Representation Learning for Visual Tasks: A Study of Attention and Information Selection

---

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
VIVEK TRIVEDY
May 2025

Examining Committee Members:

Dr. Longin Jan Latecki, Advisory chair, Computer and Information Science
Dr. Richard Souvenir, Computer and Information Science
Dr. Mindy Shi, Computer and Information Science
Dr. Yiwei Chen, External Member, Fox School of Business and Management

Abstract

This dissertation investigates methods for improving visual representation learning by optimizing attention mechanisms and information selection strategies within deep learning models. Standard approaches often process images independently and compress them into single global descriptors, limiting performance on tasks requiring contextual understanding or fine-grained detail, and can be susceptible to shortcut learning. This work proposes and evaluates techniques that address these limitations by leveraging inter-example context, developing efficient multi-vector representations, and explicitly controlling attention. The research utilizes Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Graph Neural Networks (GNNs) and targets improvements in image classification (single and multi-label) and fine-grained image retrieval.

Four primary contributions are detailed: (1) **CNN2Graph**, a hybrid CNN-GNN framework using cross-attention over a bipartite graph connecting image batches to learnable proxies and fixed anchors, designed to integrate dataset-level context into image classification efficiently and inductively. (2) **DMCAC**, a self-supervised image retrieval method that aligns representation learning with the retrieval task by conditioning training on database interactions, employing distributional divergence minimization between augmented query views relative to the database and a cross-attention classification mechanism. (3) **Using Register Tokens** as an efficient multi-vector image representation method for fine-grained retrieval that supplements the ViT '[CLS]' token with specialized register tokens. This allows us to internally discover Region-of-Interest (ROI) tokens derived from attention patterns. We otimizine performance versus computational cost using a late-interaction framework. (4) **Object-Focused Attention (OFA)**, a training technique for ViTs that adds an auxiliary loss based on semantic segmentation masks to penalize attention to non-object regions, aiming to reduce shortcut learning, improve out-of-distribution robustness, and enhance object shape representation without increasing inference complexity.

The results demonstrate that managing attention and information flow—through context integration, multi-vector feature selection, and explicit object focus—yields visual representations with improved performance, robustness, and efficiency. This research provides methodologies and principles for advancing visual representation learning, particularly for complex models and tasks.

I would like to thank my family, friends, advisor, and the amazing Temple community that played a huge part in shaping my educational journey.

Contents

List of Figures

List of Tables

Introduction

The ability of machines to perceive and interpret the visual world is a cornerstone of modern artificial intelligence, underpinning transformative technologies from autonomous navigation and medical diagnostics to robotics and large-scale content retrieval. At the heart of this capability lies *visual representation learning*—the process of transforming raw, high-dimensional pixel data into compact, structured, and semantically meaningful feature representations, or embeddings. The quality of these representations is paramount; effective embeddings capture the essence of visual content, invariant to nuisance factors like lighting or viewpoint, and directly dictate the performance ceiling for downstream tasks such as classification, detection, segmentation, and retrieval [1, 2]. While the deep learning revolution has enabled the automatic discovery of powerful hierarchical features, superseding traditional hand-crafted approaches, significant challenges remain in developing representations that are simultaneously accurate, robust, efficient, and semantically grounded.

This dissertation addresses several critical limitations inherent in contemporary visual representation learning methodologies. A prevalent paradigm involves processing images as independent and identically distributed (i.i.d.) samples, often compressing the entirety of an image's complex visual information into a single global feature vector, such as the ubiquitous '[CLS]' token in Vision Transformers (ViTs) [2]. While computationally convenient, this approach suffers from two major drawbacks. First, the i.i.d. assumption ignores the rich contextual information present across examples within datasets or specific databases, failing to leverage inter-example relationships that could enhance discrimination [3, 4]. Second, the single-vector representation acts as an information bottleneck, particularly detrimental for fine-grained tasks that demand sensitivity to subtle, localized details [5]. While using denser representations, like the full set of patch tokens from a ViT, can alleviate the bottleneck, it introduces prohibitive computational and storage costs, rendering it impractical for large-scale applications [5].

