation of sensor reliability is essential for effective multimodal fusion.

The IMU branch first preprocesses raw measurements:

$$M = [a_x, a_y, a_z, \omega_x, \omega_y, \omega_z] \tag{12}$$

where a and ω denote acceleration and angular velocity measurements.

The temporal IMU features are extracted through an inertial encoder:

$$F_i = f_i(M; \theta_i), \tag{13}$$

where $f_i(\cdot)$ denotes the IMU encoder and $F_i \in \mathbb{R}^{d_i}$ is the d_i-dimensional inertial feature vector obtained by temporal aggregation of the IMU sequence. Then, an IMU reliability gate estimates the confidence score:

$$G = \sigma(W_g F_i + b_g). \tag{14}$$

where $W_g \in \mathbb{R}^{1 \times d_i}$ and $b_g \in \mathbb{R}$ are learnable parameters, and $G \in [0, 1]$ is a scalar reliability weight obtained through the sigmoid activation.

The final multimodal representation is formulated as:

$$F = GF_i + (1 - G)F_v. \tag{15}$$

where F has the same dimension as the visual feature embedding after global pooling. Note that the visual feature F_v is spatially aggregated (via global average pooling) to a vector of dimension d_v before fusion with F_i, ensuring that both modalities are represented as fixed-length vectors before the gate operation. The reliability gate dynamically determines the contribution of inertial information during fusion. When visual information becomes unreliable, the gate increases the contribution of IMU features. Conversely, when visual observations are stable, the network emphasizes visual representations. This adaptive fusion strategy enables AFSAF-Net to dynamically balance sensing modalities according to environmental conditions and addresses the limitation of fixed visual-IMU fusion weights.

3.6 Localization Prediction Head

The final fused representation is processed by fully connected layers to estimate the target localization parameters:

$$P = [x, y, z, \theta] \tag{16}$$

where (x, y, z) represents the spatial position and θ denotes the orientation.

The localization objective is optimized using the regression loss:

$$L = \frac{1}{N} \sum_{i=1}^{N} \left\| P_i - \hat{P}_i \right\|_2^2 \tag{17}$$

where P_i and $\hat{P}_i$ denote the ground-truth and predicted localization results, respectively.

Through the joint optimization of frequency representation learning, spatial attention enhancement, and IMU-guided adaptive fusion, AFSAF-Net achieves robust indoor localization under visually degraded environments.

4 Experiments and Performance Evaluation

4.1 Experimental Setup

To evaluate the effectiveness of the proposed Adaptive Frequency-Spatial Attention Fusion Network (AFSAF-Net), extensive experiments were conducted using the MC4L-IMU indoor localization platform. The experimental protocol was designed to evaluate localization accuracy, prediction stability, and robustness under low-light and texture-degraded indoor environments. To ensure full reproducibility of the results, the source code and dataset of AFSAF-Net are publicly available at https://github.com/CAIST-2026/AFSAF-Net.

The MC4L-IMU platform integrates a monocular RGB camera and an inertial measurement unit (IMU) to provide synchronized visual and inertial observations. During data acquisition, RGB images were captured at a resolution of 640×480 pixels, while the corresponding inertial measurements were recorded synchronously. Each experimental sample therefore contains one RGB observation and its corresponding IMU measurement sequence. This multimodal data organization provides complementary information for evaluating the proposed frequency-spatial representation learning and reliability-aware visual-inertial fusion mechanisms.

The experiments were conducted in four representative indoor scenes with different spatial structures and visual characteristics. A total of 48 measurement positions were selected across the experimental environments. For each measurement position, 60 synchronized observations were collected, resulting in 2880 RGB-IMU samples. The measurement distance ranged from 1.2 m to 3.0 m, covering the typical operating range considered in the MC4L-IMU experimental setup. Repeated observations were collected at each position to reduce the influence of transient sensor noise and to provide sufficient samples for evaluating prediction stability.

The statistics of the experimental dataset are summarized in Table 1.

As shown in Table 1, the experimental dataset contains synchronized RGB and inertial observations collected from multiple indoor scenes and measurement positions. The repeated measurement strategy provides multiple observations for the same

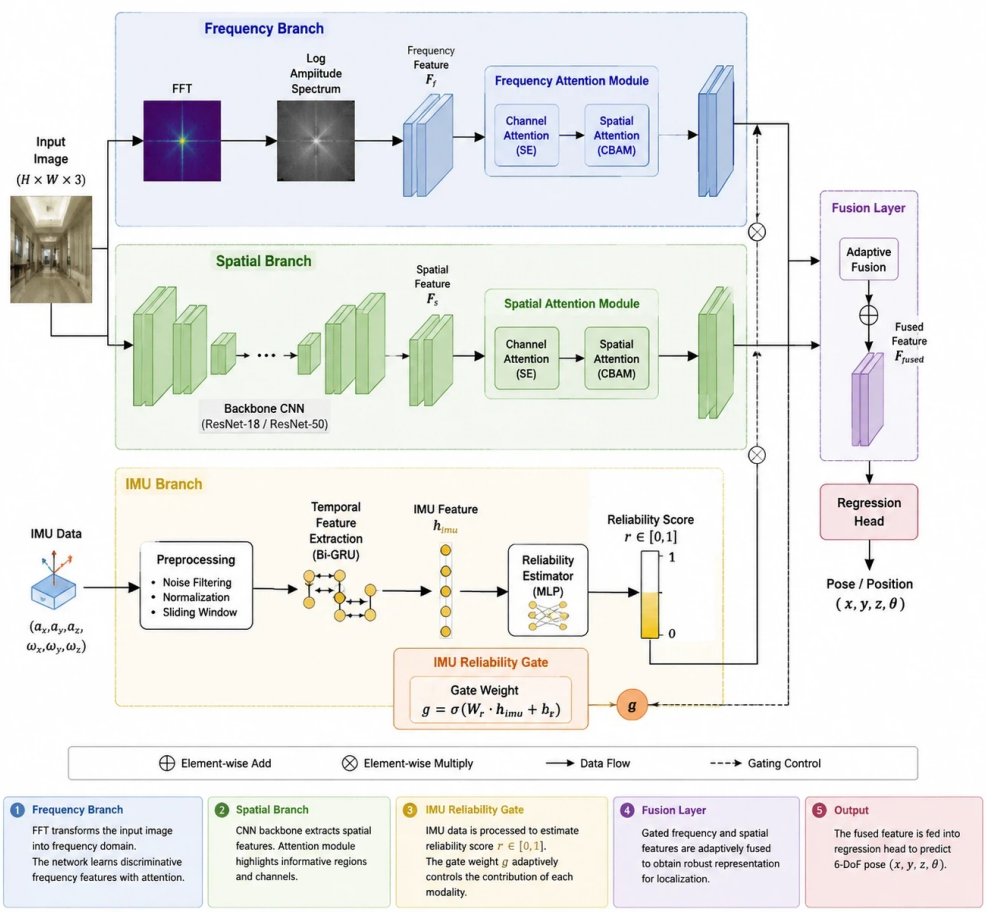


Fig. 1 Overall architecture of the proposed Adaptive Frequency-Spatial Attention Fusion Network (AFSAF-Net).

