E-ISSN:3108-1711
INTERNATIONAL RESEARCH JOURNAL OF SCIENTIFIC REPORTS AND REVIEWS
IRJSRR.ORG
Open Access | Peer Reviewed | Online Journal

Robust and Structure-Aware Visual Representation Learning for Reliable Deep Neural Networks

[1]Mayank Sharma*, [2]Dimpal Agrawal, [3]Kirti Sharma
[1,2,3]Research Associate, Department of Computer Science, Kalp Laboratories, Mathura, Uttar Pradesh.
*Correspondence E-mail: mynkshrm0@gmail.com

Abstract: The focus of this study's strong and structure-aware visual representation learning framework is medical picture analysis, which aims to make deep neural networks more dependable, resilient, and easy to understand. To transcend accuracy-focused evaluation, edge-guided structural oversight, corruption-sensitive robustness assessment, and calibration-oriented reliability analysis are introduced. The structure-aware MobileNetV3 does well on the Chest X-Ray dataset, with an accuracy of 0.8574, a high average confidence of 0.9284, and a controlled Expected Calibration Error (ECE) of 0.0710. The structure-aware ResNet-18 achieved an accuracy of 0.9071 and a low ECE of 0.0160. DenseNet121 had an accuracy of 0.8894 and an ECE of 0.0319. A robustness study reveals that performance trends remain consistent with ROC-AUC values exceeding 0.92, even after multiple changes, including the presence of Gaussian noise and occlusion. Grad-CAM explainability analysis demonstrates an anatomically directed emphasis on pulmonary regions, reinforcing structural priors. To evaluate the system's ability to work outside of medical imaging, it is tested on CIFAR-10. The robust model gets 71.92% clean accuracy, and this number goes up a lot when noise and blur corruptions are included. The results indicate that structure-aware and reliability-driven learning enhances the behaviour of trustworthy models, making the proposed framework appropriate for real-world, safety-critical visual recognition systems.

Keywords: Structure-Aware Representation Learning, Robust Deep Neural Networks, Reliability and Calibration Analysis, Explainable Medical Image Classification, Cross-Domain Robustness Evaluation.

Received: 25-Jul-2025 **Revised:** 24-Aug-2025 **Accepted:** 25-Sep-2025
Citation for the Paper: Sharma Mayank, Agrawal Dimpal, and Sharma Kirti. "Robust and Structure-Aware Visual Representation Learning for Reliable Deep Neural Networks." *International Research Journal of Scientific Reports and Reviews* 1, no. 1 (September 2025): 139–154.

1. Introduction

Visual representation learning plays a central role in the development of current computer vision systems, which allow deep neural networks to filter out the noise and to detect the significant aspects in visual raw data. Self-supervised learning has been one of the main contributors to this area; its recent developments have greatly diminished the dependency on annotated datasets while at the same time increasing the generalisation capability of the representations to be learned. One of the most important pieces of research work on self-supervised visual representation learning supplements the discussion by asserting that since 2020, contrastive and predictive learning strategies have completely transformed the paradigm of feature learning [1]. However, even with such breakthroughs, deep visual models are often very weak and unreliable in their performance when presented with noise, adversarial perturbation, or real-world distribution shift. Consequently, robust deep learning has become an important area of research, where the main focus is on increasing the stability and reliability of learned representations under difficult conditions. Recent research papers are conducting a systematic evaluation of robustness failures in vision models and pointing out the urgent need for robustness-oriented representation learning frameworks [2].

Robustness aside, the other major shortcoming of conventional deep neural networks is their lack of sensitivity to the structural information that exists in the visual scenes. Visual data structure comprises spatial relationships, object interactions, and geometric consistency, which are usually

neglected by data-driven learning methods. In this regard, recent advancements in the development of generalised robust vision toolkits and benchmarks have drawn attention to the need for integrating structure into representation learning workflows [3]. Graph-based visual learning techniques have come into the limelight as a powerful method to explicitly represent the structural relationships. Vision graph neural networks divide images into regions and simulate their interactions, thus allowing for drawing global conclusions based on the information received from local receptive fields only. Cluster-based vision graph models prove to be more efficient and structurally aware due to the merging of image partitioning with graph reasoning [4].

New advancements in vision graph neural networks have provided the introduction of graph construction mechanisms that learn and thus could be used to capture relational dependencies within the visual data adaptively. The models of this kind perform a dynamic optimisation of graph topology during training, which results in the generation of visual representations more expressive and aware of the structure, thereby making them suitable for complex recognition tasks [5]. The concept of structure awareness has also been examined through multimodal constraints, where visual representations are indirectly supported by auxiliary modalities, such as language. Vision-language constraint graph learning has shown a significant positive impact in the case of unsupervised vehicle re-identification through the embedding of semantic structure into the visual representations [6]. In the field of self-supervised learning, local structure-aware contrastive learning methods have been proposed by researchers, which have the ability to retain neighbourhood relationships during the process of learning the invariant features. These methods increase the robustness of the models by making sure that the regions of similar structure are kept close in the representation space under different augmentations [7].

Recent studies inspired by neuroscience are returning to the conclusion that visual representations preserving structures are more in line with human perception and communication. Learning representation frameworks that focus on semantic and structural coherence have proved to be very interpretable and robust in the difficult visual tasks and complex visuals [8]. Moreover, it is the case that with the development of large visual foundation models, the focus has moved towards comprehending their properties of robustness. Recent studies show that even though foundation models manage to reach an amazing level of accuracy, their representations are still likely to be fragile when there is a shift in the distribution, which in turn calls for robustness-aware learning strategies [9]. Similarly, structure-aware correspondence learning has been used more in geometrical vision tasks like relative pose estimation. This kind of learning has higher resilience to changes in viewpoints and occlusions since it incorporates structural constraints into the learning process [10]. Geometric priors have also been part of representation learning through masked pre-training methods utilising 3D structural information. Learning masked 3D priors during pre-training does enable vision transformers to develop representations that are shape-aware and spatially consistent [11].

