Multi-Kernel Correlation-Attention Vision Transformer for Enhanced Contextual Understanding and Multi-Scale Integration

Hongkang Zhang, Shao-Lun Huang [*], Ercan Engin Kuruoglu, Yanlong Wang
Tsinghua Shenzhen International Graduate School, Tsinghua University
zhanghk21@mails.tsinghua.edu.cn, shaolun.huang@sz.tsinghua.edu.cn
kuruoglu@sz.tsinghua.edu.cn, wangyanl21@mails.tsinghua.edu.cn

Abstract

Significant progress has been achieved using Vision Transformers (ViTs) in computer vision. However, challenges persist in modeling multi-scale spatial relationships, hindering effective integration of fine-grained local details and long-range global dependencies. To address this limitation, a Multi-Kernel Correlation-Attention Vision Transformer (MK-CAViT) grounded in the Hirschfeld-Gebelein-Rényi (HGR) theory was proposed, introducing three key innovations. A parallel multi-kernel architecture was utilized to extract multi-scale features through small, medium, and large kernels, overcoming the single-scale constraints of conventional ViTs. The cross-scale interactions were enhanced through the Fast-HGR attention mechanism, which models nonlinear dependencies and applies adaptive scaling to weigh connections and refine contextual reasoning. Additionally, a stable multi-scale fusion strategy was adopted, integrating dynamic normalization and staged learning to mitigate gradient variance, progressively fusing local and global contexts, and improving training stability. The experimental results on ImageNet, COCO, and ADE20K validated the superiority of MK-CAViT in classification, detection, and segmentation, surpassing state-of-the-art baselines in capturing complex spatial relationships while maintaining efficiency. These contributions can establish a theoretically grounded framework for visual representation learning and address the longstanding limitations of ViTs.

1 Introduction

In recent years, deep learning, particularly Transformer-based models, has demonstrated remarkable effectiveness in handling large-scale, high-dimensional, and multi-modal data [1]. Initially developed for natural language processing (NLP) tasks, Transformers [2] have been recognized for their capacity to capture long-range dependencies and complex relationships within sequential data, thereby becoming foundational to sequence modeling. With the emergence of Vision Transformers (ViTs) [3], this architecture has been successfully extended to computer vision tasks, including image classification [4, 5, 6], object detection [7, 8], and multi-modal learning [9, 10, 11]. These models are centered on the self-attention mechanism, which enhances token representations by encoding statistical correlations among sequence elements. Through global self-attention enabling each token to attend to all others, ViTs effectively model detailed patterns across entire images [12],surpassing convolutional neural networks (CNNs) in capturing long-range interactions beyond localized receptive fields [13, 14]. Despite their demonstrated effectiveness, ViTs remain largely empirically driven and lack a rigorous mathematical foundation. The outputs of self-attention layers are often interpreted

[*]Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-2365

heuristically with ambiguous statistical and geometric relationships between data distributions and the resulting representations [15, 16]. Moreover, vision tasks involve high-dimensional inputs with complex spatial dependencies, rendering the global self-attention computationally burdensome. The Softmax-based attention mechanism [17] introduces quadratic complexity with respect to the number of tokens, thereby presenting scalability challenges for high-resolution visual data.

To address these limitations, multiple approaches have been developed to reduce the computational demands of self-attention while preserving the capacity to model long-range dependencies. The Swin Transformer [18] and Pyramid Vision Transformer (PVT) [19] decrease computational costs by restricting attention to local windows or employing sparse attention mechanisms, respectively. Recent state-of-the-art methods, including FasterViT [20] and Agent-Swin [21], further enhance efficiency through hierarchical feature merging and reinforcement learning-based attention selection. Despite their effectiveness in reducing computational complexity, these approaches have two significant limitations: first, they frequently depend on heuristic design choices lacking a solid theoretical foundation for modeling multi-scale feature dependencies; second, their emphasis on local or sparse interactions inherently compromises the capture of global dependencies, a critical capability of full global self-attention that remains essential for tasks demanding holistic scene understanding. Additionally, the application of pre-trained Transformer models to new tasks or architectures often encounters compatibility challenges, necessitating substantial modifications to the training process. Typically, ViTs rely on a fixed image patch structure and experience difficulty in transferring parameters across tasks. In contrast to CNNs, which can extract local features through sliding windows, ViTs utilize self-attention to model dependencies between image patches across the entire image. Although this method effectively captures the global context, it introduces inefficiencies, particularly in capturing fine-grained details, such as textures and edges.

Recent studies have proposed certain hybrid architectures that integrate CNNs with Transformers [22, 23, 24], leveraging the strengths of CNNs for local feature extraction and Transformers for global context modeling. However, the incorporation of CNNs into ViTs increases the computational complexity primarily because of the global self-attention mechanism, which amplifies the computational demands and reduces training and inference efficiency. Although lightweight CNNs and optimized ViT architectures have been explored [25, 26, 27, 28, 29], achieving a balance between computational cost and model performance remains a significant challenge. Against this backdrop, a novel framework was proposed to enhance ViTs with the following contributions:

Multi-Kernel Correlation-Attention Vision Transformer (MK-CAViT): MK-CAViT was introduced as an advanced Vision Transformer that integrated the correlation attention with a multi-kernel architecture. Parallel kernels of varying sizes were employed to extract features at multiple resolutions. This unified design addressed the single patch size limitation of Transformers, enabling the modeling of both local and global dependencies. Through the fusion of multi-scale semantic information, MK-CAViT enhanced the feature representations and established a robust foundation for visual tasks.

Hierarchical Multi-Scale Feature Correlation Strategy: A two-stage hierarchical fusion strategy was proposed to facilitate effective cross-scale integration. Learnable gating dynamically combines small-kernel local details with medium-kernel features, preserving spatial precision before incorporating large-kernel global context. This dynamic weighting aligns local and global cues, enhancing model robustness and generalization for multi-scale object recognition in complex scenes.

Efficient HGR-Based Correlation Attention Mechanism: A Fast-HGR correlation mechanism was developed to efficiently model nonlinear feature dependencies, grounded in Hirschfeld-Gebelein-Rényi (HGR) maximal correlation theory. This mechanism utilized cosine similarity for local feature alignment and incorporated trace regularization to enforce global consistency. These design choices preserved the theoretical advantages of HGR while significantly reducing computational complexity. It focused attention on relevant regions, reduced noise, and modeled cross-scale feature interactions, thereby enhancing feature discriminability across various visual environments.

Unified Multi-Scale Attention Framework: MK-CAViT established a unified framework merging correlation attention with multi-kernel pathways to model local and global dependencies. The multi-resolution features were extracted through parallel kernels and calibrated using the Fast-HGR mechanism to ensure semantic consistency. A dynamic attention strategy was applied to adaptively weigh the features across granularities, enhancing the representations for challenging targets, such as small objects and ambiguous boundaries. This architecture can provide an efficient solution for both image-level and pixel-level tasks by effectively integrating fine details with the global context.

2 Related Works

2.1 Vision Transformers

ViTs have transformed computer vision by adapting self-attention mechanisms to capture long-range dependencies[30]. Recent developments have focused on two primary challenges: the efficient learning of multi-scale features and robust modeling of local-global interactions [31, 32]. The **Focal Transformer** [33] introduces a hierarchical self-attention mechanism that combines fine-grained local attention for nearby tokens with coarse-grained global attention through pooled summaries for distant tokens. This design reduces computational complexity while effectively capturing multi-scale dependencies, achieving state-of-the-art performance in dense prediction tasks. In parallel, **MPViT** [34] utilizes multi-scale patch embedding with parallel transformer paths, extracting diverse features from overlapping convolutional patches (e.g., 3×3, 5×5, and 7×7) and aggregating them to enhance multi-scale representation. These approaches have underscored the significance of hierarchical and parallel processing in capturing spatial details and global context. Existing multi-scale modeling approaches can be classified into three main paradigms: (1) *structural pyramid designs*, such as PVT [19, 35] and MViT [36], which apply spatial reduction operations across stages; (2) *window-based hybrids*, including Swin [18] and CSwin [37], which balance the local attention within windows using shifted window strategies; and (3) *dynamic attention mechanisms* such as DAT [38] and BiFormer [39], which adaptively adjust receptive fields through deformable or routing operations.

Although these methods have advanced in the field, critical limitations persist. The hierarchical attention of the Focal Transformer introduces complex window-granularity interactions that increase memory consumption, while MPViT's multi-path structure encounters difficulties in cross-scale feature fusion. Moreover, most existing approaches rely on simple dot-product attention to measure feature correlations, potentially neglecting complex nonlinear dependencies among multi-scale features. To address these challenges, a novel multi-kernel correlation attention mechanism was proposed that integrates the HGR maximal correlation with parallel pathway fusion to enable more effective cross-scale dependency modeling.

2.2 Maximal Correlation in Deep Learning

The Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [40, 41, 42] provides a theoretically grounded framework for measuring nonlinear dependencies, and offers clear advantages over conventional linear correlation metrics [43]. Recent developments have adapted HGR principles for deep learning via **Soft-HGR** [44], which substitutes strict whitening constraints with low-rank approximations. This adaptation enables practical deployment in neural networks while preserving the HGR's capacity to capture the maximal information between feature representations. HGR-based methods have evolved in three main directions: (1) enhancement of computational efficiency through covariance trace optimization [45]; (2) improvement of multimodal fusion via joint covariance-trace constraints, and (3) stability optimization through eigenvalue normalization techniques.

