CLUENet: Cluster Attention Makes Neural Networks Have Eyes

Xiangshuai Song[1*], **Jun-Jie Huang**[1*], **Tianrui Liu**[1†], **Ke Liang**[1†], **Chang Tang**[2]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China
[2]School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China
sxs@nudt.edu.cn, jjhuang@nudt.edu.cn, trliu@nudt.edu.cn, liangke200694@126.com, tangchang@hust.edu.cn

Abstract

Despite the success of convolution- and attention-based models in vision tasks, their rigid receptive fields and complex architectures limit their ability to model irregular spatial patterns and hinder interpretability, thereby posing challenges for tasks requiring high model transparency. Clustering paradigms offer promising interpretability and flexible semantic modeling, but suffer from limited accuracy, low efficiency, and gradient vanishing during training. To address these issues, we propose the CLUster attEntion Network (CLUENet), a transparent deep architecture for visual semantic understanding. Specifically, we introduce three key innovations, including (i) a Global and Soft Feature Aggregation with a Temperature-Scaled Cosine Attention for capturing long-range dependencies and a Gated Fusion Mechanism for enhanced local modeling, (ii) Hard and Shared Feature Dispatching, and (iii) an Improved Cluster Pooling Block. These enhancements significantly improve both classification performance and visual interpretability. Experiments on CIFAR-100 and Mini-ImageNet demonstrate that CLUENet outperforms existing clustering methods and mainstream visual models, offering a compelling balance of accuracy, efficiency, and transparency.

Code — https://github.com/52KunKun/CLUENet

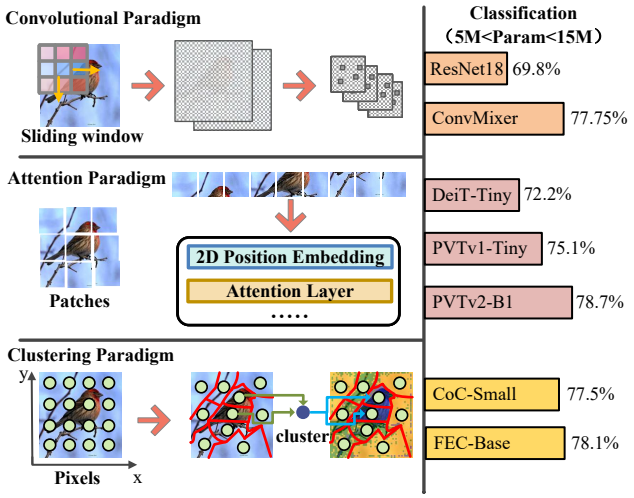


Figure 1: Different paradigms of visual models. Top: convolutional paradigms using convolution for local information modeling. Middle: attention paradigms leveraging positional embeddings and attention for global feature modeling. Bottom: clustering paradigms employing coordinate-guided clustering for semantic modeling. (Representative models are shown on the right, w/ 5M to 15M parameters.)

Introduction

Deep Neural Networks have become the primary approach for image analysis, and various deep architectures have been proposed and differ in their modeling of intrinsic image structures. Early deep learning methods predominantly relied on the convolutional paradigm (LeCun et al. 2002; Simonyan and Zisserman 2014; He et al. 2016; Huang et al. 2017; Trockman and Kolter 2022; Ma et al. 2018; Sandler et al. 2018). These Convolutional Neural Networks (CNNs) extract features by applying shared-weight kernels locally via sliding windows, inherently offering translation equivariance and local context modeling (LeCun et al. 2002). With the rise of Transformers, the attention-based paradigm (Dosovitskiy et al. 2020; Touvron et al. 2021; Wang et al. 2021, 2022b; Liu et al. 2021) has emerged as a powerful alternative for image analysis. Transformers leverage global modeling capabilities to extract features through adaptive weight allocation. However, this comes at the cost of significant computational complexity.

Despite significant progress in vision tasks, deep networks remain limited in modeling the workings of human visual system (Biederman 1987; Beutter and Stone 2000; Bill et al. 2020). Unlike artificial neural network systems, the human visual system dynamically clusters scenes by grouping semantically consistent parts (Liang et al. 2023; Wilson 2003; Ahissar and Hochstein 2004). Conversely, convolutional and attention paradigms rely on fixed geometric windows (convolutional kernels or patches), constraining their ability to adaptively capture irregular objects in images (Chen et al. 2024). Moreover, both paradigms prioritize performance through escalating architectural complex-

*These authors contributed equally.
†Corresponding author.

ity, generating opaque abstractions that lack human interpretability (Guidotti et al. 2018).

To bridge this gap, the clustering paradigm (Liang et al. 2023; Chen et al. 2024; Ma et al. 2023) has emerged as a transparent alternative inspired by clustering algorithms. This approach explicitly models images as sets of semantically coherent pixel groups, directly mirroring human perceptual organization. Pioneering this direction, Ma *et al.* (Ma et al. 2023) introduced Context-Cluster (CoC) for general vision tasks, leveraging unsupervised clustering at multiple feature levels to enable intrinsic interpretability via cluster visualization. Building on this, Chen *et al.* (Chen et al. 2024) proposed Feature Extraction with Clustering (FEC) which integrated clustering into pooling layers, enhancing model performance and enabling receptive field tracing for each feature point. ClusterFormer (Liang et al. 2023) incorporated cross-attention within iterative Expectation-Maximization (EM) steps to refine initial clusters solely from adjacency, enhancing cluster initialization and expanding spatial receptive fields.

Despite the enhanced model transparency offered by clustering paradigms, three key challenges persist: (1) Suboptimal performance: Current clustering-based models underperform relative to state-of-the-art vision architectures (Trockman and Kolter 2022; Wang et al. 2022b), typically exceeding only pre-2022 baselines. (2) Limited receptive field: Computational constraints force methods like CoC and ClusterFormer to perform local clustering within isolated windows. This design lacks inter-window communication, compromising semantic continuity and feature quality. (3) Gradient vanishing in specialized components: FEC's similarity projection layer in cluster pooling suffers from gradient vanishing on small datasets, freezing during training and causing redundancy and inefficiency.

