

Article

A Sparse Attention Mechanism Based Redundancy-Aware Retrieval Framework for Power Grid Inspection Images

Wei Yang ^{1,*}, Zhenyu Chen ^{1,*} , Xiaoguang Huang ², Ming Li ³, Hailu Wang ³ and Shi Liu ¹

¹ Big Data Center, State Grid Corporation of China, Beijing 100031, China; weiyang@sgcc.com.cn (W.Y.); shi-liu@sgcc.com.cn (S.L.)

² State Grid Information & Telecommunication Group Co., Ltd., Beijing 102211, China; huangxiaoguang@sgitg.sgcc.com.cn

³ Information & Telecommunication Branch State Grid Anhui Electric Power Company, Hefei 230061, China; lim0977@ah.sgcc.com.cn (M.L.); wanghl102x@ah.sgcc.com.cn (H.W.)

* Correspondence: czy9907@gmail.com

Abstract

Driven by the rapid advancement of smart grid frameworks, the volume of visual data collected from power system diagnostic equipment has surged exponentially. A substantial portion of these images (30–40%) are redundant or highly similar, primarily due to periodic monitoring and repeated acquisitions from multiple angles. Traditional redundancy removal methods based on manual screening or single-feature matching are often inefficient and lack adaptability. In this paper, we propose a two-stage redundancy removal paradigm for power inspection imagery, which integrates abstract semantic priors with fine-grained perceptual details. The first stage combines an improved discrete cosine transform hash (DCT Hash) with the multi-scale structural similarity index (MS-SSIM) to efficiently filter redundant candidates. In the second stage, a Vision Transformer network enhanced with a hierarchical sparse attention mechanism precisely determines redundancy via cosine similarity between feature vectors. Experimental results demonstrate that the proposed method achieves an algorithm sensitivity of 0.9243, surpassing ResNet and VGG by 5.86 and 8.10 percentage points, respectively, highlighting its robustness and effectiveness in large-scale power grid redundancy detection. These results underscore the paradigm's capability to balance efficiency and precision in complex visual inspection scenarios.

Keywords: image retrieval; sparse attention mechanism; redundancy elimination; power grid inspection images; vision transformer



Academic Editors: Prayag Tiwari, Guang Yang, Changshuo Wang, Gang Wang and Ruiping Wang

Received: 28 July 2025

Revised: 19 August 2025

Accepted: 21 August 2025

Published: 10 September 2025

Citation: Yang, W.; Chen, Z.; Huang, X.; Li, M.; Wang, H.; Liu, S. A Sparse Attention Mechanism Based Redundancy-Aware Retrieval Framework for Power Grid Inspection Images. *Electronics* **2025**, *14*, 3585. <https://doi.org/10.3390/electronics14183585>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As power systems and smart grids continue to advance, power equipment inspection is progressively transitioning from traditional manual methods toward intelligent and automated visual monitoring. Leveraging high-resolution cameras, unmanned aerial vehicles (UAVs), and mobile inspection robots, modern power systems continuously generate massive volumes of image data. These images provide critical visual information for equipment condition analysis, fault detection, and defect identification. However, the exponential growth of visual data also introduces a major challenge: data redundancy. In particular, power image datasets often contain a large proportion of redundant or highly similar samples [1,2]. Such redundancy primarily arises from periodic inspections that repeatedly capture the same equipment under similar viewpoints, as well as derivative images generated during preprocessing steps such as format conversion, resizing, and

rotation. Redundant data not only impose significant storage burdens but also degrade the quality of downstream tasks, for instance, by causing deep learning models to overfit redundant features and reducing generalization performance. Further, some kinds of various deep learning methods are applied to vision and language applications [3–9] and other learning-based applications [10–16].

Existing techniques for redundancy reduction in power images can be broadly categorized into three types: conventional content-driven feature matching, perceptual hashing for rapid comparison, and deep learning models for capturing image similarity. Conventional feature matching offers fast computation but struggles with complex backgrounds, lighting variations, and geometric structures typical of power equipment images. Perceptual hashing improves efficiency and provides some robustness, yet often fails to capture deeper semantic differences. Deep learning models provide stronger representational capacity but require substantial computational resources and large labeled datasets, and may still lack sufficient semantic understanding of power-specific content. More importantly, existing approaches frequently cannot distinguish between genuine equipment state changes and superficial imaging variations, limiting their effectiveness in real-world inspection scenarios. Consequently, existing methods are limited in real-world power inspection scenarios, motivating the need for a more robust and semantically aware redundancy removal framework.

To address these challenges, we propose a dual-phase redundancy removal framework that integrates both perceptual cues and high-level semantic information. In the first stage, an improved Discrete Cosine Transform (DCT)-based hashing algorithm combined with multi-scale structural similarity (MS-SSIM) performs fast and lightweight initial filtering, emphasizing visual similarity with low computational cost. In the second stage, a Vision Transformer enhanced with a hierarchical sparse attention mechanism ($\text{top-}k = 50\%$) extracts semantic representations, and cosine similarity is employed for precise redundancy judgment. The framework introduces several key innovations: (1) a disturbance-resilient hashing algorithm that maintains stable responses under $\pm 15^\circ$ rotations and $\pm 30\%$ brightness variations; (2) a dynamic threshold-gated sparse attention mechanism within the Transformer that reduces computation by approximately 42%; and (3) a three-level feature processing pipeline that combines low-level MS-SSIM filtering, mid-level perceptual hashing, and high-level Transformer-based semantic analysis to balance efficiency and accuracy.