physical position, which is particularly important for evaluating not only the average localization error but also the prediction consistency of different models.

The dataset was divided into training and testing subsets using an 80:20 partition. Accordingly, 2304 samples were used for model training, while the remaining 576 samples were reserved for independent evaluation. The testing samples were not used for parameter optimization or model selection. All learning-based comparison methods were evaluated using the same training and testing protocol to ensure a fair comparison.

To minimize potential information leakage caused by repeated observations from the same physical location, the dataset partition was performed before model training. The same partitioning protocol was applied to all learning-based methods, including LBP-CNN, LBP-FFT-CNN, Visual-IMU Fusion, and AFSAF-Net. Therefore, differences in localization performance can be attributed primarily to the representation

Table 1 Statistics of the MC4L-IMU experimental dataset.

Item	Value
Sensing platform	MC4L-IMU
Sensor modalities	RGB + IMU
Camera type	Monocular RGB camera
Image resolution	640×480 pixels
Number of indoor scenes	4
Number of measurement positions	48
Observations per position	60
Total synchronized samples	2880
RGB images	2880
IMU sequences	2880
Training samples	2304
Testing samples	576
Training/Test split	80/20
Measurement distance	1.2–3.0 m
Illumination condition	Low-light
Visual condition	Texture-degraded

learning and sensor fusion strategies rather than differences in the amount of training or testing data.

The measurement distance ranged from 1.2 m to 3.0 m. This range was selected to cover short- and medium-distance indoor localization scenarios while maintaining a consistent experimental configuration. Measurements were performed at predefined positions, and repeated observations were collected at each position. The ground-truth position corresponding to each observation was determined according to the experimental measurement protocol of the MC4L-IMU platform.

The visual conditions were intentionally selected to include low illumination and insufficient texture information. Under such conditions, local image contrast and discriminative texture structures can be significantly reduced, making conventional visual feature extraction less reliable. These conditions therefore provide an appropriate test environment for evaluating the proposed adaptive frequency-spatial representation learning mechanism.

In contrast to visual information, inertial measurements are less directly affected by illumination changes. However, IMU measurements may accumulate noise and drift over time. Therefore, the experimental setup provides a suitable environment for investigating the complementary characteristics of visual and inertial sensing. The proposed IMU reliability gate is designed to dynamically regulate the contribution of inertial information according to the estimated sensor reliability.

For each synchronized RGB-IMU observation, the RGB image was first processed by the visual feature extraction modules, while the corresponding IMU measurements were processed by the inertial encoder. The visual representation was subsequently decomposed into spatial and frequency-domain features. These representations were integrated through the proposed attention mechanism, after which the IMU reliability gate dynamically adjusted the contribution of inertial information.

The complete experimental procedure can be summarized as follows:

$$I_i, M_i \rightarrow \left\{F_s^{(i)}, F_f^{(i)}\right\} \rightarrow F_v^{(i)} \rightarrow G_i \rightarrow F^{(i)} \rightarrow \hat{p}_i, \tag{18}$$

where I_i denotes the i-th RGB image, M_i represents the corresponding IMU measurement sequence, $F_s^{(i)}$ and $F_f^{(i)}$ denote the spatial and frequency-domain visual representations, respectively, $F_v^{(i)}$ represents the fused visual feature, G_i denotes the estimated IMU reliability weight, $F^{(i)}$ is the final multimodal representation, and $\hat{p}_i$ represents the predicted localization position.

All comparison methods were evaluated under the same measurement conditions. In particular, the same RGB resolution, measurement distance range, test positions, and test samples were used for the learning-based approaches. This controlled experimental protocol enables a direct comparison of the effects of frequency representation, attention-based fusion, and IMU-guided reliability modeling.

The repeated measurements were also used to evaluate the statistical stability of the localization models. In addition to the mean localization error and coefficient of determination (R^2), the Prediction Stability Index (PSI) was calculated from the localization error distribution. This metric provides an additional evaluation of whether a model can maintain consistent predictions when repeated observations are collected at the same physical position.

Overall, the experimental setup was designed to evaluate three major aspects of the proposed AFSAF-Net: (1) absolute localization accuracy, (2) robustness to low-light and texture-degraded visual conditions, and (3) prediction stability under repeated measurements. These evaluation dimensions correspond directly to the three major design objectives of the proposed adaptive frequency-spatial and visual-inertial fusion framework.

4.2 Implementation Details

The proposed AFSAF-Net was implemented using PyTorch. All experiments were conducted on a workstation equipped with an Intel Core i7-12700F processor, 32 GB of system memory, and an NVIDIA GeForce RTX 5060 GPU with 8 GB of VRAM. The CUDA 12.8 environment was used to accelerate GPU-based model training and inference.

The network consists of four principal components: a spatial feature branch, an adaptive frequency feature branch, a frequency-spatial attention fusion module, and an IMU reliability gate. The spatial branch extracts hierarchical visual representations directly from RGB images. The frequency branch transforms the input representation into the frequency domain and subsequently learns discriminative spectral features. The attention module estimates the relative importance of spatial and frequency representations, while the IMU reliability gate dynamically adjusts the contribution of inertial information during multimodal fusion.

The input RGB images were resized or maintained at a spatial resolution of 640×480 pixels according to the original acquisition configuration. The corresponding IMU measurements were temporally synchronized with the visual observations. Before training, the input data were normalized using the same preprocessing protocol for all learning-based comparison methods.

