Attention Aware Dual-Path Autoencoder with Asymmetric Loss for Recognition in Complex Scenes

Hashim Rosli, Rozniza Ali*, Muhamad Suzuri Hitam, Ashanira Mat Deris, Noor Hafhizah Abd Rahim
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu (UMT),
Kuala Nerus, 21030 Terengganu, Malaysia

***Abstract*—Object recognition in complex scenes is challenging due to cluttered backgrounds, overlapping objects, and degraded image quality. Another difficulty arises from sparse label presence, as most images contain only one to three active labels despite the dataset being balanced across 20 object classes. This intra-sample sparsity complicates binary classification by exposing models to a high proportion of inactive classes. This work aims to improve recognition accuracy, robustness under sparse multi-label conditions, and interpretability in visually complex environments. The objective is to help models focus on relevant visual features, suppress background noise, and better distinguish objects that are rare or overlapping. To address these challenges, we introduce an attention aware dual-path autoencoder that enhances image features while learning to classify multiple objects. The model uses asymmetric loss to reduce the influence of easy negatives and emphasize rare or difficult labels. It also integrates an attention mechanism in the reconstruction path to improve object clarity. The proposed model achieves 96.72 percent accuracy, 0.0328 Hamming Loss, 0.9809 macro ROC-AUC, and 0.8925 macro mAP, along with 0.9372 SSIM and 7.1012 dB PSNR in reconstruction. These results confirm its effectiveness for robust classification and enhanced visual understanding in complex scenes.**

Keywords*—*Component; autoencoder; attention aware; feature fusion; image enhancement; multi-label classification

I. Introduction

In real-world applications, object recognition systems often must contend with complex scenes which are characterized by varied backgrounds, overlapping objects, and degraded image quality such as noise, blur, and low brightness [1]. These conditions reduce the effectiveness of conventional convolutional neural networks, which often assume clean, well-segmented inputs. Cluttered environments with sensor noise or motion blur, for example, can lower classification accuracy by over 10% in multi-label recognition tasks [2]-[3]. While previous work has explored pre-processing techniques such as denoising and deblurring prior to recognition [4]-[6], these modular approaches introduce computational overhead and suffer from inefficient feature usage, as enhancement and recognition are handled separately. Although the dataset is balanced across 20 object categories, each image includes only 1 to 3 active labels, creating label sparsity during training. This, combined with frequent object overlaps, makes classification more challenging, especially with standard binary loss functions. This work proposes a unified framework that enhances image quality and improves multi-label recognition in complex scenes with sparse, overlapping labels. The model is designed to operate in cluttered environments while learning effectively from limited active labels per image.

We propose an end-to-end architecture that combines image reconstruction and classification into a single unified pipeline. A DenseNet-based convolutional autoencoder, enhanced with a Convolutional Block Attention Module (CBAM), is used to suppress noise and focus on salient object regions during image enhancement [7]. The reconstructed output is processed by a second DenseNet classifier, while latent features from the autoencoder are fused with classification features using multiplicative fusion to integrate both structural and semantic information [8], [9]. To address the challenge of label sparsity, the model is trained using a composite loss function consisting of Mean Squared Error (MSE) for reconstruction and either Binary Cross-Entropy (BCE) or Asymmetric Loss (ASL) for classification. ASL is particularly effective in multi-label settings with few active classes per image, as it down-weights easy negatives and emphasizes difficult positive enhancing discrimination in overlapping or infrequent categories.

The proposed method is evaluated against two baselines, a standard convolutional autoencoder without any transfer learning, CBAM or feature fusion and a standalone DenseNet-121 classifier without image enhancement. Results show that the attention-aware dual-path model trained with ASL outperforms both reconstruction quality and classification accuracy, proving its robustness in complex scene recognition. The rest of the paper is organized as follows. Section II reviews related work. Section III outlines the methodology. Section IV presents the experimental setup. Results and discussions are given in Section V. Section VI discusses the results. Section VII concludes and suggests future directions.

II. Related Work

A. Complex Scene Object Recognition

Object recognition in complex scenes has been widely studied, especially under clutter, occlusion, and image degradation. Cheng et al. reviewed deep learning methods for remote-sensing scene classification, highlighting robust feature extraction in noisy settings [10]. Fu et al. demonstrated that classification accuracy in astronomical imaging can be improved using denoising autoencoders to recover from noise and blur [11]. In medical imaging, autoencoder-based architectures have also been employed to extract salient features from visually noisy data [12]. Lu et al. introduced a co-attention Siamese network for video segmentation, utilizing co-attention mechanisms to better segment overlapping objects in complex backgrounds [13].

In more domain-specific research which using kitchen utensils dataset, Yusro et al. compared Faster R-CNN and YOLOv5 on kitchen datasets with overlapping utensils, reporting YOLOv5's superior accuracy of 89.12% compared to 83.92% [14]. Building on this, Hashim applied YOLOv5 with different deep learning backbones to manage image degradation, achieving strong performance under noisy conditions [15]. Gallego et al. developed the Kurcuma dataset, comprising over 6,800 annotated images ranging from isolated objects to cluttered domestic scenes, designed for evaluating domain-adaptive recognition methods [16].

B. Multitask Image Enhancement with Object Recognition

Integrating image enhancement and classification in unified architecture has improved recognition in challenging visual environments. Hami and JameBozorg showed that using a denoising autoencoder as preprocessing significantly increased accuracy when paired with VGG16 and InceptionV3 for defect detection tasks [17]. Liu et al. proposed a DenseNet-based denoising self-encoder for image dehazing, which enhanced both image clarity and object detection in outdoor scenes [18]. In medical applications, CNN autoencoders have been paired with classification heads for tasks such as brain tumor diagnosis, achieving strong results on degraded images [19]. Rahimzadeh and Attar combined DenseNet and ResNet features via a shared latent space, improving classification performance on COVID-19 chest X-ray datasets [9].

Attention-aware enhancement mechanisms have also gained increasing traction in improving object recognition tasks across various domains. Fu [11] demonstrated that incorporating a denoising autoencoder significantly enhanced classification accuracy on noisy galaxy imagery. Praharsha and Poulose [20] introduced CBAM-VGG16 for distracted driver detection, achieving over 98% accuracy on the AUCD2 dataset. In remote sensing, Hu et al. [21] developed a CBAM-integrated hybrid network combining residual and dilated convolutions, which improved hyperspectral image classification in spectrally cluttered conditions.

C. Transfer Learning Based Autoencoders

Transfer learning has been widely adopted in autoencoder-based systems to improve generalization and representation learning across domains. Fu [11] utilized a denoising convolutional autoencoder built on DenseNet-121 to enhance galaxy images, resulting in notable classification improvements. In histopathology, Lee and Lah [22] fine-tuned a DenseNet encoder within an autoencoder-classifier pipeline for multi-task slide analysis, demonstrating increased learning efficiency. Praharsha and Poulose [20] embedded CBAM into VGG16 to support attention-based classification in distracted driving scenarios. Tran et al. [23] evaluated multiple encoder backbones, including ResNet-50, EfficientNet-B3, and DenseNet-121, for lung cancer detection using DICOM images.