To enhance robustness under real-world conditions, we incorporate a Retinex-theory-based preprocessing module to mitigate lighting variability and an attention-guided feature enhancement mechanism to focus on critical structural regions of power equipment. The approach is evaluated on a large-scale dataset covering six categories of power equipment: substations, transmission lines, towers, switches, junction boxes, and insulators. Experimental results demonstrate that our method surpasses state-of-the-art alternatives, notably improving retrieval sensitivity, mitigating spurious detections, and enhancing responsiveness to subtle apparatus perturbations. Specifically, sensitivity to minor visual anomalies—such as small cracks or corrosion—is improved by approximately 15% over the best-performing baseline, while overall processing speed is increased by over $4\times$ due to the two-stage design. These findings underscore the proposed framework's strong potential for real-world deployment in engineering applications.

2. Related Work

2.1. Image Retrieval

The exigency for expeditious image retrieval across voluminous datasets has catalyzed the advent of diversified hashing paradigms, wherein intricate high-dimensional visual descriptors are transmuted into succinct binary encodings. These methods aim to preserve

semantic similarity while significantly reducing computational cost. Hashing methodologies are conventionally bifurcated into two principal classes: data-agnostic schemes and data-adaptive algorithms that iteratively infer the underlying data manifold.

Data-independent hashing does not rely on training data to define its transformation functions. A classic example is Locality Sensitive Hashing (LSH) [17], which ensures that similar inputs are more likely to be mapped to the same hash code by using random projections. Despite its efficiency in retrieving data from high-dimensional feature representations, LSH exhibits certain limitations. Its effectiveness is constrained to specific similarity metrics and lacks adaptiveness to complex data distributions.

In contrast, data-driven hashing methods derive hash functions from the training data itself, enabling improved conformity to the intrinsic structure of the data. Contingent upon the availability of annotated supervision during the training phase, these learning-centric methodologies bifurcate into unsupervised and supervised taxonomies.

In unsupervised hashing, the algorithms exploit statistical patterns without requiring class annotations. One prominent method, Spectral Hashing [18], formulates the hash learning task as an eigen-decomposition problem based on graph Laplacians. Despite its elegance, it assumes uniform data distribution, limiting its generalization. Another variant, Kernelized LSH [19], incorporates nonlinear transformations using kernel functions, enhancing the capacity to capture more complex structures without supervision.

Supervised hashing methods, in contrast, leverage semantic labels to guide the learning of hash codes that reflect class-level similarity. For instance, Binary Reconstructive Embedding [20] endeavors to attenuate the reconstruction disparity between the intrinsic feature manifold and its corresponding representation within the Hamming space. Similarly, Minimum Loss Hashing (MLH) [21] reduces quantization loss to better approximate semantic distances. Kernel Supervised Hashing (KSH) [22] further introduces kernel-based metrics to emphasize separation between dissimilar categories while clustering similar ones. These methods often achieve higher retrieval precision but incur increased training cost and require sufficient labeled samples.

Nevertheless, supervised approaches face a major obstacle due to their reliance on large-scale annotations, which require considerable manual effort and expense. While some semi-supervised variants attempt to reduce this burden by utilizing partial labels [23], their performance can degrade with small labeled subsets or under distribution shifts. Moreover, many conventional hashing models rely on linear projections, which are insufficient to model the intricate nonlinear relationships inherent in image data [24].

Recently, deep learning has significantly propelled image retrieval forward by combining feature extraction and hash code creation within a single framework. Convolution-based neural networks are extensively utilized to encode semantic content into binary representations [25,26]. These networks are typically trained using label supervision, with the learned embeddings binarized via thresholding mechanisms applied to fully connected layers.

To overcome the dependence on annotated data, unsupervised deep hashing methods have also emerged [27]. These approaches often employ techniques such as autoencoders, contrastive learning, or generative models to learn effective hash functions by exploiting the inherent data structure. While still an active area of research, they offer promising alternatives for large-scale retrieval scenarios where labeled data is scarce.

In summary, image hashing techniques have evolved from randomized projections to sophisticated deep models. Although supervised deep hashing attains leading accuracy, issues related to scalability, annotation expenses, and generalization persist. As such, developing lightweight, robust, and label-efficient hashing frameworks continues to be a crucial focus for modern image retrieval tasks.

2.2. Vision Transformer

Originally developed to tackle challenges in natural language processing tasks, the architecture known as the Transformer has progressively gained prominence as a powerful model in computer vision. Its self-attention mechanism facilitates flexible modeling of long-range relationships, benefiting visual representation learning significantly. Nonetheless, the computational burden increases quadratically with input size, making direct application to high-resolution images challenging.

To mitigate these computational limitations, a range of vision-specific adaptations have been proposed. The Vision Transformer (ViT) [28] represents one of the earliest efforts in this direction. It segments images into non-overlapping patches and processes them as a sequence of tokens, effectively lowering input dimensionality. This approach has subsequently motivated numerous subsequent studies [29,30], focusing on improving tokenization techniques or tailoring the architecture for downstream vision applications like detection and segmentation [31].

An auxiliary design tenet embraced by numerous architectures entails the gradual diminution of spatial resolution in feature representations, thereby regulating the token count engaged in attention operations. For instance, Pyramid Vision Transformer (PVT) [32,33] introduces a hierarchical encoder and incorporates a sparsified attention mechanism to improve scalability. Deformable Attention Transformer (DAT) [34] advances this idea by leveraging a deformable attention module that learns adaptive sampling locations based on input content.

Other approaches emphasize localization of attention to limit the receptive field during early processing stages. The Swin Transformer [35], for example, applies attention within shifted non-overlapping windows, striking a balance between efficiency and expressivity. Similarly, NAT [36] adopts a query-wise strategy, wherein each query token computes attention independently using convolution-inspired local priors.