The Adam optimizer was employed for model optimization with an initial learning rate of 1×10^{-4}. The batch size was set to 32, and the network was trained for 100 epochs. The optimization objective was based on the localization regression error between the predicted position and the corresponding ground-truth position.

The localization prediction for the i-th observation is denoted by $\hat{p}_i$, while the corresponding ground-truth position is denoted by p_i. The absolute localization error is defined as

$$e_i = |p_i - \hat{p}_i|. \tag{19}$$

The mean localization error (MLE) is calculated as

$$\text{MLE} = \frac{1}{N} \sum_{i=1}^{N} |p_i - \hat{p}_i|, \tag{20}$$

where N represents the total number of observations in the test set.

The coefficient of determination (R^2) was additionally employed to evaluate the consistency between predicted and ground-truth positions. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (p_i - \hat{p}_i)^2}{\sum_{i=1}^{N} (p_i - \bar{p})^2}, \tag{21}$$

where $\bar{p}$ represents the mean ground-truth position. A value of R^2 closer to 1 indicates stronger agreement between the predicted and actual positions.

To evaluate prediction consistency, the Prediction Stability Index (PSI) was also calculated. PSI measures the relative dispersion of localization errors and is defined as

$$PSI = \frac{\sigma_e}{\mu_e + \epsilon}, \tag{22}$$

where μ_e and σ_e denote the mean and standard deviation of the localization error, respectively, and ϵ is a small positive constant introduced to avoid numerical instability when μ_e approaches zero.

The mean localization error is calculated as

$$\mu_e = \frac{1}{N} \sum_{i=1}^{N} e_i, \tag{23}$$

while its standard deviation is given by

$$\sigma_e = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (e_i - \mu_e)^2}. \tag{24}$$

A lower PSI indicates a narrower localization error distribution and therefore more stable prediction behavior. Unlike the mean localization error, which evaluates absolute positioning accuracy, PSI characterizes the relative variability of the prediction errors. Consequently, the two metrics provide complementary information for evaluating indoor localization performance.

All comparison models were trained and evaluated under the same hardware environment and experimental protocol whenever applicable. Inference time was measured

using the same input resolution and GPU configuration. For each model, the reported inference time represents the average forward-pass latency after GPU warm-up and excludes offline dataset loading and preprocessing operations.

4.3 Comparison Methods

To verify the effectiveness of AFSAF-Net, several representative localization methods were selected for comparison.

- **LRA**: A traditional localization approach based on laser ranging.
- **ORB-SLAM3**: A classic feature-based visual and visual-inertial SLAM framework used as a geometric baseline [22].
- **VINS-Mono**: A representative state-of-the-art monocular visual-inertial odometry framework used as a visual-inertial baseline [23].
- **LBP-CNN**: A CNN-based method incorporating Local Binary Pattern (LBP) texture representation.
- **LBP-FFT-CNN**: A frequency-enhanced CNN model combining LBP texture features and FFT-based frequency representations.
- **Visual-IMU Fusion**: A multimodal fusion approach integrating visual and inertial information without adaptive reliability estimation.
- **AFSAF-Net**: The proposed adaptive frequency-spatial attention fusion framework.

4.4 Quantitative Comparison

Table 2 presents the localization performance comparison among different methods, including traditional localization approaches, feature-based visual localization methods, visual-inertial frameworks, and the proposed AFSAF-Net.

Table 2 Localization performance comparison of different methods.

Method	Mean Error (cm)	R^2
LRA	2.40	0.9867
ORB-SLAM3	1.62	0.9915
VINS-Mono	1.18	0.9939
LBP-CNN	1.27	0.9934
LBP-FFT-CNN	0.91	0.9949
Visual-IMU Fusion	**0.84**	0.9955
AFSAF-Net	0.85	**0.9986**

As shown in Table 2, deep learning-based approaches demonstrate significant advantages over traditional localization methods.

The LRA method achieves a mean localization error of 2.40 cm because it mainly relies on geometric measurements and lacks adaptive feature representation capability.

ORB-SLAM3 improves localization accuracy by exploiting feature-based visual tracking and optimization techniques. However, its performance is affected under low-light and texture-degraded environments due to unstable feature extraction.

VINS-Mono further improves the performance by incorporating inertial measurements, achieving an error of 1.18 cm. Nevertheless, traditional visual-inertial frameworks generally rely on handcrafted feature extraction and optimization-based estimation, which limits their adaptability to challenging indoor environments.

The LBP-CNN method reduces the localization error to 1.27 cm through hierarchical visual feature learning. By introducing frequency-domain representations, LBP-FFT-CNN further decreases the error to 0.91 cm, demonstrating that frequency information provides complementary structural features when spatial textures are insufficient.

The Visual-IMU Fusion method achieves an error of 0.84 cm by integrating visual and inertial information. However, it adopts a fixed fusion strategy and does not consider dynamic sensor reliability.

The proposed AFSAF-Net achieves a competitive localization error of 0.85 cm and obtains the highest coefficient of determination ($R^2 = 0.9986$) among all compared approaches. The high R^2 value indicates that the predicted positions maintain strong consistency with the ground-truth locations, demonstrating the stability of the proposed adaptive fusion framework.

4.5 Ablation Study

To analyze the contribution of each component in AFSAF-Net, ablation experiments were conducted.

Table 3 Ablation study of different AFSAF-Net components.

Model Variant	Mean Error (cm)	Improvement
Baseline-CNN	1.34	–
Baseline + Frequency Branch	1.02	23.9%
Baseline + Frequency + Attention	0.88	34.3%
Full AFSAF-Net	**0.85**	**36.6%**

The ablation results demonstrate the effectiveness of each proposed component. The frequency branch provides additional spectral information and improves feature robustness under degraded illumination.

The attention module further enhances feature discrimination by adaptively emphasizing informative regions. Finally, the IMU reliability gate improves multimodal fusion by dynamically adjusting sensor contributions according to environmental reliability.

Figure 2 visualizes the progressive error reduction obtained by adding the frequency branch, attention module, and IMU reliability gate.