Xie et al. [24] introduced MEEAFusion, an architecture that integrates CBAM-like attention and multi-scale fusion within an autoencoder framework for infrared-visible image fusion. DenseNet-121 has been commonly employed as a preferred encoder across these efforts due to its densely connected layers and feature reuse capabilities, proving effective in domains such as astronomy [11] medical imaging [22], [24], and spectral fusion tasks [23].

D. Multi-label Classification and Label Sparcity Handling

Multi-label classification assigns multiple class labels to each image, with each label represented as a binary indicator of presence or absence. In this context, standard loss functions such as Binary Cross-Entropy (BCE) are commonly used, but their performance has been analyzed for scenarios involving label sparsity. For example, Wang et al. [25] discussed how BCE treats active and inactive labels equally, which may lead to suboptimal learning when most labels are negative. Yessou et al. [26] further evaluated several loss functions, identifying similar limitations when applied to satellite imagery datasets.

To address class imbalance, various alternative loss functions have been proposed. Asymmetric Loss (ASL), introduced by Ben-Baruch et al. [27], adjusts gradient contributions by down-weighting easy negatives and emphasizing hard positives, leading to improved mAP scores across datasets such as MS-COCO, Pascal-VOC, NUS-WIDE, and Open Images. Huang et al. [28] later refined this with Asymmetric Polynomial Loss (APL), incorporating polynomial coefficients for further control over gradient scaling. Ji et al. [29] applied these loss functions in remote sensing, while similar techniques have also been adopted in medical imaging to enhance detection of rare labels and performance on sparsely annotated data.

E. Gaps and Limitation

Although prior studies have advanced object recognition, image enhancement, and multi-label classification, several limitations remain. Few existing models integrate enhancement and recognition into a unified architecture, especially for visually complex scenes with occlusion, noise, and clutter. While encoder-decoder structures and attention mechanisms have been explored independently, their combined use in multitask pipelines is limited. Additionally, many approaches are domain-specific and lack generalizability across different real-world environments. Most models focus on either low-level enhancement or high-level classification, missing the performance gains of joint optimization.

Another limitation is the challenge of label sparsity in multi-label classification. Standard loss functions like Binary Cross-Entropy often struggle when most classes are inactive, making it harder to learn from rare but important labels. Although advanced losses such as Asymmetric Loss (ASL) and Asymmetric Polynomial Loss (APL) have shown success in other domains, they are still rarely applied in complex scene recognition. Additionally, while DenseNet-121 and attention modules like CBAM perform well individually, their integration into autoencoder-based multitask frameworks remains limited. Addressing these gaps requires a unified, attention-aware model that applies transfer learning and specialized loss functions for more robust performance in cluttered, degraded environments.

III. Architecture

The proposed architecture in Fig. 1 employs a pre-trained DenseNet121 as a feature extraction backbone. This backbone serves dual purposes: (1) encoding the noisy input into a latent representation for clean image reconstruction, and (2) extracting discriminative features for classification. We extract intermediate feature maps from four major transition layers of the DenseNet, capturing hierarchical spatial information. To preserve spatial resolution and semantic depth, we remove the classification head of the backbone and use the feature pyramid as input to the decoder. All batch normalization layers remain trainable to adapt to the domain distribution, while the initial training phase freezes the rest of the backbone to stabilize learning.

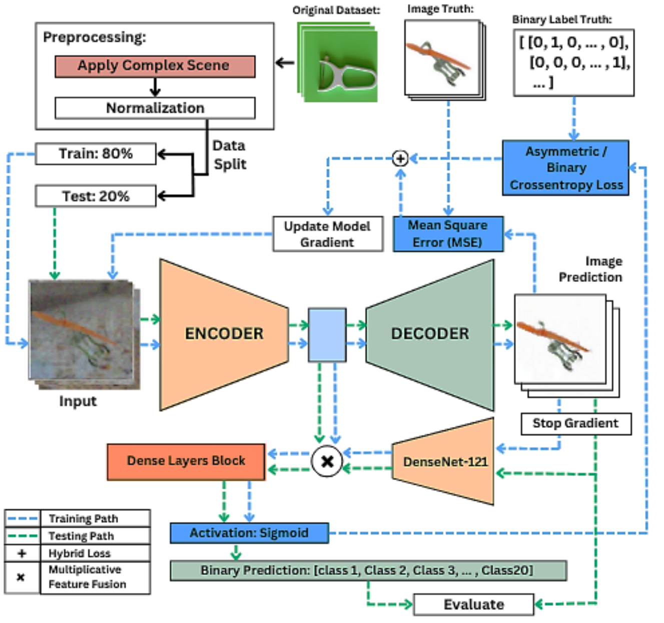


Fig. 1. Overall model pipeline showing the training objectives, autoencoder structure, classification paths, and gradient flow during training and inference.

A. Encoder Backbone

Let the input image be denoted as $\boldsymbol{x} \in \mathbb{R}^{64\times64\times3}$, representing a three-channel red, green and blue (RGB) image. We adopted a DenseNet-121 encoder ($\boldsymbol{E}$), pretrained on ImageNet and used here without its classification head to extract hierarchical representations. DenseNet-121 is well-suited for feature extraction because its layers are densely connected, allowing each layer to reuse feature maps from all previous layers in the same block, which improves feature propagation and reduces vanishing gradients [30]. As $\boldsymbol{x}$ passes through successive dense blocks and transition layers, we collect feature maps from five key stages, denoted as (1), where each $\boldsymbol{f_i}$ is the output of the i-th dense block within DenseNet-121. These feature maps are progressively deeper and semantically richer, capturing spatial context at multiple scales. The final representation $\boldsymbol{f_5}$ serves as the encoder's bottleneck output. To generate a compact semantic descriptor, a global average pooling operation (GAP) is applied to $\boldsymbol{f_5}$ in Eq. (2).

$$\boldsymbol{f_i} = \boldsymbol{E_i}(\boldsymbol{f_{i-1}}), \qquad \boldsymbol{f_0} = \boldsymbol{x}, \qquad for\ \boldsymbol{i} = 1,2,3,4,5 \tag{1}$$

$$\boldsymbol{z} = GAP(\boldsymbol{f_5}) \in \mathbb{R}^{1024} \tag{2}$$

The final output of the encoder produces a 1024-dimensional vector $\boldsymbol{z}$ that summarizes the spatial content of each channel. This latent representation is used to drive the reconstruction in the decoder, and, in parallel, informs downstream classification tasks. During the initial training phase, the encoder weights are frozen to retain robust pretrained features, allowing the decoder and classifier to adapt without perturbing the foundational representations.