Multimodal visual representation learning has looked beyond the alignment of vision and language and opened the door to the development of elaborate frameworks that process and fuse disparate types of data. The recent literature points out that structural guidance in multimodal combinations plays a crucial role in enhancing the robustness and generalisation of visual representations [12]. Vision transformers have taken an upper hand in representation learning architectures due to their capability of capturing long-range dependencies effectively. Self-supervised DINO-like transformer models are showing that attention-based structures can still produce extraordinarily strong and easily transferable visual features [13]. The so-called structure-aware multi-view contrastive learning frameworks are going one step further in their learning approach by having consistency not only among different structured views of data but also across the views themselves. These latter methods seem to highlight the need for protecting relational structure to some extent, along with the commonality [14]. Ultimately, the new research produced in the area of long-lasting and robust representation learning affirms that maintaining both the temporal and structural consistency is

a major factor in the attainment of generalisation across different environments. These conclusions further substantiate the vision of integrated, resilient, and structure-sensitive learning paradigms [15].

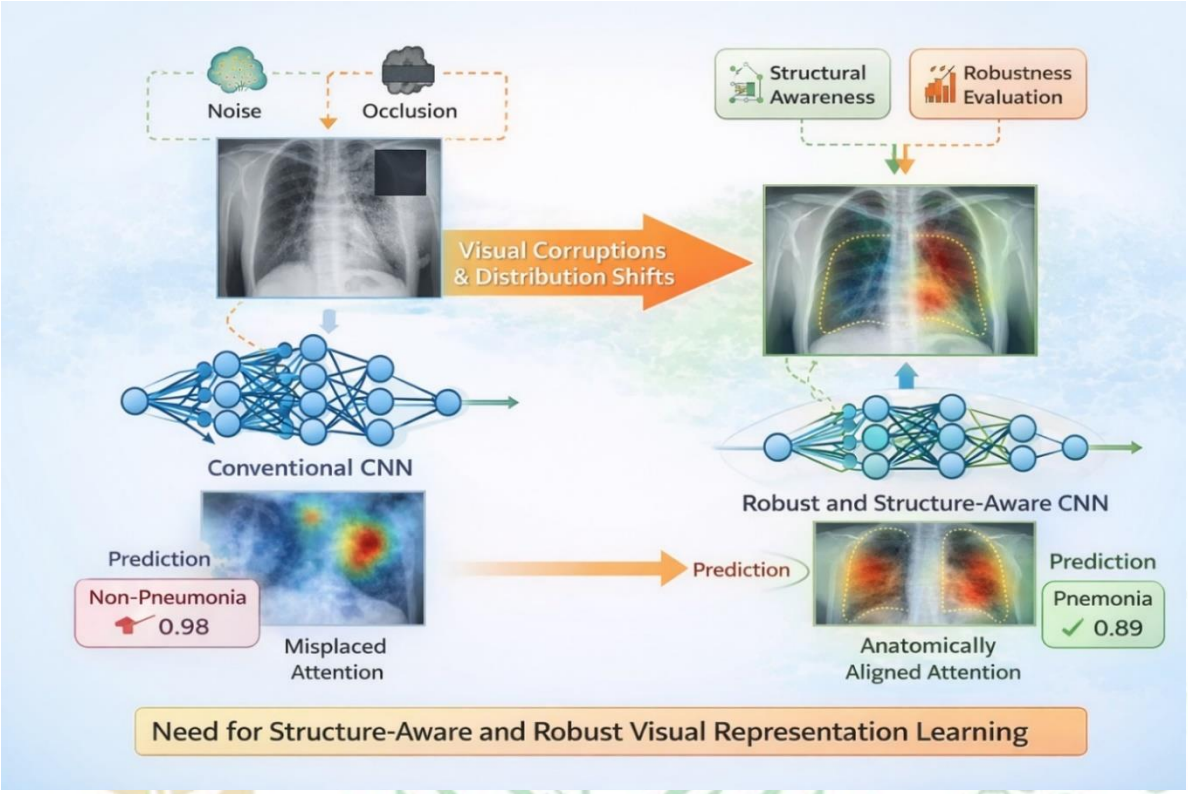

Fig. 1: Conceptual Motivation for Structure-Aware and Robust Visual Representation Learning

Figure 1 illustrates the contrast between conventional CNNs, which often exhibit misplaced attention and overconfident predictions under visual corruptions, and the proposed robust, structure-aware CNN that leverages structural awareness and robustness evaluation to achieve anatomically aligned attention and reliable predictions. The visualisation highlights how integrating structural priors enables stable and trustworthy decision-making under noise, occlusion, and distribution shifts, directly motivating the proposed learning framework.

The primary aim of this research is to construct a strong and structure-conscious visual representation learning framework that will take the reliability and interpretability of deep neural networks beyond the typical accuracy-focused optimisation. The method put forward will make use of the explicit structural priors in such a way that the anatomically and semantically meaningful representations will be the result, and at the same time, the robustness against the common visual corruptions and shifts in distribution will be increased. A dual-data evaluation strategy is adopted, wherein the reliability, calibration, and explainability of the proposed method are tested with medical imaging data, and the robustness and generalisation are shown with a benchmark dataset. The contribution of this research is in the consolidated merging of the structure-aware learning and robustness-driven training, which is the basis for the extensive evaluation through uncertainty analysis, explainability, and ablation studies. Thus, more trustworthy visual recognition in real-world applications is made possible.

2. Literature Review

2.1 Robust Visual Representation Learning in Deep Networks

The latest studies highlight robustness-centric visual representation learning as a major need for the trustworthy use of deep neural networks. Zhou et al. (2024) [16] devised a widespread visual representation learning framework that enlists multi-modal structured priors, pairing visual

embeddings with semantic knowledge so as to enhance generalisation under distribution shifts. Their method reveals excellent performance on cross-domain and corrupted image benchmarks. Çağatan et al. (2025) [17] explored adversarial robustness in discriminative self-supervised learning and proved that self-supervised representations are, by nature, more resistant to adversarial attacks than fully supervised models, mainly in the case of transfer learning. Liu et al. (2024) [18] offered a decoupled visual representation masking strategy that cuts down feature dependency and makes intra-class diversity in a way that strengthens adversarial robustness and does not affect standard accuracy negatively. Geirhos et al. (2020) [19] first pointed out that traditional vision models depend largely on non-robust texture cues rather than global shape information; hence, they become vulnerable to perturbations. Along these lines, Hendrycks et al. (2021) [20] argued that mainstream visual representations are incapable under natural distribution shifts, thus emphasising the necessity of semantically grounded and robustness-driven representation learning.