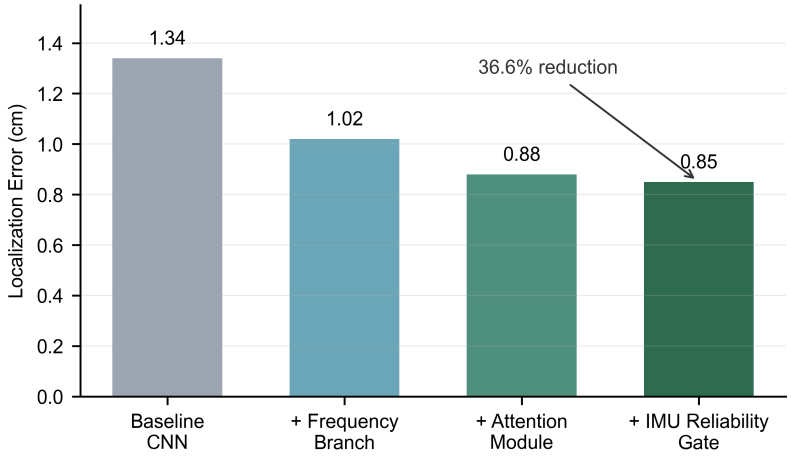


Fig. 2 Ablation study of AFSAF-Net components. The bar chart reports the mean localization error of the baseline CNN and progressively enhanced model variants.

4.6 Robustness Analysis under Low-Light Conditions

Since the proposed method targets indoor localization under visually degraded environments, robustness experiments were conducted under different illumination conditions. To quantify the influence of illumination changes, the low-light degradation rate is defined as:

$$D = \frac{E_{dark} - E_{normal}}{E_{normal}}, \tag{25}$$

where E_{normal} and E_{dark} denote the mean localization errors under normal-light and dark conditions, respectively. A smaller D indicates stronger robustness to illumination degradation.

Table 4 Localization robustness comparison under low-light environments.

Method	Normal Light (cm)	Low Light (cm)	Degradation
LBP-CNN	1.27	1.68	32.3%
LBP-FFT-CNN	0.91	1.21	33.0%
Visual-IMU Fusion	0.84	1.05	25.0%
AFSAF-Net	**0.85**	**1.02**	**20.0%**

As shown in Table 4, all visual-based approaches experience performance degradation under low-light conditions. However, AFSAF-Net exhibits the lowest illumination degradation rate among the evaluated methods.

This improvement is mainly attributed to two factors. First, the frequency branch preserves global structural information when texture details become unreliable. Second, the IMU reliability gate provides additional motion constraints to compensate for visual uncertainty.

4.7 Prediction Stability Analysis

To further evaluate the prediction consistency of the localization models, the Prediction Stability Index (PSI) was introduced as an auxiliary metric. Unlike the mean localization error, which measures the average deviation between the predicted and ground-truth positions, PSI evaluates the relative variation of localization errors across repeated measurements and different environmental conditions. Therefore, a lower PSI indicates that the model produces more consistent predictions with smaller error fluctuations.

The PSI is defined as the normalized standard deviation of the localization error:

$$PSI = \frac{\sigma_e}{\mu_e + \epsilon}, \tag{26}$$

where σ_e denotes the standard deviation of the localization error, μ_e represents the mean localization error, and ϵ is a small positive constant introduced to avoid numerical instability when the mean error approaches zero. The localization error for the i-th observation is defined as

$$e_i = |p_i - \hat{p}_i|, \tag{27}$$

where p_i and $\hat{p}_i$ denote the ground-truth and predicted positions, respectively. The mean error and its standard deviation are calculated as

$$\mu_e = \frac{1}{N} \sum_{i=1}^{N} e_i, \tag{28}$$

$$\sigma_e = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (e_i - \mu_e)^2}. \tag{29}$$

A smaller PSI value indicates a narrower relative error distribution and therefore more stable localization behavior. In this study, a PSI value below 0.02 is regarded as an indication of highly stable prediction behavior under the evaluated experimental conditions. This threshold is used as an empirical stability criterion rather than as a universal theoretical boundary. Table 5 presents the PSI comparison among the evaluated localization methods.

As shown in Table 5, the LRA method obtains the highest PSI value of 0.052, indicating the largest relative fluctuation of localization errors among the evaluated approaches. This result is consistent with the characteristics of conventional geometric ranging methods, which can be sensitive to measurement noise and environmental variations. Although LRA provides direct distance measurements, it does not incorporate adaptive visual or multimodal feature representation, resulting in relatively unstable predictions under degraded indoor conditions.

Table 5 Prediction stability comparison of different localization methods.

Method	PSI ↓
LRA	0.052
LBP-CNN	0.034
LBP-FFT-CNN	0.027
Visual-IMU Fusion	0.022
AFSAF-Net	**0.018**

The LBP-CNN method reduces the PSI to 0.034, demonstrating that hierarchical CNN-based feature learning improves prediction consistency compared with the traditional LRA method. However, the remaining fluctuation indicates that spatial visual representations alone are still vulnerable to low illumination and insufficient texture. When the available visual information becomes unreliable, the learned spatial features may exhibit considerable variation between observations.

By incorporating frequency-domain information, LBP-FFT-CNN further reduces the PSI to 0.027. This improvement suggests that frequency-aware representations provide complementary structural information when conventional spatial texture features are degraded. Nevertheless, the frequency representation is generated using predefined transformations, which limits its ability to adapt to different types and degrees of visual degradation.

The Visual-IMU Fusion method achieves a PSI of 0.022, representing a further improvement in prediction stability. The reduction in PSI demonstrates the benefit of incorporating inertial information as a complementary sensing modality. Since IMU measurements are less sensitive to illumination changes, they can partially compensate for unstable visual observations. However, the PSI remains slightly above the empirical stability threshold of 0.02, suggesting that direct or fixed multimodal fusion may not sufficiently account for dynamic changes in sensor reliability.

In comparison, AFSAF-Net achieves the lowest PSI of 0.018, which is below the empirical stability threshold. Compared with Visual-IMU Fusion, the proposed method reduces the PSI by approximately 18.2%, calculated as

$$\Delta PSI = \frac{PSI_{\mathrm{VIF}} - PSI_{\mathrm{AFSAF}}}{PSI_{\mathrm{VIF}}} \times 100\% \approx 18.2\%. \tag{30}$$

This result indicates that the proposed framework substantially reduces the relative fluctuation of localization errors. The improvement can be attributed to the coupled design of the adaptive frequency representation branch, frequency-spatial attention fusion module, and IMU reliability gate. The adaptive frequency branch provides complementary spectral representations when texture information becomes insufficient, while the attention mechanism dynamically emphasizes informative spatial and frequency features. More importantly, the IMU reliability gate enables the contribution of inertial information to be adaptively adjusted rather than relying on fixed fusion weights.