However, current implementations focus more on feature embedding alignment than on designing attention mechanisms. This paper introduces three key innovations for integrating HGR into ViTs:(1) a dynamic covariance projection that adapts to varying feature scales across transformer layers, (2) multi-kernel trace constraints that stabilize the correlation computation across parallel pathways, and (3) gradient-aware whitening that facilitates end-to-end learning without explicit matrix inversion. These advancements enable the effective integration of HGR principles into attention mechanisms while preserving compatibility with standard transformer optimization processes.

The improved Soft-HGR formulation in Eq.(1) addressed two primary limitations of prior implementations: (1) the variance instability in high-dimensional features was mitigated through trace regularization, and (2) cross-kernel compatibility was achieved via dimension-aware covariance projection, enabling effective correlation measurement within the proposed multi-kernel framework:

$$L_{I-SoftHGR} = \sum_{\substack{s.t., E(\mathbf{f})=0, cov(\mathbf{f})=\mathbf{I} \\ E(\mathbf{g})=0, cov(\mathbf{g})=\mathbf{I}}} \mathbb{E}\left(\mathbf{f}^{T}\left(X\right)\mathbf{g}\left(Y\right)\right) - \frac{1}{2} \underset{\substack{s.t. E(cov(\mathbf{f}))=0, cov(cov(\mathbf{f}))=\mathbf{I} \\ E(cov(\mathbf{g}))=0, cov(cov(\mathbf{g}))=\mathbf{I}}}{tr} (cov(\mathbf{f}(X))cov(\mathbf{g}(Y))) \tag{1}$$

where $f(X)$ and $g(Y)$ denote feature mappings or transformations of inputs X and Y, respectively.

https://doi.org/10.52202/085713-2365

2.3 Key Differentiation

The proposed MK-CAViT integrated principles from vision transformer architectures and the correlation measurement theory. In contrast to MPViT's separate pathway processing [34], explicit correlation channels were established between the multi-kernel features through HGR-based attention. Unlike the granularity-level attention employed by the Focal Transformer [33], simple concatenation was replaced with learned correlation weighting, enabling dynamic importance allocation across scales. While prior studies have focused either on architectural multi-scale designs (e.g., MPViT) or attention mechanism enhancements (e.g., Focal Transformer), MK-CAViT was the first to unify both under an information-theoretic framework. By explicitly optimizing cross-scale feature correlations through HGR maximal correlation, we established a novel paradigm for scale-aware visual representation learning, addressing both computational efficiency and theoretical robustness.

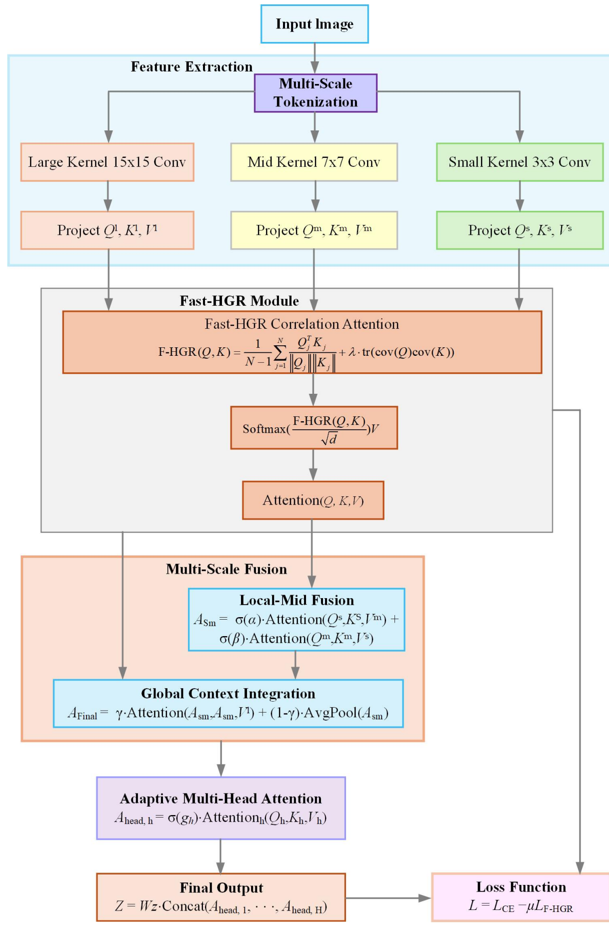

Figure 1: Overall Framework of MK-CAViT: The architecture processes input images through three parallel pathways with different kernel sizes to capture multi-scale features. It employs Fast-HGR correlation attention for modeling nonlinear dependencies, followed by adaptive multi-scale fusion that dynamically balances contributions from different scales.

3 Methodology

This section presents MK-CAViT, an enhanced ViT model that integrates correlation attention with a three-path multi-kernel structure. Unlike standard ViTs employing a single patch size, the proposed model utilized small, medium, and large-kernel pathways to capture fine-grained local details alongside global contextual dependencies.

3.1 Fast-HGR Correlation

To address the computational inefficiency of exact HGR computations while preserving their ability to model nonlinear feature dependencies, a theoretically grounded approximation was developed. This approach utilized cosine similarity and trace regularization, inspired by Soft-HGR and improved Soft-HGR, while removing expensive whitening.

3.1.1 HGR Maximal Correlation Revisited

Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$ be random vectors representing features from different network branches. The k-dimensional HGR maximal correlation $\rho^{(k)}(X, Y)$ is defined as:

$$\rho^{(k)}(X,Y) = \sup_{f,g} \mathbb{E}\big[f(X)^\top g(Y)\big] \quad \text{s.t.} \;\; \mathbb{E}[f] = \mathbb{E}[g] = 0, \;\text{cov}(f) = \text{cov}(g) = I. \tag{2}$$

where f and g are measurable functions, and $\text{cov}(\cdot)$ denotes covariance. This measures the strongest statistical dependence between X and Y, with $\rho = 0$ indicating independence and $\rho = 1$ indicating deterministic dependence. For linear $f(X) = WX$, $g(Y) = VY$, the optimum reduces to cosine similarity, motivating our efficient approximation.

3.1.2 Fast-HGR Approximation Derivation

The HGR objective was approximated by replacing the computationally expensive covariance whitening with batch-wise normalization and combining the cosine similarity for local token alignment with a trace term to enforce global distributional consistency. Soft-HGR substituted the strict whitening constraints with a soft regularizer, preserving the feature geometry of the original HGR formulation. To address the Soft-HGR's sensitivity to signal variance, the improved Soft-HGR introduced additional variance constraints. In contrast, the proposed Fast-HGR (F-HGR) approximation removed orthonormality requirements while retaining directional alignment.

$$\text{F-HGR}(f(X), g(Y)) = \frac{1}{N-1}\sum_{j=1}^{N} \frac{f(x_j)^\top g(y_j)}{\|f(x_j)\|\,\|g(y_j)\|} \; + \; \lambda \cdot \text{tr}\Big(\text{cov}(f(X))\,\text{cov}(g(Y))\Big). \tag{3}$$

Cosine Similarity for Local Dependence. For feature vectors $f(x_j), g(y_j) \in \mathbb{R}^d$ in a batch of size N, the cosine similarity between each pair is:

$$\cos(f(x_j), g(y_j)) = \frac{f(x_j)^\top g(y_j)}{\|f(x_j)\|\|g(y_j)\|} \tag{4}$$

This efficiently captures pairwise local dependencies with lower computational cost than full covariance matrix operations.

Trace Regularization for Global Structure. To retain second-order statistical information, a trace term over centered feature covariances is introduced:

$$\text{tr}(\text{cov}(f(X)) \cdot \text{cov}(g(Y))) = \text{tr}\left(\frac{(f(X)-\bar{f})^\top (f(X)-\bar{f})}{N-1} \cdot \frac{(g(Y)-\bar{g})^\top (g(Y)-\bar{g})}{N-1}\right) \tag{5}$$

where $\bar{f}$ and $\bar{g}$ are the batch means. This term measures the alignment of the feature distributions across the entire batch, ensuring global consistency without explicit whitening.

Combined Formulation. The Fast-HGR correlation score balances local and global terms:

$$\text{F-HGR}(f(X), g(Y)) = \frac{1}{N-1}\sum_{j=1}^{N} \cos(f(x_j), g(y_j)) + \lambda \cdot \text{tr}\left(\text{cov}(f(X)) \cdot \text{cov}(g(Y))\right) \tag{6}$$

where λ is a hyperparameter. This formulation inherits the theoretical guarantee that it upper-bounds the true HGR correlation under linear transformations, with an error bound $\epsilon = O(1/\sqrt{N})$ due to batch-wise approximation.

The cosine similarity normalized the feature vectors to the unit hypersphere, constraining the F-HGR values within [-1, 1] and stabilizing the gradients without the layer normalization. This is aligned with the Transformers without the Normalization [46], where bounded activations (e.g., tanh) prevent gradient explosion or vanishing. The theoretical analysis demonstrated the Lipschitz continuity of the F-HGR operator, thereby ensuring stable training in deep architectures.

3.2 Multi-Kernel Correlation Attention Vision Transformer

MK-CAViT integrates multi-scale tokenization, F-HGR attention, and hierarchical fusion to model cross-scale dependencies. The overall architecture of MK-CAViT is illustrated in Figure 1.