Furthermore, powerful models like ViTs exhibit vulnerabilities to *shortcut learning* [6]. They

can achieve high performance on training data by exploiting spurious correlations—such as associating objects with specific background textures—rather than learning the intrinsic, causal properties of the objects themselves [7, 8, 9, 10]. This reliance on superficial cues leads to brittle models that fail to generalize reliably to out-of-distribution (OOD) scenarios where these correlations break down [11]. Compounding these issues, a disconnect often exists between the objectives used during representation learning (e.g., generic metric learning losses) and the specific requirements of the downstream task, particularly evident in image retrieval where models are seldom trained with direct interaction with the target database [4].

This dissertation contends that a central pathway to overcoming these limitations lies in the **strategic selection and utilization of information**, primarily orchestrated through the principled application and control of **attention mechanisms**. We hypothesize that significant improvements in representation quality, robustness, and efficiency can be achieved by carefully managing what information a model attends to, how this attention is structured across different scopes (within images, across examples), and how the attention process itself can be explicitly guided towards semantically meaningful content.

To investigate this hypothesis, this work presents a cohesive suite of four research studies, each developing and evaluating novel frameworks centered on attention and information selection:

1. **Leveraging Inter-Example Context:** We first challenge the independent processing paradigm by introducing methods that incorporate broader context. The **CNN2Graph** framework [3] integrates dataset-level structural information into image classification using a hybrid CNN-GNN architecture, learnable proxies, and cross-attention over a dynamically constructed bipartite graph. Subsequently, the **DMCAC** framework [4] aligns representation learning specifically for image retrieval by conditioning a self-supervised objective on interactions with a target database during training, using divergence minimization and cross-attention classification.

2. **Enhancing Intra-Image Detail with Efficiency:** Addressing the single-vector bottleneck for fine-grained tasks, the **Augmenting CLS** approach [5] develops an efficient

multi-vector representation. It augments the standard ViT ‘[CLS]‘ token with specialized register tokens (found to capture part-level features) and novel Region-of-Interest (ROI) tokens discovered automatically by leveraging the ViT’s internal self-attention patterns, enabling richer representations without the cost of dense methods.

3. **Explicitly Controlling Attention for Robustness:** To combat shortcut learning and enhance semantic grounding, the **Object-Focused Attention (OFA)** framework [11] introduces an auxiliary loss during ViT training. Guided by semantic segmentation masks, this loss explicitly penalizes attention allocated to non-object regions, encouraging object-centric processing, improving OOD robustness, and fostering better shape understanding, all without added inference cost.

Through rigorous empirical evaluation on standard benchmarks for image classification (single and multi-label) and image retrieval, these studies collectively demonstrate the efficacy of principled attention and information management. The subsequent chapters detail these contributions: Chapter 1 provides a review of the relevant literature and foundational concepts. Chapter 2 presents the CNN2Graph and DMCAC frameworks focusing on attention across examples. Chapter 3 details the Augmenting CLS approach for attention within images. Chapter 4 describes the OFA framework for controlling attention. Finally, Chapter 5 synthesizes the findings, discusses overarching principles, acknowledges limitations, outlines future research directions, and offers concluding remarks on the significance of this work for advancing visual representation learning.

Chapter 1

Literature Review

1.1 Motivating Representation Learning

1.1.1 Importance of Visual Representation Learning

The automated interpretation of visual information stands as a fundamental pillar of modern artificial intelligence, enabling machines to perceive, understand, and interact with the world in ways previously confined to biological systems. The capacity to distill meaningful patterns from raw visual data—ranging from simple object recognition in photographs to the intricate analysis of dynamic scenes in videos—forms the bedrock for transformative applications across diverse domains. Autonomous vehicles rely on visual understanding for navigation and safety, medical imaging analysis leverages it for diagnostics, robotics depends on it for manipulation and interaction, and content-based search engines utilize it to organize and retrieve vast visual archives. Central to these advancements is the field of visual representation learning, a discipline focused on developing methods to transform high-dimensional, raw pixel data into lower-dimensional, structured, and semantically rich feature representations, often termed embeddings.