2.2 Structure-Aware and Graph-Based Visual Learning Methods

The learning of visual representations based on a graph has become a strong competitor to the classical methods based on convolutional and Transformer architectures, just by purposing to depict the structural relationships in images. Han et al. (2022) [21] put forward the Vision GNN (ViG), which gives similarity to the images in the form of graphs with node connections that are dynamically learned, thus leading to local and long-range dependencies being accurately grabbed while at the same time competing with the Transformers and having a boost in interpretability. After that, Munir et al. (2024) [22] produced GreedyViG using Dynamic Axial Graph Construction, to establish feature-relevant connections rather than fixed neighbourhoods; this caused a substantial increase in scalability and representation quality based on ImageNet benchmarks. Also, Gedik et al. (2025) [23] took this research thread one step further with AttentionViG, where the integration of cross-attention into graph neighbour aggregation provided adaptive structural reasoning, thus leading to improved robustness and generalisation across classification, detection, and segmentation tasks. Previously, Wu et al. (2021) [24] had shown that the structure-aware graph representations could successfully code the spatial and semantic relationships for visual recognition. Adding on to these investigations, Senior et al. (2024) [25] conducted a literature review on GNN applications in the field of vision, underscoring their advantages overall for tasks that explicitly require spatial and semantic reasoning.

2.3 Self-Supervised and Multimodal Visual Representation Learning

The latest developments in the fields of multimodal and self-supervised visual representation learning emphasise the need for strong universal visual features that can support reasoning and generalisation. Caffagni et al. (2025) [26] experimented with self-supervised visual learning methods on multimodal large language models and found that the updating of visual representations without the aid of textual supervision considerably strengthens the reliability of visual grounding and reasoning. Fan et al. (2025) [27] went a step further and proved that massive language-free and vision-only self-supervised models trained with contrastive objectives could equal language-supervised systems in terms of performance on multimodal downstream tasks, further providing an argument for scalability and robustness. Dave et al. (2024) [28] suggested a self-supervised multimodal framework that learns visual and tactile representations simultaneously, resulting in enhanced robustness and transferability of robotic perception in real-world scenarios. AlSuwat et al. (2025) [29] did a comprehensive review of audio-visual self-supervised learning and pointed out the role of cross-modal correspondence in boosting semantic comprehension without the need for manual labelling. To start with, Radford et al. (2021) [30] brought forth CLIP, which connected visual and textual representations utilising large-scale contrastive pretraining, thus ensuring exceptional zero-shot generalisation, and forming the basis of contemporary multimodal representation learning paradigms. The recent developments in the robust, structure-sensitive, and multimodal visual representation learning are all captured in Table 1, which also indicates the key contributions, focus areas, and challenges, along with the progress, limitations,

and the opening of new gaps for future research.

Table 1: Approach to literature review

Author(s) & Year	Focus	Main Contribution	Key Challenge
Zhou et al. (2024)	Robust Visual Representation Learning	Integrates multimodal priors to improve generalisation under distribution shifts	Handling cross-domain and corrupted images
Çağatan et al. (2025)	Adversarial robustness in SSL	Self-supervised visual representations show stronger resilience to attacks than supervised models.	Ensuring transferability across tasks
Liu et al. (2024)	Decoupled representation masking	Enhances adversarial robustness and intra-class diversity	Balancing robustness with accuracy
Han et al. (2022)	Structure-aware graph learning	ViG models images as node graphs, capturing local and long-range dependencies	Scaling graph-based models efficiently
Gedik et al. (2025)	Attention in graph-based learning	AttentionViG uses cross-attention for adaptive structural reasoning	Maintaining robustness-generalisation balance
Wu et al. (2021)	GNNs for visual recognition	Captures spatial and semantic relationships for improved interpretability	Computational complexity
Senior et al. (2024)	GNN survey	Highlights the benefits of explicit spatial/semantic modelling in visual understanding	Integrating graph reasoning into pipelines
Fan et al. (2025)	Large-scale self-supervised learning	Vision-only contrastive models achieve comparable performance to language-supervised models.	Scaling SSL for robustness
Dave et al. (2024)	Multimodal self-supervision	Joint visual-tactile learning improves robustness in robotics	Transferring representations to real environments
Radford et al. (2021)	CLIP: Contrastive image-language pretraining	Learns aligned visual-text representations; strong zero-shot generalisation.	Learning robust representations from noisy web data

2.4 Research Gap

Although there have been significant advances in robust, structure-aware, and multimodal visual representation learning, the field still has some gaps. The majority of methods that focus on robustness first deal with analyses of adversarial or corruption resilience, but do not consider structural consistency in the case of safety-critical domains like medical imaging, for instance. On the other hand, graph-based and structure-aware models aid in interpretability and spatial reasoning but are seldom combined with robustness-oriented goals. Self-supervised and multimodal frameworks improve generalisation but often depend on exceptionally large datasets, thus restricting their use in specific domains. Moreover, there are very few studies that simultaneously assess the three concepts of reliability, robustness, and explainability over dual datasets. Hence, there is a need for a unifying

framework that will bestow the combined properties of robustness, structure-awareness, and cross-domain generalisation.

3. Research Methodology

3.1 Proposed Framework Overview

This investigation puts forward a visual representation learning framework that is strong and aware of structure, which is aimed at improving the deep neural networks for reliability, robustness, and interpretability. The proposed system does not follow the traditional vision models that solely focus on maximising the classification accuracy, but rather it incorporates structural prior knowledge and robustness-driven learning objectives as part of the model to eliminate the difference between the real-world applications' predictor performance and their trustworthiness. The end-to-end pipeline that combines robustness perturbations and structure extraction for learning reliable visual representations, evaluated through accuracy, robustness, and explainability metrics across datasets, is shown in Figure 2.