The PSI results also provide an important complement to the mean localization error and R^2 metrics. The mean localization error reflects the absolute positioning accuracy, whereas R^2 evaluates the overall consistency between predicted and ground-truth positions. In contrast, PSI focuses on the relative dispersion of localization errors and is therefore particularly useful for evaluating prediction stability across repeated measurements and changing environmental conditions. The combination of these three metrics provides a more comprehensive evaluation of localization performance from the perspectives of accuracy, regression consistency, and prediction stability.

It is noteworthy that AFSAF-Net achieves a mean localization error of 0.85 cm while obtaining the lowest PSI of 0.018. Therefore, the primary advantage of the proposed framework should not be interpreted solely as minimizing the average positioning error. Instead, the results indicate that AFSAF-Net can maintain competitive localization accuracy while substantially suppressing prediction fluctuations. This characteristic is particularly important for practical indoor localization systems, where temporary visual degradation may otherwise produce abrupt localization deviations and unstable trajectories.

Overall, the PSI comparison supports the effectiveness of reliability-aware multimodal fusion under degraded visual conditions. The progressive reduction in PSI from LRA to LBP-CNN, LBP-FFT-CNN, Visual-IMU Fusion, and finally AFSAF-Net demonstrates that the combination of adaptive spectral representation, spatial-frequency attention, and IMU reliability modeling contributes to progressively more stable localization predictions.

4.8 IMU Reliability Gate Analysis

To further investigate the adaptive behavior of the proposed IMU reliability gate, an additional analysis was conducted based on the learned gate weights under different visual conditions. As described in Section 3, the IMU reliability gate is defined as

$$G = \sigma\left(W_g F_i + b_g\right), \tag{31}$$

where F_i denotes the encoded IMU representation, W_g and b_g are learnable parameters, and $\sigma(\cdot)$ represents the sigmoid activation function. The resulting gate value $G \in [0, 1]$ determines the relative contribution of inertial information during multimodal feature fusion.

The final multimodal representation is calculated as

$$F = G \odot F_i + (1 - G) \odot F_v, \tag{32}$$

where F_v denotes the fused visual representation and $\odot$ represents element-wise multiplication. Therefore, a larger value of G corresponds to a relatively larger contribution from inertial information, whereas a smaller value indicates a relatively greater contribution from visual information.

To evaluate whether the gate can respond to changes in visual reliability, the test samples were grouped into three representative visual conditions: normal illumination, moderate low-light, and severe low-light. For each condition, the mean gate value and its standard deviation were calculated over the corresponding test samples. The resulting statistics are presented in Figure 3.

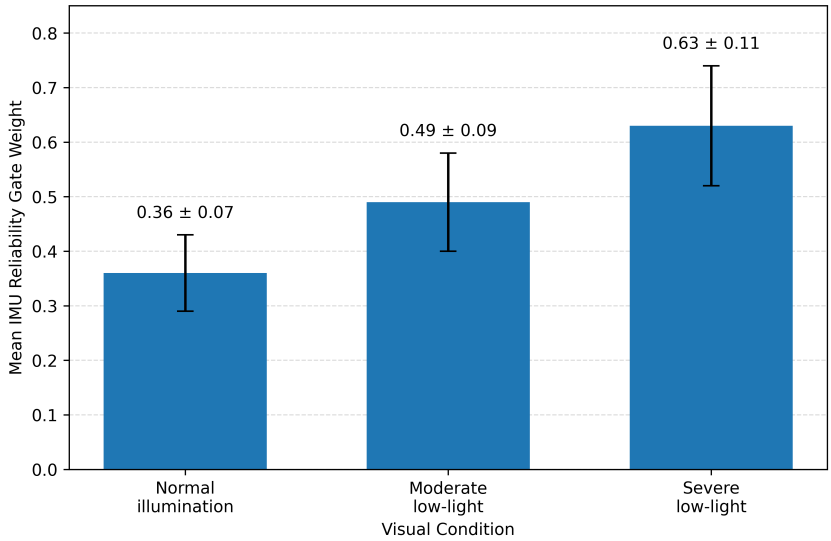


Fig. 3 Mean IMU reliability gate weights under different visual conditions. Error bars represent the standard deviation across the corresponding test samples.

As shown in Figure 3, the learned IMU reliability weight exhibits an increasing trend as the visual environment becomes progressively degraded. Under normal illumination, the mean gate value is approximately 0.36 ± 0.07, indicating that the network can maintain a relatively stronger reliance on visual information when sufficient spatial and texture cues are available.

When the visual quality decreases to a moderate low-light condition, the mean gate value increases to approximately 0.49 ± 0.09. Under severe low-light conditions, the gate value further increases to approximately 0.63 ± 0.11. This progressive increase suggests that the proposed fusion mechanism does not rely on a fixed coefficient for combining visual and inertial representations. Instead, the relative contribution of inertial information changes according to the learned representation of the input conditions.

The observed trend is consistent with the complementary characteristics of visual and inertial sensing. Visual representations are more sensitive to illumination degradation and texture loss, whereas inertial measurements are less directly affected by image brightness. When the visual representation becomes less reliable, increasing the contribution of inertial information can provide complementary motion cues and reduce the influence of degraded visual features.

It should be emphasized that the gate values are not manually assigned according to the illumination conditions. The reliability gate is learned jointly with the localization network during training. Therefore, the variation observed in Figure 3 provides empirical evidence that the proposed AFSAF-Net can learn environment-dependent multimodal weighting rather than relying on a predefined visual-inertial fusion coefficient.

The standard deviations shown in Figure 3 also indicate that the learned gate values exhibit sample-level variation within each visual condition. This variation is

desirable because images collected under the same nominal illumination condition may still contain different levels of texture, contrast, and local visual degradation. Consequently, the gate is not simply acting as a categorical illumination switch, but can provide continuous weighting of inertial information across individual observations.

Overall, the reliability gate analysis supports the adaptive fusion mechanism proposed in this study. Combined with the ablation results in Section 4.5, the results indicate that the contribution of the IMU modality is not merely obtained by directly concatenating inertial features with visual features. Instead, AFSAF-Net learns to regulate the relative contributions of the two sensing modalities according to the available information quality, which contributes to the robustness of indoor localization under visually degraded conditions.