The goal is to create representations that capture the essence of the visual content, preserving critical semantic information while exhibiting invariance to nuisance variables such as

fluctuations in lighting, changes in viewpoint, scale differences, or background clutter. The efficacy of any downstream visual task, be it classification, detection, segmentation, or retrieval, is intrinsically linked to the quality of the underlying visual representations. A well-learned representation facilitates simpler, more effective decision-making by subsequent modules, whereas a poor representation can irrevocably hinder performance. Historically, visual features were often hand-crafted, requiring significant domain expertise and engineering effort. The deep learning revolution, however, ushered in an era of learned representations, where hierarchical features are automatically discovered from data, leading to unprecedented performance gains. As the volume of visual data generated continues to explode, the development of powerful, efficient, and robust methods for visual representation learning remains a critical frontier in computer science research.

1.1.2 Target Visual Tasks

The research presented in this thesis evaluates the proposed representation learning strategies primarily through the lens of several core visual tasks. A principal focus is image classification, a fundamental task concerned with assigning a categorical label to an input image from a predefined set of classes [3, 11]. This encompasses both the standard single-label scenario, where each image belongs to exactly one class, and the more complex multi-label image classification setting, where an image may contain objects or concepts belonging to multiple classes simultaneously [11]. Multi-label classification introduces challenges related to modeling label correlations and handling a potentially vast output space.

Another significant area of investigation is content-based image retrieval [4, 5]. The objective here is to search through a large-scale database and identify images that are semantically similar to a given query image. Effective retrieval often hinges on capturing subtle, fine-grained visual distinctions that differentiate closely related object subcategories or specific instances. This demands representations that encode not just coarse object categories but also nuanced details regarding shape, texture, and spatial configuration. While the principles and methods developed

have broader implications for other visual understanding problems, these specific tasks—classification (single and multi-label) and retrieval—serve as the primary experimental platforms for assessing the quality and effectiveness of the learned representations.

1.1.3 Central Role of Attention and Information Selection

Early paradigms in visual representation learning, particularly those dominated by Convolutional Neural Networks (CNNs), often adopted a holistic processing approach. Features were extracted through stacked convolutional and pooling layers, with pooling mechanisms providing a degree of spatial invariance but also potentially discarding valuable information. While highly successful, these architectures could face limitations in efficiently identifying and utilizing the most critical information within an image, especially in cluttered scenes or when dealing with tasks requiring fine-grained analysis where subtle local details matter. The fixed receptive fields and static aggregation strategies inherent in many CNN designs could restrict their ability to adaptively focus on task-relevant features.

A significant shift occurred with the adaptation of the Transformer architecture [12] for visual tasks, leading to the development of Vision Transformers (ViTs) [2]. A key element of the Transformer's success is its attention mechanism. In the context of vision, attention allows the model to dynamically compute the relevance of different input components—typically image patches—when constructing a representation. Instead of treating all parts of the input equally, the model learns to assign varying degrees of importance, effectively focusing its computational resources on the most informative parts of the visual scene relative to the task at hand. This dynamic, context-dependent weighting provides a powerful mechanism for information filtering and selection.

This thesis posits that the strategic selection and utilization of information, primarily orchestrated through attention mechanisms, is a central theme for advancing visual representation learning. We contend that significant improvements in representation quality, robustness, and efficiency can be achieved by carefully controlling what information a model attends to and how

this attention is deployed. Our research explores this theme across different scopes: investigating attention within individual images to pinpoint salient regions or discriminative features, examining attention across different examples to harness contextual information from datasets or databases, and developing methods to explicitly control the attention process itself to align with semantic priors or task objectives. By optimizing the flow and focus of information, we aim to build representations that are not only accurate but also more interpretable and resilient.