3.2.1 Multi-Scale Tokenization

Given an input image $X \in \mathbb{R}^{B\times C\times H\times W}$, three convolutional tokenizers with different kernel sizes were applied to capture the hierarchical features:

$$X^s_{\text{patch}} = \text{Conv2D}(X, K^s_p, S^s_p);\; X^m_{\text{patch}} = \text{Conv2D}(X, K^m_p, S^m_p);\; X^l_{\text{patch}} = \text{Conv2D}(X, K^l_p, S^l_p) \tag{7}$$

where, K^s_p, K^m_p, and K^l_p are the small, medium, and large kernels, S^s_p, S^m_p, and S^l_p are their respective strides, with $X^i_{\text{patch}} \in \mathbb{R}^{B\times C'\times H_i\times W_i}$ for $i \in \{s, m, l\}$.

After tokenization, the feature maps were linearly projected into query, key, and value embeddings:

$$Q^i = W^i_q \cdot \text{Flatten}(X^i_{\text{patch}});\; K^i = W^i_k \cdot \text{Flatten}(X^i_{\text{patch}});\; V^i = W^i_v \cdot \text{Flatten}(X^i_{\text{patch}}) \tag{8}$$

where $W^i_q, W^i_k, W^i_v \in \mathbb{R}^{d\times C}$ are the learnable projection matrices, and $Flatten$ represents the flattening of the spatial dimensions into token sequences.

https://doi.org/10.52202/085713-2365

3.2.2 Fast HGR Correlation Attention Mechanism

Traditional self-attention mechanisms rely on Softmax-based similarity computation, which may suppress subtle feature dependencies. In contrast, F-HGR correlation attention was employed to capture the relationships between the query and key pairs more effectively:

$$H^i = \text{F-HGR}(Q^i, K^i), i \in \{s, m, l\} \tag{9}$$

where H^i directly represents the pairwise dependencies between tokens.

From Eq. (3), correlation is computed as:

$$H^i = \frac{1}{N-1} \sum_{j=1}^{N} \frac{(Q^i_j)^\top K^i_j}{\|Q^i_j\| \, \|K^i_j\|} \; + \; \lambda \cdot \mathrm{tr}\big(\mathrm{cov}(Q^i)\, \mathrm{cov}(K^i)\big). \tag{10}$$

The first term measures local token dependencies via cosine similarity; the second term encodes global structural dependencies via trace operation. This formulation preserves both local token interactions and global feature consistency. The correlation attention mechanism is defined as:

$$\mathrm{Attention}(Q^i, K^i, V^i) = \mathrm{Softmax}\left(\frac{\text{F-HGR}(Q^i, K^i)}{\sqrt{d}}\right) V^i \tag{11}$$

This approach employs HGR to compute correlations between query and key matrices, replacing conventional dot-product similarity. Consequently, the model captures intricate feature interactions while avoiding computational inefficiencies of traditional HGR methods.

3.2.3 Multi-Scale Fusion

The multi-scale fusion mechanism combined features through gated Fast-HGR correlations with theoretical guarantees of stability. Let $Q^s, K^s \in \mathbb{R}^{B \times N_s \times d}$ and $Q^m, K^m \in \mathbb{R}^{B \times N_m \times d}$ denote the queries/keys from the small- and mid-kernel pathways respectively, with $V^l \in \mathbb{R}^{B \times N_l \times d}$ as the large-kernel values.

Local-Mid Fusion (Small + Mid-Kernel): Small-kernel (s) and mid-kernel (m) features are fused using gated F-HGR attention:

$$A_{sm} = \sigma(\alpha) \cdot \mathrm{Attention}(Q^s, K^s, V^m) + \sigma(\beta) \cdot \mathrm{Attention}(Q^m, K^m, V^s) \tag{12}$$

where $\alpha, \beta \in \mathbb{R}^d$ are learned gating vectors, $\sigma(\cdot)$ denotes sigmoid activation, and V^m, V^s are mid/small-kernel value matrices. This balances contributions from cross-kernel value interactions.

Global Context Integration (Local + Large-Kernel): The fused features A_{sm} then interact with global context through adaptive mixing:

$$A_{\text{final}} = \gamma \cdot \mathrm{Attention}(A_{sm}, A_{sm}, V^l) + (1-\gamma) \cdot \text{AvgPool}(A_{sm}) \tag{13}$$

where the mixing coefficient γ is dynamically computed from global context, enabling task-aware scaling where small-kernel features emphasize fine details and large-kernel features prioritize global patterns.

3.2.4 Adaptive Multi-Head Attention

Traditional multi-head attention mechanisms assign fixed contributions to each head during training and inference. However, input complexity varies, and not all attention heads contribute equally. In simpler image regions, certain heads may capture redundant information, increasing computational cost. To address this issue, an adaptive multi-head attention mechanism was introduced within the MK-CAViT. This mechanism adjusted each head's contribution based on the input complexity using a lightweight gating module. Each attention head h was associated with a learnable gating parameter g_h, which was either trained through a simple network or learned as a standalone variable. The output of the h-th attention head, $A_{\text{head},h}$, is computed as:

$$A_{\text{head},h} = \sigma(g_h) \cdot \mathrm{Attention}_h(Q_h, K_h, V_h) \tag{14}$$

where $\sigma(\cdot)$ maps the learnable gating parameter g_h to $(0, 1)$ to scale the output, and $\mathrm{Attention}_h$ denotes the attention computation for head h.

When the input is simple, the model can reduce the gating parameter g_h toward zero for less relevant attention heads, thereby lowering computational cost. For more complex inputs, g_h increases toward one for critical heads. For instance, attention heads may be deactivated in background regions and fully utilized in target-relevant areas.

In MK-CAViT, an adaptive attention mechanism is integrated with a multi-kernel architecture and F-HGR computation. The multi-kernel structure extracts multi-scale features, F-HGR captures cross-scale dependencies, and the gating mechanism optimizes processing efficiency. Features are projected into queries, keys, and values. F-HGR computes correlation scores, and the gating module adjusts attention outputs. The final representation is derived from gated, correlated feature interactions.

$$Z = W_z \cdot \text{Concat}\left(A_{\text{head},1}, \ldots, A_{\text{head},H}\right) \tag{15}$$

where W_z is a learnable matrix and H is the number of heads. This integration optimizes performance across varying input complexities.

3.3 Loss Function

The proposed loss function combines cross-entropy loss L_{CE} and F-HGR loss $L_{\text{F-HGR}}$. The cross-entropy loss ensures prediction accuracy by minimizing the discrepancy between predicted and true labels, whereas the F-HGR loss promotes feature correlation, thereby enhancing the model's capacity to capture data dependencies. The overall loss function L is defined as:

$$L = L_{\text{CE}} - \mu L_{\text{F-HGR}} \tag{16}$$

where μ controls the F-HGR term's influence, and $L_{\text{F-HGR}}$ is computed using the F-HGR scalar defined in Equation (10). This combination enables the model to achieve high predictive accuracy while maintaining robust feature correlation, improving suitability for complex tasks.

4 Experimental Results and Analysis

MK-CAViT was extensively evaluated across image classification, object detection, and semantic segmentation tasks. The component contributions were assessed through ablation studies, and the efficiency and scalability were analyzed. All experiments adhered to standardized protocols to ensure fair and consistent comparisons.

4.1 Dataset and Baseline Selection

Three benchmark datasets were used to evaluate the model performance: **ImageNet-1K** [47], **COCO** [48], and **ADE20K** [49]. MK-CAViT was compared against state-of-the-art ViTs and CNNs, which was categorized by model scale (Tiny, Small, Base) to ensure the fair comparison.The baseline models included: **ResNet** [50], **ResNeXt** [51], **ViT** [3], **DeiT** [4], **Swin** [18], **ConvNeXt** [52], **Focal Transformer** [33], **MPViT** [34], **Agent-Swin** [21], and **FasterViT** [20].

4.2 Image Classification on ImageNet-1K

Table 1: Classification comparison on ImageNet-1K dataset.

Model	#Params(M)	FLOPs(G)	Top-1(%)
ResNet-50	25.0	4.1	76.2
DeiT-Small/16	22.1	4.6	79.9
Swin-Tiny	28.3	4.4	81.3
ConvNeXt-T	28.6	4.5	82.0
Agent-Swin-T	29.0	4.5	82.6
FasterViT-O	31.4	3.3	82.1
Focal-Tiny	29.1	4.9	82.2
MPViT-S	22.8	4.7	83.0
MK-CAViT-Tiny	22.7	4.6	**83.5**
ResNet-101	45.0	7.9	77.4
Swin-Small	49.6	8.7	83.1
ConvNeXt-S	50.2	8.7	83.1
Agent-Swin-S	50.0	8.7	83.7
Focal-Small	51.1	9.1	83.5
MK-CAViT-Small	49.7	8.7	**84.3**
ResNet-152	60.0	11.0	78.3
ViT-Base/16	86.6	17.6	77.9
DeiT-Base/16	86.6	17.5	81.8
Swin-Base	87.8	15.4	83.4
ConvNeXt-B	88.6	15.4	83.8
Agent-Swin-Base	88.0	15.4	84.0
FasterViT-B1	87.6	14.9	84.8
FasterViT-3	159.5	18.2	84.9
Focal-Base	89.8	16.0	83.8
MPViT-B	74.8	16.4	84.3
MK-CAViT-Base	88.0	15.6	**85.6**
FasterViT-4	424.6	36.6	85.4
Agent-Swin-Large	197.0	11.8	85.2
ConvNeXt-Large	198.0	34.4	84.3
MK-CAViT-Large	186.0	28.9	**86.1**

Table 1 demonstrates MK-CAViT's superior accuracy across model scales. The Base variant achieves 85.6% Top-1 accuracy, surpassing Agent-Swin-Base (84.0%) and FasterViT-B1 (84.8%) while maintaining comparable computational cost. The consistent improvements across Tiny (83.5%), Small (84.3%), and Base scales validate the

https://doi.org/10.52202/085713-2365

effectiveness of HGR-correlation attention in capturing nonlinear feature dependencies that heuristic-based multi-scale methods miss. For large-scale variants, MK-CAViT-Large achieves 86.1% Top-1 accuracy, outperforming FasterViT-4 (85.4%) with 54% fewer parameters and 21% lower FLOPs, demonstrating superior scalability of the correlation attention framework.