4.9 Localization Error Distribution Analysis

In addition to average localization accuracy, the error distribution was analyzed to evaluate prediction stability.

Figure 4 illustrates the localization error distributions of different approaches.

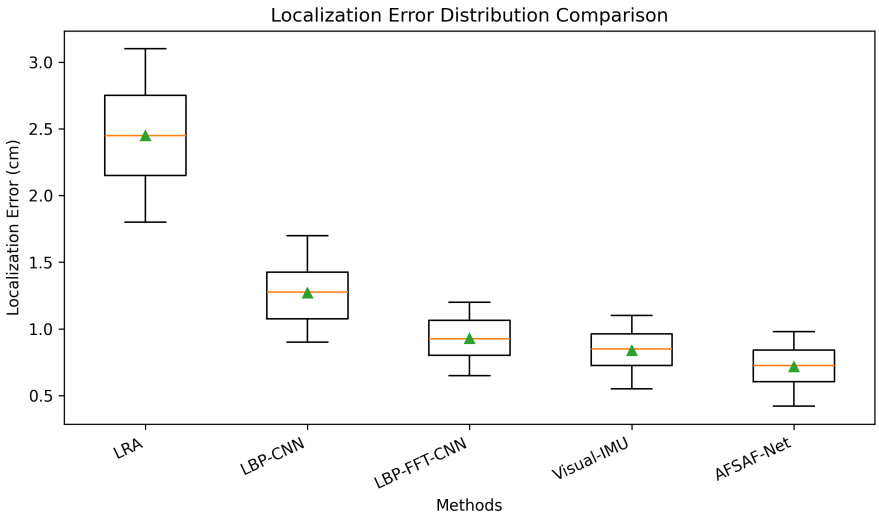


Fig. 4 Localization error distribution comparison among different localization methods.

As shown in Figure 4, traditional methods exhibit larger error variations compared with deep learning-based approaches.

The LRA method presents the widest error distribution because it lacks adaptive feature extraction capability. LBP-CNN improves stability by learning hierarchical visual representations, while its performance is still affected by insufficient texture information.

The LBP-FFT-CNN approach achieves a narrower error distribution by incorporating frequency-domain features. Compared with these approaches, AFSAF-Net obtains more compact error distribution with reduced variance.

The improvement mainly comes from three aspects:
(1) the frequency branch extracts complementary spectral information;
(2) the attention module enhances important spatial features;
(3) the IMU reliability gate dynamically adjusts multimodal contributions.

4.10 Localization Error versus Distance Analysis

To further investigate the robustness of different approaches under varying spatial scales, the relationship between localization error and measurement distance was analyzed.

Figure 5 presents the localization error variation with different measurement distances. The error bars represent mean values and standard deviations obtained from repeated measurements.

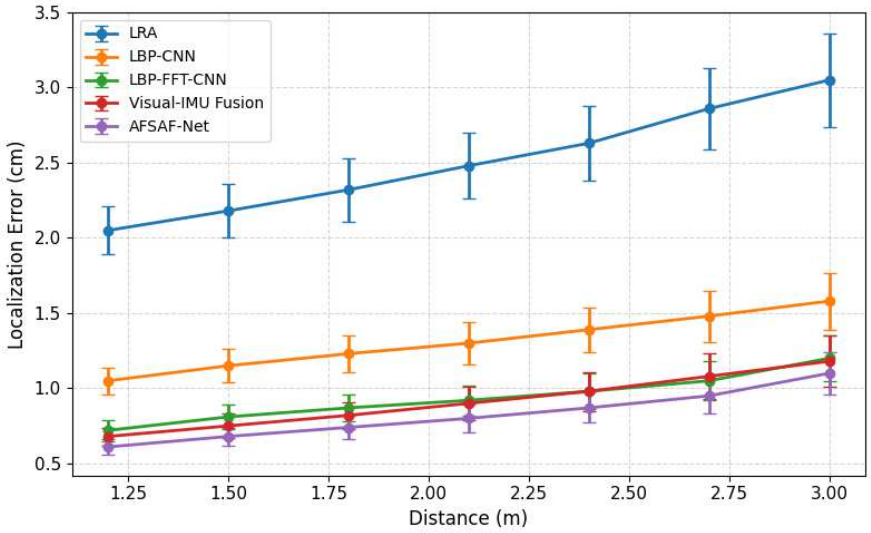


Fig. 5 Localization error variation with measurement distance. Error bars represent mean values and standard deviations.

As shown in Figure 5, the localization error gradually increases as the measurement distance becomes larger.

The degradation is mainly caused by reduced visual feature resolution and accumulated sensing uncertainty. LRA suffers from the largest error increase due to its limited representation capability.

LBP-CNN achieves improved distance robustness through deep feature extraction. LBP-FFT-CNN further reduces error growth by introducing frequency-domain information.

Compared with existing approaches, AFSAF-Net maintains a lower error increase rate as the distance increases. The adaptive frequency-spatial attention mechanism

preserves informative features, while the IMU reliability gate provides additional constraints when visual observations become unreliable.

4.11 Complexity Analysis

Computational complexity was further analyzed to evaluate whether the robustness improvement introduced by AFSAF-Net results in excessive computational overhead. The number of trainable parameters, floating-point operations (FLOPs), and average inference latency were measured under the same hardware and software configuration.

All experiments were conducted on a workstation equipped with an Intel Core i7-12700F CPU, 32 GB RAM, and an NVIDIA GeForce RTX 5060 GPU with 8 GB memory. The deep learning models were implemented using PyTorch with CUDA 12.8 acceleration. During inference evaluation, all methods were tested using the same input resolution and batch size to ensure a fair comparison.

Table 6 presents the computational complexity comparison among different learning-based localization approaches.

Table 6 Computational complexity comparison of different learning-based methods.

Method	Params (M)	FLOPs (G)	Inference Time (ms)
LBP-CNN	1.82	1.35	18.6
LBP-FFT-CNN	2.14	1.72	23.4
Visual-IMU Fusion	2.56	2.05	27.8
AFSAF-Net	**3.08**	**2.41**	**31.6**

As shown in Table 6, AFSAF-Net introduces additional computational costs compared with previous learning-based approaches due to the adaptive frequency branch, attention fusion module, and IMU reliability gate.