In this paper, we propose a novel CLUster attEtion Network (CLUENet) to address the above discussed limitations. The contributions of CLUENet are summarized as follows:

- Global Soft Aggregation and Hard Assignment allows CLUENet to compute global similarities between cluster centers and all pixels to form soft clusters via weighted fusion, while incorporating a gated residual module to supplement local context and a global query head to enable precise hard assignment for each pixel.
- Efficient Aggregation with Shared Assignment employs cosine attention with learnable temperature in half-precision via FlashAttention for fast, memory-efficient aggregation, and shares hard assignment matrices across blocks within each stage to reduce redundancy and enhance stability.
- Improved Cluster Pooling performs clustering and pooling in similarity space and projects results back to feature space via a perceptron, effectively alleviating gradient vanishing and enhancing performance.
- CLUENet achieves Top-1 accuracies of 76.55% on CIFAR-100 and 82.44% on Mini-ImageNet, outperforming existing clustering paradigm models and showing superior performance across paradigms.

Related Work

Model interpretability. Deep Neural Networks (DNNs) have achieved remarkable success in computer vision, yet their black-box nature raises concerns, especially in safety-critical domains such as medical diagnosis and autonomous driving (Ribeiro, Singh, and Guestrin 2016). Convolutional Neural Networks (CNNs) inherently lack intrinsic interpretability, making it difficult to intuitively reveal their decision rationale. Model-inspired deep architectures (Huang and Dragotti 2020, 2022; Huang et al. 2025; Pu et al. 2022) improve model interpretability by embedding the formation model and priors into the model. Existing visualization methods mainly focus on class activation maps (Zhou et al. 2016), feature visualization (Selvaraju et al. 2017), and input sensitivity analysis (Sundararajan, Taly, and Yan 2017). While these approaches improve interpretability to some extent, they are largely post-hoc and do not fully reveal the true internal decision mechanisms. With the rise of Vision Transformers (ViTs), attention-based interpretability has become a new research hotspot, including attention visualization (Dosovitskiy et al. 2020) and emergent segmentation phenomena in self-supervised learning (Caron et al. 2021). However, the interpretability of ViTs remains limited and uncertain. On the one hand, attention weights may be influenced by input noise or randomness during training, questioning their reliability as explanations (Jain and Wallace 2019). On the other hand, most ViT interpretability still relies on post-hoc methods rather than uncovering internal decision processes. Recently, research focus has been shifting towards enhancing model transparency through intrinsically interpretable model architectures. Sparse Transformers (Child et al. 2019) introduce sparse attention to highlight decision-critical features. Deep Nearest Centroids (Wang et al. 2022a) offers analogy-based explanations by learning feature-to-centroid distances. And white-box Transformers like CRATE (Yu et al. 2023) are designed for observable and traceable attention paths and feature flows.

Clustering-based Model. The clustering paradigm proposed in 2023 (Ma et al. 2023) innovatively applies the Simple Linear Iterative Clustering (Achanta et al. 2012) to general visual representation tasks. This approach achieves strong performance in classification, detection, segmentation, and 3D point cloud processing while generating semantic grouping and visual explanations during inference, offering intuitive decision cues for users. As the first systematic clustering paradigm model, CoC (Ma et al. 2023) represents images as spatially structured pixel groups and extracts features through a deep architecture inspired by a simplified clustering algorithm. Recently, ClusterFormer (Liang et al. 2023) integrates clustering with cross-attention in iterative EM steps. However, it relies on local windows, limiting its receptive field. FEC (Chen et al. 2024) supports global clustering but simplifies feature aggregation by averaging pixels per cluster without similarity weighting, reducing memory usage but risking gradient vanishing and limiting semantic complexity due to fewer clusters. Other recent works extend clustering and semantic aggregation to broader settings, highlighting the generality of cluster-based semantic modeling (Ju et al. 2023; Liang et al. 2025; Guo et al. 2024).

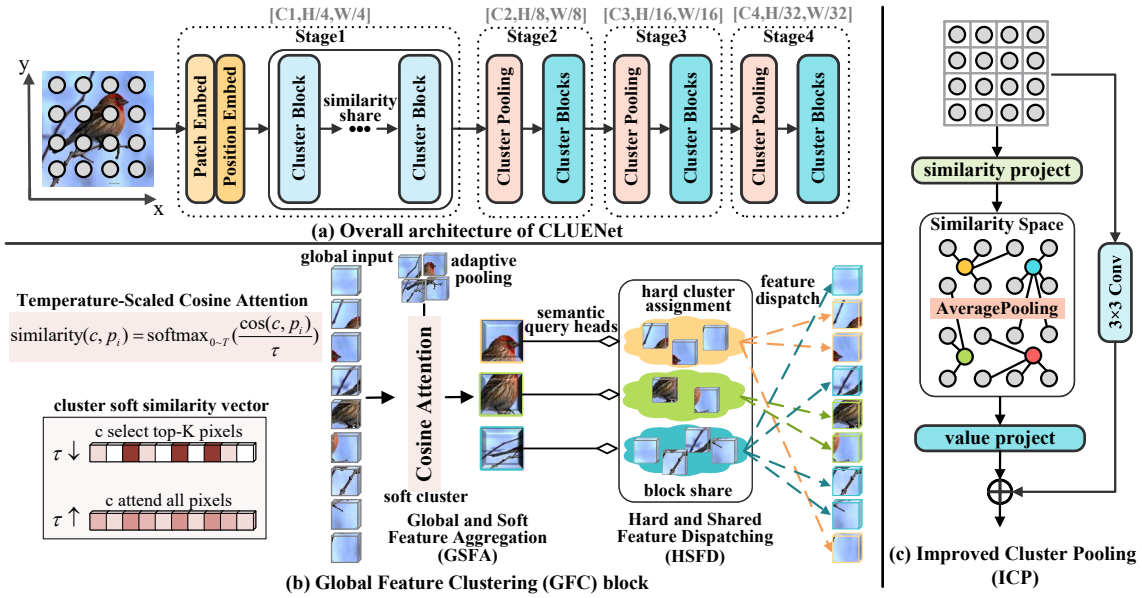


Figure 2: (a) Overall architecture of the CLUENet with a four-stage pyramid network; (b) The key components within the Global Feature Clustering (GFC) block, illustrating Global and Soft Feature Aggregation (GSFA) that updates cluster centers from all pixels, and Hard and Shared Feature Dispatching (HSFD) that updates pixel features according to their assigned cluster centers; (c) The Improved Cluster Pooling (ICP) block, depicting how pixel features are grouped into clusters in similarity space while preserving hierarchical structure.

