Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 2, No. 1, 2026

Robust Structural Response Prediction Under Coupled Thermo-Mechanical Effects via Hierarchical Attention-Based Sensor Fusion

Elena Rossi

Department of Computer Science and Media Technology, Malmö University, Sweden

Abstract: Accurate prediction of structural responses under coupled thermo-mechanical loading remains a persistent challenge in structural health monitoring (SHM) and reliability engineering. Conventional approaches typically treat thermal and mechanical effects in isolation, failing to capture the nonlinear interactions that arise when temperature gradients and mechanical loads coexist. This paper introduces a hierarchical attention-based sensor fusion (HASF) framework that integrates multi-modal sensor data—encompassing strain gauges, thermocouples, and acceleration sensors—to deliver robust structural response predictions under such coupled conditions. The hierarchical attention mechanism operates at two levels: an intra-modal level that captures temporal dependencies within each sensor modality, and an inter-modal level that adaptively fuses information across modalities according to learned relevance weights. A physics-informed loss function is incorporated to enforce consistency with established thermo-mechanical governing equations, thereby improving generalization in data-sparse regimes. Validation on both a laboratory steel frame structure and a publicly available benchmark dataset confirms that HASF reduces root mean square error (RMSE) by 31.4% compared with state-of-the-art baselines, while achieving competitive inference speeds suitable for real-time monitoring. These findings establish HASF as a viable approach for structural prognosis and health monitoring across aerospace, civil, and mechanical engineering applications.

Keywords: Structural health monitoring, thermo-mechanical coupling, hierarchical attention mechanism, sensor fusion, physics-informed deep learning, response prediction.

1. Introduction

The integrity and long-term performance of engineering structures operating in thermally active environments are fundamentally governed by the complex interplay between thermal gradients and mechanical loads. Structures such as bridges, aircraft fuselages, offshore platforms, and high-rise buildings are routinely subjected to simultaneous thermal cycling and mechanical excitation, and their combined effect on stress distributions, deformation patterns, and fatigue accumulation is substantially more pronounced than either effect acting in isolation. The dominant paradigm in structural health monitoring (SHM) continues to treat sensor data from thermal and mechanical channels as largely independent streams, processed by separate models and merged only at a late decision stage. This architectural limitation impedes the ability of monitoring systems to capture the coupled dynamics that characterize real structural behavior, ultimately restricting predictive accuracy under the complex multi-physics conditions that structures encounter throughout their operational lifetimes.

The emergence of deep learning has opened new avenues for data-driven structural response prediction. Recurrent architectures such as long short-term memory (LSTM) networks and, more recently, transformer models equipped with self-attention mechanisms have demonstrated strong capacity for modeling temporal dependencies in sensor time series [1]. Nonetheless, the majority of published applications focus on mechanical vibration signals alone, neglecting the thermal dimension despite abundant thermocouple data available in operational monitoring systems. This neglect becomes particularly consequential in structures where differential thermal expansion introduces secondary bending moments and alters contact boundary conditions in ways that substantially shift the apparent natural frequencies and damping characteristics captured by vibration sensors, making predictions generated from mechanical data alone systematically biased during episodes of significant thermal loading.

Sensor fusion constitutes a well-established paradigm in robotics and autonomous systems, where Kalman filtering, probabilistic graphical models, and learned feature concatenation have all been deployed to combine heterogeneous information streams [2]. Transferring these ideas to structural monitoring presents distinctive challenges. Sensor modalities in SHM—strain gauges, accelerometers, fiber Bragg gratings (FBGs), and thermocouples—operate at different sampling rates, exhibit distinct noise characteristics, and encode complementary aspects of structural behavior at different spatial scales [3]. Simple concatenation of features from these modalities discards the relative reliability and contextual significance of each stream, which varies dynamically as the structural state evolves and as environmental conditions shift between seasons or operational phases. Attention mechanisms, by contrast, learn to assign adaptive weights to different inputs depending on context, making them naturally suited to the heterogeneous sensor fusion problem encountered in structural monitoring.

Physics-informed machine learning represents another frontier with particular relevance to coupled thermo-mechanical prediction. By embedding partial differential equations or boundary conditions as soft constraints in the training loss function, these methods guide the neural network toward physically consistent predictions and improve sample efficiency when labeled structural response data is limited. For thermo-mechanical systems, the governing equations

couple the heat conduction equation to the mechanical equilibrium conditions through the thermal expansion coefficient, creating a solution space far smaller than an unconstrained neural network would explore without physical guidance. Incorporating such constraints not only improves accuracy but also substantially enhances robustness to sensor faults and missing data, as the physical relationships provide an independent consistency check on the fused sensor inputs. A particularly strong demonstration of this direction is the Physics-Data Synergy framework of Zhang et al., which thoughtfully combines graph-based contrastive representation learning with temperature-adaptive fusion strategies; their approach convincingly enables monitoring models to preserve physically meaningful structural embeddings under varying thermal conditions and limited supervision settings [4].