In addition to architectural innovations, some researchers explore hybrid designs that fuse convolutional operations with self-attention to enhance both performance and efficiency. For example, CMT [37] integrates depthwise convolutions [38] into Transformer blocks, capitalizing on their lightweight characteristics. ACmix [39] further streamlines computation by sharing operations between convolution and attention modules. Such hybridization is motivated by the complementary nature of local convolution and global attention, leading to enhanced feature modeling.

Efforts have also been made to optimize training efficiency and reduce inference latency. For instance, MobileFormer [40] utilizes dual processing streams—one for convolution and one for Transformer—to combine the benefits of both paradigms. Dynamic Perceiver [41] implements early exit mechanisms, enabling adaptive computation paths based on input complexity. MobileViT [42] extends the MobileNet [38] backbone by integrating Transformer modules into a compact convolutional architecture, rendering it highly suitable for environments with limited resources.

Overall, Vision Transformers have rapidly evolved through innovations in tokenization, attention mechanisms, and hybrid architectures. Notwithstanding these advancements, harmonizing precision, computational efficiency, and scalability persists as a formidable challenge, especially concerning deployment on edge hardware and latency-sensitive real-time systems.

3. Methodology

3.1. Overview

This investigation delineates a redundancy abatement paradigm for power inspection imagery, which amalgamates semantic saliencies with perceptual cues, as schematically

depicted in Figure 1. The approach targets large-scale power image datasets and aims to efficiently filter out visually redundant samples, thereby improving data processing efficiency and enhancing the accuracy of downstream recognition tasks.

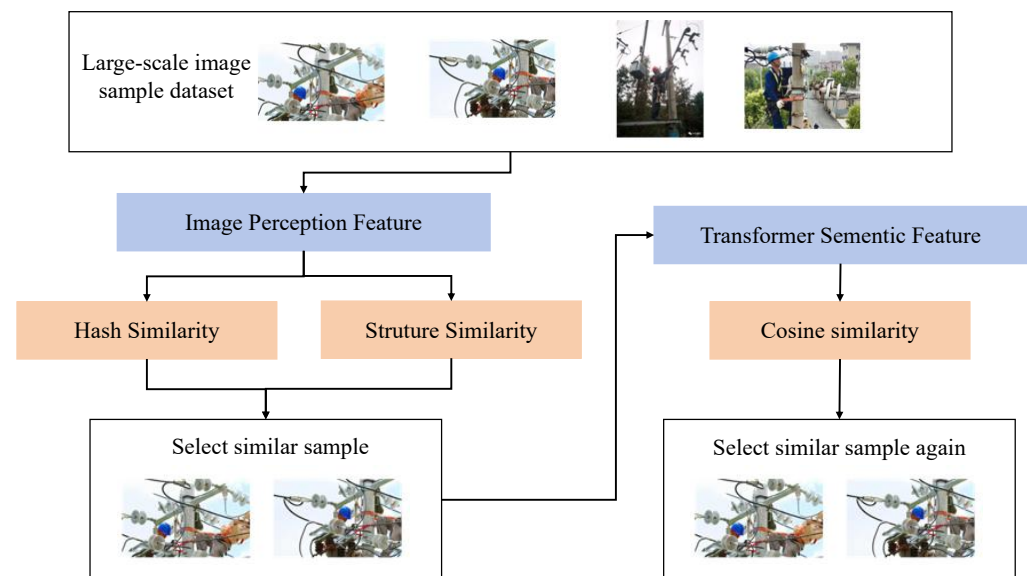


Figure 1. Outline of the proposed pipeline.

Initially, low-level visual attributes like texture, edges, and structural details are extracted from each image at the perceptual stage. Based on these features, both hash similarity and structural similarity between images are computed. Hash similarity provides a rapid assessment of overall visual similarity, while structural similarity captures spatial arrangement and detailed differences. A threshold-based filtering is then applied, where the perceptual similarity threshold T_p is empirically set to 0.85 for hash similarity and 0.90 for structural similarity, as determined from validation experiments balancing recall and precision. Only image pairs exceeding both thresholds are retained as candidate redundant samples.

In the next stage, a sparse Transformer network is utilized to derive high-dimensional semantic representations from the selected candidates. These features represent abstract semantic information, including object-level understanding, contextual relationships, and scene semantics, which complement the limitations of perceptual features in capturing high-level meaning. Cosine similarity is then calculated between semantic feature vectors, and a semantic similarity threshold T_s of 0.92 is applied to distinguish samples that are perceptually similar but semantically different. Finally, a second-stage filtering based on this semantic similarity ensures that only image pairs with both high perceptual and semantic consistency are identified as truly redundant.

By explicitly defining the threshold parameters and their selection rationale, the proposed dual-path filtering mechanism combining perceptual and semantic analysis is both precise and reproducible. This not only augments the discriminability and informational richness of the preserved instances but also furnishes high-fidelity data for downstream applications, including power apparatus identification and anomaly localization.

3.2. Visual Perception Branch

To measure the low-level visual similarity between power inspection images, we employ a perceptual hashing technique based on the Discrete Cosine Transform (DCT). This approach encodes each image into a compact 64-bit fingerprint by extracting essential

luminance and structural patterns, and then compares image pairs using Hamming distance. The process consists of several stages, as described below.

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we first resize it to a fixed resolution of 32×32 pixels to standardize the input and eliminate variations caused by image size and aspect ratio. The resized image is denoted by

$$\tilde{I} = \text{Resize}(I, 32 \times 32). \quad (1)$$

Next, the resized image is converted into a single-channel grayscale image $I_g \in \mathbb{R}^{32 \times 32}$, which discards chromatic components and focuses solely on the structural and brightness information:

$$I_g = \text{Gray}(\tilde{I}). \quad (2)$$

We then apply the two-dimensional Discrete Cosine Transform (DCT) to I_g to obtain the frequency-domain representation $D \in \mathbb{R}^{32 \times 32}$:

$$D = \text{DCT}(I_g). \quad (3)$$

The DCT decomposes the image into a set of cosine basis functions of varying frequency, capturing both global and local visual patterns. Low-frequency components in the top-left region of D preserve most of the essential structural information, while higher frequencies correspond to finer details and noise.