Proposed Method

In this paper, we present CLUster attEntion Network (CLUENet), a transparent deep architecture for visual semantic understanding built on the clustering paradigm. As shown in the overview of CLUENet in Fig. 2(a), the proposed CLUENet uses a four-stage pyramid structure and consists of three novel components: (i) A Positional-aware Feature Embedding (PFE) Block encodes spatial relationships while preserving semantic locality. (ii) Global Feature Clustering (GFC) Blocks (see Fig. 2(b)) enable interpretable and flexible semantic modeling through global-soft aggregation paired with efficient hard dispatching, overcoming the fixed-window limitations. (iii) Improved Cluster Pooling (ICP) Blocks (see Fig. 2(c)) perform semantic information aggregation by grouping pixels into clusters in similarity space, maintaining hierarchical structure while mitigating gradient vanishing. In the following, we introduce the key components of CLUENet in detail.

Positional-aware Feature Embedding (PFE) Block

For visual semantic clustering, positional information plays a crucial role in distinguishing structured visual regions. Conventional methods rely on fixed grid coordinates for positional embedding. However, this rigid approach lacks adaptability to scale variations. This limitation becomes particularly detrimental in scale-sensitive tasks such as instance and semantic segmentation (Strudel et al. 2021).

To improve robustness to scale variations, we introduce a learnable convolutional positional embedding inspired by PosCNN (Strudel et al. 2021) and PVTv2 (Wang et al. 2022b). It employs lightweight depth-wise convolutions (DWConv) to encode spatial positional information. Formally, given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and fixed grid coordinates $\mathbf{G} \in \mathbb{R}^{H\times W\times 2}$ with $\mathbf{G}_{ij} = [\frac{i}{W} - 0.5, \frac{j}{H} - 0.5]$, the positional embedding is then performed as follows:

$$\begin{aligned} \mathbf{X} &= \text{PatchEmbed}(\text{Concat}[\mathbf{I}, \mathbf{G}]), \\ \mathbf{X} &= \mathbf{X} + \text{DWConv}(\mathbf{X}), \end{aligned} \tag{1}$$

where $\text{DWConv}(\cdot)$ denotes a depth-wise convolution operation with kernel size k, and PatchEmbed refers to a non-overlapping convolutional layer with a patch size of 4×4. The embedding is inserted after the initial patch embedding and within the feed-forward network of each attention block.

Global Feature Clustering (GFC) Block

The GFC block serves as the core visual semantic clustering component in CLUENet. As shown in Fig. 3, the GFC block integrates a Global and Soft Feature Aggregation (GSFA) for cluster center updating and a Hard and Shared Feature Dispatching (HSFD) for pixel feature updating. They jointly enable the effective capture of semantic relationships while maintaining high computational efficiency.

Global and Soft Feature Aggregation (GSFA). Global aggregation enables each cluster center to dynamically attend to all pixels across the image, capturing long-range dependencies beyond local neighborhoods and promoting continuous clusters instead of disjoint window-based fragmented segments. And soft aggregation provides flexible, weighted attention over multiple pixels, enhancing feature aggregation and improving training stability.

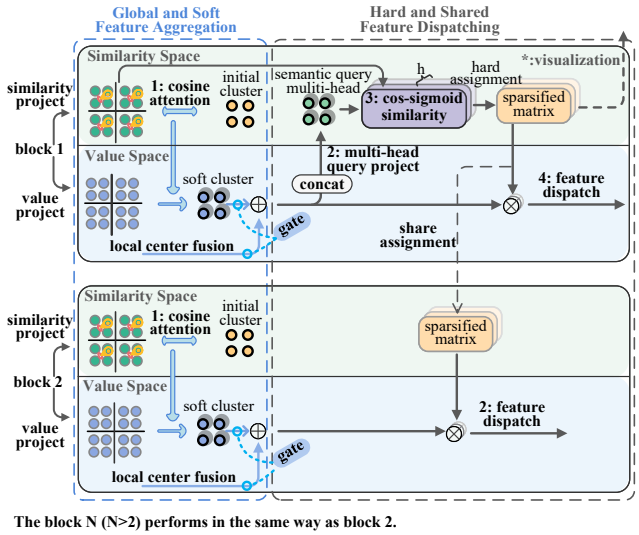


Figure 3: The details of the Global Feature Clustering (GFC) block. Global and Soft Feature Aggregation (GSFA) includes center initialization, Temperature-Scaled Cosine Attention, and Gated Fusion Mechanism. Hard and Shared Feature Dispatching (HSFD) includes multi-head query projection, hard clustering, and assignment shared across blocks.

As illustrated in Fig. 3, the input feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times d}$ is first projected into similarity and value spaces via two linear transformations $\mathbf{W}_i \in \mathbb{R}^{d'\times 1\times 1\times d}$, producing feature maps $\mathbf{P}_i = \mathbf{W}_i \otimes \mathbf{F} \in \mathbb{R}^{H\times W\times d'}$ for $i \in \{s, v\}$. The cluster centers $\mathbf{C}_i \in \mathbb{R}^{h\times w\times d'}$ are then obtained by applying 2D adaptive pooling on $\mathbf{P}_i$, partitioning the feature map into an $h \times w$ grid, yielding $m = h \times w$ cluster centers:

$$\mathbf{C}_i = \text{AdaptivePool}(\mathbf{P}_i, h, w). \tag{2}$$

The cluster centers and pixel features are reshaped into 1D sequences $\mathbf{c}_i \in \mathbb{R}^{m\times d'}$, $\mathbf{p}_i \in \mathbb{R}^{n\times d'}$, where $n = H \times W$.

Temperature-Scaled Cosine Attention: We propose a Temperature-Scaled Cosine Attention in which cosine similarity with a learnable temperature parameter and softmax operator is computed between cluster centers and all pixels, which is then used to perform weighted aggregation in the value space:

$$\begin{aligned}\mathbf{S}_C &= \text{softmax}\left(\frac{\mathbf{c}_s \cdot \mathbf{p}_s^\top}{\tau\|\mathbf{c}_s\|\|\mathbf{p}_s\|}\right),\\ \mathbf{c'}_v &= \mathbf{S}_C \cdot \mathbf{p}_v,\end{aligned} \tag{3}$$

where $\mathbf{S}_C \in \mathbb{R}^{m\times n}$ is the similarity matrix between cluster centers and all pixels, $\mathbf{c}'_v \in \mathbb{R}^{m\times d'}$ represents the soft cluster centers obtained by weighted aggregation, and τ is a learnable temperature parameter.