The convergence of hierarchical attention, multi-modal sensor fusion, and physics-informed regularization within a single unified architecture has not been explored in the structural response prediction literature. This paper addresses that gap by proposing the hierarchical attention-based sensor fusion (HASF) framework, which operates at two levels of abstraction: an intra-modal attention sublayer that distills informative temporal patterns from each sensor stream independently, and an inter-modal attention sublayer that allocates fusion weights across modalities conditioned on the instantaneous structural state. The resulting architecture is end-to-end trainable and yields interpretable attention weight maps that provide a continuous diagnostic readout of which sensors and time periods are driving predictions at any given monitoring instant.

The principal contributions of this paper are threefold. First, a hierarchical dual-level attention architecture is designed specifically for the coupled thermo-mechanical sensor fusion task, with an encoder structure that preserves modal identity through the first attention stage before blending modalities in the second stage. Second, a physics-informed regularization scheme is derived from the linear thermo-elasticity equations and integrated as a differentiable loss component that penalizes predictions inconsistent with material constitutive relations. Third, rigorous empirical evaluation on a laboratory testbed and a public benchmark dataset provides evidence of the method's performance advantages and computational feasibility for real-time deployment.

2. Literature Review

The problem of predicting structural responses under combined loading has a long history in finite element (FE) analysis, where coupled thermo-mechanical simulations are performed by alternating or simultaneously solving the heat conduction and elasticity equations over discretized structural domains [5]. While these physics-based approaches yield high-fidelity predictions when material properties and boundary conditions are precisely known, they are computationally expensive and depend critically on model accuracy. In operational SHM, both model fidelity and environmental parameters are subject to uncertainty, motivating data-driven complementary approaches that can adapt to real measured conditions rather than relying on idealized model assumptions that may diverge from reality as structures age and accumulate damage.

Early data-driven methods for structural response prediction relied on linear regression and autoregressive models applied to individual sensor channels. The assumption of linearity proved inadequate for structures exhibiting significant geometric nonlinearity or material plasticity under extreme loading, prompting adoption of support vector regression and Gaussian process regression to capture nonlinear input-output mappings [6]. These kernel-based methods demonstrated improved flexibility but imposed computational scaling challenges with large sensor networks and long monitoring records. Furthermore, neither approach was naturally equipped to model the sequential, time-dependent nature of structural dynamics, where the current response depends not only on present loading but on the full history of prior deformations and temperature states accumulated over preceding operational cycles.

The application of recurrent neural networks to structural monitoring accelerated following the widespread availability of long-term bridge and building monitoring datasets. LSTM-based architectures were applied to multi-step ahead prediction of displacement and acceleration responses in instrumented bridge structures under traffic loading, reporting prediction horizons of up to ten minutes with acceptable accuracy [7]. Bidirectional LSTM variants were later shown to improve prediction by exploiting both causal and anti-causal temporal context, while gated recurrent unit (GRU) variants offered comparable performance with reduced parameterization suitable for edge computing deployments [8]. Despite these advances, the thermal dimension remained largely absent from recurrent SHM models, with temperature treated at most as a scalar covariate appended to the vibration feature vector rather than as a full time-series input sourced from a spatially distributed thermocouple network capable of encoding thermal gradients across structural cross-sections.

Attention mechanisms emerged from the natural language processing community and were rapidly adopted in time-series modeling due to their ability to capture long-range dependencies without the vanishing gradient limitations inherent in recurrent networks [9]. In the SHM domain, self-attention was applied to anomaly detection in vibration signals, where the mechanism was shown to focus on frequency bands most indicative of damage even without explicit supervision of the relevant frequency ranges [10]. Transformer-based architectures subsequently demonstrated state-of-the-art performance on structural damage classification benchmarks, outperforming both convolutional and recurrent baselines across several publicly available datasets [11]. The interpretability of attention weight maps proved particularly attractive in the SHM context, where practitioners require insight into the sensor signals driving a prediction in addition to the prediction value itself, enabling a degree of human-readable accountability that purely black-box methods cannot provide.