To isolate these dominant features, we extract the upper-left 8×8 submatrix from the DCT coefficient matrix:

$$D_{8 \times 8} = D[0:8, 0:8]. \quad (4)$$

We compute the mean of all elements in $D_{8 \times 8}$, which will serve as a threshold to binarize the coefficients:

$$\mu = \frac{1}{64} \sum_{i=0}^7 \sum_{j=0}^7 D_{i,j}. \quad (5)$$

The perceptual hash code is generated by comparing each coefficient in $D_{8 \times 8}$ with the mean value μ . For each position (i, j) , the binary value is assigned as

$$H_{i,j} = \begin{cases} 1, & \text{if } D_{i,j} \geq \mu \\ 0, & \text{otherwise} \end{cases}. \quad (6)$$

The final hash code $H \in \{0, 1\}^{64}$ is obtained by flattening the 8×8 binary matrix in row-major order:

$$H = \text{Flatten}_{\text{row-major}}(H_{i,j}). \quad (7)$$

For any two images I_a and I_b , with respective hash codes H_a and H_b , we compute their Hamming distance d_H to measure perceptual dissimilarity:

$$d_H(H_a, H_b) = \sum_{k=1}^{64} \mathbb{I}[H_a^{(k)} \neq H_b^{(k)}], \quad (8)$$

where $\mathbb{I}[\cdot]$ is the indicator function. A smaller d_H indicates higher perceptual similarity between the two images.

To normalize the distance into a similarity score, we define:

$$S_{\text{hash}} = 1 - \frac{d_H(H_a, H_b)}{64}. \quad (9)$$

A larger S_{hash} implies stronger visual resemblance. If S_{hash} exceeds a predefined threshold θ , the images are considered visually redundant:

$$\text{Redundant}(I_a, I_b) = \begin{cases} \text{True}, & \text{if } S_{\text{hash}} \geq \theta \\ \text{False}, & \text{otherwise} \end{cases}. \quad (10)$$

This DCT-based perceptual hashing method provides several advantages. First, it yields highly compact binary descriptors that enable fast pairwise comparisons. Second, it is robust to small variations in image scale, illumination, and compression. Third, the computational cost is minimal, making it well-suited for large-scale redundancy analysis. By integrating this hashing mechanism into the visual perception branch, we ensure efficient and scalable filtering of redundant samples prior to downstream semantic processing.

3.3. Semantic Feature Extraction Branch

To extract high-level semantic representations from power grid images, we employ a Transformer-based feature extractor enhanced with a sparse attention mechanism. The process begins by converting input images into a sequence of patch embeddings, followed by semantic modeling through attention-based encoding. The resulting feature vectors are used to compute semantic similarities among images. The detailed procedure is described below.

Let the input image be denoted as $I \in \mathbb{R}^{H \times W \times 3}$. Initially, the image is segmented into multiple non-overlapping patches of size $P \times P$, yielding a total of $N = \frac{H \cdot W}{P^2}$ patches. Each patch is then reshaped into a one-dimensional vector and transformed into a d -dimensional feature embedding via a trainable linear layer. We refer to the resultant sequence of embedded patches as:

$$X = [x_1, x_2, \dots, x_N]^T \in \mathbb{R}^{N \times d}, \quad (11)$$

where $x_i = \text{Proj}(\text{Flatten}(\text{Patch}_i))$ represents the embedded vector of the i -th patch. To encode positional information, we add a learnable position embedding $E_{\text{pos}} \in \mathbb{R}^{N \times d}$:

$$X' = X + E_{\text{pos}}. \quad (12)$$

The resulting patch sequence X' is input into a Transformer encoder that leverages sparse attention. As delineated in Figure 2, each stratum integrates a multi-head sparse self-attention paradigm to encapsulate protracted contextual interdependencies among visual tokens. For every attention head, queries, keys, and values are produced through separate linear projections:

$$Q = X'W_Q, \quad K = X'W_K, \quad V = X'W_V, \quad (13)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable weight matrices. The full attention score matrix $A \in \mathbb{R}^{N \times N}$ is computed as:

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right). \quad (14)$$

To enhance the interpretability and robustness of attention, sparsity is instilled by conserving exclusively the foremost k prominent entries in every row of the attention matrix. Concretely, a binary mask $M \in \{0, 1\}^{N \times N}$ is defined as

$$M_{ij} = \begin{cases} 1, & \text{if } A_{ij} \in \text{Top-}k(A_{i*}), \\ 0, & \text{otherwise} \end{cases}, \quad (15)$$

where $\text{Top-}k(A_{i*})$ denotes the top- k entries in the i -th row of A . The sparse attention matrix is then defined as:

$$\tilde{A}_{ij} = M_{ij} \cdot A_{ij}. \quad (16)$$

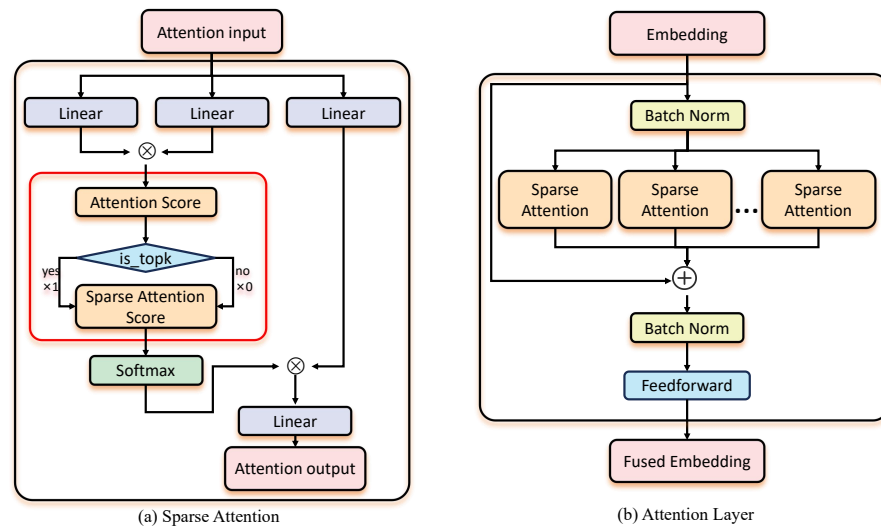


Figure 2. Illustration of the sparse attention and the transformer layer.

The output of a single sparse attention head is obtained as:

$$\text{head}_h = \tilde{A}^{(h)} V^{(h)}, \quad (17)$$

and outputs from each attention head are combined by concatenation and subsequently transformed via a linear projection to yield the multi-head sparse attention result:

$$\text{MultiHead}(X') = \text{Concat}(\text{head}_1, \dots, \text{head}_H) W^O, \quad (18)$$

where $W^O \in \mathbb{R}^{Hd_k \times d}$ is the output projection matrix. To ensure training stability and expedite convergence, batch normalization is applied pre- and post-attention as well as feedforward sublayers. The resultant output of the Transformer encoder layer is:

$$Z = \text{FFN}(\text{BN}(X' + \text{MultiHead}(\text{BN}(X')))), \quad (19)$$

where FFN denotes a two-layer position-wise feedforward network with non-linear activation, and BN is batch normalization.

After passing through multiple such Transformer encoder blocks, the sequence Z contains contextually enriched representations for each image patch. To consolidate these into one semantic feature vector representing the entire image, average pooling is performed:

$$z = \frac{1}{N} \sum_{i=1}^N Z_i, \quad (20)$$

where $z \in \mathbb{R}^d$ is the final semantic embedding of the input image. Consequently, all images are projected into a unified d -dimensional latent feature manifold.

To quantify the semantic similarity between two images I_a and I_b , represented by embeddings z_a and z_b , we calculate their cosine similarity as follows:

$$S_{\cos}(z_a, z_b) = \frac{z_a \cdot z_b}{\|z_a\|_2 \cdot \|z_b\|_2}. \quad (21)$$

A value close to 1 indicates that the semantic contents of the two images are highly similar. If the cosine similarity exceeds a predefined threshold τ , the image pair is considered semantically redundant:

$$\text{Redundant}(I_a, I_b) = \begin{cases} \text{True,} & \text{if } S_{\cos}(z_a, z_b) \geq \tau \\ \text{False,} & \text{otherwise} \end{cases}. \quad (22)$$

This semantic feature extraction pipeline, from patch embedding to sparse attention and vector-based comparison, enables the model to robustly filter redundant images based on their high-level meaning, complementing the low-level perception branch and enhancing the overall effectiveness of image redundancy governance.

4. Experiments

4.1. Implementation Details

To improve model robustness and mitigate overfitting, several data augmentation techniques are applied to the training set, including the following:

- Random rotations within $\pm 15^\circ$;
- Horizontal and vertical flips;
- Random brightness and contrast adjustments up to $\pm 30\%$;
- Random cropping and resizing.

These augmentations increase the diversity of the training data and ensure that the model is resilient to common variations encountered in real-world power grid imagery.

The semantic feature extraction branch is based on a Vision Transformer with sparse attention. The network consists of 12 Transformer encoder blocks, each with 8 attention heads and a hidden dimension of 512. Patch size is set to 16×16 , yielding a sequence length of $N = (H \cdot W)/16^2$ per image. Each encoder block contains the following:

- Multi-head sparse self-attention layer;
- Two-layer feedforward network with GELU activation;
- Batch normalization before and after attention and feedforward layers.

The final image embedding dimension is 512. The total number of trainable parameters in the semantic branch is approximately 21.3 million.

The model is trained using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-2} . A cosine annealing scheduler gradually reduces the learning rate over 100 epochs. The batch size is set to 32, and dropout with a rate of 0.1 is applied in each Transformer block to prevent overfitting. All models are implemented in PyTorch 2.7.1 and trained on an NVIDIA RTX 4090 GPU.

4.2. Dataset Construction

To appraise the efficacy of our sparse attention-infused semantic abstraction paradigm in power image analytics, we curated a bespoke dataset carefully designed for this study. A total of 295 high-resolution images were collected from multiple real-world power grid operation scenarios, including three substations and two transmission line segments. These images cover a variety of typical elements such as electric towers, maintenance personnel, operation tools, and the surrounding environmental context. The dataset is primarily intended for evaluating redundancy detection and semantic representation in power inspection imagery, supporting tasks such as anomaly identification and equipment condition assessment.

To emulate authentic redundancy scenarios within training datasets and to meticulously evaluate the model's capacity for redundancy detection while maintaining semantic distinctiveness, we augmented the original images via a suite of transformation operations:

- **Cropping:** Random regions of interest (ROI) were selected and cropped to introduce spatial variation.
- **Scaling:** Images were rescaled using random scale factors to simulate different zoom levels.
- **Filtering:** Image filters such as Gaussian blur, edge enhancement, and brightness contrast adjustment were applied to mimic sensor-induced variations.
- **Compression:** JPEG compression with varying quality factors was applied to simulate storage or transmission artifacts.