1.2 Foundational Concepts

1.2.1 Key Architectures (CNNs, ViTs, GNNs)

The methodologies explored in this thesis are situated within the landscape of modern deep learning architectures for vision, primarily engaging with three key architectural families:

Convolutional Neural Networks (CNNs): For a significant period, CNNs [1] stood as the de facto standard for a wide array of computer vision tasks. Their architecture, characterized by layers of learnable convolutional filters applied across spatial dimensions, excels at capturing local patterns and spatial hierarchies. Key operations like convolution (for feature extraction), activation functions (introducing non-linearity), and pooling (for downsampling and spatial invariance) allow CNNs to learn increasingly abstract features, from edges and textures in early layers to object parts and complete objects in deeper layers. Although ViTs have challenged their dominance in some areas, CNNs remain highly relevant due to their efficiency, strong inductive biases for visual data, and proven effectiveness. They often serve as robust backbone networks for feature extraction or are integrated into hybrid models, as demonstrated in our CNN2Graph framework which utilizes a CNN encoder [3].

Vision Transformers (ViTs): Representing a paradigm shift, ViTs [2] adapt the Transformer architecture [12], originally designed for sequence processing in NLP, to visual inputs. The core idea involves partitioning an image into a sequence of non-overlapping patches, linearly embedding these patches, adding positional information (positional embeddings) to retain spatial

awareness, and then processing this sequence through standard Transformer encoder blocks. The cornerstone of the Transformer block is the self-attention mechanism, which allows every patch to attend to every other patch, enabling the model to capture global dependencies and long-range interactions within the image from the outset. While ViTs have achieved state-of-the-art results, often surpassing CNNs on large datasets, they typically lack the strong spatial inductive biases of CNNs, requiring substantial training data or sophisticated regularization techniques. A significant portion of this thesis focuses on dissecting, leveraging, and refining the attention mechanisms within ViTs to enhance their representational power and robustness [11, 4, 5].

Graph Neural Networks (GNNs): GNNs constitute a class of neural networks specifically designed to operate on data structured as graphs [13, 14, 15]. They learn representations for nodes (and potentially edges or entire graphs) by iteratively aggregating information from their local neighborhoods through a process often referred to as message passing. In each layer, a node updates its representation based on its own current state and the aggregated representations of its neighbors. This iterative aggregation allows GNNs to capture complex relational information and dependencies within the graph structure. In the context of computer vision, GNNs provide a natural framework for moving beyond the traditional assumption of independent data points. They can be employed to explicitly model relationships between images within a dataset, potentially capturing similarities, differences, or other structural patterns that are ignored by methods processing images in isolation. Our CNN2Graph research leverages GNNs precisely for this purpose, constructing a graph to integrate dataset-level context into the classification process [3].

1.2.2 Attention Mechanisms

Attention mechanisms are not merely components but fundamental computational primitives enabling the dynamic information routing central to many modern architectures, including those explored in this thesis:

Self-Attention: This mechanism, the linchpin of the Transformer architecture [12] and thus

ViTs [2], allows elements within a single sequence (e.g., image patches) to interact and influence each other's representations. Computationally, for each element (patch), self-attention calculates three vectors: a Query (Q), a Key (K), and a Value (V), typically through linear projections of the input representation. The attention score between two elements is computed based on the similarity (often dot product) between the Query vector of the attending element and the Key vector of the attended-to element. These scores are then normalized (usually via softmax) to form attention weights, which are used to compute a weighted sum of the Value vectors of all elements. The result is an updated representation for each element that incorporates information from across the entire sequence, weighted by learned relevance. This allows ViTs to model global context and intricate spatial relationships within an image effectively. Our work explores both leveraging the emergent properties of self-attention for feature discovery [5] and explicitly guiding it for improved robustness [11].

Cross-Attention: While self-attention operates within a single set of inputs, cross-attention models interactions between two distinct sets of inputs. One set provides the Query vectors, while the other provides the Key and Value vectors. Similar to self-attention, attention weights are computed based on Q-K similarities, and these weights are used to aggregate the Value vectors from the second set. This mechanism allows information to be selectively transferred or integrated from one modality or source to another. For instance, in our CNN2Graph framework [3], cross-attention relates input image representations (Queries) to learnable class proxy vectors (Keys/Values) to integrate class-level information. In our DMCAC retrieval method [4], cross-attention relates query image representations to retrieved database item representations, enabling the query representation to be conditioned on the database context and facilitating classification based on retrieved items.