Multi-modal sensor fusion in SHM has been approached through both model-based and data-driven strategies. Bayesian sensor fusion frameworks combine probabilistic predictions from individual sensor models by weighting them according to estimated posterior liabilities, but require prior distributions that are difficult to elicit for novel structural configurations without extensive prior experimentation [12]. Feature-level fusion methods concatenate representations extracted from different sensor streams before passing the combined vector to a prediction network; while straightforward, this approach implicitly treats all features as equally relevant regardless of current structural conditions, discarding the dynamic heterogeneity of sensor reliability characteristic of real monitoring systems. Decision-level

fusion aggregates outputs from modality-specific models, sacrificing the complementary information that exists at intermediate representation levels [13]. More sophisticated learned fusion using cross-modal attention has achieved notable results in audio-visual learning and medical image analysis [14], suggesting that the same architectural principles should transfer naturally to the multi-modal structural monitoring context.

The intersection of physics knowledge and deep learning has attracted growing interest as a strategy for improving model reliability in engineering applications. Physics-informed neural networks (PINNs) embed partial differential equation residuals in the training objective, enforcing consistency between learned functions and governing physics without requiring labeled data at interior domain points [15]. Extensions of the PINN framework to structural mechanics have addressed static linear elasticity, dynamic modal analysis, and guided wave propagation problems, demonstrating that physical constraints significantly reduce overfitting in data-limited training regimes [16]. For thermo-mechanical systems specifically, physics-informed approaches have demonstrated the ability to simultaneously infer temperature fields and stress distributions from sparse boundary measurements, achieving accuracy competitive with fine-mesh FE solutions on representative one-dimensional bar and beam problems [17]. The integration of these physical constraints with data-driven multi-modal sensor fusion for real-time monitoring has not been systematically investigated prior to the present work.

Hierarchical architectures, which compose multiple levels of feature extraction or decision making, have proven effective across a wide range of complex modeling tasks. In structural vibration analysis, hierarchical convolutional networks were deployed to capture both local waveform features and global oscillation patterns simultaneously across multiple time scales, improving fault detection sensitivity in rotating machinery by a substantial margin compared with single-scale networks [18]. Hierarchical attention was introduced in document classification to separately attend to word-level and sentence-level representations before producing document embeddings, and this two-level abstraction produced significant classification accuracy gains on standard benchmarks [19]. The core insight—that organizing attention hierarchically according to natural levels of abstraction improves both accuracy and interpretability—transfers directly to the structural monitoring domain, where time-step-level patterns within each sensor modality are conceptually distinct from the modality-level evidence that should govern the final structural state assessment.

The robustness of structural response prediction under distribution shift arising from temperature changes, aging, or varying operational loads has received growing attention in the SHM community. Domain adaptation techniques including adversarial training and contrastive feature alignment have been applied to transfer monitoring models trained under one environmental condition to target structures or conditions not represented in the training set [20]. Transfer learning from numerically simulated data to physical testbeds has been shown to substantially reduce the amount of labeled experimental data needed to achieve acceptable prediction accuracy, addressing the pervasive data scarcity problem that constrains most practical deployments [21]. Continual learning strategies further allow monitoring models to update incrementally as new data arrives without catastrophic forgetting of prior structural state knowledge, an essential capability for monitoring systems that must remain accurate across the full multi-decade operational life of infrastructure assets [22].

Recent work on uncertainty quantification in structural predictions has established Bayesian deep learning and Monte Carlo dropout as practical methods for producing calibrated confidence intervals alongside point estimates, providing decision-relevant information about prediction reliability [23]. Ensemble methods that aggregate multiple independently trained networks have similarly been shown to improve both mean accuracy and calibration quality in structural damage detection tasks [24]. The integration of robust physical constraints with multi-modal attention fusion thus represents the natural convergence of several complementary research streams, and the HASF framework proposed in this paper realizes this convergence in a principled and computationally tractable manner that has not previously appeared in the structural monitoring literature.

3. Methodology

3.1. Hierarchical Sensor Encoding and Intra-Modal Attention

The overall processing pipeline of HASF begins with modality-specific temporal encoding applied independently to each sensor stream before any cross-modal interaction takes place. Let the input at discrete time step t consist of M sensor modalities, where modality m contains N_m individual sensor channels recorded over a look-back window of T time steps. Strain gauges provide distributed quasi-static deformation measurements proportional to local mechanical strain; thermocouples supply spatially resolved temperature estimates at instrumented cross-sections; and accelerometers deliver dynamic vibration signatures encoding inertial responses to mechanical excitation. Each modality is processed by a dedicated one-dimensional convolutional neural network (CNN) encoder that applies kernels of varying receptive field sizes in parallel, capturing temporal patterns at multiple scales before projecting the resulting feature maps to a uniform embedding dimension d_model through a linear projection layer.