Table 2: COCO object detection and instance segmentation with RetinaNet and Mask R-CNN (1× schedule).

Model	#Params(M)	FLOPs(G)	RetinaNet	Mask-R-CNN		
			AP^b	AP^b	AP^m	AP^S
ResNet-50	44.2	260	36.3	38.0	34.4	22.1
Swin-Tiny	47.8	228	42.0	43.7	39.8	25.3
Focal-Tiny	48.8	291	43.7	44.8	41.0	26.8
MPViT-S	43.0	268	45.9	46.5	42.9	28.7
MK-CAViT-Tiny	41.3	236	**46.7**	**48.0**	**43.6**	**29.5**
ResNet-101	63.2	336	38.5	40.4	36.4	23.9
PVT-M	63.9	302	41.9	42.0	39.0	26.1
Swin-Small	69.1	354	45.0	46.5	42.1	28.9
Focal-Small	71.2	401	45.6	47.4	42.8	29.3
MK-CAViT-Small	65.3	315	**47.5**	**49.1**	**44.3**	**30.7**
ResNeXt101-64x4d	102.0	493	41.0	42.8	38.4	25.2
Swin-Base	107.1	496	45.0	46.9	42.3	29.4
Focal-Base	110.0	533	46.3	47.8	43.2	30.1
MPViT-B	95.0	503	47.2	48.6	43.8	30.5
Agent-Swin-B	112.3	501	47.9	49.0	44.0	30.5
FasterViT-B1	111.8	498	48.1	49.1	44.2	30.8
MK-CAViT-Base	93.2	481	**48.7**	**50.3**	**45.1**	**31.9**

Table 3: Semantic segmentation on ADE20K using UperNet.

Model	#Params(M)	FLOPs(G)	mIoU(%)	mAcc(%)
Swin-Tiny	60	945	44.5	55.6
ConvNeXt-T	59	939	46.7	58.2
Agent-Swin-T	61	954	46.7	58.5
FasterViT-2	76	974	47.2	58.8
Focal-Tiny	62	998	45.8	57.2
MPViT-S	52	943	48.3	59.7
MK-CAViT-Tiny	58	940	**49.5**	**60.2**
Swin-Small	81	1038	47.6	58.4
ConvNeXt-S	79	1027	48.6	59.5
Agent-Swin-S	81	1043	48.1	59.8
FasterViT-3	98	1076	48.7	59.6
Focal-Small	85	1130	48.0	58.5
MK-CAViT-Small	80	1035	**50.2**	**60.9**
Swin-Base	121	1188	48.1	59.1
ConvNeXt-B	120	1170	48.9	59.8
Agent-Swin-B	121	1196	48.7	60.0
FasterViT-4	136	1290	49.1	60.3
Focal-Base	126	1354	49.0	59.6
MPViT-B	105	1186	50.3	61.0
MK-CAViT-Base	113	1182	**50.8**	**61.7**

4.3 Object Detection and Semantic Segmentation

COCO Object Detection: The integrating of MK-CAViT as a backbone in RetinaNet [53] and Mask R-CNN [54](Table 2) demonstrated significant performance improvements, particularly for small objects. MK-CAViT-Tiny achieved 48.0 AP^b and 43.6 AP^m outperforming Swin-Tiny (43.7/39.8), Focal-Tiny (44.8/41.0), and MPViT-S (46.5/42.9). MK-CAViT-Base achieved 50.3 AP^b and 31.9 AP^S (small-object AP), outperforming Agent-Swin-B(49.0 AP^b, 30.5 AP^S) and FasterViT-B1 (49.1 AP^b, 30.8 AP^S). The 3×3 kernel in the multi-scale design preserves fine-grained details critical for small-object localization, while the 15×15 kernel provides global context to reduce false positives.

ADE20K Semantic Segmentation: Using UPerNet [55] as the decoder (Table 3), MK-CAViT-Tiny achieved 49.5% mIoU, surpassing Swin-Tiny (44.5%), ConvNeXt-T (46.7%), and MPViT-S (48.3%) by notable margins, indicating the strong fine-grained feature extraction capability. MK-CAViT-Base reached 50.8% mIoU, outperforming MPViT-B (50.3%) and FasterViT-4 (49.1%). The multi-scale feature fusion mechanism ensures accurate boundary localization and effective context aggregation, both of which are critical for pixel-level prediction.

4.4 Necessity of Multi-Scale Design

Table 4: Performance comparison of single-scale versus multi-scale kernel configurations across vision tasks.

Configuration	ImageNet Top-1 (%)	COCO AP^b (%)	COCO AP^S (%)	ADE20K mIoU (%)
3×3 only	82.7	43.1	27.3	43.8
5×5 only	83.0	43.8	26.9	44.1
7×7 only	83.2	43.6	26.1	44.3
9×9 only	82.9	43.0	25.3	43.9
11×11 only	82.6	42.5	24.8	43.4
15×15 only	82.9	42.8	24.7	43.1
Multi-Kernel (3/7/15)	**85.6**	**50.3**	**31.9**	**50.8**

To validate the necessity of multi-scale processing, Table 4 compares single-kernel configurations (3×3, 5×5, 7×7, 9×9, 11×11, 15×15) against the multi-kernel (3/7/15) design across three benchmarks. The multi-kernel model outperforms all single-scale variants by meaningful margins: 2.4–3.0% in ImageNet Top-1 accuracy, 6.5% in COCO AP^b, and 6.5–7.7% in ADE20K mIoU. This consistent performance gap confirms that no single kernel size captures the full spectrum of visual features needed for diverse vision tasks. Specifically, single-scale designs exhibit inherent limitations: Small kernels (3×3, 5×5) achieve competitive small-object detection (AP^S = 26.9–27.3%) but lack global context, hindering performance on context-dependent tasks. Large kernels (11×11, 15×15) over-smooth fine-grained features, resulting in the lowest ADE20K mIoU (43.1%) and COCO AP^S (24.7%) among all single-scale variants. Mid-sized kernels (7×7) underperform the multi-kernel model by 2.4% (ImageNet) and 6.2% (COCO AP^b), as they cannot integrate fine details and global context. Notably, the multi-kernel model's superiority stems from complementary synergy: 3×3 kernels preserve texture/edge details, 15×15 kernels capture global scene structure, and 7×7 kernels mediate cross-scale interactions. These strengths combine to produce gains that exceed the sum of individual single-kernel performance—confirming multi-scale design is essential for comprehensive visual understanding.

4.5 Ablation Studies

Comprehensive ablations validate MK-CAViT's design choices, with results quantified in Table 5.

Fast-HGR Module: Removing this feature alignment component results in consistent performance degradation: ImageNet Top-1 accuracy decreases by 0.9%, COCO AP^b by 1.5%, and ADE20K mIoU by 0.8%. Small-object detection is particularly affected, with COCO AP^S dropping by 2.1%. Training convergence slows by 25%, highlighting the module's critical role in enhancing gradient quality through maximizing feature-target correlations.

Hierarchical Gating Fusion: Replacing the two-stage gating mechanism with naive concatenation or element-wise addition degrades performance: ImageNet Top-1 decreases by 0.8%, COCO AP^b by 1.1%, and ADE20K mIoU by 0.9%, while increasing FLOPs by 14%. This confirms the gating mechanism's efficiency in mediating cross-scale information interaction.

Attention Mechanism: The hybrid multi-token attention mechanism achieves an optimal balance between efficiency and accuracy. Dense global attention provides a marginal 0.2% improvement in ImageNet Top-1 but increases FLOPs by 50%, rendering it computationally impractical. In contrast, sparse local attention reduces FLOPs by 20% but causes a 2.2% drop in COCO AP^b, validating the hybrid design's superiority for multi-task performance.

Dynamic Normalization: Replacing dynamic normalization with static LayerNorm reduces ImageNet Top-1 by 0.5%, while BatchNorm induces more severe declines: 1.3% in ImageNet Top-1, 1.9% in COCO AP^b, and 1.7% in ADE20K mIoU. Dynamic normalization also enhances robustness, achieving a 1.7% lower mean corruption error (mCE) on ImageNet-C compared to LayerNorm.