B. Attention Aware Decoder

The decoder's primary objective is to reconstruct a clean version of the input image, denoted as $\hat{\boldsymbol{x}} \in \mathbb{R}^{64\times64\times3}$, from the latent vector ($\boldsymbol{z}$). This step is essential for removing noise and enhancing image quality before classification. Structurally, the decoder is paired with a DenseNet-121 encoder, where skip connections, $\boldsymbol{f_i}$ are extracted by encoder to paired with decoder blocks. These skip connections significantly improve their ability to preserve fine spatial details and reduce reconstruction errors. Notably, Ham et al. [31] theoretically and empirically analyzed linear denoising autoencoders with skip connections, showing that bypass pathways stabilize reconstruction performance. To further refine reconstruction, the decoder integrates the Convolutional Block Attention Module (CBAM) [7] which enhances feature selection through two types of attention, channel and spatial located after the last decoder block in Fig. 2.

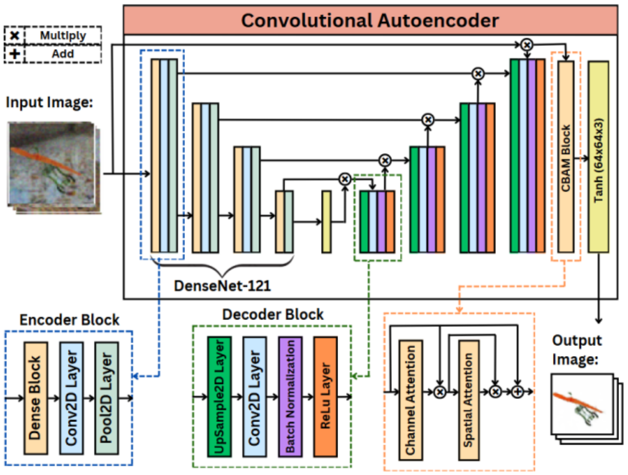


Fig. 2. Detailed architecture of the DenseNet-based encoder used in the autoencoder. Features from various depths are extracted and passed to the decoder through skip connections.

Channel attention learns a map $\boldsymbol{M_c} \in \mathbb{R}^{1\times1\times C}$ to assign importance weights to each feature channel, emphasizing those that contribute most to accurate reconstruction. Spatial attention computes a map $\boldsymbol{M_s} \in \mathbb{R}^{H\times W\times 1}$ to focus on the most relevant spatial regions of the image, such as edges and textures. These attention mechanisms are applied to the intermediate feature map $\boldsymbol{F} \in \mathbb{R}^{H\times W\times C}$, producing a refined feature map $\boldsymbol{F_{att}}$ through element-wise multiplication calculated in Eq. (3). To preserve the original feature representation while enriching it with attention aware enhancements, a residual connection is applied between the input feature map $\boldsymbol{F}$ and the refined output $\boldsymbol{F_{att}}$, yielding the final attended feature map $\boldsymbol{F_{out}}$ in Eq. (4). This combined representation is then passed through the final convolutional layer with a $tanh$ activation to produce the decoded image $\boldsymbol{\hat{x}}$. Finally, the output image $\boldsymbol{\hat{x}}$ is generated using a convolutional layer followed by a $tanh$ activation, ensuring that pixel values fall between –1 and 1, and match the original input's shape and color space for loss computation in Eq. (5).

$$\boldsymbol{F_{att}} = \boldsymbol{F} \odot \boldsymbol{M_c} \odot \boldsymbol{M_s} \tag{3}$$

$$\boldsymbol{F_{out}} = \boldsymbol{F_{att}} + \boldsymbol{F} \tag{4}$$

$$\boldsymbol{\hat{x}} = tanh(Conv(\boldsymbol{F_{out}})) \tag{5}$$

C. Classification from Reconstructed Inputs

The classification module uses the reconstructed image $\boldsymbol{\hat{x}}$, instead of the original noisy input, to benefit from improved visual quality during prediction. To prevent classification gradients from affecting the decoder, a stop-gradient operation is applied, treating $\boldsymbol{\hat{x}}$ as a fixed input during backpropagation and isolating the reconstruction learning. This detached image, $\boldsymbol{\hat{x}_{stop\_gradient}}$ is then passed through a new DenseNet-121 as decoded image's feature extractor, identical in architecture but with independent weights to extract a classification feature vector, $\boldsymbol{z_c}$ as described in Eq. (6). To enhance robustness, feature fusion is performed by element-wise multiplication of the latent vectors, $\boldsymbol{z}$ from the autoencoder and decoded image's extracted feature, $\boldsymbol{z_c}$ as shown in Eq. (7). This operation amplifies shared information while suppressing noise and inconsistencies.

$$\boldsymbol{z_c} = GAP(\boldsymbol{E}(\boldsymbol{\hat{x}_{stop\_gradient}})), \qquad \boldsymbol{z_c} \in \mathbb{R}^{1024} \tag{6}$$

$$\boldsymbol{z_{fused}} = \boldsymbol{z} \odot \boldsymbol{z_c} \tag{7}$$

The fused feature vector $\boldsymbol{z_{fused}}$ is then fed into fully connected layers with nonlinear activations and dropout regularization to generate the final multi-label classification prediction at Eq. (8) where W and B are learnable parameters of the classifier, and σ is the sigmoid activation function applied elementwise, mapping the outputs to probabilities between 0 and 1 for each class. The output $\hat{y} \in [0,1]^C$ represents predicted probabilities across C classes.

$$\hat{y} = \sigma(W^T \boldsymbol{z_{fused}} + b) \tag{8}$$

D. Multitask Hybrid Loss

The proposed model is optimized through a multitask learning framework that jointly addresses image reconstruction and multi-label classification. A hybrid loss function is formulated to guide the network in learning both low-level structural details and high-level discriminative features for accurate classification. This is achieved by minimizing reconstruction and classification losses simultaneously, each weighted within the final objective. As a result, the network preserves semantic content while improving its ability to differentiate object categories. This integration of learning signals helps the model build richer, more generalizable feature representations essential for robust recognition in complex visual scenes.

$$\boldsymbol{\mathcal{L}_{total}} = (\lambda_{recon} \cdot \boldsymbol{\mathcal{L}_{recon}}) + (\lambda_{class} \cdot \boldsymbol{\mathcal{L}_{class}}) \tag{9}$$

In Eq. (9) formulation, $\mathcal{L}_{total}$ denotes the complete training loss used to optimize the network parameters. The term $\mathcal{L}_{recon}$ corresponds to the image reconstruction loss, which guides the autoencoder decoder to regenerate the clean image from the latent representation. The term $\mathcal{L}_{class}$ represents the classification loss, used to supervise the semantic categorization based on the reconstructed features. The scalar coefficients λ_{recon} and λ_{class} are hyperparameters used to weigh the influence of each loss. In our experiments, we set λ_{recon} = 1.5 and λ_{class} = 0.5, which prioritizes structural reconstruction while maintaining a significant emphasis on classification.