This design allows each cluster center to retain similarity weights for all pixels across the global field while emphasizing differences in similarity, enabling selective and nuanced aggregation.

Gated Fusion Mechanism: To flexibly incorporate local contextual information, we further introduce a gated fusion mechanism to replace the uniform local residual fusion used in prior work (Ma et al. 2023). Specifically, grid-based local centers are first aggregated from the original features via grid partitioning, and then fed together with the soft cluster centers into the gating module:

$$\begin{aligned}\mathbf{g} &= \sigma\left(f([\mathbf{c}_v, \mathbf{c}'_v])\right),\\ \widetilde{\mathbf{C}}_v &= (1-\mathbf{g}) \circ \mathbf{c}'_v + \mathbf{g} \circ \mathbf{c}_v,\end{aligned} \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, $f(\cdot)$ represents the gating network implemented as a two-layer perceptron, and $\mathbf{g} \in [0, 1]^m$ is the gating weight vector which is broadcast along the feature dimension.

Hard and Shared Feature Dispatching (HSFD). We propose a hard and shared dispatching strategy to enable discrete and interpretable semantic modeling. Hard assignment enhances semantic distinctiveness by enforcing exclusive pixel-to-cluster associations, while shared dispatch rules across blocks within the same stage ensure consistency and reduce computational cost.

The obtained cluster centers are projected into semantic query heads $\mathbf{q} \in \mathbb{R}^{m\times d'}$ (here we assume a single-head configuration for clarity) via an additional mapping layer. The query head then computes the cosine similarities with pixel features $\mathbf{p}_s$ in the similarity space, followed by a sigmoid activation (referred to as cos-sigmoid in Fig. 3).

$$\mathbf{S}_P = \sigma\left(\alpha\left(\frac{\mathbf{p}_s \cdot \mathbf{q}^\top}{\|\mathbf{p}_s\|\|\mathbf{q}\|}\right) + \beta\right), \tag{5}$$

where α and β are learnable scalars, initialized to 1 and 0, respectively.

A hard assignment strategy is then applied, where each pixel selects only its most similar cluster center. The resulting sparse similarity matrix $\widetilde{\mathbf{S}}_P$ is reused in the subsequent blocks:

$$\begin{aligned}\widetilde{\mathbf{S}}_P[i,j] &= \mathbf{S}_P[i,j] \cdot \mathbf{1}\left(j = \arg\max_k \mathbf{S}_P[i,k]\right),\\ \mathbf{P}' &= \mathbf{P} + \text{FC}\left(\widetilde{\mathbf{S}}_P \cdot \widetilde{\mathbf{C}}_v\right),\end{aligned} \tag{6}$$

where $\mathbf{S}_P \in \mathbb{R}^{n\times m}$ denotes the similarity between each pixel and the query heads, $\widetilde{\mathbf{S}}_P$ is the sparsified matrix after hard assignment, and FC$(\cdot)$ denotes a fully connected layer used to project features from hidden dimension d' back to the original dimension d.

Multi-Head Clustering: Inspired by multi-head attention (Vaswani et al. 2017), we split semantic queries and pixel features along the feature dimension into M heads, enabling parallel cluster assignment and improving semantic modeling accuracy. Specifically, $\mathbf{p}_i$ is first partitioned into M heads $\mathbf{p}_i = \{\mathbf{p}_i^{(k)} \in \mathbb{R}^{n\times d'/M}\}_{k=1}^M, i \in \{s, v\}$. These multi-head pixel features induce multi-head cluster centers in the value space $\widetilde{\mathbf{C}}_v = \{\widetilde{\mathbf{C}}_v^{(k)} \in \mathbb{R}^{m\times d'/M}\}_{k=1}^M$. Then, the multi-head cluster centers $\widetilde{\mathbf{C}}_v$ are concatenated along the channel dimension and linearly projected via a learnable matrix $\mathbf{W}_q \in \mathbb{R}^{d'\times d'}$ to form query features $\mathbf{q}' =$

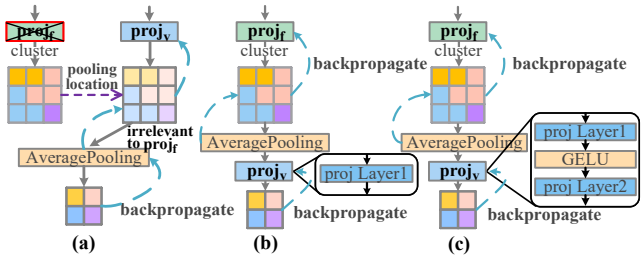


Figure 4: Cluster pooling configurations. (a) In FEC, proj_f only guides pixel selection and does not update, (b) connecting proj_f and proj_v avoids gradient issues but offers limited performance gains, (c) The proposed cluster pooling adopts a two-layer perceptron for proj_v, enabling effective training and improved performance.

$\text{concat}\left(\widetilde{\mathbf{C}}_v^{(1)}, \ldots, \widetilde{\mathbf{C}}_v^{(M)}\right) \cdot \mathbf{W}_q \in \mathbb{R}^{m \times d'}$. The projected queries $\mathbf{q}'$ are also split back into M heads $\mathbf{q} = \{\mathbf{q}^{(k)} \in \mathbb{R}^{m \times d'/M}\}_{k=1}^{M}$ which serve as semantic query heads for the hard assignment in each subspace.

Improved Cluster Pooling (ICP) Block

We identify a key limitation in the cluster pooling design of FEC (Chen et al. 2024). As shown in Fig. 4, the projection layer proj_f which maps features into the similarity space, merely marks pixel-cluster associations during the forward pass but does not participate in the actual aggregation. This leads to a disconnection between the forward data flow and the learnable parameters of proj_f, resulting in ineffective gradient backpropagation and a near-identity behavior during training.

To address this, we propose to perform average pooling directly in the similarity space, making the aggregation explicitly dependent on proj_f's output and ensuring effective gradient updates. Meanwhile, the proj_v mapping layer is redefined to map the aggregated similarity information back to the feature space. Notably, we adopt a two-layer perceptron for proj_v instead of a linear layer, as experiments show that single-layer designs offer limited performance gains. With these modifications, both proj_f and proj_v are effectively optimized during training, leading to significant performance improvements.

Experiment

This section presents a comprehensive evaluation of the proposed CLUENet, conducts visualization analysis and ablation studies[1].