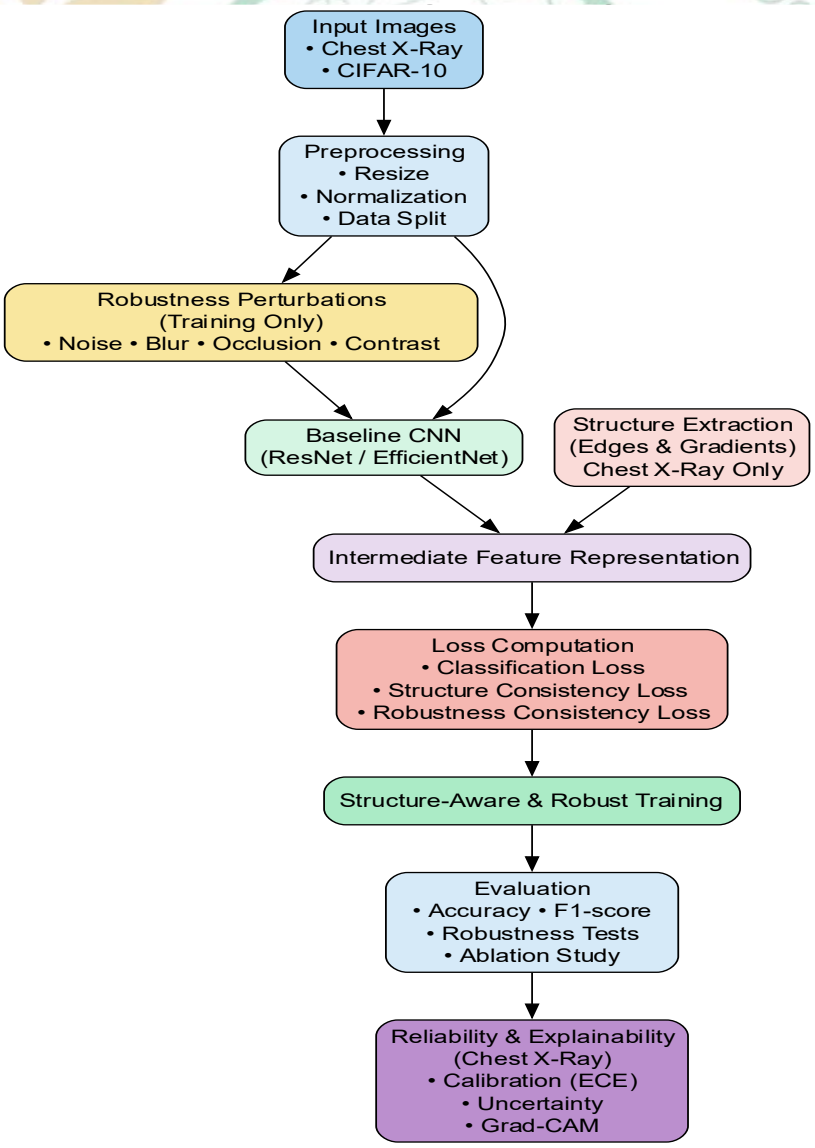

Fig. 2: Proposed Structure-Aware and Robust Visual Representation Learning Framework

A dual-dataset experimental strategy is decided upon to guarantee rigorous methodology and generalisation. The Chest X-Ray Pneumonia dataset is utilised as the major dataset for examining structure-aware learning, reliability, and explainability within a safety-critical medical domain. At the

same time, CIFAR-10 is applied as a standard dataset to check for robustness and cross-domain generalisation under controlled perturbations. This design ensures that the suggested framework is not restricted to a single domain and, therefore, is widely applicable.

3.2 Datasets, Preprocessing, and Experimental Protocol

The Chest X-Ray Pneumonia dataset is a great choice because of its clear anatomical structure and high clinical significance, which make it a good candidate for testing structure-aware and reliability-oriented learning. CIFAR-10 is the existing standard benchmark for robustness analysis and comparative evaluation. In the case of both datasets, images are scaled to the same resolution and normalised to the dataset-specific statistics in order to provide stable training.

Given an input image $x \in R^{H\times W\times C}$, normalisation is performed as:

$$x' = \frac{x - \mu}{\sigma}$$

where μ and σ denote the dataset-specific mean and standard deviation, respectively.

To exhibit reproducibility and a fair comparison among all trials standard train–validation–test splits are maintained. Data preprocessing is consistent for both the baseline and proposed models so that the effects of structure-aware and robustness-oriented learning can be isolated. For good quality strength evaluation, different kinds of noise, like Gaussian, blur, occlusion, and contrast variations, are used in a controlled manner according to the accepted robustness assessment protocols.

3.3 Baseline Architecture and Learning Setup

A widely adopted convolutional neural network architecture from the ResNet or EfficientNet family is used as the baseline model, given its strong performance and acceptance in both medical imaging and general vision tasks. Let the network be represented as a function. $f_\theta(\cdot)$ parameterised by θ.

For a given input image x and ground-truth label y, the baseline model is trained using standard supervised learning with cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{K} y_k \log(\widehat{y_k})$$

where $\hat{y} = \text{softmax}\big(f_\theta(x)\big)$ and K denotes the number of classes.

Baseline performance is determined through the use of accuracy, precision, recall, F1-score, and confusion matrix analysis, among others. These outcomes are regarded as benchmarks by which all further advancements brought about by the proposed framework are quantitatively assessed, thus ensuring that the performance improvements can be attributed unambiguously to the proposed methodological contributions.

3.4 Structure-Aware and Robust Representation Learning

Structure-aware representation learning is introduced exclusively for the Chest X-Ray dataset to leverage anatomical consistency. Structural priors are derived using edge- and gradient-based cues that emphasise lung boundaries and spatial organisation. Let $S(x)$ denote the extracted structural map corresponding to the image x, and $\phi(x)$ represent intermediate feature maps of the network.

A structure consistency loss is defined to encourage alignment between learned representations and anatomical structure:

$$\mathcal{L}_{\text{struct}} = \| \phi(x) - S(x) \|_2^2$$

This regularisation discourages the model from relying on spurious background correlations and promotes anatomically grounded representations.

To further enhance real-world reliability, robustness-oriented training is applied to both datasets. During training, corrupted samples $\tilde{x} = \mathcal{T}(x)$ are generated using corruption functions $\mathcal{T}(\cdot)$ such as noise injection, blur, occlusion, and contrast shifts. The robustness objective ensures consistency between predictions on clean and corrupted inputs:

$$\mathcal{L}_{\text{robust}} = \| f_\theta(x) - f_\theta(\tilde{x}) \|_2^2$$

The final training objective is formulated as a weighted combination:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{struct} + \lambda_2 \mathcal{L}_{robust}$$

where λ_1 and λ_2 control the influence of structure-aware and robustness components, respectively.

3.5 Reliability, Explainability, and Evaluation Protocol

Reliability and trust analysis is conducted only on the Chest X-Ray dataset, reflecting its safety-critical nature. Model calibration is evaluated using Expected Calibration Error (ECE), defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$$

where B_m denotes the set of samples in confidence bin m, N is the total number of samples, and $\text{acc}(\cdot)$ and conf $(\cdot)$ represent accuracy and confidence, respectively.

Uncertainty estimation is further analysed using predictive entropy:

$$\mathcal{H}(\hat{y}) = -\sum_{k=1}^{K} \widehat{y_k} \log(\widehat{y_k})$$

with particular attention given to high-confidence misclassifications.