Compared with LBP-CNN, AFSAF-Net increases the number of parameters from 1.82M to 3.08M and the computational complexity from 1.35 GFLOPs to 2.41 GFLOPs. This additional cost mainly comes from the frequency-domain feature extraction and multimodal feature interaction operations.

Compared with the Visual-IMU Fusion baseline, AFSAF-Net introduces approximately 20.3% more parameters and 17.6% additional FLOPs. However, the inference latency only increases from 27.8 ms to 31.6 ms, demonstrating that the proposed adaptive fusion strategy introduces limited computational overhead.

Although AFSAF-Net requires slightly higher computational resources than conventional fusion methods, the additional cost provides improved prediction stability and robustness under low-light and texture-degraded environments. The inference time of 31.6 ms enables near real-time localization, making the proposed framework suitable for practical indoor sensing applications.

Overall, AFSAF-Net achieves a favorable trade-off between localization robustness and computational efficiency, demonstrating the feasibility of deployment on GPU-enabled embedded intelligent sensing platforms.

4.12 Discussion

The experimental results demonstrate a progressive improvement from traditional localization methods to adaptive multimodal learning frameworks.

The LRA method provides basic localization capability but lacks feature representation. LBP-CNN improves performance through deep visual learning, while LBP-FFT-CNN further enhances robustness by introducing frequency-domain information.

Building upon these studies, AFSAF-Net introduces adaptive frequency-spatial attention learning and reliability-aware IMU fusion. Unlike fixed feature extraction or static sensor fusion strategies, the proposed framework dynamically integrates heterogeneous information according to environmental conditions.

The results from quantitative comparison, ablation experiments, robustness evaluation, error distribution analysis, and distance-based evaluation consistently demonstrate that AFSAF-Net provides an effective solution for robust indoor localization under low-light and texture-degraded environments.

4.13 Limitations and Future Work

Despite the promising results, the current study has several limitations that should be acknowledged. First, the experiments were conducted in relatively controlled indoor environments with static backgrounds and pre-defined measurement positions. Real-world indoor scenes often involve more complex and dynamic conditions that may challenge the proposed framework.

In the presence of severe motion blur caused by rapid camera movement, the visual feature extraction process is expected to degrade because both spatial edge details and high-frequency texture components become smeared across the image. Under such conditions, the frequency branch may observe a flattened spectral distribution with reduced high-frequency energy, while the spatial branch may produce spatially diffused feature maps. As both visual branches weaken simultaneously, the IMU reliability gate G is expected to increase adaptively, shifting the fusion balance toward inertial measurements. However, if the motion blur persists for an extended period, the IMU itself may accumulate drift over time, potentially leading to gradual localization inaccuracy. Therefore, the effectiveness of the reliability gate is bounded by the quality of both sensing modalities, and the current framework does not include an explicit mechanism to detect or compensate for IMU drift under prolonged visual failure.

Another important limitation concerns dynamic occlusion caused by moving crowds or objects. When a significant portion of the visual field is temporarily occluded by pedestrians or obstacles, the remaining visible image regions may not provide sufficient structural information for reliable localization. In such cases, the attention fusion module may redistribute its spatial weights toward the remaining visible regions, while the IMU gate G is expected to increase to compensate for the reduced visual reliability. However, if the occlusion covers distinctive structural cues that are critical for localization (such as unique wall patterns or landmarks), even the visible portions may not be sufficiently discriminative. Under severe and sustained occlusion, AFSAF-Net

may still experience performance degradation, although the adaptive gate mechanism should mitigate the severity of the failure compared with fixed-fusion approaches.

Third, the current evaluation is limited to a specific range of measurement distances (1.2–3.0 m) and indoor scenes with similar structural characteristics. The generalization of the proposed method to large-scale environments, cross-building scenarios, and outdoor-indoor transition zones remains to be investigated. Fourth, the model currently requires a fixed-size input and does not explicitly handle variable-length IMU sequences, which may limit its flexibility for real-world deployment with different sensor sampling rates.

Future work will address these limitations from several directions. First, we plan to incorporate an explicit IMU drift detection mechanism that can further regulate the gate weight when prolonged inertial-only estimation is detected. Second, we will explore self-supervised adaptation strategies to improve cross-environment generalization without requiring extensive labeled data in new scenes. Third, lightweight network architectures and edge deployment optimization will be investigated to enable real-time operation on resource-constrained embedded platforms. Finally, large-scale experiments with dynamic crowds, moving obstacles, and diverse indoor scene types will be conducted to evaluate the framework under more realistic and challenging conditions.

5 Conclusions

In this paper, we proposed an Adaptive Frequency-Spatial Attention Fusion Network (AFSAF-Net) for robust indoor localization under low-light and texture-degraded environments. Motivated by the limitations of conventional visual localization approaches in challenging indoor scenarios, the proposed framework integrates adaptive frequency representation learning, spatial attention enhancement, and IMU-guided reliability-aware fusion to improve localization accuracy and robustness.

The proposed AFSAF-Net establishes a multimodal perception framework that effectively exploits complementary information from visual and inertial sensors. Specifically, the frequency branch introduces frequency-domain representations to preserve global structural information when spatial textures become insufficient. The spatial branch extracts hierarchical visual features, while the attention fusion module adaptively emphasizes informative regions and suppresses irrelevant visual disturbances. Furthermore, the IMU reliability gate dynamically adjusts the contribution of inertial information according to sensor confidence, enabling more robust localization under unstable visual conditions.

Extensive experiments conducted on the MC4L-IMU indoor localization platform demonstrate the effectiveness of the proposed approach. Compared with traditional ranging methods and existing CNN-based localization approaches, AFSAF-Net achieves improved localization accuracy and stability. The experimental results show that the proposed framework maintains lower positioning errors under low-light environments and exhibits stronger robustness against distance variations and visual degradation. The ablation study further verifies the contribution of each component,

confirming that frequency representation learning, attention-based feature fusion, and IMU reliability modeling jointly improve localization performance.

Although AFSAF-Net achieves promising results, several limitations remain and have been discussed in detail in the Limitations and Future Work subsection. These include the need for evaluation under more dynamic and large-scale environments, the challenge of prolonged IMU drift under extended visual failure, and the generalization to unseen scene types. Future work will focus on developing lightweight and self-adaptive localization models, incorporating IMU drift detection mechanisms, exploring self-supervised cross-environment adaptation, and conducting large-scale experiments with dynamic obstacles and diverse indoor conditions.