Experiment Settings

Datasets. CIFAR-100 and Mini-ImageNet are used for the experiment. CIFAR-100 contains 50K training and 10K validation images of size 32×32. Mini-ImageNet is a subset of ImageNet-1K with 100 classes and 600 images per class.

[1]Further experimental details, visualizations, and failure case analyses are available in the supplementary material.

	Method	#Param	FLOPs	Top-1	FPS	Memory
Convolution	ResNet	14.16	186.21	69.84	$16858_{\pm 431}$	1.1
	ConvMixer	2.78	175.88	67.21	**20433**$_{\pm 908}$	**0.6**
	ShuffleNet	5.55	186.83	**71.09**	$14660_{\pm 290}$	1.0
	MobileNet	2.37	68.43	66.62	$17241_{\pm 862}$	0.7
Attention	ViT	3.22	224.12	56.93	**16809**$_{\pm 570}$	**0.8**
	PVTv2	3.43	172.88	**70.77**	$12711_{\pm 158}$	1.2
	CPVT	3.12	155.27	66.09	$14885_{\pm 215}$	1.2
	Swin	5.15	245.84	65.33	$11496_{\pm 181}$	1.6
Cluster	CoC	2.72	161.11	71.92	**10712**$_{\pm 98}$	**1.4**
	FEC	2.83	197.12	69.73	$9663_{\pm 362}$	1.4
	ClusterFormer	2.92	173.47	66.05	$8041_{\pm 46}$	1.6
	CLUENet (ours)	3.02	188.88	**76.55**	$7807_{\pm 280}$	1.4

Table 1: Comparison with representative backbones on CIFAR-100 benchmark.

We split Mini-ImageNet 4:1 into training and validation sets, ensuring balanced class distribution.

We follow the standard data augmentation procedure during training. For CIFAR-100, random horizontal flip and normalization are applied. For Mini-ImageNet, random crop, horizontal flip, color jitter, random erasing, and normalization are applied for training.

Training Details. All models are trained using the AdamW optimizer with momentum 0.9 and cosine decay. Weight decay is set to 0.05. The learning rate is auto-tuned using `findLR` (Smith 2018), with a warmup of 5 epochs. The batch size for CIFAR-100 is 256, and the batch size for Mini-ImageNet is 128. All experiments are conducted on a computer with an NVIDIA RTX 4090 (24GB) GPU.

Evaluation Protocol. Following common practice (Liang et al. 2023; Chen et al. 2024; Ma et al. 2023), all models are evaluated on the validation set after completing 100 training epochs, rather than using the model with the best validation performance, to ensure a fair comparison. Each model is evaluated 7 times; the first 2 runs are discarded to avoid cold-start effects. We report the average of the last 5 runs for classification accuracy (Top-1 and Top-3), throughput (Frames Per Second, FPS), number of parameters (#Param), FLOPs, and inference memory usage, while all metrics except FPS show no significant change.

Image Classification Results

Results on CIFAR-100. As shown in Table 1, in the convolution paradigm, ShuffleNetv2 achieves the best Top-1 accuracy of 71.09%. For the attention paradigm, PVTv2 leads with 70.77%. Among clustering paradigm, CLUENet stands out with a Top-1 accuracy of 76.55%, surpassing existing clustering model CoC by 4.63% and outperforming PVTv2 and ShuffleNetv2 by 5.78% and 5.46%, respectively.

Results on Mini-ImageNet. Table 2 summarizes the classification results on the Mini-ImageNet validation set. Under similar parameter scales, CLUENet (base) achieves the best performance among clustering-based models, surpassing CoC (medium), FEC (large), and ClusterFormer (tiny)

	Method	#Param	FLOPs	Top-1	Top-3	FPS	Memory
Convolution-based method	ResNet18 (He et al. 2016)	14.17	2.38	76.95	89.88	1150.75$_{\pm 27.15}$	2.7
	ShuffleNetv2 (x1.5) (Ma et al. 2018)	2.58	0.31	**78.39**	**90.40**	**1194.72**$_{\pm 23.26}$	**1.1**
	ShuffleNetv2 (x2.0) (Ma et al. 2018)	6.52	0.64	**79.63**	**90.93**	**1205.17**$_{\pm 7.59}$	**1.3**
	ConvNeXtv2 (A) (Woo et al. 2023)	3.42	0.55	71.15	84.97	1172.04$_{\pm 34.80}$	1.6
	ConvNeXtv2 (F) (Woo et al. 2023)	4.89	0.79	73.14	86.33	1177.66$_{\pm 21.15}$	1.8
	ConvNeXtv2 (N) (Woo et al. 2023)	15.05	2.46	75.04	87.48	1148.19$_{\pm 28.45}$	2.8
Attention-based method	PVTv2 (b0) (Wang et al. 2022b)	3.44	0.54	75.34	88.52	1195.32$_{\pm 31.63}$	1.7
	PVTv2 (b1) (Wang et al. 2022b)	13.55	2.06	77.57	89.86	1154.58$_{\pm 39.86}$	2.9
	EfficientFormer (s0) (Li et al. 2023)	3.26	0.40	**77.92**	**89.71**	1166.09$_{\pm 21.46}$	1.5
	EfficientFormer (s1) (Li et al. 2023)	5.76	0.66	**78.97**	**90.32**	1128.59$_{\pm 0.75}$	1.5
	EfficientFormer (s2) (Li et al. 2023)	12.16	1.27	**79.75**	**90.42**	695.41$_{\pm 1.71}$	1.5
	EfficientViT (m2) (Liu et al. 2023)	3.99	0.20	73.59	86.78	**1173.13**$_{\pm 22.48}$	**0.9**
	EfficientViT (m3) (Liu et al. 2023)	6.61	0.26	73.88	87.16	**1132.69**$_{\pm 24.25}$	**1.0**
	EfficientViT (m5) (Liu et al. 2023)	12.13	0.52	75.32	88.03	**1160.99**$_{\pm 44.32}$	**1.0**
Clustering-based method	CoC (tiny) (Ma et al. 2023)	5.28	1.12	74.90	88.29	**887.11**$_{\pm 18.92}$	**2.2**
	CoC (small) (Ma et al. 2023)	14.20	2.80	76.56	89.13	**883.46**$_{\pm 29.31}$	**3.2**
	CoC (medium) (Ma et al. 2023)	28.83	5.96	78.39	90.29	605.67$_{\pm 3.18}$	**4.2**
	ClusterFormer (Liang et al. 2023)†	5.63	1.24	70.73	85.73	669.00$_{\pm 4.52}$	9.3
	ClusterFormer (Liang et al. 2023)†	15.01	3.02	73.93	87.29	694.91$_{\pm 5.99}$	12.0
	ClusterFormer (tiny) (Liang et al. 2023)	30.27	5.58	73.99	88.26	588.56$_{\pm 5.11}$	9.6
	FEC (small) (Chen et al. 2024)	5.44	1.38	76.74	89.34	854.25$_{\pm 37.49}$	5.2
	FEC (base) (Chen et al. 2024)	14.63	3.37	77.81	90.26	859.63$_{\pm 16.23}$	5.3
	FEC (large) (Chen et al. 2024)	29.26	6.55	79.33	90.55	550.69$_{\pm 2.58}$	7.6
	CLUENet (micro)	3.02	0.65	**78.75**	**90.83**	871.89$_{\pm 13.84}$	4.2
	CLUENet (tiny)	5.68	1.30	**80.51**	**91.29**	867.76$_{\pm 17.89}$	5.5
	CLUENet (small)	15.05	3.16	**81.49**	**92.06**	878.60$_{\pm 21.68}$	6.5
	CLUENet (base)	30.20	6.40	**82.44**	**92.47**	**679.05**$_{\pm 7.12}$	7.7