1.2.3 Relevant Learning Paradigms

The process of learning effective visual representations is guided by the choice of learning paradigm and associated objective functions. Several paradigms are pertinent to the work in this

thesis:

Supervised Learning: This remains the most established paradigm, relying on datasets where inputs (images) are paired with explicit ground-truth labels (e.g., object categories). The model learns a mapping from input to output by minimizing a loss function that measures the discrepancy between its predictions and the true labels, typically using cross-entropy for classification tasks. Supervised learning can achieve high performance when large labeled datasets are available. However, the cost and effort involved in acquiring such large-scale annotations represent a significant bottleneck. Several methods explored in this thesis incorporate supervised components: CNN2Graph [3] and DMCAC [4] use cross-entropy loss for classification aspects, and the Object-Focused Attention (OFA) framework employs an auxiliary supervised loss derived from semantic segmentation masks to guide attention learning [11].

Metric Learning: Instead of directly predicting labels, metric learning focuses on learning an embedding space where the distance between representations reflects the semantic similarity of the corresponding inputs. The goal is to structure the embedding space such that similar items are clustered together while dissimilar items are pushed apart. This is particularly relevant for tasks like image retrieval [5, 4]. Common techniques include:

- Contrastive Loss: Operates on pairs of examples, pulling positive pairs (similar items, e.g., same class or different augmentations of the same image) closer in the embedding space and pushing negative pairs (dissimilar items) farther apart [16].

- Triplet Loss: Considers triplets of examples: an anchor, a positive (similar to anchor), and a negative (dissimilar to anchor). The loss enforces that the distance between the anchor and the positive is smaller than the distance between the anchor and the negative by at least a predefined margin [17].

Effective metric learning often requires careful sampling strategies to select informative pairs or triplets.

Self-Supervised Learning (SSL): This paradigm offers a powerful alternative for learning

representations from large amounts of unlabeled data. SSL methods generate their own supervisory signals directly from the input data. One major family of SSL techniques relies on data augmentation, training the model to be invariant to certain transformations. For example, different augmented views of the same image are generated, and the model is trained to produce similar embeddings for these views, often using a contrastive loss objective [4]. Another family involves pretext tasks, where the model is trained to solve an auxiliary task that requires understanding the data structure, such as predicting masked or corrupted portions of the input (e.g., Masked Autoencoders [18]) or predicting the relative spatial location of patches. By solving these self-generated tasks, the model learns rich, transferable visual features without manual labels. Our DMCAC method leverages an augmentation-based SSL approach, uniquely conditioning the invariance objective on interactions with a database [4].

1.3 Motivation and Thesis Structure

1.3.1 Identifying Research Gaps

The pursuit of more effective visual representations is driven by limitations inherent in existing methodologies. Despite remarkable progress, several key challenges persist, forming the primary motivation for the research presented in this thesis.

Firstly, many conventional approaches, especially those rooted in early CNNs, adhere to an independent example processing paradigm. Images are typically fed through the network one by one, assuming they are independent and identically distributed (i.i.d.). This overlooks the rich contextual information that might exist across examples within a dataset or a specific database relevant to a task like retrieval [3, 4]. Learning representations in isolation prevents the model from exploiting inter-example relationships, relative comparisons, or global dataset statistics that could enhance understanding and discrimination.

Secondly, the prevalent use of single global descriptors for representing complex visual inputs poses a significant limitation. Compressing the entirety of an image's semantic content into a single

vector, such as the [CLS] token commonly extracted from ViTs, inevitably leads to information loss [5]. While sufficient for coarse categorization, this bottleneck hinders performance on tasks demanding fine-grained understanding or localization, where subtle details, object parts, and their spatial arrangements are critical.

Thirdly, while using denser representations, such as the full set of patch tokens from a ViT, can mitigate the single-vector bottleneck, it introduces substantial computational and storage costs. Storing and performing similarity comparisons on potentially hundreds of high-dimensional vectors per image becomes intractable for large-scale applications like web-scale image retrieval [5]. This necessitates the development of methods that can efficiently select or construct a compact yet highly informative set of vectors that captures the richness of the visual content without prohibitive overhead.