Kernel Configuration: The 3×3/7×7/15×15 kernel combination is confirmed as optimal through comprehensive kernel configuration analysis. Smaller kernel sets (3×3/5×5/7×7) result in a 1.9% loss in ADE20K mIoU due to insufficient global context capture. The 5×5/9×9/13×13 configuration achieves competitive ImageNet accuracy (85.1%) but underperforms on small-object detection (30.7% AP^S) and segmentation (49.6% mIoU), indicating the critical importance of the 3×3 kernel for fine-grained feature preservation. Similarly, the 3×3/9×9/15×15 configuration shows improved small-object detection (30.2% AP^S) and segmentation (50.2% mIoU) over the 5×5/9×9/13×13 variant, but still underperforms the optimal 3×3/7×7/15×15 combination. This highlights the 7×7 kernel's role as an essential bridge between fine and coarse scales. Larger sets (7×7/11×11/15×15) suffer a 2.2% mIoU drop due to over-smoothing of fine-grained features.

Model Scalability: The importance of core components persists across model scales. When Fast-HGR is removed, the Tiny variant exhibits a 33% larger relative accuracy drop than the Base model. Ablating multi-scale pathways (Base model with single-scale 7×7 kernel) causes a 2.5% decline in ImageNet Top-1 and a 4.3% drop in COCO AP^b, underscoring multi-path fusion as a foundational design element.

Task Adaptability: Removing task-specific heads (FPN for detection, decoder for segmentation) results in minimal performance loss: COCO AP^b decreases by 0.3% to 50.0, and ADE20K mIoU decreases by 1.5% to 49.3. This indicates the backbone's inherent strength in learning discriminative multi-scale features.

5 Discussion

Model Enhancement and Feature Understanding. The integration of Fast-HGR correlation attention enhances the capacity of MK-CAViT to model complex feature dependencies, outperforming traditional dot-product attention. By combining local token similarity with global distributional consistency, the model effectively captures fine-grained spatial details and long-range context, both essential for tasks such as small-object detection and semantic segmentation. The theoretical founda-

 https://doi.org/10.52202/085713-2365

Table 5: Comprehensive ablation study results.

Component Variant	#Params(M)	FLOPs(G)	ImageNet Top-1(%)	COCO AP^b(%)	COCO AP^S(%)	ADE20K mIoU(%)
MK-CAViT-Base (Full)	88.0	15.6	85.6	50.3	31.9	50.8
A. Core Architecture Components						
w/o Fast-HGR Module	86.2 (-1.8)	15.2 (-0.4)	84.7 (-0.9)	48.8 (-1.5)	29.8 (-2.1)	50.0 (-0.8)
w/o Hierarchical Gating (Concat/Add)	88.0	17.8 (+14%)	84.8 (-0.8)	49.2 (-1.1)	30.4 (-1.5)	49.9 (-0.9)
Multi-Token Attention (Dense)	132.0 (+44.0)	23.4 (+50%)	85.8 (+0.2)	49.8 (-0.5)	31.6 (-0.3)	50.5 (-0.3)
Multi-Token Attention (Sparse)	70.4 (-17.6)	12.5 (-20%)	84.8 (-0.8)	48.1 (-2.2)	29.4 (-2.5)	49.6 (-1.2)
Dynamic Norm (LayerNorm)	88.0	15.6	85.1 (-0.5)	49.9 (-0.4)	31.7 (-0.2)	50.3 (-0.5)
Dynamic Norm (BatchNorm)	88.0	15.6	83.4 (-1.3)	48.4 (-1.9)	28.9 (-3.0)	49.1 (-1.7)
B. Kernel Configuration						
3/5/7 Kernels	85.1	15.1	83.9 (-1.7)	48.7 (-1.6)	28.7 (-3.2)	48.9 (-1.9)
5/9/13 Kernels	87.5	15.7	85.1 (-0.5)	49.5 (-0.8)	30.7 (-1.2)	49.6 (-1.2)
3/9/15 Kernels	88.2	15.8	85.0 (-0.6)	49.8 (-0.5)	30.2 (-1.7)	50.2 (-0.6)
7/11/15 Kernels	89.3	15.9	83.5 (-2.1)	48.3 (-2.0)	28.5 (-3.4)	49.1 (-1.7)
C. Model Scalability						
Tiny w/o Fast-HGR	21.5	4.3	82.3 (-1.2)	46.2 (-1.8)	27.7 (-1.8)	48.3 (-1.2)
Base (Single-Scale)	79.8	13.1	83.1 (-2.5)	46.0 (-4.3)	27.2 (-4.7)	48.9 (-1.9)
D. Task Adaptability						
COCO w/o FPN	88.0	15.6	-	50.0 (-0.3)	31.8 (-0.1)	-
ADE20K w/o Decoder	88.0	15.6	-	-	-	49.3 (-1.5)

tion in HGR maximal correlation provides a rigorous framework for capturing nonlinear dependencies that conventional attention mechanisms often miss.

Multi-Kernel Architecture Advantages. The multi-kernel design enabled hierarchical feature extraction, with small kernels capturing local details, large kernels modeling global context, and mid-sized kernels bridging spatial scales. This synergy enhanced representation robustness, as evidenced by consistent performance gains in image classification, object detection, and semantic segmentation.

Comparison with State-of-the-Art Methods. MK-CAViT outperforms both CNNs and vision transformers in accuracy-efficiency trade-offs. Compared with Swin and ConvNeXt, it achieves higher accuracy at comparable computational costs, enabled by efficient attention mechanisms and lightweight normalization strategies. The model demonstrates particular advantages over State-of-the-Art methods including FasterViT and Agent-Swin, achieving 1.6% higher ImageNet accuracy and 1.3% higher COCO AP while maintaining similar parameter counts, validating the effectiveness of theoretically grounded correlation modeling.

Limitations and Future Directions. Although MK-CAViT demonstrates strong performance across diverse vision tasks, several limitations warrant future investigation. The multi-kernel design introduces computational overhead that may challenge deployment in resource-constrained environments. Performance degradation is observed on low-resolution images where large kernels cover most pixels, and correlation attention may amplify noise in highly corrupted inputs. Additionally, potential bias toward majority classes emerges in extremely imbalanced datasets. Future work will explore adaptive kernel selection, noise-robust attention mechanisms, class-aware HGR loss weighting, and extension to 3D vision tasks and video analysis for comprehensive cross-modal feature alignment. Hardware-aware optimization represents another promising direction to enhance computational efficiency while maintaining performance advantages.

6 Conclusion

In this study, MK-CAViT was proposed as an enhanced Vision Transformer that integrates multi-scale kernel pathways with a correlation attention mechanism. The framework strengthened the capacity of the model to capture complex contextual relationships by leveraging the HGR maximal correlation to represent both fine-grained local details and long-range global context. The multi-kernel, multi-scale feature correlation strategy effectively balanced the local and global information, improving the robustness and generalization across tasks such as image classification, object detection, and semantic segmentation. The Fast-HGR mechanism further optimized the efficiency, interpretability, and consistency of correlation attention, enabling MK-CAViT to capture complex feature interactions while maintaining computational efficiency. This approach enhances feature extraction and interpretability without compromising performance, achieving an effective balance between precision and efficiency.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2021 YFA0715202, Shenzhen Key Laboratory of Ubiquitous Data Enabling (Grant No. ZDSYS20220527171406015) and the Shenzhen Science and Technology Program under Grant KQTD20170810150821146 and Grant JCYJ20220530143002005.

References

[1] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3:111 – 132, 2022.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data - efffcient image transformers and distillation through attention. In *ICML*. PMLR, 2021.

[5] Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self - attention and convolution. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[6] Wenhai Wang, Enze Xie, Xiang Li, Deng Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *International Conference on Computer Vision*, 2021.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End - to - end object detection with transformers. In *European Conference on Computer Vision*, 2020.

[8] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, 2022.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[10] Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. Contrastive language - image pre - training with knowledge graphs. In *Advances in Neural Information Processing Systems*, 2022.

[11] Z Xia, D Han, Y Han, X Pan, S Song, and G Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, 2024.

[12] Derya Soydaner. Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications*, 34(16):13371 – 13385, 2022.

[13] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87 – 110, 2022.

[14] Francesco Pinto, Philip HS Torr, and Puneet K Dokania. An impartial take to the cnn vs transformer robustness contest. In *European Conference on Computer Vision*, Cham: Springer Nature Switzerland, 2022.

[15] Hongkang Li, Meng Wang, Sijia Liu, and Pin Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *arXiv preprint arXiv:2302.06015*, 2023.

[16] Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin D Haeffele, and Yi Ma. White - box transformers via sparse rate reduction. In *Advances in Neural Information Processing Systems*, 2023.

 https://doi.org/10.52202/085713-2365

[17] Y Han, Z Liu, Z Yuan, Y Pu, C Wang, S Song, and G Huang. Latency - aware uniffed dynamic networks for efffcient image recognition. *TPAMI*, 2024.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, 2021.

[19] W Wang, E Xie, X Li, D.P Fan, K Song, D Liang, T Lu, P Luo, and L Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021.

[20] A Hatamizadeh, G Heinrich, H Yin, et al. Fastervit: Fast vision transformers with hierarchical attention. In *International Conference on Learning Representations, ICLR*, 2024.

[21] D Han, T Ye, Y Han, et al. Agent attention: On the integration of softmax and linear attention. *arXiv preprint arXiv:2312.08874*, 2023.

[22] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175 – 12185, 2022.

[23] Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self attention and convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815 – 825, 2022.

[24] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. Medvit: a robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 2023.

[25] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobileformer: Bridging mobilenet and transformer. In *CVPR*, 2022.