To quantify the reconstruction performance, we employ the Mean Squared Error (MSE) loss [32], a standard measure for evaluating image similarity in pixel space. This loss is defined as Eq. (10). Here, x_i denotes the i-th original input image in a batch, while $\hat{x}_i$ is its corresponding reconstructed output produced by the decoder. Both are assumed to be RGB images with dimensions 64×64×3. The notation $\| \cdot \|^2{}_2$ refers to the squared L2 norm computed over all pixel values, which penalizes deviations between each input pixel and its reconstruction. The term N represents the batch size. This formulation encourages the decoder to generate outputs that are visually and numerically close to the original clean image inputs.

$$\boldsymbol{\mathcal{L}_{recon}}(x, \hat{x}) = (1/N) \sum \|x_i - \hat{x}_i\|^2{}_2 \tag{10}$$

For the classification task, we evaluate two loss functions, the Asymmetric Loss (ASL) by Ridnik et al. [27] and the conventional Binary Cross-Entropy (BCE). The ASL is designed for multi-label classification with severe label imbalance, where it emphasizes hard negatives while suppressing easy ones through focusing parameters and a threshold margin. It is defined as in (11). Here, $y \in \{0,1\}^C$ and $\hat{y} \in [0,1]^C$ denote the ground truth and predicted probability vectors across C classes. γ^+ and γ^- are focusing factors for positives and negatives, m is the margin that suppresses low-confidence negatives, and ε ensures numerical stability. We use $\gamma^+ = 0$, $\gamma^- = 2$, $m = 0.01$, and $\varepsilon = 1e{-8}$ as recommended in [27]. In contrast, the BCE loss [32], given by Eq. (12) treats each class independently and penalizes incorrect predictions equally, regardless of imbalance. We compare both losses to evaluate their influence on classification performance within our multitask framework.

$$\boldsymbol{\mathcal{L}_{class\_ASL}}(y, \hat{y}) = -[y \cdot \log(\hat{y}) \cdot (1 - \hat{y})^{\gamma^+} + (1 - y) \cdot log(1 - \hat{y} + \varepsilon) \cdot (\hat{y} - m)^{\gamma^-}] \tag{11}$$

$$\boldsymbol{\mathcal{L}_{class\_BCE}}(y, \hat{y}) = -\sum[y_k \log(\hat{y}_k) + (1 - y_k) \; log(1 - \hat{y}_k)] \tag{12}$$

By combining the two loss components in Eq. (9), the model learns to perform both tasks within a unified process. A key implementation detail is the use of a stop-gradient operation between the reconstruction and classification branches. This ensures the classification loss does not backpropagate through the decoder, keeping its training focused solely on reconstruction. Meanwhile, the encoder is shared across both branches, enabling learning of joint features useful for both semantic discrimination and visual reconstruction. This separation of gradient flow, along with shared encoder supervision, promotes stable, task-specific optimization and allows both branches to converge effectively while reinforcing the shared representations.

IV. Experimental Setup

A. *Dataset and Simulation of Complex Scenes*

The dataset used in this work originates from the Edinburgh Kitchen Utensil Database (EKUD), which consists of isolated images of kitchen objects with plain backgrounds. The original EKUD dataset, however, presents two major limitations which are class imbalance, where some classes have significantly fewer instances than others, and lack of scene complexity, with only single-object images and minimal background variation. To address these limitations and generate a more realistic, complex training dataset, we developed a Complex Scene Generator pipeline in Fig. 3. First, all object images from EKUD were preprocessed by removing their backgrounds and converting them to transparent PNGs. This enabled spatial transformations and compositing onto new backgrounds.

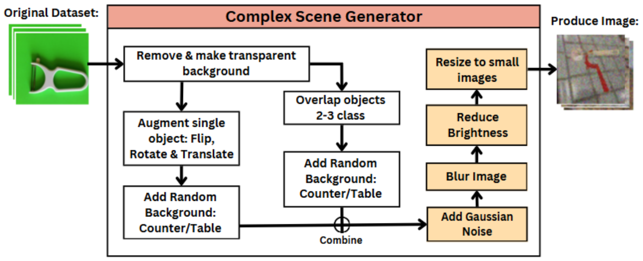

Original Dataset:
Complex Scene Generator
Produce Image:
Remove & make transparent background
Augment single object: Flip, Rotate & Translate
Add Random Background: Counter/Table
Overlap objects 2-3 class
Add Random Background: Counter/Table
Combine
Resize to small images
Reduce Brightness
Blur Image
Add Gaussian Noise

Fig. 3. Complex scene generator.

Each object was augmented using geometric transformations such as flipping, rotation, and translation to increase diversity. To simulate realistic environments, objects were overlaid on random background images of counters or tabletops, mimicking domestic kitchen scenes. To add complexity, 2–3 objects from different classes were randomly combined with spatial overlap, creating occlusions typical of real-world clutter. These composites then underwent image degradation, including resizing to 64×64 pixels, brightness reduction, Gaussian blurring, and added noise to mimic conditions like low light or motion blur. This process produced a balanced dataset of 6,000 images with equal class distribution. Each image contained one to three overlapping objects from different classes, introducing intra-class variation and inter-class spatial relationships vital for robust classification and reconstruction. Finally, the dataset was split into training and testing sets using an 80:20 ratio with stratified class distribution.

B. *Model Training*

The proposed dual-branch architecture was optimized using a two-phase training strategy to enhance image reconstruction and multi-class classification. To improve evaluation reliability and reduce overfitting, 5-fold StratifiedKFold cross-validation was applied to the training set, generating five training runs with preserved class distribution. In the first phase, the DenseNet encoder was frozen, allowing the decoder and classifier to learn task-specific features without altering backbone representations. A composite loss combined Mean Squared Error (MSE) for reconstruction and Sparse Categorical Crossentropy for classification, weighted 1.5 and 0.5, respectively. The model trained for 30 epochs using the Adam optimizer with a 0.001 learning rate. This phase aimed to stabilize reconstruction and extract semantically rich latent features. Structural Similarity Index (SSIM) assessed reconstruction quality, while standard metrics tracked classification accuracy.

In the second phase, the full model was unfrozen for end-to-end fine-tuning to enable joint optimization of low-level and high-level features across both branches. The learning rate was reduced to 0.0001 to support stable convergence and prevent disruption of previously learned features. This phase refined the encoder-decoder synergy and strengthened the alignment between reconstructed structures and class-relevant semantics. Following the completion of training with Asymmetric Loss, a parallel experiment was conducted using Binary Cross Entropy as the classification loss, while retaining the same optimizer settings, loss weights, and training schedule. This allowed a direct performance comparison under controlled conditions, highlighting the impact of loss design on multi-label recognition in cluttered scenes.