Fourthly, Vision Transformers, despite their representational power, exhibit a vulnerability to shortcut learning and a lack of robustness [11]. They can learn to rely on spurious correlations within the training data—associating objects with specific background textures or artifacts, for instance—rather than learning the intrinsic properties and shape of the objects themselves [6]. This reliance on superficial cues leads to brittle models that fail to generalize well to out-of-distribution (OOD) data or adversarial perturbations where these spurious correlations are broken. Enhancing the robustness and semantic grounding of ViTs requires mechanisms that explicitly encourage attention towards meaningful object regions and discourage reliance on background or texture shortcuts.

Finally, a disconnect often exists between the training objectives used for representation learning and the downstream task, particularly in image retrieval [4]. Models are frequently trained using generic metric learning losses on curated datasets, without direct interaction with the target database they will be deployed against. This mismatch can lead to suboptimal performance, as the learned representations may not be perfectly aligned with the characteristics and distribution of the specific database used at test time. Bridging this gap by incorporating database context during training yields significant performance improvements.

Chapter 2

Attention Across Examples: Leveraging Dataset and Database Context

2.1 Motivation: Beyond Independent Processing

The dominant paradigm in visual representation learning has long relied on processing images as independent entities, often assuming they are drawn i.i.d. from some underlying distribution. As discussed in Chapter 1, this approach, while simplifying model design and training, fundamentally overlooks the rich tapestry of contextual information and inter-example relationships present within any large collection of visual data, such as a training dataset or a target retrieval database [3, 4]. Ignoring this context imposes significant limitations. Models trained under the i.i.d. assumption may struggle to capture relative similarities and differences, fail to adequately model the underlying class distributions or dataset biases, and learn representations that are suboptimal for downstream tasks inherently involving comparison or interaction, such as few-shot learning or large-scale retrieval. For instance, understanding the subtle distinctions between bird species might be enhanced by comparing an image not just to an abstract class prototype, but also to other specific bird images within the dataset. Similarly, optimizing a representation for retrieving items from a specific e-commerce database might

benefit from learning features that are discriminative within the context of that particular database during training, rather than relying on generic features learned in isolation.

This section details two distinct research thrusts from this thesis that directly confront the limitations of independent processing by introducing mechanisms for attending to information across examples. The *unifying hypothesis* is that enriching an image's representation with information gleaned from its surrounding context—whether that context is derived from curated representatives of the dataset or from dynamically relevant items within a database—yields representations that are more discriminative, robust, and better aligned with specific downstream objectives. We first explore the CNN2Graph framework, which leverages static dataset context via graph structures and learnable proxies to enhance image classification. Subsequently, we delve into the DMCAC framework, which focuses on dynamic database conditioning during training to learn representations specifically tailored for the demands of image retrieval. Both methodologies prominently feature cross-attention as the core mechanism facilitating the crucial information exchange between individual examples and their broader context, demonstrating the power of looking beyond the single image.

2.2 Incorporating Dataset Context via Graphs (CNN2Graph)

The CNN2Graph framework [3] was conceived to explicitly leverage inter-example relationships for improving standard image classification, moving beyond the limitations of traditional CNNs that process images independently. A key motivation was to overcome persistent challenges in applying graph-based methods to unstructured image data. Prior attempts often relied on constructing k-Nearest Neighbor (KNN) graphs based on initial image features, a process that is typically non-differentiable, hindering end-to-end learning. Furthermore, such methods often operated transductively, requiring test examples to be present during training, or faced significant computational hurdles for inductive inference on new examples, as adding a node required costly comparisons to the entire training set. CNN2Graph aimed to provide an end-to-end differentiable

framework capable of inductive inference while effectively integrating dataset-level context.