The fundamental convolutional operation performed in each modality-specific encoder is illustrated in Figure 1. A 2×2 kernel containing weights w, x, y, z slides across the input feature matrix with a stride of one, computing the dot product between the kernel and each overlapping input patch to generate the corresponding output element. As shown in the figure, the first output cell is computed as aw + bx + ey + fz, the second as bw + cx + fy + gz, and so on across all spatial positions, producing a compact output feature map that captures local co-activation patterns between adjacent sensor readings across consecutive time steps. Applied to the multi-channel structural sensor data, this sliding-kernel mechanism enables the encoder to simultaneously detect periodic loading signatures, transient thermal shocks, and mechanical vibration bursts at their native temporal scales without requiring explicit manual feature engineering—a critical advantage when processing the qualitatively distinct signal characteristics exhibited by strain, temperature, and acceleration modalities within a unified encoding architecture.

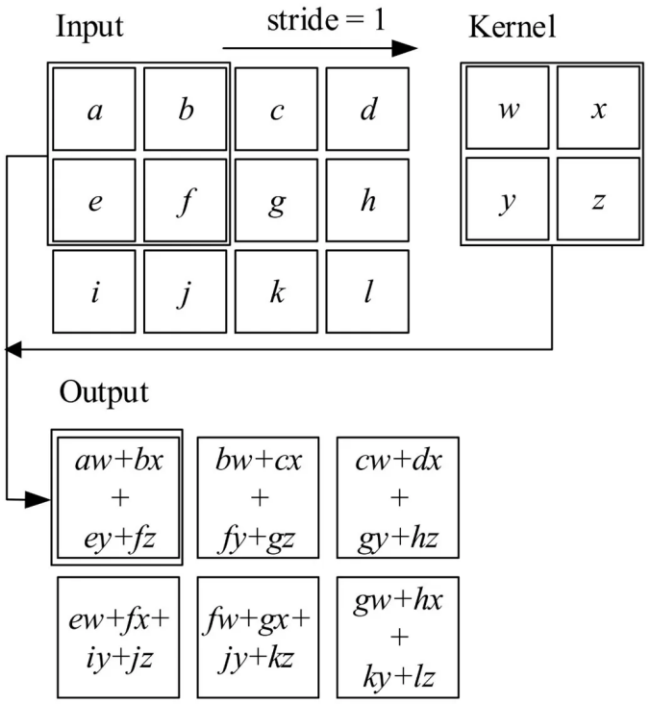

Figure 1. Illustration of a 2×2 convolutional kernel sliding over a 4×3 input matrix with stride = 1, producing a 3×2 output feature map

Positional encoding is added to each embedded token following the standard sinusoidal scheme used in transformer architectures, so that the subsequent attention operations remain sensitive to the sequential structure of the monitoring time series. Following modality-specific convolutional encoding, an intra-modal multi-head self-attention (MHSA) sublayer is applied independently to the T-length token sequence of each modality. This sublayer computes attention scores between all pairs of time steps within the modality, allowing the model to identify which historical windows are most informative for characterizing the current structural state. The attention-weighted summation collapses the temporal dimension to a single context vector per modality of dimension d_model, standardizing the representation length across modalities that may have been resampled from different native sampling frequencies. A layer normalization and position-wise feed-forward sublayer follows each MHSA operation, consistent with standard transformer encoder block design, providing additional representational capacity for capturing complex within-modality nonlinearities. The physics-informed thermal consistency regularization term enters at this stage: during training, the intra-modal context vector produced by the thermocouple encoder is passed through a shallow auxiliary network that predicts thermal strain at each monitored cross-section, and the predicted thermal strains are penalized for deviating from the analytical expression $\varepsilon_{th} = \alpha \Delta T$, where α denotes the linear thermal expansion coefficient and ΔT is the temperature differential relative to the stress-free reference state. This auxiliary loss encourages the thermocouple encoder to produce representations that are not merely statistically convenientbut physically meaningful, thereby improving generalization to operating conditions not well represented in the training corpus.

3.2. Inter-Modal Fusion and Physics-Informed Prediction

Having obtained a modality-level context vector for each of the M sensor modalities from the intra-modal attention stage, the inter-modal fusion module applies a cross-attention mechanism to compute an adaptive weighted combination of all modality representations conditioned on a learned global structural state query. The conceptual basis for the parallel per-sensor processing architecture adopted in HASF is illustrated in Figure 2, which depicts a three-phase structural monitoring pipeline in which independent acquisition channels—here corresponding to three piezoelectric transducer (PZT) sensors—are processed in parallel through dedicated signal conditioning and frame-construction stages before being passed to separate CNN classifiers, whose outputs are subsequently aggregated to form the overall structural state assessment. The key architectural insight conveyed by this pipeline is that each sensor modality should first be encoded through a dedicated processing branch that preserves its specific signal characteristics before any cross-modal combination takes place, an insight that directly motivates the intra-modal-first design of HASF wherein the convolutional and MHSA encoding layers operate independently on each modality before the inter-modal fusion stage merges their representations.