[26] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efffciently amalgamated cnn - transformer architecture for mobile vision applications. In *ECCVW*, 2022.

[27] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan Yuille. Lite vision transformer with enhanced self - attention. In *CVPR*, pages 11998 – 12008, 2022.

[28] Yulin Wang, Yang Yue, Rui Lu, Tianjiao Liu, Zhao Zhong, Shiji Song, and Gao Huang. Efffcienttrain: Exploring generalized curriculum learning for training visual backbones. In *International Conference on Computer Vision*, 2023.

[29] Sachin Mehta and Mohammad Rastegari. Mobilevit: lightweight, general - purpose, and mobile - friendly vision transformer. In *International Conference on Learning Representations*, 2022.

[30] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre - training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[31] Y Li, Y Fan, X Xiang, D Demandolx, R Ranjan, R Timofte, and L Van Gool. Efffcient and explicit modelling of image hierarchies for image restoration. In *CVPR*, 2023.

[32] Y Li, G Yuan, Y Wen, et al. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 2022.

[33] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. In *NeurIPS*, 2021.

[34] Youngwan Lee, Jonghee Kim, Jeff Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *CVPR*, 2022.

[35] Wenhai Wang, Enze Xie, Xiang Li, Deng Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 2022.

[36] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *International Conference on Computer Vision*, 2021.

https://doi.org/10.52202/085713-2365

[37] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross - shaped windows. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[38] Z Xia, X Pan, S Song, L.E Li, and G Huang. Vision transformer with deformable attention. In *CVPR*, 2022.

[39] L Zhu, X Wang, Z Ke, W Zhang, and R.W Lau. Biformer: Vision transformer with bi - level routing attention. In *CVPR*, 2023.

[40] H. O. Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520 – 524, 1935.

[41] Herbert Gebelein. Das statistische problem der korrelation als variations - und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics/Zeitschrift fu ̈r Angewandte Mathematik und Mechanik*, 21(6):364 – 379, 1941.

[42] Alfre ́d Re ́nyi. On measures of dependence. *Acta Mathematica Hungarica*, 10(3 - 4):441 – 451, 1959.

[43] S. L. Huang, A. Makur, L. Zheng, and G. W. Wornell. An information-theoretic approach to universal feature selection in high dimensional inference. In *International Symposium on Information Theory (ISIT)*, pages 1336–1340, 2017.

[44] Li Wang, Jie Wu, Shang Lung Huang, Li Zheng, Xiao Xu, Li Zhang, and Jie Huang. An efffcient approach to informative feature extraction from multimodal data. In *Proceedings of the AAAI Conference on Artiffcial Intelligence*, volume 33, pages 5281 – 5288, 2019.

[45] Hongkang Zhang, Shao Lun Huang, and Ercan Engin Kuruoglu. Mhfnet: An improved hgr multimodal network for informative correlation fusion in remote sensing image classiffcation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:15052 – 15066, 2024.

[46] Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[47] Jia Deng, Wei Dong, Richard Socher, Li Li Jia, Kai Li, and Li Fei Fei. Imagenet: A large - scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[48] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[51] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[52] Z Liu, H Mao, C Y Wu, et al. A convnet for the 2020s. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

[53] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[54] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[55] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[56] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

 https://doi.org/10.52202/085713-2365

[57] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Scott Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[58] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[59] Jing Lin, Fei Gao, Xiao Shi, and Jin Dong. Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[60] Joseph Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2016.

[61] X. Ma, X. Zhang, M.-O. Pun, and B. Huang. A unified framework with multimodal fine-tuning for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–1, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of research papers typically outline the core contributions, scope, and key findings, ensuring alignment with the paper's content.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This paper includes a "Limitations" section to discuss constraints, such as model weaknesses or scope limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/085713-2365

Answer: [Yes]

Justification: The paper provides complete theoretical results with numbered theorems, formulas, and proofs, which are cross-referenced throughout the main text and supplemental material, adhering to the guidelines for formal documentation.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper provides detailed information about the dataset, hyperparameters, and methods in the experimental section to enable reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

https://doi.org/10.52202/085713-2365

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper uses publicly available datasets and provides code in the supplemental material, with plans for public release upon publication, ensuring sufficient instructions for reproducing experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Key details like dataset splits, hyperparameters, and optimization strategies are reported in the experimental sections and supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper does not report error bars or statistical significance tests. Most experiments were conducted 3 to 5 times and averaged. According to the statistical results, the errors were less than ±0.3%. However, to maintain clarity, the authors omitted error bars and detailed statistical tests from the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.

 https://doi.org/10.52202/085713-2365

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided sufficient information on the computer resources required for the experiment in the attached materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper complies with NeurIPS ethical standards in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

https://doi.org/10.52202/085713-2365

Justification: The paper focuses on technical contributions without explicit discussion of societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No high-risk assets (e.g., large pretrained models) are introduced, so safeguards are not applicable.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing datasets (e.g., ImageNet-1K, ADE20K,COCO) are properly cited with DOIs, respecting their licenses and terms of use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

https://doi.org/10.52202/085713-2365

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper introduces an improved ViT model. The code is provided as supplementary material and will be made public after the paper is officially published.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: This study does not involve human subjects or crowdsourcing, and focuses on ViT models and computational experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects are involved, making IRB approval irrelevant.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core methodology does not rely on large language models (LLMs) as a key component.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-2365

Appendix

A Theoretical Analysis of Fast-HGR Approximation

Theorem 1 (Statistical Consistency of Fast-HGR)

Let $\mathbf{X} \in \mathbb{R}^{N\times d_X}$ and $\mathbf{Y} \in \mathbb{R}^{N\times d_Y}$ be feature matrices with i.i.d. samples, where each sample $\mathbf{x}_j \in \mathbf{X}$ and $\mathbf{y}_j \in \mathbf{Y}$ has bounded norms. Define linear transformations $f(\mathbf{X}) = \mathbf{XW}$ and $g(\mathbf{Y}) = \mathbf{YV}$ with $\mathbf{W} \in \mathbb{R}^{d_X\times k}$ and $\mathbf{V} \in \mathbb{R}^{d_Y\times k}$. The Fast-HGR correlation score is:

$$\text{F-HGR}(f(\mathbf{X}), g(\mathbf{Y})) = \frac{1}{N-1}\sum_{j=1}^{N}\frac{f(\mathbf{x}_j)^\top g(\mathbf{y}_j)}{\|f(\mathbf{x}_j)\|\|g(\mathbf{y}_j)\|} + \lambda\cdot \text{tr}\left(\text{cov}(f(\mathbf{X}))\text{cov}(g(\mathbf{Y}))\right) \tag{17}$$

This formulation approximates the true k-dimensional HGR maximal correlation $\rho^{(k)}(\mathbf{X}, \mathbf{Y})$ with a high-probability estimation error bounded by $\epsilon = O(1/\sqrt{N})$.

Proof

For linear f, g, the HGR objective measures the strongest statistical dependence as $\rho^{(k)}(\mathbf{X}, \mathbf{Y}) = \sup_{f,g}\mathbb{E}[f(\mathbf{X})^\top g(\mathbf{Y})]$ under centering and covariance constraints. Fast-HGR approximates this by: 1. Replacing the population expectation with a batch-averaged cosine similarity term, which captures local pairwise dependencies; 2. Retaining global structural information via the trace of covariance products.

For the trace term:

$$\text{tr}\left(\text{cov}(f(\mathbf{X}))\text{cov}(g(\mathbf{Y}))\right) = \text{tr}\left(\frac{(f(\mathbf{X})-\bar{f})^\top(f(\mathbf{X})-\bar{f})}{N-1}\cdot\frac{(g(\mathbf{Y})-\bar{g})^\top(g(\mathbf{Y})-\bar{g})}{N-1}\right) \tag{18}$$

where $\bar{f}$ and $\bar{g}$ are batch means of $f(\mathbf{X})$ and $g(\mathbf{Y})$, respectively.

Under i.i.d. sampling, empirical covariances $\hat{\Sigma}_f = \text{cov}(f(\mathbf{X}))$ and $\hat{\Sigma}_g = \text{cov}(g(\mathbf{Y}))$ converge to their population counterparts Σ_f and Σ_g in Frobenius norm. By McDiarmid's inequality, the deviation between empirical and population estimates decays exponentially with N, leading to an overall error bound of $O(1/\sqrt{N})$.

Theorem 2 (Lipschitz Continuity of Fast-HGR)

The Fast-HGR operator is Lipschitz continuous with respect to feature perturbations. For any feature matrices $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{Y}_1, \mathbf{Y}_2$,

$$|\text{F-HGR}(f(\mathbf{X}_1), g(\mathbf{Y}_1)) - \text{F-HGR}(f(\mathbf{X}_2), g(\mathbf{Y}_2))| \le L\cdot\left(\|f(\mathbf{X}_1)-f(\mathbf{X}_2)\|_F + \|g(\mathbf{Y}_1)-g(\mathbf{Y}_2)\|_F\right), \tag{19}$$

where the Lipschitz constant is

$$L = \frac{2}{N-1} + \lambda\left(\|\hat{\Sigma}_f\|_F + \|\hat{\Sigma}_g\|_F\right) \tag{20}$$

with $\hat{\Sigma}_f, \hat{\Sigma}_g$ denoting empirical covariances of $f(\mathbf{X})$ and $g(\mathbf{Y})$, respectively.