Table 2: Comparison with representative backbones on Mini-ImageNet benchmark. († indicates architectures adjusted to reach the specified parameter scale.)

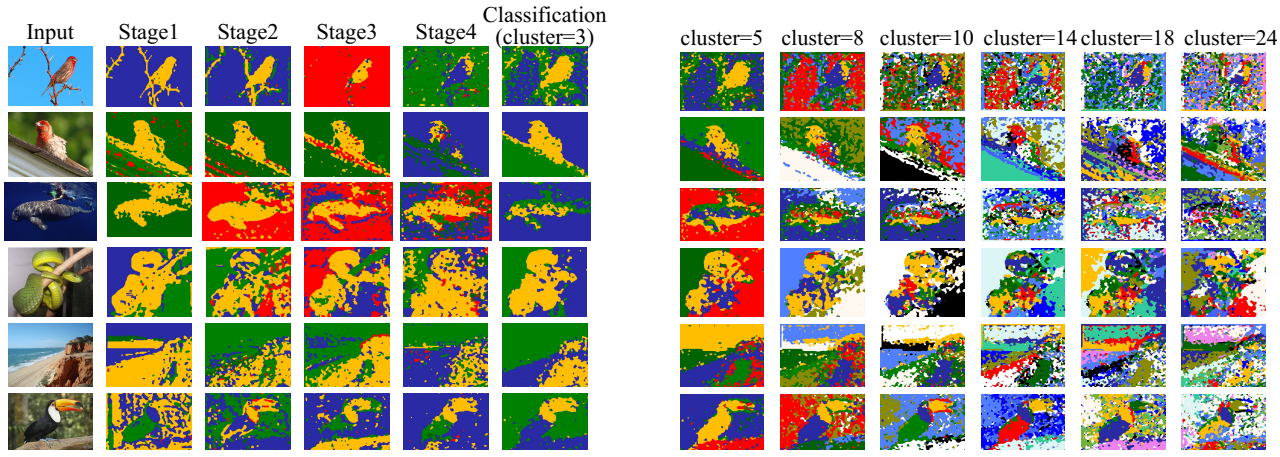
by 4.05%, 3.11%, and 8.45% in Top-1 accuracy, and by 2.18%, 1.92%, and 4.21% in Top-3 accuracy, respectively. Without using windowing mechanisms while adopting a relatively dense cluster count, CLUENet (base) achieves an inference speed of 679.05 img/s, which is comparable to that of FEC (550.69 img/s) and CoC (605.67 img/s), demonstrating its efficient architecture and strong feature representation.

Compared with the Attention-based models, CLUENet (micro) outperforms the attention-based PVTv2 (b0) by 3.41% in Top-1 accuracy and 2.31% in Top-3 accuracy while using 0.42M fewer parameters. It also surpasses EfficientFormer (s0) by 0.83% in Top-1 and 1.12% in Top-3 accuracy with a similar parameter budget. CLUENet (tiny) outperforms EfficientFormer (s1) by 1.54% in Top-1 and 0.97% in Top-3 accuracy. Furthermore, CLUENet (small) surpasses PVTv2 (b1) and EfficientFormer (s2) by 3.92% and 1.74% in Top-1, respectively.

Compared with the convolution-based models, CLUENet (micro) achieves slightly higher Top-1 and Top-3 accuracy than ShuffleNetv2 (x1.5) (+0.36%, +0.43%), and CLUENet (tiny) outperforms ShuffleNetv2 (x2.0) and ResNet18 by 0.88% and 3.56% in Top-1 accuracy, and by 0.36% and 1.41% in Top-3 accuracy, respectively. Moreover, CLUENet (small) surpasses both ResNet18 and ConvNeXtv2 (N) by 4.54% and 6.45% in Top-1 accuracy, respectively.

Visualization Analysis

Model Visualization. The proposed CLUENet offers intrinsic interpretability through its architecture. We visualize the cluster receptive fields of CLUENet (tiny) across all four stages on the Mini-ImageNet classification task. As shown in Fig. 5, the model captures multi-level semantic structures at different stages, reflecting a hierarchical feature extraction process similar to human visual perception. For clarity, only one semantic head is visualized per stage. Additionally, we visualize the receptive field of each position on the 7×7 feature map from the fourth stage to highlight the semantic areas most influential for classification. Since directly displaying all 49 clusters can be overwhelming, we adopt the K-Means-based cluster merging method proposed by Chen *et al.* (Chen et al. 2024) to reduce the number of visualized clusters. To preserve the interpretability purity of the network, we visualize the fourth-stage 7×7 feature map with increasing cluster numbers by adjusting the K-Means merge parameter.



(a) The semantic clustering.

(b) The impact of cluster numbers.

Figure 5: Visualization of (a) the clustering results of semantic heads at each of the four stages, along with the global receptive field map *w.r.t.* the final classification decision, and (b) the global receptive field map *w.r.t.* different cluster numbers.