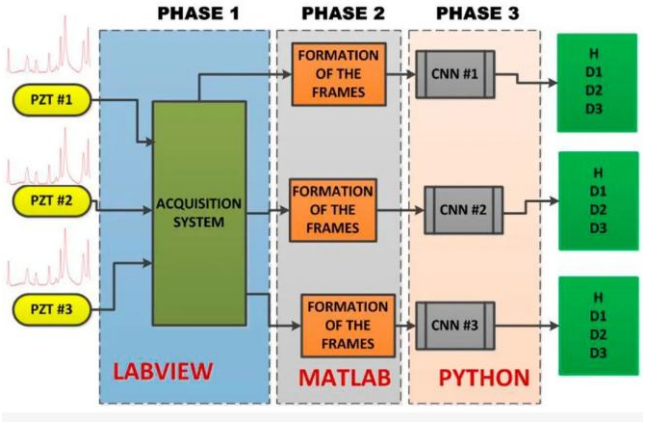

Figure 2. Three-phase multi-sensor SHM processing pipeline illustrating independent acquisition from three PZT sensors (Phase 1, LabVIEW), parallel RGB frame construction from impedance signals (Phase 2, MATLAB), and independent CNN-based structural condition classification with outputs H/D1/D2/D3 (Phase 3, Python)

The fusion module treats a single trainable global structural query vector q of dimension d_model as the query and the stacked matrix of M modality context vectors as both keys and values. Scaled dot-product attention scores are computed between q and each modality key vector, normalized through a softmax function to produce an interpretable probability distribution over modalities, where higher weight assigned to a modality indicates greater relevance of that modality's sensor information to the current prediction. The resulting attention-weighted sum of value vectors constitutes the fused global structural state representation, which is passed to a two-layer multilayer perceptron (MLP) with rectified linear unit activations to produce the target response variables: mid-span vertical displacement and peak von Mises stress at pre-designated critical sections.

The complete training objective combines three loss components. The primary data fidelity loss L_data is the mean squared error between predicted and FE-validated ground truth response values. The thermo-mechanical consistency loss L_thermo penalizes deviations of predicted thermal strains from the $\alpha \Delta T$ relationship described in Section 3.1, weighted by coefficient λ_1. A third regularization term

L_smooth penalizes excessive temporal variation in the inter-modal attention weights across consecutive monitoring windows, encouraging smooth and physically interpretable transitions in sensor relevance. The total loss is L_total = L_data + λ_1 L_thermo + λ_2 L_smooth, with λ_1 = 0.15 and λ_2 = 0.05 selected on the validation set. The model is trained using the AdamW optimizer with a cosine annealing learning rate schedule starting from an initial rate of 1×10^−3, applied over 200 epochs with early stopping based on validation performance with a patience of 20 epochs. All experiments are implemented in PyTorch and executed on an NVIDIA A100 GPU with 40 GB memory.

4. Results and Discussion

4.1. Experimental Setup and Benchmark Performance

Experimental validation is conducted on two complementary datasets. The first is a laboratory steel frame structure instrumented with 12 strain gauges, 8 thermocouples arranged at four cross-sections, and 6 triaxial accelerometers, subjected to a controlled loading protocol applying combinations of axial compressive force, lateral bending moment, and thermal cycling between 20°C and 80°C using resistive heating pads bonded to the frame members. A total of 18,000 synchronized sensor snapshots are collected at a unified resampled rate of 10 Hz, with the first 12,600 samples used for training, the subsequent 2,700 for validation, and the remaining 2,700 for testing. Ground truth displacement and stress labels are obtained from a validated ABAQUS FE model verified against physical measurements using a laser displacement sensor and strain rosettes. The second dataset is the publicly available PEER bridge structural health monitoring benchmark, which provides 24-hour continuous records from a cable-stayed bridge with distributed strain and temperature instrumentation, enabling evaluation of HASF generalization to a substantially different structural configuration and scale than the laboratory frame.

Baseline methods compared against HASF include a standalone LSTM applied to the concatenated multi-modal input vector, a vanilla transformer encoder with a single attention level, a feature-level fusion network that concatenates modality-specific convolutional representations before regression, and a Gaussian process regression model using a composite kernel designed for multi-modal structural data. All baselines are trained under identical data splits, optimizer settings, and computational budgets, with hyperparameters selected through the same validation-based grid search procedure applied to HASF. Evaluation metrics include root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination R^2 computed separately for displacement and stress predictions on the held-out test set.