Pseudocode 1: Dual-Phase Training

Input: Noisy images I_{noisy}, (Clean Target Image I_{clean}, Clean labels L_{true})

Output: Trained model with decoder output D and classifier C

```
Initialize encoder with DenseNet121 (frozen), decoder blocks, and classification branch
Compile model with:
    - Losses: MSE, ASL
    - Loss weights: D: 1.5, C: 0.5}
    - Optimizer: Adam (lr = 0.001)
Train model for $N_1$ (30) epochs, freezing encoder
Save best weights based on highest validation accuracy
Unfreeze all layers of model
Compile model with:
    - Updated optimizer: Adam (lr = 0.0001)
For epoch = $N_1$ to $N_2$(100) do
    For each batch ($I_{noisy}$, $L_{true}$) in training set do
        Compute decoder output $D_{pred}$ and classifier output $C_{pred}$
        Compute:
            - MSE loss between $D_{pred}$ and $I_{clean}$
            - ASL loss between $C_{pred}$ and $L_{true}$
            - SSIM metric for $D_{pred}$
        Backpropagate total loss = 1.5 × MSE + 0.5 × ASL
        Update weights
    Evaluate on validation set
    Save model if validation accuracy improves
End for
```

The dual-phase training strategy in Pseudocode 1 reflects a common practice in transfer learning and multi-task deep learning. Freezing the pretrained encoder early allows the decoder and classifier to learn task-specific features without modifying core representations. Once stabilized, full fine-tuning with a reduced learning rate improves alignment between reconstructed outputs and classification. Deepankan and Agarwal found that freezing the backbone initially and unfreezing later boosted accuracy and reduced overfitting on limited data [33]. Likewise, Taha et al. employed staged fine-tuning with pretrained language models, showing that freezing phases improved learning stability and generalization in imbalanced tasks [34]. In remote sensing, Khotimah et al. used a dual-stage framework combining masked autoencoder pretraining and few-shot learning to enhance hyperspectral image classification with limited supervision [35]. These studies support phased training as a reliable strategy for stabilizing early optimization and enhancing global feature refinement.

C. Evaluation Metrics

To assess model performance, five-fold Stratified K-Fold cross-validation was applied to the training data to ensure balanced class representation and generalizability. All metrics are reported as averages across the five folds. Since the model performs both tasks jointly, each output was evaluated using suitable criteria. For reconstruction, evaluation focused on perceptual similarity and signal fidelity between input and reconstructed images. Structural Similarity Index Measure (SSIM) was used to compare luminance, contrast, and structural details, as defined in Eq. (13), where μ_x and μ_y are the means, σ_x and σ_y their standard deviations, σ_{xy} the cross-covariance, and C_1, C_2 are constants for stability. Additionally, Peak Signal-to-Noise Ratio (PSNR) was used to measure reconstruction quality based on the ratio of maximum signal to noise power. As shown in Eq. (14), PSNR uses MAX_I as the maximum pixel value and MSE as the mean squared error. Combined, these metrics evaluate perceptual quality and pixel-level accuracy.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{13}$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{14}$$

For the multi-label classification component, an adaptive top-T_1 binarization scheme was applied. Each output prediction vector was binarized by retaining the top-T_1 elements with the highest predicted probabilities, where T_1 corresponds to the number of positive ground truth labels in the input, as shown in Eq. (15). This approach ensured prediction-ground truth cardinality alignment, enabling fair comparisons across samples with varying label sparsity. After binarization, classical metrics such as precision, recall, and F1-score were computed using Eq. (16), Eq. (17), and Eq. (18), respectively. These metrics were evaluated under both macro and micro averaging. Macro-averaging calculates per-class metrics and averages them, while micro-averaging aggregates all class contributions before computing the metric, capturing overall performance. To assess label-wise correctness, we also reported standard accuracy Eq. (19) and Hamming Loss (20), which quantify the fraction of misclassified labels.

$$T_1 = \| y_{true} \|_0 \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{18}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

$$Hamming\ Loss = \frac{1}{N \cdot C}\sum_{i=1}^{N}\sum_{j=1}^{C} 1[y_{ij} \neq \hat{y}_{ij}] \tag{20}$$

To assess ranking quality, which is important in multi-label tasks with overlapping label distributions, we computed the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) and Mean Average Precision (mAP) in micro settings. These metrics, based on raw prediction scores without thresholding, evaluate the model's ability to rank relevant labels above irrelevant ones, following best practices in recent multi-label research [36]. Macro metrics average performance across classes, while micro metrics aggregate true and false positives globally. A multi-label confusion matrix was generated using the MLCM (Multi-Label Confusion Matrix) library [37], enabling per-class analysis of co-occurrence between true and predicted labels. The matrix was visualized as a heatmap to expose common misclassifications and confusion patterns. MLCM offers sparse-aware, normalized outputs well suited for complex multi-label data, following protocols from [38]. Finally, a classification report summarized per-class precision, recall, and F1-scores for detailed performance interpretation.

V. Results and Discussion

A. Multi-Labeled Classification

The classification performance of the proposed model in Table I was compared with two baselines: a standard convolutional autoencoder and a standalone DenseNet-121 classifier. All models used binary cross-entropy loss, while the proposed model was also evaluated with asymmetric loss to test sensitivity to label imbalance. Metrics included accuracy, Hamming Loss, macro-averaged ROC-AUC, and macro mean average precision (mAP). The autoencoder achieved 88.99% accuracy, 0.1101 Hamming Loss, 0.8803 ROC-AUC, and 0.5278 mAP. The standalone DenseNet-121 reached 95.37%

accuracy, 0.0463 Hamming Loss, 0.9678 ROC-AUC, and 0.8366 mAP. The proposed model, combining reconstruction and classification, reached 96.41% accuracy, 0.0359 Hamming Loss, 0.9805 ROC-AUC, and 0.8862 mAP.

The highest performance was achieved when the proposed model was trained using the asymmetric loss function. Accuracy increased to 96.72%, and Hamming Loss decreased to 0.0328. The macro ROC-AUC reached 0.9809, while the macro mAP peaked at 0.8925. To further evaluate classification performance, Table II reports macro and micro averages of precision, recall, and F1-score. The standard convolutional autoencoder yielded a macro F1-score of 0.4296 and a micro F1-score of 0.4496. The DenseNet-121 baseline achieved a macro F1-score of 0.7637 and a micro F1-score of 0.7683. The proposed model with binary cross-entropy achieved a macro F1-score of 0.8196. When trained with asymmetric loss, the model attained a macro F1-score of 0.8350 and a micro F1-score of 0. 8362.

TABLE I. EVALUATION OF CLASSIFICATION RESULT

Model	**Classification Loss Function**	**Evaluation**			
		Accuracy	***Hamming Loss***	***Macro ROC-AUC***	***Macro mAP***
Standard Convolutional Autoencoder (CAE)	BCE	88.99%	0.1101	0.8803	0.5278
DenseNet-121 (Alone)	BCE	95.37%	0.0463	0.9678	0.8366
Proposed Model	BCE	96.41%	0.0359	0.9805	0.8862
	ASL	96.72%	0.0328	0.9811	0.8925

TABLE II. PERFORMANCE MATRICS OF CLASSIFICATION RESULT

Model		**Performance Matrics**					
		Macro Precision	***Macro Recall***	***Macro F1***	***Micro Precision***	***Micro Recall***	***Micro F1***
Standard CAE	BCE	0.5658	0.4496	0.4296	0.4496	0.4496	0.4496
DenseNet -121 (Alone)	BCE	0.7798	0.7683	0.7637	0.7683	0.7683	0.7683
Proposed Model	BCE	0.8311	0.8204	0.8196	0.8204	0.8204	0.8204
	ASL	0.8435	0.8362	0.8350	0.8362	0.8362	0.8362