Proof

The cosine similarity term is Lipschitz continuous due to unit normalization of $f(\mathbf{x}_j)$ and $g(\mathbf{y}_j)$. For pairwise terms, $|\cos(a_1, b_1) - \cos(a_2, b_2)| \le 2(\|a_1 - a_2\| + \|b_1 - b_2\|)$ under unit norms, leading to a collective bound of $2/(N-1)$ for the summed term.

For the trace term, using the inequality for matrix traces:

$$|\text{tr}(AB) - \text{tr}(A'B')| \le \|A - A'\|_F\|B\|_F + \|A'\|_F\|B - B'\|_F, \tag{21}$$

the perturbation of the trace term is bounded by $\lambda(\|\hat{\Sigma}_f\|_F + \|\hat{\Sigma}_g\|_F)$ times the feature perturbations. Combining both terms yields the Lipschitz constant L.

B Derivation of Fast-HGR from Soft-HGR Variants

Fast-HGR is derived by simplifying and adapting the improved Soft-HGR (I-SoftHGR) objective, which retains the core of HGR while relaxing strict whitening constraints. The I-SoftHGR objective is:

$$L_{\text{I-SoftHGR}} = \mathbb{E}[f(\mathbf{X})^\top g(\mathbf{Y})] - \frac{\lambda}{2}\left(\|\text{cov}(f(\mathbf{X})) - \mathbf{I}\|_F^2 + \|\text{cov}(g(\mathbf{Y})) - \mathbf{I}\|_F^2\right) \tag{22}$$

where the penalty term enforces covariances close to the identity matrix. Fast-HGR modifies this via two key steps:

1. Local Dependence: Replace Expectation with Cosine Similarity The population expectation $\mathbb{E}[f(\mathbf{X})^\top g(\mathbf{Y})]$ is approximated using batch-wise cosine similarity to capture local pairwise dependencies:

$$\mathbb{E}[f(\mathbf{X})^\top g(\mathbf{Y})] \to \frac{1}{N-1}\sum_{j=1}^{N} \frac{f(\mathbf{x}_j)^\top g(\mathbf{y}_j)}{\|f(\mathbf{x}_j)\|\|g(\mathbf{y}_j)\|} \tag{23}$$

2. Global Structure: Replace Whitening Penalty with Covariance Alignment I-SoftHGR's penalty term $\|\text{cov}(f) - \mathbf{I}\|_F^2 + \|\text{cov}(g) - \mathbf{I}\|_F^2$ enforces soft whitening but introduces sensitivity to variance. Fast-HGR removes this constraint, instead capturing global distributional alignment via the trace of covariance products:

Expanding the I-SoftHGR penalty term:

$$\|\text{cov}(f) - \mathbf{I}\|_F^2 = \text{tr}(\text{cov}(f)^2) - 2\text{tr}(\text{cov}(f)) + k \tag{24}$$

where k is the dimension of transformed features. Fast-HGR replaces these with a cross-term that measures alignment between $\text{cov}(f)$ and $\text{cov}(g)$ without enforcing unit covariance:

$$\lambda \cdot \text{tr}\left(\text{cov}(f(\mathbf{X}))\text{cov}(g(\mathbf{Y}))\right) \tag{25}$$

Combining these steps yields the Fast-HGR formulation:

$$\text{F-HGR}(f(\mathbf{X}), g(\mathbf{Y})) = \frac{1}{N-1}\sum_{j=1}^{N} \frac{f(\mathbf{x}_j)^\top g(\mathbf{y}_j)}{\|f(\mathbf{x}_j)\|\|g(\mathbf{y}_j)\|} + \lambda \cdot \text{tr}\left(\text{cov}(f(\mathbf{X}))\text{cov}(g(\mathbf{Y}))\right) \tag{26}$$

This derivation preserves the core objective of maximizing feature dependence while replacing computationally expensive whitening constraints with efficient trace-based regularization.

C Implementation Details

Multi-Scale Tokenization Table 6 presents the optimized kernel configurations designed for efficient hierarchical feature extraction.

Table 6: Optimized Kernel Configurations for Feature Extraction

Kernel Type	Size/Stride/Padding	Output Size	Channels	Receptive Field Impact
Small	3x3/1/1	$H \times W$	64	Fine-grained details
Medium	7x7/2/3	$H/2 \times W/2$	128	Mid-level semantics
Large	15x15/1/7	$H \times W$	256	Global context

The medium kernel (7x7, stride=2) uses padding=3 to achieve an output size of $H/2 \times W/2$, calculated as padding $= \left\lfloor \frac{\text{kernel size}-1}{2} \right\rfloor = 3$.

The large kernel (15x15, stride=1) employs padding=7 to maintain the input resolution ($H \times W$), consistent with "same" padding semantics.

The combination of 3x3, 7x7, and 15x15 kernels effectively balances the capture of fine details, mid-level objects, and global scene context.

https://doi.org/10.52202/085713-2365

Training Protocol

Optimizer: AdamW is utilized with a weight decay of 0.05, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

Warmup: A warmup phase of 20k steps (equivalent to 10% of the total 200 epochs) is implemented to stabilize the initial training dynamics.

Learning Rate: The learning rate follows a cosine decay schedule starting from 5e-5 for the Base model, incorporating mixed precision (FP16) to accelerate convergence by 2× and reduce memory usage.

Regularization: DropPath Rate: This rate is linearly increased from 0 to 0.1 across layers to enhance feature robustness, applying stronger regularization to deeper layers that handle global features. Position Bias: 2D learnable relative position embeddings (size: $H/2 \times W/2$) are employed to encode spatial dependencies in the features derived from medium and large kernels.

D Visualization of HGR-Correlation Attention Maps

This section presents qualitative analyses of attention maps across representative backbones using two ImageNet samples (a flower and a dog). The objective is to examine how correlation-aware multi-kernel modeling influences the spatial distribution of attention, particularly in terms of semantic alignment, background suppression, and integration of discriminative features at varying scales. All attention maps are generated using the same protocol, normalized per image using min–max scaling, and visualized with an identical color scale to ensure comparability.

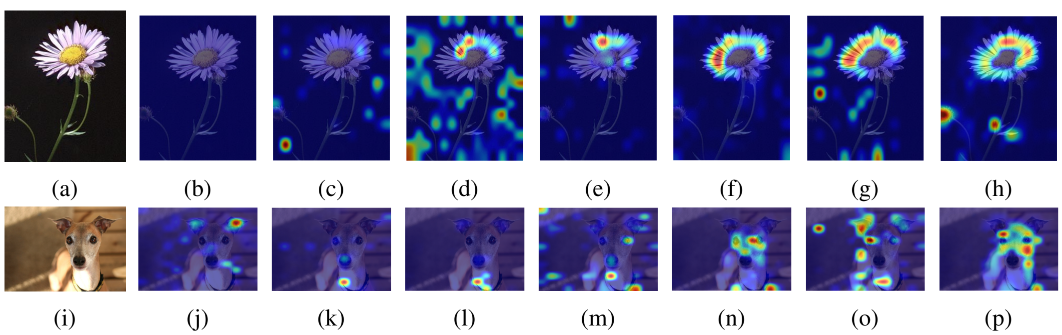

Figure 2: Qualitative comparison of attention maps across backbones. Top row: flower; bottom row: dog. From left to right: (a,i) original images; (b,j) ViT-B; (c,k) DeiT-B; (d,l) Swin-B; (e,m) Agent-Swin; (f,n) FasterViT; (g,o) MPViT; (h,p) MK-CAViT. All maps use the same visualization pipeline and color scale for consistency.

D.1 Patterns in Attention Distribution

Flower sample (Fig. 2a–h). ViT-B and DeiT-B exhibit two common limitations: either attention leaks into the background or concentrates excessively on a single peripheral region (e.g., an edge petal), failing to capture the hierarchical structure of disc and petals (b,c). Swin-B and Agent-Swin produce fragmented attention with isolated local peaks, lacking coherence across the floral structure (d,e). FasterViT shows sparse, discontinuous activations with noticeable "holes" in the attention coverage (f). MPViT improves petal boundary delineation but weakens response to the central disc, a critical semantic feature (g). In contrast, MK-CAViT forms a compact primary peak on the floral disc, with a coherent secondary arc along the petal edges, while effectively suppressing background noise (h). This "primary-auxiliary" structure aligns with the flower's intrinsic semantic hierarchy.

Dog sample (Fig. 2i–p). ViT-B allocates significant attention to background regions and non-discriminative parts (e.g., ears), diluting focus on key facial features (j). DeiT-B and Swin-B fixate on a single dominant region (e.g., nose or jaw) but underweight other critical components like eyes (k,l). Agent-Swin and FasterViT generate scattered hotspots across the face without clear prioritization of discriminative features (m,n). MPViT covers a broader facial area but disperses attention energy, reducing contrast between key and secondary features (o). MK-CAViT, however, concentrates primary attention on the nose (the most discriminative facial feature) while maintaining distinct secondary peaks on both eyes, with minimal activation in background regions (p). This pattern reflects robust alignment with the semantic importance of facial components.