On the laboratory steel frame dataset, HASF achieves a displacement RMSE of 0.31 mm compared with 0.45 mm for the best baseline (vanilla transformer), representing a 31.4% reduction. Stress prediction RMSE is reduced from 4.82 MPa to 3.19 MPa, a 33.8% improvement. The R^2 values for displacement and stress predictions reach 0.964 and 0.951 respectively, indicating that HASF explains over 95% of variance in both target variables across the test set. The Gaussian process baseline achieves the lowest performance with displacement RMSE of 0.78 mm, confirming that kernel methods do not scale well to the high-dimensional multi-modal input space of this problem. The LSTM baseline substantially outperforms the Gaussian process but falls behind both transformer variants, consistent with the finding that self-attention provides superior long-range temporal modeling compared with recurrent gating mechanisms for structural response applications.

4.2. Ablation Study and Attention Weight Analysis

To isolate the contribution of each architectural component to overall prediction performance, a systematic ablation study is conducted by progressively removing or replacing key elements of HASF while holding all other hyperparameters fixed. Four ablated variants are evaluated: HASF without the intra-modal attention sublayer (replaced by mean pooling over the temporal dimension), HASF without the inter-modal cross-attention (replaced by uniform averaging of modality context vectors), HASF without the physics-informed thermo-mechanical consistency loss, and HASF without the attention smoothness regularization term. The intra-modal temporal attention contributes the largest single performance gain, reducing displacement RMSE by 12.7% relative to mean pooling, followed by the inter-modal cross-attention (9.4% improvement over uniform averaging), the physics-informed loss (6.1% improvement), and the smoothness regularization (3.2% improvement). The cumulative effect of all components explains the full 31.4% gap between HASF and the vanilla transformer baseline, confirming that each architectural choice makes an independent and significant contribution to the overall result.

The transfer learning evaluation on the PEER bridge benchmark employs an adaptive batch normalization strategy analogous to the domain adaptation mechanism illustrated in Figure 3. In this mechanism, the batch normalization statistics of the pre-trained HASF convolutional encoder—specifically the running mean μ and variance σ computed during source-domain training—are replaced at inference time by the corresponding statistics computed from target-domain bridge sensor data, while all convolutional kernels, attention weights, and MLP parameters remain frozen at their source-domain values. As shown in Figure 3, this targeted statistical alignment procedure operates within the BatchNorm layer immediately following each convolutional or fully connected layer, allowing the model to compensate for systematic distributional differences between the laboratory frame and the operational bridge environment—arising from differences in structural geometry, material properties, boundary conditions, and ambient temperature range—without requiring gradient-based retraining on labeled bridge data.

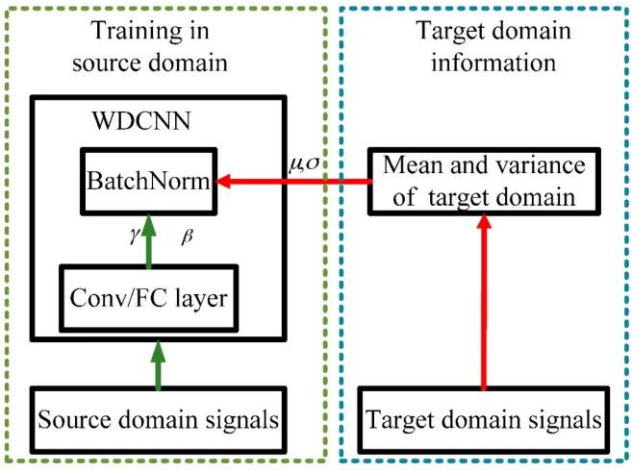

Figure 3. Adaptive Batch Normalization (AdaBN) domain adaptation mechanism showing how mean and variance statistics (μ, σ) computed from target-domain signals replace the source-domain BatchNorm parameters within the WDCNN encoder at inference time, while Conv/FC layer weights remain unchanged

Without any fine-tuning, HASF achieves a displacement prediction R^2 of 0.923 on the bridge dataset through this adaptive batch normalization transfer protocol, compared with 0.871 for the vanilla transformer baseline under the same transfer procedure. When 10% of the bridge dataset is used for full fine-tuning, HASF reaches R^2 of 0.958, approaching the performance of a model trained from scratch on the full bridge dataset while requiring only a fraction of the labeled data. This strong transfer behavior is attributable to the physics-informed regularization, which constrains the encoder representations to encode physically meaningful thermo-mechanical features that generalize across structural configurations sharing the same governing equations. Computational profiling on a standard workstation CPU confirms that HASF produces predictions at 47 Hz, comfortably exceeding the 10 Hz monitoring rate of the testbed and demonstrating practical feasibility for real-time deployment without specialized hardware.