Explainability is assessed using Grad-CAM, where class-discriminative localisation maps are computed as:

$$\text{Grad-CAM} = \text{ReLU}\left(\sum_c \alpha_c A^c\right)$$

Here, A^c denotes feature maps of the last convolutional layer, and α_c are importance weights derived from gradient information.

The framework that is being proposed has undergone validation via a series of comparative experiments and ablation studies, where systematically the structure-aware and robustness components have been completely removed. The performance of the models is evaluated on both datasets under clean and corrupted conditions, which has confirmed the robustness, reliability, and generalisation resulting from individual and combined contributions of each component.

Algorithm 1: Proposed Structure-Aware and Robust Visual Learning Framework

Input:
Image dataset $\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}$ (Chest X-Ray, CIFAR-10)

Output:
Trained a robust and structure-aware model f_θ

Steps:

1.Start

2.Input Acquisition

i. Load input images $x \in R^{H \times W \times C}$

3.Preprocessing

i. Resize images to a fixed resolution

ii. Normalise images $x' = \frac{x-\mu}{\sigma}$

iii. Split data into training, validation, and test sets

4.Baseline Model Initialisation

i. Initialise CNN $f_\theta(\cdot)$

5.Structure Extraction (Chest X-Ray only)

i. Extract structural map using edges and gradients $S(x) = \text{EdgeGrad}(x)$

6.Robust Sample Generation

i. Generate corrupted samples $\tilde{x} = \mathcal{T}(x)$

7.Loss Computation

i. **Classification loss** $\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{K} y_k \log(\widehat{y_k})$

ii. **Structure consistency loss** $\mathcal{L}_{\text{struct}} = \| \phi(x) - S(x) \|_2^2$

iii. **Robustness consistency loss** $\mathcal{L}_{\text{robust}} = \| f_\theta(x) - f_\theta(\tilde{x}) \|_2^2$

8.Model Optimisation

i. Compute total loss $\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{struct} + \lambda_2 \mathcal{L}_{robust}$

ii. Update parameters θ

9.Evaluation

i. Measure accuracy and F1-score

ii. Compute calibration error ECE $= \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$

iii. Generate Grad-CAM maps for explainability

10. Return Output

i. Output trained model f_θ

11. End

4. Results And Analysis

This section presents a comprehensive evaluation of the proposed framework through baseline performance analysis, robustness assessment under input perturbations, and reliability examination using calibration and confidence-based metrics. The results are further supported by explainability-driven visualisations that analyse attention behaviour and structural alignment in medical images. Finally, structure-aware learning and cross-dataset validation are discussed to highlight architecture-dependent reliability improvements and generalisation beyond accuracy-centric evaluation.

Table 2: Comparative Performance of Baseline CNN Architectures on the Chest X-Ray Test Set

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	ROC-AUC
ResNet18	0.7404	0.8469	0.6547	0.6508	0.9237
DenseNet121	0.8413	0.8821	0.7936	0.8130	0.9480
MobileNetV3	0.8862	0.9052	0.8560	0.8718	0.9557

As summarised in Table 2, MobileNetV3 exhibits the most balanced and effective performance among the evaluated baseline models, achieving the highest accuracy (0.8862), macro F1-score (0.8718), and ROC-AUC

(0.9557). DenseNet121 follows with competitive discriminative capability but slightly reduced macro-level consistency, while ResNet18 shows limited recall for the NORMAL class, resulting in lower overall balance despite a strong ROC-AUC value. These findings indicate that MobileNetV3 provides a robust baseline for subsequent robustness, reliability, and structure-aware investigations, rather than serving as an endpoint for accuracy-centric comparison.

Table 3: Robustness Comparison of Baseline CNN Models under Clean and Corrupted Conditions on the Chest X-Ray Dataset

Model	Condition	Accuracy	Recall	F1-Score	ROC-AUC
ResNet18	Clean	0.7404	0.9974	0.8277	0.9237
ResNet18	Gaussian Noise	0.6282	1.0000	0.7708	0.8117
ResNet18	Occlusion	0.7308	1.0000	0.8228	0.9052
DenseNet121	Clean	0.8413	0.9846	0.8858	0.9481
DenseNet121	Gaussian Noise	0.6250	1.0000	0.7692	0.7387
DenseNet121	Occlusion	0.8029	0.9872	0.8623	0.9346
MobileNetV3	Clean	0.8766	0.9667	0.9073	0.9526
MobileNetV3	Gaussian Noise	0.6234	0.9949	0.7676	0.5845
MobileNetV3	Occlusion	0.8494	0.9641	0.8889	0.9344

As shown in Table 3, all baseline models experience a notable degradation in performance under Gaussian noise and occlusion, despite maintaining consistently high recall values close to unity, indicating a strong bias toward positive class detection under perturbations. While MobileNetV3 achieves the highest clean-condition accuracy (0.8766) and F1-score (0.9073), its ROC-AUC drops sharply to 0.5845 under Gaussian noise, revealing substantial confidence–discrimination mismatch. Overall, the results highlight that robustness degradation is model-agnostic and emphasise that high recall under corruption does not necessarily translate to dependable or well-calibrated predictions, motivating the need for reliability-aware and structure-guided learning strategies.

Fig. 3: Grad-CAM Visualisation Highlighting Discriminative Regions for Pneumonia Prediction in Chest X-Ray Imaging

Figure 3 illustrates the Grad-CAM–based attention map corresponding to a test chest X-ray predicted as PNEUMONIA with a confidence score of 0.6192. The highlighted regions are primarily concentrated within anatomically relevant lung areas, indicating that the model's decision is guided by clinically meaningful pulmonary patterns rather than background artefacts. Although the prediction confidence is moderate, the spatial alignment between activation regions and lung structures supports the interpretability of the model's decision-

making process while also reflecting inherent uncertainty, reinforcing the need for reliability-aware evaluation in medical imaging applications.

Fig. 4: Effect of Input Perturbations on Model Attention and Prediction Consistency Visualised Using Grad-CAM

Figure 4 compares Grad-CAM images of the same chest X-ray with Gaussian noise, clean, and blocked. When given clean input, the model predicts PNEUMONIA with low confidence (0.510) and generally activates anatomically suitable lung regions. Adding Gaussian noise and occlusion helps people focus on non-discriminative regions and changes the expected class to NORMAL. Higher but false confidence levels (0.781 and 0.543) accompany this. This illustrates that input changes may destabilise spatial attention and promote confidence. This illustrates that regular deep neural networks do not always forecast well.