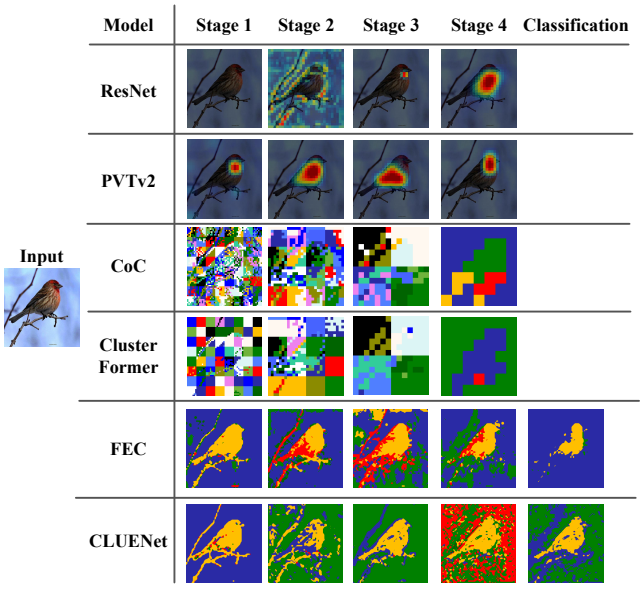


Figure 6: Visualization results of different paradigm models. ResNet shows Grad-CAM activation maps at four stages. PVTv2 presents attention maps based on the center query token in the first attention head of the last block at each stage. CoC, ClusterFormer and FEC visualize one attention head from one block per stage. FEC and CLUENet additionally provide the final-stage semantic clustering maps used for classification.

As illustrated in Fig. 5a, CLUENet exhibits a progressive semantic understanding. Early stages (1–3) focus on the object's overall contour and location, separating it from the background to form clear object boundaries. The deeper stages (3–4) progressively shift attention to finer and more discriminative regions critical for classification. For example, in the first row of Fig. 5a, the model initially treats the bird and branch as a whole, but by Stage 3, it distinguishes the branch as background and focuses on the bird's head and torso in Stage 4. As shown in Fig. 5b, when the number of clusters is further increased, the same semantic object is split into multiple semantic parts, capturing finer details of the object. Importantly, all visualized samples are correctly classified, and the final-stage feature maps (the last column of Fig. 5a) consistently exhibit clear cluster structures that concentrate on the target object and effectively distinguish it from the background.

Cross-Paradigm Visualization Comparison. Fig. 6 presents a comparative visualization analysis of different paradigms on the Mini-ImageNet classification task. For clustering-based models (CoC, ClusterFormer, FEC and CLUENet), we directly visualize the semantic clusters at each stage. For FEC and our proposed CLUENet, we additionally show the final-stage clustering maps to illustrate their global semantic representations. For convolution-based models (*e.g.*, ResNet), we use Grad-CAM (Selvaraju et al. 2017) to highlight class-discriminative regions. For attention-based models (*e.g.*, PVTv2), we visualize the attention response to the center query in the first attention head of the last block at each stage. It is observed that ResNet converges on the head, while PVTv2 shifts focus from the torso to the head. Both CoC and ClusterFormer produce fragmented clusters due to local window constraints, thereby lacking coherent global semantics. In contrast, FEC and CLUENet consistently exhibit clear foreground-background separation across stages, indicating stronger interpretability. Moreover, CLUENet combines high accuracy, efficiency, and transparency, surpassing FEC in overall performance.

Ablation Study

We validate the effectiveness of model components on the Mini-ImageNet image classification task.

Key Components: As shown in Table 3(a), we conducted ablation studies on five core components of the

FA	TCosAttn	Gate	Shared	PosEmb	Param	FLOPs	Top-1	Top-3
✗	✗	✗	✓	✓	2.87	0.518	76.97	89.62
✓	✗	✗	✓	✓	2.98	0.649	77.87	90.27
✓	✓	✗	✓	✓	2.98	0.649	78.46	90.38
✓	✓	✓	✗	✓	3.12	0.663	78.32	90.05
✓	✓	✓	✓	✗	2.97	0.621	78.21	90.17
✓	✓	✓	✓	✓	3.02	0.649	**78.75**	**90.83**

(a) Key Component

Design		Top-1	Top-3	FPS	FLOPs
w/o Cluster Pooling		77.38	89.78	899.03	0.539
Parallel	FEC	78.07	89.96	846.61	0.769
	FEC ($\mathrm{proj_f}$ = identity)	78.09	89.96	893.92	0.635
Sequential	Ours (Single Layer)	78.10	90.07	879.83	0.641
	Ours (2-layer MLP)	**78.75**	**90.83**	871.89	0.649
	Ours (3-layer MLP)	78.71	90.41	856.19	0.667

(b) Cluster Pooling Configurations

Design	Top-1	Top-3
w/o Query	77.72	90.06
AvgPool Query	78.02	90.11
Single-head center Query	78.41	90.27
Multi-head center Query	**78.75**	**90.83**

(c) Multi-Head Semantic Query

Design	FLOPs	Mem	FPS	Top-1	Top-3
+Fewer Clusters ($\leq$16)	0.637	4.1	882.63	78.06	90.18
+Window Partition	0.636	4.1	881.40	78.30	90.37
Ours (64)	0.653	4.2	870.36	78.70	90.62
Ours (49)	0.649	4.2	871.89	**78.75**	**90.83**

(d) Model Design Comparison

Table 3: Ablation Studies of Core Components, Cluster Pooling, Query Design, and Model Variants

model to evaluate their individual contributions. Specifically, these components include: Feature Aggregation (FA), Temperature-Scaled Cosine Attention (TCosAttn), Gated Residual (Gate), Shared Dispatching (Shared), and Learnable Positional Embedding (PosEmb). Since both TCosAttn and Gated Residual are embedded within the global and soft feature aggregation module, removing FA also disables the other two. Disabling FA leads to a significant drop in accuracy (Top-1: –1.78%, Top-3: –1.21%), highlighting the importance of global and soft feature aggregation in semantic integration. Replacing TCosAttn with standard attention reduces Top-1 and Top-3 accuracy by 0.59% and 0.11%, respectively, indicating its advantage in modeling semantic hierarchy and improving cluster-to-pixel alignment. Removing the Gated Residual and using fixed weights results in a smaller drop (–0.29%, –0.45%), showing that dynamic gating better adapts to local feature injection. Disabling shared dispatching causes accuracy to decrease by 0.43% (Top-1) and 0.78% (Top-3), along with increased parameters and computation. Finally, removing the learnable positional embedding leads to a 0.54% drop in Top-1 and 0.66% in Top-3 accuracy, suggesting its role in enhancing spatial awareness. Overall, all five components contribute meaningfully to performance and work together to improve the model's representation capacity.