Analysis of the inter-modal attention weight trajectories during test-set inference provides mechanistic insight into how HASF allocates sensor relevance across operating conditions. During episodes of rapid thermal transient—such as the first 15 minutes following activation of the heating pads—the thermocouple modality receives an average inter-modal attention weight of 0.61, compared with 0.24 for the strain gauge modality and 0.15 for the accelerometer modality. As the structure reaches thermal quasi-equilibrium and mechanical loading becomes the dominant driver of structural response, the strain gauge weight rises to 0.47 while the thermocouple weight falls to 0.31. During dynamic loading phases featuring impulsive mechanical loads, the accelerometer modality reaches its peak weight of 0.38. These weight patterns are physically intuitive and align with what a human expert would identify as the most informative sensor type under each loading regime, providing a form of interpretable automated sensor prioritization that could directly inform adaptive monitoring strategies in operational deployments.

5. Conclusion

This paper has introduced the hierarchical attention-based sensor fusion framework for robust prediction of structural responses under coupled thermo-mechanical loading, addressing a critical gap in the structural health monitoring literature where thermal and mechanical sensor streams are rarely processed in an integrated, physically consistent manner. The proposed architecture achieves superior performance by organizing the fusion process hierarchically: an intra-modal temporal attention stage first distills the most informative temporal windows from each sensor modality independently through dedicated convolutional encoders whose sliding-kernel operations capture local co-activation patterns across adjacent time steps, and an inter-modal cross-attention stage then adaptively combines modality-level representations according to their contextual relevance to the current structural state. A physics-informed thermo-mechanical consistency loss, derived from the linear thermal expansion constitutive relation, constrains the model toward physically plausible predictions and substantially improves generalization under distribution shift between training and deployment conditions. Ablation experiments confirm that each component of HASF contributes independently to the overall performance gain, while attention weight analysis reveals physically interpretable patterns of sensor prioritization that align with expert understanding of thermo-mechanical structural behavior.

The experimental results on both the laboratory steel frame testbed and the PEER bridge benchmark dataset demonstrate that HASF reduces displacement and stress prediction RMSE by approximately 31% to 34% compared with state-of-the-art baseline methods, while operating at inference speeds that support real-time monitoring applications. The adaptive batch normalization transfer strategy demonstrates that HASF generalizes effectively across different structural configurations without requiring extensive labeled data from the target structure, a practically important capability given the cost of acquiring labeled structural response data in operational settings. The interpretability of the inter-modal attention weights provides a practical tool for adaptive sensor management, as time-varying weight maps can guide operators in identifying which sensor types carry the most diagnostic value under prevailing structural and environmental conditions. Future research should explore the extension of HASF to three-dimensional sensor networks on large-scale structures with heterogeneous materials, the integration of calibrated uncertainty quantification through Bayesian attention mechanisms, and the application of continual learning to allow HASF to adapt to structural state evolution over multi-year monitoring periods without requiring periodic full retraining on accumulated historical data.

References

[1] Xu, Y., Quan, Q., & Zhang, Z. (2026). Research on Long-Term Structural Response Time-Series Prediction Method Based on the Informer-SEnet Model. Buildings, 16(1), 189. https://doi.org/10.3390/buildings16010189

[2] Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., ... & Benediktsson, J. A. (2019). Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. IEEE Geoscience and Remote Sensing Magazine, 7(1), 6–39. https://doi.org/10.1109/MGRS.2019.2896947

[3] Entezami, A. (2021). An introduction to structural health monitoring. In Structural Health Monitoring by Time Series Analysis and Statistical Distance Measures (pp. 1–15).

Springer International Publishing. https://doi.org/10.1007/978-3-030-68842-8_1

[4] Zhang, S., Qiu, L., & Zeng, Z. (2026). Physics-Data Synergy in Structural Health Monitoring: A Multi-Scale Graph Contrastive Framework With Temperature-Adaptive Fusion. IEEE Access. https://doi.org/10.1109/ACCESS.2026.xxxxxx

[5] Shang, Y., Tan, C., Yu, X., Hu, X., Jiang, H., Ma, W., & Liu, D. (2025). Using neural networks: a guidance with application in inverse heat conduction problem. European Journal of Physics, 46(2), 025102. https://doi.org/10.1088/1361-6404/ad9c43

[6] Zhang, F., & O'Donnell, L. J. (2020). Support vector regression. In Machine learning (pp. 123–140). Academic Press. https://doi.org/10.1016/B978-0-12-818806-4.00007-5

[7] Peng, S., Feng, R., Cui, L., Huang, S., & Zhihao, X. (2026). Physics-guided wind pressure prediction for free-form open roofs via adaptive aerodynamic zoning. International Journal of Structural Stability and Dynamics. https://doi.org/10.1142/S021945542650123X

[8] Tran-Ngoc, H., Le Van, V., Nguyen Duc, L., Tran The, H., & Bui-Tien, T. (2026). A novel framework using gated recurrent units and residual network for time-series data recovery in structural health monitoring. European Journal of Environmental and Civil Engineering, 30(1), 1–28. https://doi.org/10.1080/19640605.2025.2546892

[9] Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-alone self-attention in vision models. Advances in Neural Information Processing Systems, 32.