Fig. 5: Prediction Stability and Confidence Variation under Clean and Corrupted Inputs for a Normal Chest X-Ray Case

Figure 5 demonstrates how the model predicts a NORMAL chest X-ray with clean Gaussian noise and occlusion. The anticipated class is normal for all three variations, although confidence ratings vary greatly. The probability of the NORMAL class increasing under Gaussian noise (0.647) and decreasing under occlusion (0.541). This result implies that consistent predictions do not necessarily imply steady confidence, and that even correct classifications may vary in uncertainty as input changes. Such activity highlights the necessity for reliability-aware evaluation since stable labels alone cannot guarantee clinical decision-making.

Table 4: Performance–Reliability Trade-off of MobileNetV3 under Clean and Corrupted Conditions on the Chest X-Ray Dataset

Condition	Accuracy	F1-Score	ROC-AUC	Avg Confidence	ECE ↓
Clean	0.8862	0.9148	0.9557	0.9349	0.0539
Gaussian Noise	0.6058	0.7490	0.5644	0.9473	0.3497
Occlusion	0.8125	0.8663	0.9292	0.9156	0.0989

Table 4 illustrates that MobileNetV3 distinguishes effectively in clean surroundings. Accuracy is 0.8862, F1-score 0.9148, and ROC-AUC 0.9557. Also, its calibration error is modest (ECE = 0.0539). Even

with high average confidence (0.9473), Gaussian noise lowers performance (ROC-AUC = 0.5644). This causes significant miscalibration (ECE = 0.3497), indicating wide dependability differences. In contrast, occlusion induces modest performance loss with a lower ECE (0.0989), indicating that various perturbations influence confidence and precision alignment differently. They need reliability-aware, structure-guided assessment methodologies in addition to accuracy-focused criteria.

Table 5: Performance and Calibration Characteristics of Structure-Aware CNN Models on the Chest X-Ray Test Set

Model Variant	Accuracy	Avg Confidence	ECE ↓
MobileNetV3 (Structure-Aware)	0.8574	0.9284	0.0710
ResNet-18 (Structure-Aware)	0.9071	0.9022	0.0160
DenseNet121 (Structure-Aware)	0.8894	0.9014	0.0319

Table 5 indicates that structure-aware supervision improves confidence calibration for all architectures examined. ResNet18 achieved the greatest accuracy (0.9071) and lowest ECE (0.0160). The greater ResNet18 gains illustrate that explicit structural priors offer shallow networks a higher inductive bias. Due to representational regularisation, deeper or more optimised models like DenseNet121 and MobileNetV3 have fewer accuracy changes. These findings suggest that structure-aware learning improves dependability based on architecture rather than accuracy ranking. This highlights how crucial medical imaging model performance and calibration evaluation are.

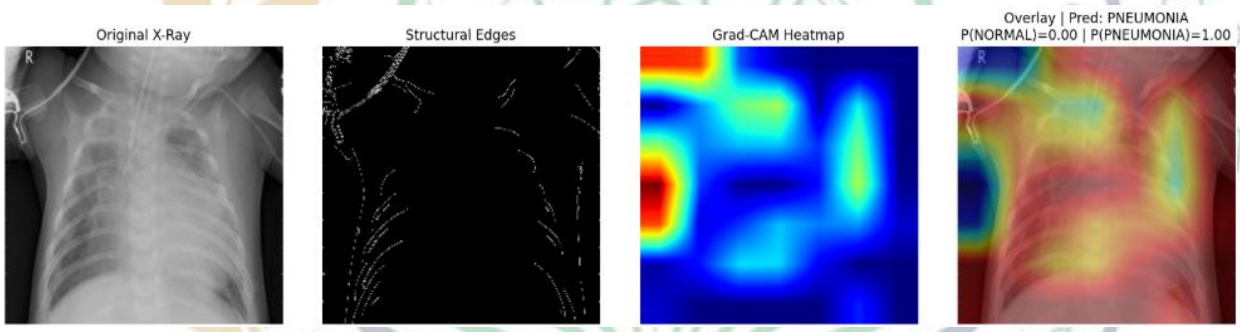

Fig. 6: Structure-Aware Explainability of Pneumonia Detection Using Edge-Guided Grad-CAM Visualisation

Figure 6 displays a structure-aware visualisation framework that includes the original chest X-ray, extracted structural edges, Grad-CAM heatmap, and pneumonia-positive prediction overlay. Ribs and lung outlines are seen on the structural edge map. Weak structural antecedent, not an apparent medical annotation. Grad-CAM activation is predominantly detected in the lungs and matches these structural signals; therefore, the model's decision is based on anatomically appropriate representations rather than background noise. The forecast's high confidence indicates that the learnt attention and classification output match. This suggests that edge-guided representation learning may improve medical imaging deep neural network predictions.

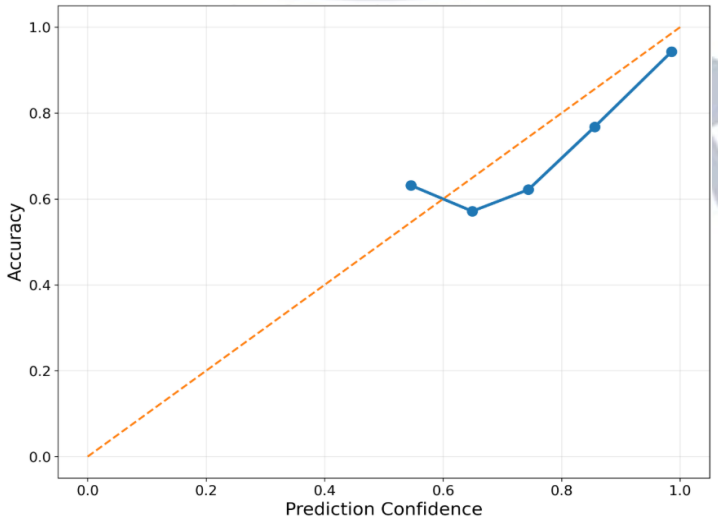

Fig. 7: Reliability Diagram Illustrating Confidence–Accuracy Alignment of the Model on the Chest X-Ray Test Set

Figure 7 depicts a reliability diagram between forecast confidence and empirical accuracy. The diagonal line calibrates well. The curve is not optimal, particularly in medium confidence zones. It seems that confidence and accuracy are mismatched. Higher confidence bins normally equal higher accuracy; however, non-linear deviations reveal that confidence estimates are not always correct across ranges. Even if it distinguishes data types effectively, this behaviour indicates calibration gaps in the model. This suggests that medical imaging applications should assess reliability beyond accuracy-based criteria.