The evaluation of classification performance using Asymmetric Loss (ASL) and Binary Cross-Entropy (BCE) was conducted with multi-label ROC-AUC metrics. Although the dataset was globally balanced across 20 categories, many binary labels had just 1 to 3 positives, causing localized imbalance. The micro-average AUC under ASL reached 0.9811, slightly higher than 0.9802 with BCE. ASL produced higher per-class AUCs in 12 of 19 overlapping categories. For example, Potato Peeler, Serving Spoon, and Wooden Spoon had stronger AUCs under ASL. The ROC curves shown in Fig. 4 and Fig. 5 further illustrate classifier performance across label distributions. Fig. 4, derived from the ASL-trained model, exhibits tighter and more consistent ROC curves, especially for low-prevalence labels. In contrast, Fig. 5, based on BCE training, presents slightly flatter and more variable ROC curves for low-activity classes such as Bread Knife (0.9720 ASL vs. 0.9786 BCE) and Dinner Fork (0.9714 ASL vs. 0.9768 BCE). BCE outperformed ASL in a few high-density labels like Bottle Opener.

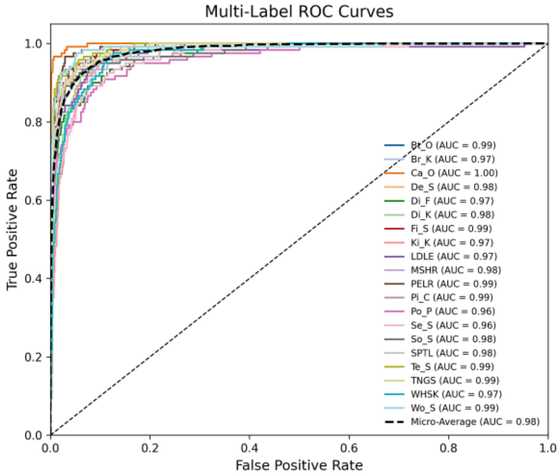


Fig. 4. ROC-AUC Graph of proposed model that uses asymmetric loss function.

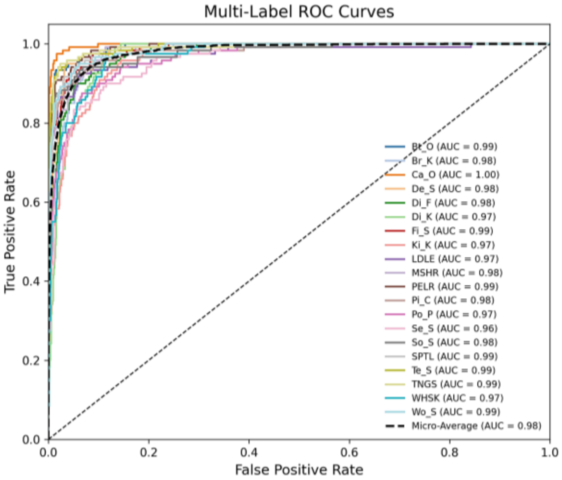


Fig. 5. ROC-AUC Graph of proposed model that uses binary crossentropy loss function.

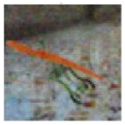
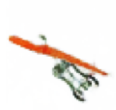
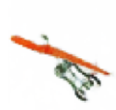
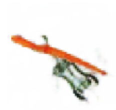
B. *Enhanced Image Reconstruction*

The reconstruction performance was evaluated using SSIM and PSNR, as detailed in Table III. Input images contained complex scenes with diverse backgrounds, object occlusion, and visual clutter. The decoder consistently generated clean object representations over simplified backgrounds, as also illustrated in Table IV. The baseline autoencoder, trained using mean squared error and binary cross-entropy loss, achieved a reconstruction score of 0.9175 SSIM and 6.0670 dB PSNR. The proposed dual-path model improved these scores to 0.9213 SSIM and 7.0028 dB PSNR. When asymmetric loss was applied during classification, the model achieved 0.9372 SSIM and 7.1012 dB PSNR. The model’s ability to reconstruct clean outputs from noisy, cluttered scenes was enhanced by integrating the Convolutional Block Attention Module (CBAM) into the decoder. DenseNet-121 was excluded, as it lacks a decoding pathway.

TABLE III. IMAGE RECONSTRUCTION PERFORMANCE EVALUATION

Model	Loss Function	Evaluation	
		SSIM	***PSNR (dB)***
Standard Convolutional Autoencoder (CAE)	MSE+BCE	0.9175	6.0670
Proposed Model	MSE+BCE	0.9213	7.0028
	MSE+ASL	0.9372	7.1012

TABLE IV. IMAGE RECONSTRUCTION BASED ON LOSS FUNCTIONS

Input	MSE+BCE	MSE+ASL	Standard CAE
			

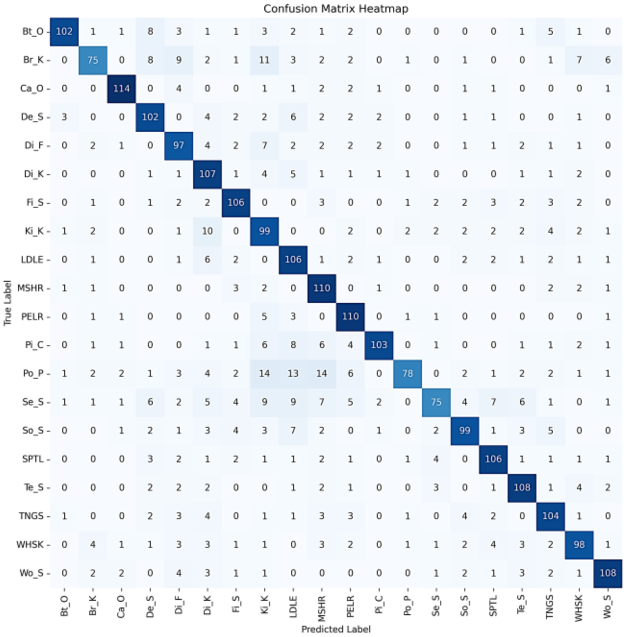


Fig. 6. Confusion matrix from proposed model with loss function MSE+ASL.

The confusion matrix in Fig. 6 shows the best classification results of the proposed model using Mean Squared Error (MSE) for reconstruction and Asymmetric Loss (ASL) for classification. Most object classes exhibit strong diagonal dominance, indicating high true positive rates, especially for Can Opener, Masher, Peeler, and Wooden Spoon, each surpassing 100 correct predictions. These results demonstrate the model's robustness against complex scene conditions such as overlapping objects, image degradation, and varied backgrounds. Some confusion remains between visually or semantically similar classes, like Bread Knife misclassified as Can Opener or Peeler, and Serving Spoon overlapping with Soup Spoon, underscoring ongoing challenges with intra-class similarity in noisy settings. Overall, the matrix supports the effectiveness of the multi-task architecture, where combining image enhancement and classification via latent fusion improves discriminative performance.