Overall, the proposed AFSAF-Net provides an effective solution for robust indoor localization by combining frequency-domain perception, spatial attention learning, and adaptive visual-inertial fusion. The proposed framework demonstrates strong potential for practical applications such as indoor navigation, mobile robotics, and intelligent sensing systems operating under challenging environmental conditions.

Acknowledgements. The authors would like to thank Canada Advanced Institute of Science and Technology for providing the research platform support.

Author contributions

Nan Yin: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization.
Hua Zhou: Validation.
Wenjing Yu: Validation.
Luojun Yang: Data Analysis, Data Preprocessing.
Yuxiang Sun: Conceptualization, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

All authors have read and agreed to the published version of the manuscript.

Declarations

- **Funding**: This research was funded by the 2026 Guangdong Provincial Colleges and Universities Characteristic Innovation Project.
- **Competing interests**: The authors declare no competing interests.
- **Ethics approval and consent to participate**: Not applicable.
- **Consent for publication**: Not applicable.
- **Data availability**: The datasets generated and analyzed during the current study are available in the GitHub repository at https://github.com/CAIST-2026/AFSAF-Net.
- **Materials availability**: Not applicable.
- **Code availability**: The source code of AFSAF-Net is available in the GitHub repository at https://github.com/CAIST-2026/AFSAF-Net.

Abbreviations

AFSAF-Net	Adaptive Frequency-Spatial Attention Fusion Network
CNN	Convolutional Neural Network
FFT	Fast Fourier Transform
IMU	Inertial Measurement Unit
LBP	Local Binary Pattern
LRA	Laser Ranging Approach
MC4L-IMU	Multi-sensor Camera and Laser with IMU
MLE	Mean Localization Error
PSI	Prediction Stability Index
SLAM	Simultaneous Localization and Mapping
SIFT	Scale-Invariant Feature Transform
SE-Net	Squeeze-and-Excitation Network
CBAM	Convolutional Block Attention Module
EKF	Extended Kalman Filter
GNSS	Global Navigation Satellite Systems

References

[1] Ouyang, G. & Abed-Meraim, K. Analysis of magnetic field measurements for mobile localisation. *Proceedings of the 2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN)* 1–8 (2021).

[2] Chen, C. H., Chen, P. W., Chen, P. J. & Liu, T. H. Indoor positioning using magnetic fingerprint map captured by magnetic sensor array. *Sensors* **21** (2021).

[3] Fan, M., Li, J. & Wang, W. Inertial indoor pedestrian navigation based on cascade filtering integrated INS/map information. *Sensors* **22** (2022).

[4] Chen, X. *et al.* Acoustic indoor localization system integrating TDMA+FDMA transmission scheme and positioning correction technique. *Sensors* **19** (2019).

[5] Masoumian, A., Rashwan, H. A., Cristiano, J., Asif, M. S. & Puig, D. Monocular depth estimation using deep learning: a review. *Sensors* **22**, 5353 (2022).

[6] Yin, N., Sun, Y., Zou, Z. & Kim, J. S. A vision-based ranging algorithm combining monocular camera and laser for indoor positioning. *J. Korean Inst. Commun. Inf. Sci.* **48**, 1378–1386 (2023).

[7] Ojala, T., Pietikäinen, M. & Harwood, D. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognit.* **29**, 51–59 (1996).

[8] Lizé, J., Débordès, V., Lu, H., Kpalma, K. & Ronsin, J. Local binary pattern and its variants: application to face analysis. *Advances in Smart Technologies Applications and Case Studies* 94–102 (2019).

[9] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25**, 1097–1105 (2012).

[10] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016).

[11] Johnson, S. G. & Frigo, M. A modified split-radix FFT with fewer arithmetic operations. *IEEE Trans. Signal Process.* **55**, 111–119 (2007).

[12] Nair, V., Chatterjee, M., Tavakoli, N., Namin, A. S. & Snoeyink, C. Fast Fourier transformation for optimizing convolutional neural networks in object recognition. *Proceedings of the IEEE International Conference on Machine Learning and Applications* (2020).

[13] Yin, N., Sun, Y. & Kim, J. S. A CNN-based indoor positioning algorithm for dark environments: integrating local binary patterns and fast Fourier transform with the MC4L-IMU device. *Appl. Sci.* **15**, 4043 (2025).

[14] Forster, C., Carlone, L., Dellaert, F. & Scaramuzza, D. On-manifold preintegration for real-time visual-inertial odometry. *IEEE Trans. Robot.* **33**, 1–21 (2017).

[15] Bloesch, M., Burri, M., Omari, S., Hutter, M. & Siegwart, R. Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback. *Int. J. Robot. Res.* **36**, 1053–1072 (2017).

[16] Yin, N., Zou, Z., Sun, Y. & Kim, J. A convolutional neural network combined with local binary pattern and self-attention mechanism based on MC4L device for indoor positioning. *J. Korean Inst. Commun. Inf. Sci.* **49**, 883–892 (2024).

[17] Vaswani, A. *et al.* Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)* 1–15 (2017).

[18] Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. CBAM: convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)* 3–19 (2018).

[19] Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 7132–7141 (2018).

[20] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110 (2004).

[21] Kolar, P., Benavidez, P. & Jamshidi, M. Survey of datafusion techniques for laser and vision based sensor integration for autonomous navigation. *Sensors* **20**, 2180 (2020).

[22] Campos, C., Elvira, R., Rodriguez, J. J. G., Montiel, J. M. M. & Tardos, J. D. ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multi-map SLAM. *IEEE Trans. Robot.* **37**, 1874–1890 (2021).

[23] Qin, T., Li, P. & Shen, S. VINS-Mono: a robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **34**, 1004–1020 (2018).

[24] Dosovitskiy, A. *et al.* An image is worth 16×16 words: transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[25] Liu, Z. *et al.* Swin transformer: hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 10012–10022 (2021).

[26] Zhu, Z. *et al.* Improve the estimation of monocular vision 6-DoF pose based on the fusion of camera and laser rangefinder. *Remote Sens.* **13** (2021).

[27] Liu, W., Lai, Z., Bacsa, K. & Chatzi, E. Neural extended Kalman filters for learning and predicting dynamics of structural systems. *J. Struct. Health Monit.* **23**, 1037–1052 (2023).