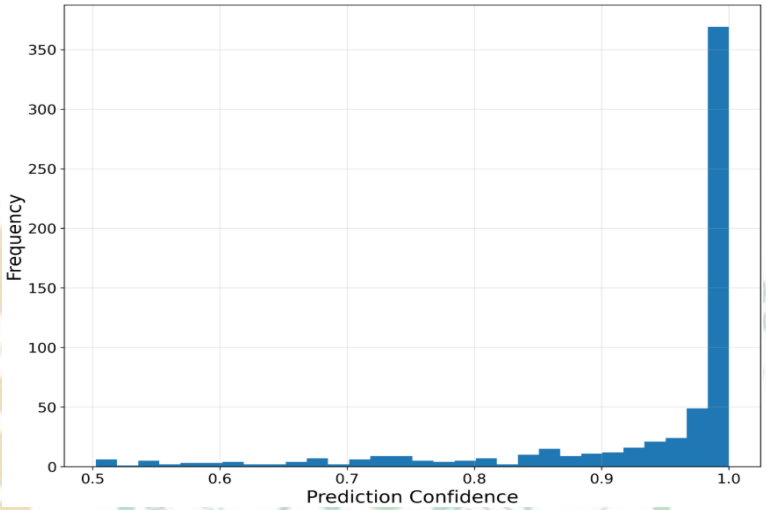

Fig. 8: Distribution of Prediction Confidence Scores on the Chest X-Ray Test Set

Figure 8 illustrates the model's Chest X-ray test set prediction confidence. Many projections are near one, and confidence ratings are typically high. This distribution indicates that the model is sure, but the calibration errors suggest that it may be excessively confident. This behaviour shows that high confidence does not necessarily mean good prediction reliability, emphasising the necessity for calibration-aware and reliability-centred evaluation in clinical decision-support systems.

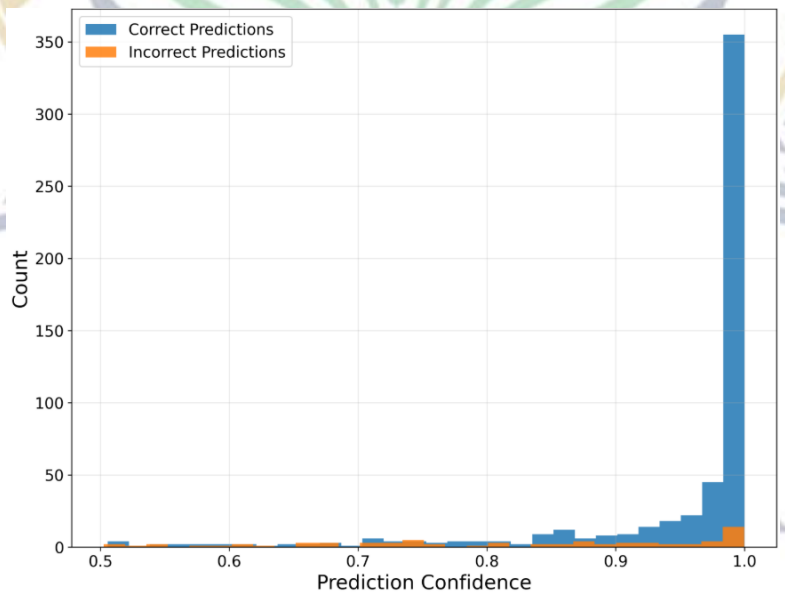

Fig. 9: Comparison of Prediction Confidence for Correct and Incorrect Classifications on the Chest X-Ray Test Set

Figure 9 demonstrates how correctly and incorrectly categorised data affect prediction confidence. Most correct projections are in the higher confidence level, but many erroneous ones are too. This overlap illustrates that high confidence alone does not indicate correctness and that too many misclassifications are confident. This tendency is concerning in clinical decision-support systems because it reveals how harmful confident but erroneous forecasts are. This makes explicit calibration and reliability-aware model evaluation increasingly crucial.

Table 6: Comparative Robustness Evaluation of Baseline and Robust CNN Models on CIFAR-10

Model	**Clean**	**Noise**	**Blur**
Baseline CNN	58.26	24.04	43.68
Robust CNN (Ours)	**71.92**	**34.59**	**62.47**

Table 6 presents a controlled robustness comparison of the Baseline CNN and the proposed Robust CNN on the CIFAR-10 dataset in clean and corrupted situations. Clean data is good for the Baseline CNN, while Gaussian noise and blur reduce its accuracy. This indicates that it is susceptible to frequent visual alterations. The recommended Robust CNN improves clean data by 13.66% and always has greater accuracy with noise (+10.55%) or blur (+18.79%).

The low blur corruption degradation suggests that the recommended training strategy promotes representations that are less vulnerable to low-level distortions and more compatible with structural visual signals. Significantly, these resilience improvements correspond with improved performance on clean data, showing that the recommended strategy does not sacrifice accuracy or robustness. These results demonstrate the framework's generalizability beyond medical imaging and its ability to improve generic visual recognition tasks.

5. Conclusion

This comprehensive examination of structure-aware and reliability-driven visual representation learning demonstrates that reliable deep learning systems need evaluation frameworks that extend beyond mere accuracy metrics. The suggested medical image classification system makes accurate predictions across neural architectures by employing edge-guided structural supervision, robustness, and calibration analysis. Structure-aware models exhibit low Expected Calibration Error values and consistent performance amongst typical disturbances in the Chest X-Ray dataset, indicating improved confidence control and robust diagnostic efficacy. The subsequent explainability investigation illustrates that structural priors enhance anatomically relevant attention, hence augmenting model interpretability and clinical validity. Cross-dataset validation on CIFAR-10 demonstrates that the proposed method is generalisable, exhibiting resilience benefits in noisy and blurred contexts without sacrificing performance on clean data. These findings show how important it is to be aware of the structure, test the resilience, and check the calibration of deep neural networks. The platform offers a practical and flexible means to implement dependable and comprehensible AI systems in safety-sensitive visual identification contexts, particularly in medical imaging.