VI. DISCUSSIONS

A. Interpretation of Multi-Label Classification Results

The multi-label classification results highlight the clear advantage of integrating reconstruction and classification within a unified model, particularly when dealing with complex data distributions and label sparsity. While the standard convolutional autoencoder lacked explicit label supervision and performed poorly in classification, the standalone DenseNet-121 improved predictive quality through deeper, more expressive features. However, DenseNet-121 operated in a single-task setup and did not benefit from auxiliary learning signals provided by image reconstruction. In contrast, the proposed dual-path architecture, which combines visual reconstruction with classification, consistently outperformed both baselines. This outcome suggests that reconstruction plays a crucial role in shaping more robust latent representations, which in turn support better discrimination, especially for less frequent object classes.

An equally important factor in the model's success is the choice of loss function. The shift from Binary Cross-Entropy to Asymmetric Loss introduced a more nuanced training dynamic, where the model learned to emphasize informative, minority class signals while reducing overfitting to dominant negatives. This led to higher macro-averaged metrics and stronger per-class AUCs for rare categories. ROC curve comparisons further confirmed this behavior, showing tighter, more stable curves under ASL for low-prevalence labels. Meanwhile, BCE slightly favored common classes due to its uniform treatment of all errors. Taken together, these observations demonstrate that the model's strength arises not solely from architectural complexity but from a careful alignment between task formulation, loss function design, and the inherent demands of multi-label learning in imbalanced and cluttered visual contexts.

B. Interpretation of Reconstruction Performance

The reconstruction results demonstrate that the proposed dual-path architecture is not only effective for classification but also capable of producing structurally coherent and visually clean reconstructions in complex scenes. Compared to the baseline autoencoder, the improved SSIM and PSNR scores reflect the model's strong capacity to abstract essential features while filtering out background noise and irrelevant details. The use of a shared encoder that supports both tasks allows the network to capture richer semantic representations, which benefit the decoder's ability to reconstruct meaningful content even in visually degraded or cluttered environments. This integration shows how multi-task learning encourages a more organized and informative latent space that serves both enhancement and recognition.

An unexpected but significant observation was the positive influence of asymmetric loss, applied only to the classification task, on the quality of reconstructed outputs. This implies that learning to better separate hard-to-classify examples also leads to more structured and semantically aligned feature embeddings, indirectly supporting the reconstruction process. Additionally, the integration of CBAM into the decoder reinforced the model's ability to selectively attend to important spatial and channel-specific regions. As a result, reconstructions were not only cleaner but also better aligned with the salient components of the original input. This attention-driven refinement is particularly beneficial in real-world scenarios where accurate reconstruction supports interpretability and aids in downstream decision-making tasks.

C. Bridging Gaps Through Multi-Task Representation

The proposed model directly addresses several unresolved challenges in existing research by unifying classification and reconstruction within a shared encoder-decoder architecture. Most prior approaches treat enhancement and recognition as separate tasks, limiting their ability to generalize in cluttered or degraded scenes. In contrast, the proposed model leverages joint optimization to build more expressive and spatially consistent latent features. The reconstruction task helps maintain structure and denoise complex scenes, while the classification branch, guided by asymmetric loss, highlights category-specific regions, especially for rare labels. This dual influence supports a more semantically aware encoder that improves both predictive accuracy and visual clarity in reconstructed images.

By integrating DenseNet-121 and attention modules like CBAM into a multi-task framework, the model closes a critical gap identified in previous studies where such components were rarely combined. Through shared learning, reconstruction helps the model preserve spatial structure and object layout, while classification encourages it to focus on distinctive features that separate one object class from another. Consistent performance gains across low-prevalence classes and complex visual backgrounds indicate that the model benefits from this interdependence. Rather than relying on handcrafted decoupling or task-specific heuristics, this approach uses end-to-end learning to harness the complementary strengths of each task. As a result, the model becomes more robust, interpretable, and suitable for real-world deployment where visual data is often noisy, sparse, and heterogeneous.

VII. Conclusion and Future Work

In this study, we proposed a multi-task deep learning architecture that integrates a convolutional autoencoder and DenseNet backbone with latent space fusion to tackle object classification challenges in complex scenes. These include cluttered backgrounds, overlapping objects, and degraded image quality due to noise, blur, or low brightness. The architecture was designed to jointly enhance visual clarity and extract discriminative features within a unified pipeline. Trained in two phases and evaluated using diverse reconstruction and classification metrics, the model outperformed baseline methods. The proposed architecture is particularly designed for image data containing overlapping objects and visual noise, where conventional classification models often underperform. Results showed notable improvements in accuracy, macro and micro F1-scores, and reduced class confusion, validating the benefit of integrated image enhancement and feature learning for robust multi-class object recognition in complex scenes.

Despite improvements, some limitations remain. The model still exhibits misclassifications in scenes with visually similar or overlapping objects, for example Dessert Spoon versus Dinner Spoon, indicating difficulty in distinguishing fine-grained object boundaries under occlusion. This reflects a broader challenge in handling complex scenes involving clutter and degraded visual quality, which the proposed architecture is specifically designed to address. In addition, the reliance on fully supervised learning requires both labeled class data and clean reconstruction targets, resources that may be limited or noisy in real-world deployments. To improve scalability, future work could investigate semi-supervised or weakly supervised strategies that reduce dependence on exhaustive annotation. Incorporating spatial-aware attention or feature disentanglement methods may further improve robustness in overlapping object conditions. Finally, expanding the dataset or extending the architecture to video inputs could strengthen recognition under temporal and contextual variation.

Acknowledgment

Funding Statement and Conflicts of Interest: This work was funded by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UMT/03/2). The also authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Rosli, R. Ali, M. S. Hitam, A. Mat Deris, N. H. Abd Rahim, and U. Haruna, "Recognizing objects in complex scenes: A Recent Systematic Review," J. Adv. Res. Appl. Sci. Eng. Technol., pp. 208–226, Oct. 2024, doi: 10.37934/araset.61.2.208226.

[2] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020. doi: 10.1109/CVPR42600.2020.00223.

[3] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 2020. doi: 10.1609/aaai.v34i07.6865.

[4] N. Rashid, M. A. F. Hossain, M. Ali, M. Islam Sukanya, T. Mahmud, and S. A. Fattah, "AutoCovNet: Unsupervised feature learning using autoencoder and feature merging for detection of COVID-19 from chest X-ray images," Biocybern. Biomed. Eng., vol. 41, no. 4, pp. 1685–1701, 2021, doi: 10.1016/j.bbe.2021.09.004.

[5] Z. Fei, J. Wang, K. Liu, E. Attahi, and B. Huang, "Deep feature fusion-based stacked denoising autoencoder for tag recommendation systems," IET Cyber-systems Robot., 2023, doi: 10.1049/csy2.12095.

[6] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," Proc. - 2021 IEEE Winter Conf. Appl. Comput. Vision, WACV 2021, pp. 3559–3568, Sep. 2021, doi: 10.1109/WACV48630.2021.00360.