These augmentations resulted in over 6000 redundant images with different degrees of similarity to the original set. Each original image and its augmented variants form a similar image group, yielding a total of 295 groups. The images within each group are visually and semantically similar but exhibit minor content and quality discrepancies, reflecting common scenarios in real-world power image datasets. A sample group of similar images is shown in Figure 3, highlighting the high intra-group similarity and subtle inter-instance differences.



Figure 3. Examples from the power grid dataset.

This dataset provides a solid foundation for evaluating both semantic feature extraction and redundancy detection tasks.

4.3. Evaluation Metrics

To holistically appraise the model, we employ evaluative criteria encompassing both regression and classification paradigms.

Recall assesses the degree to which the model effectively retrieves the entirety of pertinent (positive) targets. It is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (23)$$

Here, TP , TN , FP , and FN represent the frequencies of empirically verified positive detections, valid negative classifications, erroneously flagged positives, and undetected positive instances, respectively.

Precision quantifies the proportion of accurately identified redundant images among all instances classified as redundant by the model, reflecting its efficacy in redundancy discernment. Formally, it is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (24)$$

The F1-score provides an equilibrated evaluation by synthesizing the model's fidelity with its thoroughness in identifying redundant instances. The F1-score is calculated as:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (25)$$

ROC-AUC quantifies the model's ability to distinguish between redundant and non redundant images by evaluating its performance at all possible classification decision thresholds, where the ROC curve plots the relationship between True Positive Rate and False Positive Rate at different thresholds, and the AUC (area under the curve) value range is between 0.5 and 1, which is commonly used to quantify the quality of the ROC curve.

4.4. Results

4.4.1. Redundant Image Detection

We further test the model's capability in identifying similar images. Given the vectorized representations produced by the sparse attention semantic branch, we compute the cosine similarity between pairs of images. When the similarity exceeds a predefined threshold, the pair is marked as redundant.

The performance in this classification task is shown in Table 1. The model achieves a precision of 0.9961 and a recall of 0.8605, indicating its robust ability to capture visual redundancy even when differences are subtle.

Table 1. Detection efficacy of redundant images using the proposed framework.

Evaluation Metric	Result
Precision	0.9961
Recall	0.8605
F1-Score	0.9248
ROC-AUC	0.9303

We compare the algorithm sensitivity between the proposed method and other methods, including ResNet, VGG, which are commonly used for image similarity detection based on semantic features. Note that the sensitivity is the percentage of targets that the algorithm can identify out of the total number of targets.

Through comparison from Table 2, it can be found that the image similarity detection model based on the proposed method is better than other mainstream models, mainly because the self-attention mechanism of Transformer can build global dependencies and capture image features better.

Table 2. Algorithm sensitivity of different models for the redundant image detection.

Model	Algorithm Sensitivity
ResNet [43]	0.8657
VGG [44]	0.8433
Ours	0.9243

4.4.2. Ablation Study

This study shows the effect of sparsity level in sparse attention. To analyze the role of attention mechanism sparsity in image similarity calculations, we designed a comparative ablation experiment to examine how varying sparsity levels affect computational outcomes. The topk parameter was set at 25%, 50%, and 75%, respectively, representing different degrees of sparsity. Through four experimental groups, we compared the impacts of sparsity and its degree on computational results. As shown in Table 3, the attention mechanism with sparsity-enhanced modifications showed a 5–6 percentage point improvement in algorithm sensitivity. This demonstrates that sparsity filtering effectively removes distracting semantic features, allowing semantic expressions to concentrate on critical information. Notably, setting the sparsity parameter at 50% demonstrated slightly better performance compared to 75%, indicating that excessive sparsity may lead to information loss. Therefore, practical implementations typically adopt a 50% sparsity parameter configuration.

Table 3. Algorithm sensitivity under different levels of sparse attention mechanism.

Model	Algorithm Sensitivity
No sparse attention mechanism	0.9243
25% sparse attention mechanism	0.9786
50% sparse attention mechanism	0.9875
75% sparse attention mechanism	0.9816

4.4.3. Visualization

To substantiate the efficacy of the proposed approach, as shown in Figure 4, we show the similarity score and determination result of some redundant image pairs and non-redundant image pairs. The cosine similarity between the two images in the first row is 0.874, and they are judged as redundant image pairs. The cosine similarity between the two images in the second row is 0.153, and it is determined as a non-redundant image pair. The cosine similarity between the two images in the third row is 0.419, which is determined as a non-redundant image pair and belongs to a missed judgment. The main reason is that some enhanced data are overexposed, resulting in distortion of image feature extraction and misjudgment.



Figure 4. The visualization of retrieval results.

5. Discussion

The findings substantiate that our proposed model proficiently evaluates image quality and identifies redundant images within power grid visual data. The use of sparse attention improves the discriminative capability of feature representations, making them more resilient to minor perturbations introduced by augmentations. Furthermore, the cosine congruity between semantic embeddings serves as a potent heuristic for affinity quantification, demonstrating strong alignment with human cognitive appraisal of pictorial similarity. These characteristics are essential for ensuring data diversity and minimizing overfitting during training on real-world datasets.

Nevertheless, the current dataset size is relatively limited, which may constrain the assessment of the model's generalization capacity. Future work will involve validating the proposed method on larger and more diverse datasets, encompassing various power grid environments, imaging conditions, and equipment types, to further ascertain its robustness and applicability in broader real-world scenarios.

In addition, although the dual-path filtering mechanism enhances redundancy detection accuracy, it inevitably introduces extra computational overhead due to the combined use of perceptual and semantic feature extraction. For deployment in real-time or resource-constrained power grid monitoring systems, trade-offs between accuracy, processing speed, and hardware requirements must be carefully considered. Potential optimizations, such as model compression, hardware acceleration, and adaptive filtering strategies, will be explored to ensure efficient integration into large-scale operational environments.