References

[1] Uelwer, Tobias, Jan Robine, Stefan Sylvius Wagner, Marc Höftmann, Eric Upschulte, Sebastian Konietzny, Maike Behrendt und Stefan Harmeling. "A survey on self-supervised methods for visual representation learning." *Machine Learning* 114, no. 4 (2025): 1-56.

[2] Liu, Jia, and Yaochu Jin. "A comprehensive survey of robust deep learning in computer vision." *Journal of Automation and Intelligence* 2, no. 4 (2023): 175-195.

[3] Li, Zhao, Yuefeng Chen, Hui Xue, and Xiaofeng Mao. "A Comprehensive Toolkit for Generalised Robust Vision." In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 6456-6460. 2025.

[4] Parikh, Dhruv, Jacob Fein-Ashley, Tian Ye, Rajgopal Kannan, and Viktor Prasanna. "ClusterViG: Efficient Globally Aware Vision GNNs via Image Partitioning." *arXiv preprint arXiv:2501.10640* (2025).

[5] Elsharkawi, Ismael, Hossam Sharara, and Ahmed Rafea. "ViG-LRGC: Vision Graph Neural Networks with Learnable Reparameterized Graph Construction." *arXiv preprint arXiv:2509.18840* (2025).

[6] Wang, Dong, Qi Wang, Zhiwei Tu, Weidong Min, Xin Xiong, Yuling Zhong, and Di Gai. "Vision-language constraint graph representation learning for unsupervised vehicle re-identification." *Expert Systems with Applications* 255 (2024): 124495.

[7] Yang, Kai, Yuan Liu, Zijuan Zhao, Peijin Ding, and Wenqian Zhao. "Local structure-aware graph contrastive representation learning." *Neural Networks* 172 (2024): 106083.

[8] Wei, YuHan, ChangWook Lee, SeokWon Han, and Anna Kim. "Enhancing visual communication through representation learning." *Frontiers in Neuroscience* 18 (2024): 1368733.

[9] Gupta, Sandeep, and Roberto Passerone. "An investigation of visual foundation models' robustness." *Machine Learning* 114, no. 12 (2025): 281.

[10] Chen, Yihan, Wenfei Yang, Huan Ren, Shifeng Zhang, Tianzhu Zhang, and Feng Wu. "Structure-Aware Correspondence Learning for Relative Pose Estimation." In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11611-11621. 2025.

[11] Hou, Ji, Xiaoliang Dai, Zijian He, Angela Dai, and Matthias Nießner. "Mask3d: Pre-training 2d vision transformers by learning masked 3d priors." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13510-13519. 2023.

[12] Zhang, Dong, W. K. Wong, and I. M. Chew. "A Comprehensive Review of Multimodal Visual Representation Learning: Tracing the Evolution from CNNs to Transformers and Beyond." *International Journal of Multimedia Information Retrieval* 14, no. 4 (2025): 1-30.

[13] Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez et al. "Dinov2: Learning robust visual features without supervision." *arXiv preprint arXiv:2304.07193* (2023).

[14] Zhao, Haihong, Zhixun Li, Chenyi Zi, Aochuan Chen, Fugee Tsung, Jia Li, and Jeffrey Xu Yu. "A Survey of Cross-domain Graph Learning: Progress and Future Directions." *arXiv preprint arXiv:2503.11086* (2025).

[15] Yang, Rui, Jie Wang, Qijie Peng, Ruibo Guo, Guoping Wu, and Bin Li. "Learning Robust Representations with Long-Term Information for Generalisation in Visual Reinforcement Learning." In *The Thirteenth International Conference on Learning Representations*.

[16] Zhou, Hongkuan, Lavdim Halilaj, Sebastian Monka, Stefan Schmid, Yuqicheng Zhu, Bo Xiong and Steffen Staab. "Robust Visual Representation Learning With Multi-modal Prior Knowledge For Image Classification Under Distribution Shift." *arXiv preprint arXiv:2410.15981* (2024).

[17] Çağatan, Ömer Veysel, Ömer Faruk Tal, and M. Emre Gursoy. "Adversarial Robustness of Discriminative Self-Supervised Learning in Vision." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2313-2324. 2025.

[18] Liu, Decheng, Tao Chen, Chunlei Peng, Nannan Wang, Ruimin Hu, and Xinbo Gao. "Improving Adversarial Robustness via Decoupled Visual Representation Masking." *IEEE Transactions on Information Forensics and Security* (2025).

[19] Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* 2, no. 11 (2020): 665-673.

[20] Hendrycks, Dan, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. "Natural adversarial examples." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262-15271. 2021.

[21] Han, Kai, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. "Vision gnn: An image is worth a graph of nodes." *Advances in neural information processing systems* 35 (2022): 8291-8303.

[22] Munir, Mustafa, William Avery, Md Mostafijur Rahman, and Radu Marculescu. "Greedyvig: Dynamic axial graph construction for efficient vision gnns." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6118-6127. 2024.

[23] Gedik, Hakan Emre, Andrew Martin, Mustafa Munir, Oguzhan Baser, Radu Marculescu, Sandeep P. Chinchali, and Alan C. Bovik. "AttentionViG: Cross-Attention-Based Dynamic Neighbour Aggregation in Vision GNNs." *arXiv preprint arXiv:2509.25570* (2025).

[24] Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. "A comprehensive survey on graph neural networks." IEEE transactions on neural networks and learning systems 32, no. 1 (2020): 4-24.

[25] Senior, Henry, Gregory Slabaugh, Shanxin Yuan, and Luca Rossi. "Graph neural networks in vision-language image understanding: a survey." *The Visual Computer* 41, no. 1 (2025): 491-516.

[26] Caffagni, Davide, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Pier Luigi Dovesi, Shaghayegh Roohi, Mark Granroth-Wilding and Rita Cucchiara. "Seeing Beyond Words: Self-Supervised Visual Learning for Multimodal Large Language Models." *arXiv preprint arXiv:2512.15885* (2025).

[27] Fan, David, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat et al. "Scaling language-free visual representation learning." *arXiv preprint arXiv:2504.01017* (2025).

[28] Dave, Vedant, Fotios Lygerakis, and Elmar Rueckert. "Multimodal visual-tactile representation learning through self-supervised contrastive pre-training." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8013-8020. IEEE, 2024.

[29] Alsuwat, Manal, Sarah Al-Shareef, and Manal Alghamdi. "Audio-visual self-supervised representation learning: A survey." *Neurocomputing* (2025): 129750.

[30] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International Conference on machine learning*, pp. 8748-8763. PmLR, 2021.