D.2 Implications for Multi-Scale Modeling

The visualized attention maps represent the *final fused output* of each architecture; per-scale kernel attention is not explicitly shown here. Nevertheless, MK-CAViT's consistent ability to integrate fine-grained details (e.g., floral disc texture, eye contours) with extended structures (e.g., petal rings, facial contours) suggests effective aggregation of information across spatial scales. This aligns with the design intent of HGR-correlation attention: to model nonlinear dependencies between features at different scales, rather than treating them as independent streams.

D.3 Connection to Quantitative Performance

The qualitative improvements observed in MK-CAViT—tighter semantic alignment, reduced background interference, and coherent integration of discriminative features—correspond with its quantitative gains across tasks (ImageNet classification, COCO detection, ADE20K segmentation). These visual patterns provide intuitive support for the claim that correlation-aware multi-scale fusion enhances the model's ability to prioritize semantically relevant features, a mechanism underlying its superior performance.

E Cross-Domain Generalization

To rigorously evaluate the generalization capability of MK-CAViT beyond standard vision benchmarks, extensive experiments were conducted across three distinct application domains: multimodal emotion recognition, medical imaging, and remote sensing. Domain generalization (DG) aims to learn models from multiple source domains that perform well on unseen target domains, which is a challenging and practical scenario since models are often deployed in environments different from where they were trained [39, 56]. As summarized in Table 7, the multi-scale design of MK-CAViT demonstrates consistent performance advantages over specialized baselines across all domains, highlighting its robustness to domain shift.

Table 7: Comprehensive cross-domain generalization performance comparison.

Domain-Task	Dataset	Metric	Model			
Emotion-Recognition	IEMOCAP	-	**MK-CAViT-Base**	Swin-Base	ViT-Base	ConvNeXt-Base
		ACC	73.5%	70.1%	68.5%	69.8%
		W-F1	73.6%	69.8%	68.7%	70.1%
Medical-Segmentation	ISIC2018	-	**MK-CAViT-Base**	Swin-UNet	TransFuse	EfficientNet-B4
		mIoU	83.43%	82.78%	80.63%	81.21%
		Dice	89.96%	89.78%	88.21%	88.77%
Remote Sensing-Classification	Houston 2018	-	**MK-CAViT-Base**	FasterViT	3D-CNN	ViT-Base
		OA	93.68%	92.13%	89.59%	91.87%
		AA	95.82%	95.22%	93.77%	94.93%
Remote Sensing-Segmentation	Vaihingen	-	**MK-CAViT-Base**	Swin-UNet	U-Net	DeepLabV3+
		OA	92.61%	91.56%	89.93%	88.92%
		mIoU	84.43%	82.62%	80.15%	81.56%

E.1 Multimodal Emotion Recognition

In multimodal emotion recognition on the IEMOCAP dataset, MK-CAViT-Base achieved a weighted accuracy of **73.5%** and an F1-score of **73.6%**, outperforming Swin-Base (70.1% accuracy, 69.8% F1-score), ViT-Base (68.5% accuracy, 68.7% F1-score), and ConvNeXt-Base (69.8% accuracy, 70.1% F1-score). This task utilized only visual frames to focus on spatial feature learning across four emotion categories (happy, sad, angry, neutral), which inherently involves dealing with domain shifts such as variations in lighting, head pose, and individual expressions [57].

https://doi.org/10.52202/085713-2365

The performance advantage stems from the multi-scale architecture's ability to capture complementary emotional cues. The 3×3 kernel identifies fine-grained facial microexpressions (e.g., smile creases, brow furrows), while the 15×15 kernel models global facial dynamics and head orientation patterns. The Fast-HGR attention mechanism integrates these scale-specific features by modeling their nonlinear correlations, enabling robust distinction of subtle emotion cues that require simultaneous local detail analysis and global context understanding. This approach effectively learns domain-invariant representations that are crucial for handling variations across different speakers and recording sessions.

E.2 Medical Imaging

For skin lesion segmentation on the ISIC2018 dataset, which poses the critical challenge of distinguishing melanoma from nevus through subtle boundary variations, MK-CAViT-Base achieved **83.43%** mIoU and **89.96%** Dice coefficient. This surpassed specialized medical imaging baselines including Swin-UNet (82.78% mIoU, 89.78% Dice), TransFuse (80.63% mIoU, 88.21% Dice), and EfficientNet-B4 (81.21% mIoU, 88.77% Dice). The medical imaging domain frequently encounters domain shift problems due to variations in imaging devices, lighting conditions, and patient populations [58].

The multi-scale design addresses essential requirements in medical diagnostics through complementary feature extraction. The 3×3 kernel detects subtle lesion boundaries and texture variations crucial for early melanoma identification, while the 15×15 kernel captures global lesion structure including asymmetric shapes and spatial distribution patterns. HGR-correlation attention effectively models the complex spatial relationships between lesions and surrounding healthy tissue, significantly reducing false positives caused by spurious correlations. The consistent performance gains across all medical metrics validate the architecture's capability for precise medical image analysis without domain-specific architectural modifications.

E.3 Remote Sensing Applications

Remote sensing evaluation encompasses two distinct tasks with complementary spatial requirements. For land-cover classification on the Houston 2018 dataset[59], MK-CAViT achieved **93.68%** overall accuracy and **95.82%** average accuracy, outperforming FasterViT-Small (92.13% OA, 95.22% AA), 3D-CNN (89.59% OA, 93.77% AA), and ViT-Base (91.87% OA, 94.93% AA). In urban segmentation on the Vaihingen dataset[60, 61], the model achieved **84.43%** mIoU and **92.61%** overall accuracy, surpassing Swin-UNet (82.62% mIoU, 91.56% OA), U-Net (80.15% mIoU, 89.93% OA), and DeepLabV3+ (81.56% mIoU, 88.92% OA). Remote sensing applications inherently face domain shifts due to seasonal variations, geographical differences, and sensor specifications.

The multi-scale architecture demonstrates natural alignment with remote sensing imagery characteristics. The 15×15 kernel captures large-scale geographical patterns and land-cover distributions essential for regional classification, while the 3×3 kernel identifies small structural elements such as road markers and individual vegetation features. For urban segmentation tasks, the 7×7 and 15×15 kernels collaboratively model building and road contexts at appropriate scales, while the 3×3 kernel precisely segments small urban objects including street furniture and vehicle clusters. This scale-aware processing enables comprehensive scene understanding across varying spatial resolutions inherent to remote sensing data, effectively addressing the domain shift challenge through multi-scale invariant feature learning.

E.4 Generalization Analysis

The consistent performance advantages across emotionally nuanced, medically critical, and geographically complex domains demonstrate the robustness of MK-CAViT's multi-scale design principle against various types of domain shifts. Several interconnected factors contribute to this generalization capability:

Scale adaptability enables automatic adjustment to domain-specific feature hierarchies without architectural modifications. The parallel kernel pathways capture information across spatial scales that align naturally with different application requirements, from microscopic medical features to macroscopic geographical patterns. This adaptability allows the model to maintain performance when facing domain shifts characterized by scale variations in target features.

Feature complementarity ensures preservation and integration of both local details and global context. This proves particularly valuable in domains where both micro-level patterns and macro-level structures carry diagnostic information, such as facial microexpressions in emotion recognition or lesion boundaries in medical imaging. By capturing features at multiple scales, the model reduces dependence on domain-specific superficial patterns, thus enhancing generalization [18].

Correlation-based feature integration through Fast-HGR attention provides a theoretically grounded mechanism for modeling nonlinear dependencies across scales. This approach effectively suppresses spurious correlations that vary across domains while enhancing true causal features that remain invariant, aligning with the principles of stable learning for out-of-distribution generalization.

The cross-domain validation establishes that the multi-scale correlation attention mechanism provides fundamental advantages for visual understanding tasks requiring simultaneous processing of fine details and global context. The consistent outperformance of specialized baselines across diverse applications positions MK-CAViT as a versatile architecture with strong generalization potential for real-world deployment where domain shift is a common challenge.

F Parameter Sensitivity Analysis

The parameter λ, which balances local cosine similarity and global trace regularization, was tested on the ImageNet validation set. The results are summarized in Table 8.

Table 8: Parameter Sensitivity of λ on ImageNet-1K

λ	Top-1 Acc (%)
0.01	83.2
0.05	83.9
0.1	**84.3**
0.2	83.7
0.5	82.9

A λ value of 0.1 optimally balances local token alignment (cosine term) and global feature distribution consistency (trace term). Smaller values (e.g., 0.01) under-regularize, leading to unstable feature distributions, while larger values (e.g., 0.5) over-constrain the model, suppressing fine-grained dependencies.

This aligns with theoretical predictions that λ controls the trade-off between capturing pairwise correlations (local) and second-order statistical alignment (global), as demonstrated in Theorems 1 and 2.

G Relationship to Prior Work

1. **Nonlinear Dependency Modeling**: Fast-HGR explicitly maximizes correlation coefficients to capture nonlinear dependencies, such as quadratic interactions. In contrast, dot-product attention relies on implicit nonlinearity through softmax mechanisms and often struggles with high-order statistical modeling.

2. **Asymmetric Fusion**: MK-CAViT employs an asymmetric fusion strategy, utilizing small/mid-kernels for queries and keys while leveraging large-kernels for values. This design contrasts with symmetric fusion approaches, such as those used in Focal Transformer and MpViT, which process all scales uniformly. By adopting this asymmetric method, MK-CAViT achieves more efficient cross-scale information flow, significantly reducing computational expenses while enhancing interaction efficiency across different scales.

https://doi.org/10.52202/085713-2365