Beyond these points, other limitations should be noted. While the model demonstrates robustness to moderate rotations and lighting variations, its performance under extreme environmental conditions or severe image distortions has not been fully evaluated. Additionally, the current framework is designed for static images, and extending it to continuous video streams or multi-temporal monitoring scenarios requires further investigation. Addressing these limitations will guide future improvements and facilitate deployment in more diverse operational settings.

Future work will address these limitations by expanding the evaluation to larger and more diverse datasets, encompassing a wider variety of power grid environments, imaging conditions, and equipment categories. We also plan to investigate lightweight or real-time variants of the model, explore adaptive attention mechanisms for dynamic scene understanding, and integrate multi-modal data sources, such as infrared or LiDAR imagery, to further enhance redundancy detection accuracy and applicability in operational power grid monitoring systems.

6. Conclusions

We put forth a semantic-aware sparse attention paradigm, meticulously devised to tackle the intricate tasks of redundancy discernment and perceptual fidelity appraisal within power grid visual corpora. The proposed framework initiates by decomposing high-resolution inspection images into a structured sequence of patch-level tokens, which are then embedded into a latent space that retains both local detail and global context. This embedded representation is subsequently refined through a sparse self-attention mechanism, wherein computational emphasis is selectively allocated to the most salient interactions across token pairs. Concretely, by retaining only the top- k attention coefficients in each query row, the model circumvents the inefficiencies of dense attention and suppresses the propagation of semantically inconsequential signals.

Such a design not only mitigates the redundancy inherent in large-scale visual datasets but also augments the model's capacity to focus on discriminative regions—such as damaged components, abnormal textures, or anomalous patterns—crucial for downstream

analytic tasks. The sparse attention module thus serves a dual role: as a computational economizer and as a semantic distiller, ensuring that only the most informative visual dependencies are preserved for subsequent layers. Through this synergy of tokenization, semantic abstraction, and structured sparsity, our method delivers enhanced discriminability and robustness in scenarios demanding both precision and interpretability.

To further improve semantic discrimination, a feature extraction backbone incorporating batch normalization, multi-head sparse attention, and a feedforward module was designed. This architecture enables the network to learn robust and informative features from power operation scenes with high precision, even in the presence of noise and low-quality inputs.

We constructed a dataset of 295 original power images and generated over 6000 redundant samples using various augmentation techniques. Experimental findings on this dataset compellingly corroborate the efficacy and robustness of our proposed framework. Specifically, for binary classification tasks such as redundancy detection, the model achieves a precision of 0.9961 and a recall of 0.8605, verifying its utility in identifying redundant samples effectively.

Overall, our approach provides a practical and efficient solution for semantic representation learning and redundancy management in power image datasets. By reducing data redundancy and emphasizing high-quality, informative samples, the proposed method lays the foundation for more scalable and intelligent visual inspection systems in power grid operations. Future work will explore the integration of cross-modal attention to incorporate text or sensor metadata and the application of our approach to larger-scale industrial image analysis tasks.

Author Contributions: Conceptualization, W.Y. and Z.C.; Methodology, Z.C. and S.L.; Software, X.H. and H.W.; Formal analysis, M.L. and S.L.; Investigation, Z.C., X.H. and H.W.; Data curation, X.H. and H.W.; Writing—original draft, Z.C. and H.W.; Writing—review and editing, Z.C., X.H. and H.W.; Visualization, Z.C.; Funding acquisition, Z.C., X.H. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Project from State Grid Corporation of China (No. 5700-202390301A-1-1-ZN).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Authors Wei Yang, Zhenyu Chen and Shi Liu were employed by the company Big Data Center of State Grid Corporation of China. Author Xiaoguang Huang was employed by the company State Grid Information & Telecommunication Group Co., Ltd. Authors Ming Li and Hailu Wang were employed by the company Information & Telecommunication Branch State Grid Anhui Electric Power Company.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
DAT	Deformable Attention Transformer
DCT	Discrete Cosine Transform
KSH	Kernel Supervised Hashing
LSH	Locality Sensitive Hashing
MLH	Minimum Loss Hashing
MS-SSIM	Multi-Scale Structural Similarity index
PVT	Pyramid Vision Transformer
UAVs	Unmanned Aerial Vehicles
ViT	Vision Transformer