[7] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018. doi: 10.1007/978-3-030-01234-2_1.

[8] L. Kong and J. Cheng, "Classification and detection of COVID-19 X-Ray images based on DenseNet and VGG16 feature fusion," Biomed. Signal Process. Control, 2022, doi: 10.1016/j.bspc.2022.103772.

[9] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray

images based on the concatenation of Xception and ResNet50V2," Informatics Med. Unlocked, 2020, doi: 10.1016/j.imu.2020.100360.

[10] G. Cheng, X. Xie, J. Han, L. Guo, and G. S. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 2020, doi: 10.1109/JSTARS.2020.3005403.

[11] W. Fu, "Denoising convolutional autoencoder for improving the classification performance based on noisy galaxy images," Appl. Comput. Eng., 2023, doi: 10.54254/2755-2721/21/20231156.

[12] R. Vankayalapati and A. L. Muddana, "Denoising of Images Using Deep Convolutional Autoencoders for Brain Tumor Classification," Rev. d'Intelligence Artif., 2021, doi: 10.18280/ria.350607.

[13] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019. doi: 10.1109/CVPR.2019.00374.

[14] M. M. Yusro, R. Ali, and M. S. Hitam, "Comparison of Faster R-CNN and YOLOv5 for Overlapping Objects Recognition," Baghdad Sci. J., vol. 20, no. 3, p. 893, 2023, doi: 10.21123/bsj.2022.7243.

[15] H. Rosli, R. Ali, A. M. Deris, and M. Suzuri Hitam, "Classification of Kitchen Utensils in Noisy Condition Using YOLOv5 with Multiple Deep Learning Backbones," 2024 IEEE Int. Conf. Autom. Control Intell. Syst. I2CACIS 2024 - Proc., pp. 297–302, 2024, doi: 10.1109/I2CACIS61270.2024.10649881.

[16] A. Rosello, J. J. Valero-Mas, A. J. Gallego, J. Sáez-Pérez, and J. Calvo-Zaragoza, "Kurcuma: a kitchen utensil recognition collection for unsupervised domain adaptation," Pattern Anal. Appl., vol. 26, no. 4, pp. 1557–1569, 2023, doi: 10.1007/s10044-023-01147-x.

[17] M. Hami and M. JameBozorg, "Assessing The Impact of CNN Auto Encoder-Based Image Denoising on Image Classification Tasks," 2024, [Online]. Available: http://arxiv.org/abs/2404.10664

[18] K. Liu, Y. Yang, Y. Tian, and H. Mao, "Image Dehazing Technique Based on DenseNet and the Denoising Self-Encoder," Processes, vol. 12, no. 11, 2024, doi: 10.3390/pr12112568.

[19] Y. Tian, "Privacy Preserving Method for Image Recognition based on Denoising Autoencoder," Highlights Sci. Eng. Technol., 2023, doi: 10.54097/hset.v39i.6710.

[20] C. H. Praharsha and A. Poulose, "CBAM VGG16: An efficient driver distraction classification using CBAM embedded VGG16 architecture," Comput. Biol. Med., vol. 180, 2024, doi: 10.1016/j.compbiomed.2024.108945.

[21] Y. Hu, S. Tian, and J. Ge, "Hybrid Convolutional Network Combining Multiscale 3D Depthwise Separable Convolution and CBAM Residual Dilated Convolution for Hyperspectral Image Classification," Remote Sens., vol. 15, no. 19, 2023, doi: 10.3390/rs15194796.

[22] J. Lee and S. Lah, "Denoising AutoEncoder-based Representation Learning for Multi-Task Whole Slide Image Analysis," J. Student Res., vol. 13, no. 1, 2024, doi: 10.47611/jsrhs.v13i1.6140.

[23] V. Kumar, C. Prabha, P. Sharma, N. Mittal, S. S. Askar, and M. Abouhawwash, "Unified deep learning models for enhanced lung cancer prediction with ResNet-50–101 and EfficientNet-B3 using DICOM images," BMC Med. Imaging, 2024, doi: 10.1186/s12880-024-01241-4.

[24] Y. Xie, Z. Fei, D. Deng, L. Meng, F. Niu, and J. Sun, "MEEAFusion: Multi-Scale Edge Enhancement and Joint Attention Mechanism Based Infrared and Visible Image Fusion," Sensors, vol. 24, no. 17, 2024, doi: 10.3390/s24175860.

[25] L. Zhou, X. Zheng, D. Yang, Y. Wang, X. Bai, and X. Ye, "Application of multi-label classification models for the diagnosis of diabetic complications," BMC Med. Inform. Decis. Mak., 2021, doi: 10.1186/s12911-021-01525-7.

[26] H. Yessou, G. Sumbul, and B. Demir, "A Comparative Study of Deep Learning Loss Functions for Multi-Label Remote Sensing Image Classification," in International Geoscience and Remote Sensing Symposium (IGARSS), 2020. doi: 10.1109/IGARSS39084.2020.9323583.

[27] T. Ridnik et al., "Asymmetric Loss For Multi-Label Classification," in Proceedings of the IEEE International Conference on Computer Vision, 2021. doi: 10.1109/ICCV48922.2021.00015.

[28] Y. Huang, J. Qi, X. Wang, and Z. Lin, "Asymmetric Polynomial Loss for Multi-Label Classification," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2023. doi: 10.1109/ICASSP49357.2023.10095437.

[29] J. Ji, W. Jing, G. Chen, J. Lin, and H. Song, "Multi-label remote sensing image classification with latent semantic dependencies," Remote Sens., 2020, doi: 10.3390/rs12071110.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[31] J. Ham, M. Fleissner, and D. Ghoshdastidar, "Impact of Bottleneck Layers and Skip Connections on the Generalization of Linear Denoising Autoencoders," May 2025, [Online]. Available: http://arxiv.org/abs/2505.24668

[32] O. Elharrouss et al., "Loss Functions in Deep Learning: A Comprehensive Review," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.04242

[33] B. N. Deepankan and R. Agarwal, "A two-phase image classification approach with very less data," Adv. Intell. Syst. Comput., vol. 1108 AISC, pp. 384–394, 2020, doi: 10.1007/978-3-030-37218-7_44.

[34] T. ValizadehAslani et al., "Two-Stage Fine-Tuning: A Novel Strategy for Learning Class-Imbalanced Data," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.10858

[35] W. N. Khotimah, M. Bennamoun, F. Boussaid, L. Xu, and F. Sohel, "Dual-Phase Framework for Few-Shot Hyperspectral Image Classification with Spatiospectral Masked Autoencoder and Episode Training," IEEE Trans. Geosci. Remote Sens., 2025, doi: 10.1109/TGRS.2025.3541263.

[36] S. Huang et al., "Application of Label Correlation in Multi-Label Classification: A Survey," Appl. Sci., vol. 14, no. 19, 2024, doi: 10.3390/app14199034.

[37] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3151048.

[38] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," in Pattern Recognition, 2012. doi: 10.1016/j.patcog.2012.03.004.