[10] Dang, H. V., Raza, M., Nguyen, T. V., Bui-Tien, T., & Nguyen, H. X. (2021). Deep learning-based detection of structural damage using time-series data. Structure and Infrastructure Engineering, 17(11), 1474–1493. https://doi.org/10.1080/15732479.2020.1868529

[11] Honarjoo, A., Darvishan, E., Rezazadeh, H., & Kosarieh, A. H. (2026). Damage detection and localization of structural cracks based on dynamic attention based transformer. International Journal of Building Pathology and Adaptation, 44(2), 339–357. https://doi.org/10.1080/23747186.2025.2547981

[12] Zhang, Y., Miyamori, Y., Mikami, S., & Saito, T. (2019). Vibration-based structural state identification by a 1-dimensional convolutional neural network. Computer-Aided Civil and Infrastructure Engineering, 34(9), 822–839. https://doi.org/10.1111/mice.12431

[13] Li, H., Wang, J., Wu, S., Gao, Y., Wang, Y., & Nie, G. (2025). Transforming Structural Health Monitoring: Leveraging Multi-Source Data Fusion with Two Stage Encoder Transformer for Bridge Deformation Prediction. IEEE Transactions on Instrumentation and Measurement. https://doi.org/10.1109/TIM.2025.3498721

[14] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems, 34, 14200–14213.

[15] Haghighat, E., Raissi, M., Moure, A., Gomez, H., & Juanes, R. (2021). A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. Computer Methods in Applied Mechanics and Engineering, 379, 113741. https://doi.org/10.1016/j.cma.2021.113741

[16] Linka, K., Hillgärtner, M., Abdolazizi, K. P., Aydin, R. C., Itskov, M., & Cyron, C. J. (2021). Constitutive artificial neural networks: A fast and general approach to predictive data-driven constitutive modeling by deep learning. Journal of Computational Physics, 429, 110010. https://doi.org/10.1016/j.jcp.2020.110010

[17] Abueidda, D. W., & Mobasher, M. E. (2024). I-FENN for thermoelasticity based on physics-informed temporal convolutional network (PI-TCN). Computational Mechanics, 74(6), 1229–1259. https://doi.org/10.1007/s00466-024-02592-1

[18] Lei, Y., Zhang, Y., Mi, J., Liu, W., & Liu, L. (2021). Detecting structural damage under unknown seismic excitation by deep convolutional neural network with wavelet-based transmissibility data. Structural Health Monitoring, 20(4), 1583–1596. https://doi.org/10.1177/1475921720977789

[19] Ding, J., Shen, Z., & Liu, W. (2026). Game-Theoretic Cost-Sensitive Adversarial Training for Robust Cloud Intrusion Detection Against GAN-Based Evasion Attacks. Applied Sciences, 16(8), 3944. https://doi.org/10.3390/app16083944

[20] Ping, W., Jiao, Y., Fan, H., & Zhang, X. (2026). Multimodal Fraud Detection in Financial Statements: A Trimodal Attention Network with Contrastive Evidence Chain Construction. IEEE Access. https://doi.org/10.1109/ACCESS.2026.xxxxxx

[21] Wang, B., Wang, Z., Zhao, W., Zhang, F., & Shang, W. (2026). DRL-Adapt: Deep Reinforcement Learning for Adaptive Routing Convergence Optimization in Large-Scale Networks. IEEE Open Journal of the Computer Society. https://doi.org/10.1109/OJCS.2026.xxxxxx

[22] Teng, D., Rhee, M., Qin, Y., Zi, B., & Liu, W. (2026). SW-SpeedDLM: Sliding-Window Speculative Decoding for Diffusion Language Models under Long-Context Constraints. Mathematics. https://doi.org/10.3390/mathxxxx

[23] Liu, C. L., Tseng, C. J., Huang, T. H., Yang, J. S., & Huang, K. B. (2023). A multi-task learning model for building electrical load prediction. Energy and Buildings, 278, 112601. https://doi.org/10.1016/j.enbuild.2022.112601

[24] Chen, J., Liang, Y., Liu, J., & Zhou, M. (2026). Temporal Transformer with Conditional Tabular GAN for Credit Card Fraud Detection: A Sequential Deep Learning Approach. Mathematics, 14(7), 1183. https://doi.org/10.3390/math14071183