References

1. Wang, F.; Zou, Y.; Castillo, E.D.R.; Lim, J. Optimal UAV image overlap for photogrammetric 3d reconstruction of bridges. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1101*, 022052. [\[CrossRef\]](#)
2. Lu, Y.; Zheng, E.; Chen, Y.; Wu, K.; Yang, Z.; Yuan, J.; Xie, M. Stower-13: A Multi-View Inspection Image Dataset for the Automatic Classification and Naming of Tension Towers. *Electronics* **2024**, *13*, 1858. [\[CrossRef\]](#)
3. Zeng, Z.; Liu, C.; Tang, Z.; Chang, W.; Li, K. Training Acceleration for Deep Neural Networks: A Hybrid Parallelization Strategy. In Proceedings of the 2021 58th ACM/IEEE Design Automation Conference (DAC), Francisco, CA, USA, 5–9 December 2021; Volume 10, pp. 1165–1170.
4. Chen, C.; Lv, F.; Guan, Y.; Wang, P.; Yu, S.; Zhang, Y.; Tang, Z. Human-guided image generation for expanding small-scale training image datasets. *IEEE Trans. Vis. Comput. Graph.* **2025**, *31*, 3809–3821. [\[CrossRef\]](#)
5. Wang, M.; Pi, H.; Li, R.; Qin, Y.; Tang, Z.; Li, K. VLScene: Vision-Language Guidance Distillation for Camera-Based 3D Semantic Scene Completion. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 27 February–4 March 2025; Volume 39, pp. 7808–7816.
6. Gao, X.; Chen, Z.; Zhang, B.; Wei, J. Deep learning to hash with application to cross-view nearest neighbor search. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 3882–3892. [\[CrossRef\]](#)
7. Wang, C.; Gao, X.; Wu, M.; Lam, S.K.; He, S.; Tiwari, P. Looking Clearer with Text: A Hierarchical Context Blending Network for Occluded Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 4296–4307. [\[CrossRef\]](#)
8. Zhang, Q.; Chen, C.; Liu, Z.; Tang, Z. I-adapt: Using iou adapter to improve pseudo labels in cross-domain object detection. In Proceedings of the ECAI 2024, Santiago de Compostela, Spain, 19–24 October 2024; IOS Press: Amsterdam, The Netherlands, 2024; pp. 57–64.
9. Gao, X.; Li, Z.; Shi, H.; Chen, Z.; Zhao, P. Scribble-Supervised Video Object Segmentation via Scribble Enhancement. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 2999–3012. [\[CrossRef\]](#)
10. Chen, B.; Wu, Q.; Li, M.; Xiahou, K. Detection of false data injection attacks on power systems using graph edge-conditioned convolutional networks. *Prot. Control Mod. Power Syst.* **2023**, *8*, 1–12. [\[CrossRef\]](#)
11. Gao, X.; Wang, X.; Chen, Z.; Zhou, W.; Hoi, S.C. Knowledge enhanced vision and language model for multi-modal fake news detection. *IEEE Trans. Multimedia* **2024**, *26*, 8312–8322. [\[CrossRef\]](#)
12. Zhao, W.; Zeng, T.; Liu, Z.; Xie, L.; Xi, L.; Ma, H. Automatic generation control in a distributed power grid based on multi-step reinforcement learning. *Prot. Control Mod. Power Syst.* **2024**, *9*, 39–50. [\[CrossRef\]](#)
13. Wang, C.; Cao, R.; Wang, R. Learning discriminative topological structure information representation for 2D shape and social network classification via persistent homology. *Knowl. Based Syst.* **2025**, *311*, 113125. [\[CrossRef\]](#)
14. Zhu, S.; Ma, H.; Chen, L.; Wang, B.; Wang, H.; Li, X.; Gao, W. Short-term load forecasting of an integrated energy system based on STL-CPLE with multitask learning. *Prot. Control Mod. Power Syst.* **2024**, *9*, 71–92. [\[CrossRef\]](#)
15. Gao, X.; Chen, Z.; Wei, J.; Wang, R.; Zhao, Z. Deep mutual distillation for unsupervised domain adaptation person re-identification. *IEEE Trans. Multimedia* **2024**, *27*, 1059–1071. [\[CrossRef\]](#)
16. Chen, X.; Yu, T.; Pan, Z.; Wang, Z.; Yang, S. Graph representation learning-based residential electricity behavior identification and energy management. *Prot. Control Mod. Power Syst.* **2023**, *8*, 28. [\[CrossRef\]](#)
17. Gionis, A.; Indyk, P.; Motwani, R. Similarity search in high dimensions via hashing. In Proceedings of the VLDB, Edinburgh, UK, 7–10 September 1999; Volume 99, pp. 518–529.
18. Weiss, Y.; Torralba, A.; Fergus, R. Spectral hashing. *Adv. Neural Inf. Process. Syst.* **2008**, *21*, 1753–1760.
19. Kulis, B.; Grauman, K. Kernelized locality-sensitive hashing for scalable image search. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 2130–2137.
20. Kulis, B.; Darrell, T. Learning to hash with binary reconstructive embeddings. *Adv. Neural Inf. Process. Syst.* **2009**, *22*, 1042–1050.
21. Norouzi, M.; Fleet, D.J. Minimal loss hashing for compact binary codes. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011.
22. Liu, W.; Wang, J.; Ji, R.; Jiang, Y.G.; Chang, S.F. Supervised hashing with kernels. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2074–2081.
23. Wang, J.; Kumar, S.; Chang, S.F. Semi-supervised hashing for large-scale search. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2393–2406. [\[CrossRef\]](#)
24. Erin Liong, V.; Lu, J.; Wang, G.; Moulin, P.; Zhou, J. Deep hashing for compact binary codes learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2475–2483.
25. Lin, K.; Yang, H.F.; Hsiao, J.H.; Chen, C.S. Deep learning of binary hash codes for fast image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 27–35.
26. Zhang, R.; Lin, L.; Zhang, R.; Zuo, W.; Zhang, L. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. *IEEE Trans. Image Process.* **2015**, *24*, 4766–4779. [\[CrossRef\]](#)

27. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [[CrossRef](#)]
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
29. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
30. Yuan, L.; Hou, Q.; Jiang, Z.; Feng, J.; Yan, S. Volo: Vision outlooker for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6575–6586. [[CrossRef](#)]
31. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring plain vision transformer backbones for object detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 280–296.
32. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
33. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [[CrossRef](#)]
34. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision transformer with deformable attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4794–4803.
35. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
36. Hassani, A.; Walton, S.; Li, J.; Li, S.; Shi, H. Neighborhood attention transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 6185–6194.
37. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185.
38. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
39. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 815–825.
40. Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. Mobile-former: Bridging mobilenet and transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5270–5279.
41. Han, Y.; Han, D.; Liu, Z.; Wang, Y.; Pan, X.; Pu, Y.; Deng, C.; Feng, J.; Song, S.; Huang, G. Dynamic perceiver for efficient visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 5992–6002.
42. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
44. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.