Cluster Pooling: Table 3(b) compares various cluster pooling designs. The similarity mapping layer $\mathrm{proj_f}$ in FEC shows performance nearly identical to an identity mapping, indicating model redundancy. Even when both $\mathrm{proj_f}$ and $\mathrm{proj_v}$ layers learn jointly (Fig. 4b), accuracy gains are negligible. Our improved cluster pooling module achieves a notable Top-1 accuracy gain (78.75% vs. 78.07%) while reducing computational cost, validating the design's efficiency. Increasing $\mathrm{proj_v}$ depth from two to three layers yields no improvement, thus, the two-layer MLP is selected for balance.

Multi-Head Semantic Query: Table 3(c) shows that cluster center–guided queries significantly improve performance (78.75% vs. 77.72%), demonstrating enhanced semantic understanding. Queries based on initial average pooling lack the flexibility to capture rich semantic structures in natural images (78.75% vs. 78.02%), showing limitations compared to our cluster-guided queries. Multi-head queries outperform single-head queries (78.75% vs. 78.41%), confirming stronger semantic capture.

Design Comparison: Table 3(d) compares CLUENet with window partitioning (from CoC) and fewer clusters ($\leq$16 at all stages, from FEC). While both reduce resource consumption, they cause Top-1 accuracy drops of 0.69% and 0.45%, respectively, indicating that these simplifications trade off the model's semantic modeling and representation capabilities. Additionally, increasing the number of clusters from 49 to 64 does not provide any clear performance improvement.

Conclusions

We proposed CLUENet, a novel visual model that balances strong performance with inherent interpretability. By introducing a Global Feature Clustering Block with efficient attention and shared assignment across the global scope, we improve performance, computational efficiency, and visualization quality. The improved cluster pooling overcomes gradient vanishing while maintaining excellent performance and model interpretability. Extensive experiments on CIFAR-100 and Mini-ImageNet show that CLUENet achieves superior accuracy compared to existing clustering-based and mainstream models, while also providing clear and intuitive semantic interpretability. Future work will focus on enhancing semantic information flow across stages to improve robust recognition in complex scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Project 62572480, 62201600, 62201604, 62506371, 62522604 and 62476258.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11): 2274–2282.

Ahissar, M.; and Hochstein, S. 2004. The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10): 457–464.

Beutter, B. R.; and Stone, L. S. 2000. Motion coherence affects human perception and pursuit similarly. *Visual neuroscience*, 17(1): 139–153.

Biederman, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115.

Bill, J.; Pailian, H.; Gershman, S. J.; and Drugowitsch, J. 2020. Hierarchical structure is employed by humans during visual motion perception. *Proceedings of the National Academy of Sciences*, 117(39): 24581–24589.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, G.; Li, X.; Yang, Y.; and Wang, W. 2024. Neural clustering based visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5714–5725.

Child, R.; Gray, S.; Radford, A.; and Sutskever, I. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5): 1–42.

Guo, S.; Wang, Q.; Gao, Y.; Xie, R.; Li, L.; Zhu, F.; and Song, L. 2024. Depth-guided robust point cloud fusion NeRF for sparse input views. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9): 8093–8106.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, J.-J.; and Dragotti, P. L. 2020. Learning deep analysis dictionaries for image super-resolution. *IEEE Transactions on Signal Processing*, 68: 6633–6648.

Huang, J.-J.; and Dragotti, P. L. 2022. WINNet: Wavelet-inspired invertible network for image denoising. *IEEE Transactions on Image Processing*, 31: 4377–4392.

Huang, J.-J.; Liu, T.; Chen, Z.; Liu, X.; Wang, M.; and Dragotti, P. L. 2025. A lightweight deep exclusion unfolding network for single image reflection removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jain, S.; and Wallace, B. C. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Ju, W.; Gu, Y.; Chen, B.; Sun, G.; Qin, Y.; Liu, X.; Luo, X.; and Zhang, M. 2023. Glcc: A general framework for graph-level clustering. In *Proceedings of the AAAI conference on artificial intelligence*, 4391–4399.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, Y.; Hu, J.; Wen, Y.; Evangelidis, G.; Salahi, K.; Wang, Y.; Tulyakov, S.; and Ren, J. 2023. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16889–16900.

Liang, J.; Cui, Y.; Wang, Q.; Geng, T.; Wang, W.; and Liu, D. 2023. Clusterfomer: clustering as a universal visual learner. *Advances in neural information processing systems*, 36: 64029–64042.

Liang, K.; Meng, L.; Li, H.; Wang, J.; Lan, L.; Li, M.; Liu, X.; and Wang, H. 2025. From Concrete to Abstract: Multi-view Clustering on Relational Knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; and Yuan, Y. 2023. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14420–14430.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ma, N.; Zhang, X.; Zheng, H.-T.; and Sun, J. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

Ma, X.; Zhou, Y.; Wang, H.; Qin, C.; Sun, B.; Liu, C.; and Fu, Y. 2023. Image as set of points. *arXiv preprint arXiv:2303.01494*.

Pu, W.; Huang, J.-J.; Sober, B.; Daly, N.; Higgitt, C.; Daubechies, I.; Dragotti, P. L.; and Rodrigues, M. R. 2022. Mixed X-ray image separation for artworks with concealed designs. *IEEE Transactions on Image Processing*, 31: 4458–4473.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, L. 2018. pytorch-lr-finder: A PyTorch implementation of the learning rate range test. https://github.com/davidtvs/pytorch-lr-finder. GitHub repository, accessed 2018-05-05.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Trockman, A.; and Kolter, J. Z. 2022. Patches are all you need? *arXiv preprint arXiv:2201.09792*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Han, C.; Zhou, T.; and Liu, D. 2022a. Visual recognition with deep nearest centroids. *arXiv preprint arXiv:2209.07383*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022b. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3): 415–424.

Wilson, H. R. 2003. Computational evidence for a rivalry hierarchy in vision. *Proceedings of the National Academy of Sciences*, 100(24): 14499–14503.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16133–16142.

Yu, Y.; Buchanan, S.; Pai, D.; Chu, T.; Wu, Z.; Tong, S.; Haeffele, B.; and Ma, Y. 2023. White-box transformers via sparse rate reduction. *Advances in Neural Information Processing Systems*, 36: 9422–9457.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.