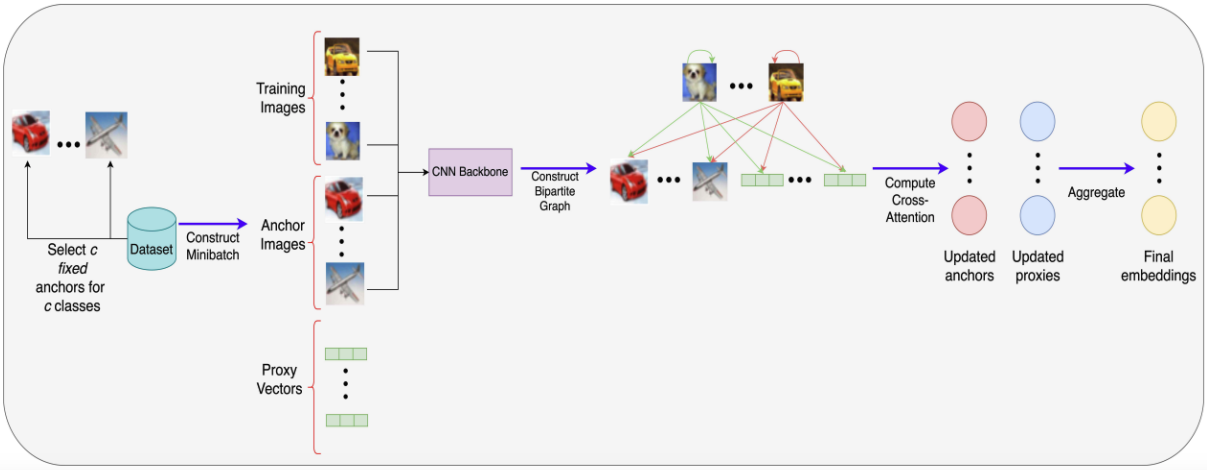

Figure 2.1: CNN2Graph Forward Pass Diagram. Illustrates the connection between batch images, the CNN encoder, the bipartite graph construction involving anchors (L) and proxies (P), and the GNN/Attention module.

2.2.1 Method: CNN-GNN Framework and Bipartite Graph Construction

At its core, CNN2Graph utilizes a hybrid architecture. A standard CNN (e.g., ResNet [1]) serves as the initial feature extractor ϕ, mapping input images to an embedding space $\mathbb{R}^F$. The novelty lies in the subsequent processing stage, where a Graph Neural Network (GNN) module operates on a dynamically constructed graph to refine these initial embeddings by incorporating dataset context.

Critically, instead of attempting to build a potentially complex and non-differentiable graph among all images, CNN2Graph constructs a simple complete bipartite graph for each mini-batch. This graph connects the embeddings of the images within the current batch ($E_{\text{batch}} = \phi(X_{\text{batch}})$) to a fixed, pre-defined proxy set. This proxy set acts as a compact, learnable summary of the entire dataset's class structure and consists of two distinct types of elements:

- Learnable Proxy Vectors (P): A collection of c learnable vectors, $P \in \mathbb{R}^{c \times F}$, where c is the total number of classes. Each vector P_k is initialized randomly and trained with the objective of becoming a representative embedding for class k. These proxies have the flexibility to

learn abstract class concepts optimized for the downstream task.

- Fixed Anchor Examples (L): A set containing c specific image examples drawn from the training set, with one example l^i uniformly sampled from each class i:

$$L = \{l^i \in_{\mathcal{U}} X^i : X^i \subseteq X\}, \quad i = 1, \ldots, c \tag{2.1}$$

(Eq. (2.1) in [3]) These anchor images, passed through the same encoder ϕ to get embeddings $E_{\text{anchors}} = \phi(L)$, provide stable, data-grounded reference points for each class. They help ground the learnable proxies and provide concrete examples for structuring the embedding space.

The nodes processed by the GNN module for a given mini-batch are thus the union of the batch image embeddings, the learnable proxy vectors, and the anchor embeddings: Nodes $= E_{\text{batch}} \cup P \cup E_{\text{anchors}}$. The graph structure is simply a complete bipartite connection between the batch nodes (E_{batch}) and the proxy set nodes ($P \cup E_{\text{anchors}}$). This design offers key advantages: the graph structure is fixed and independent of the specific image features in the batch, ensuring the construction is fully differentiable. Moreover, inductive inference is trivial: a new test image embedding is simply connected to the same trained proxy set ($P \cup E_{\text{anchors}}$) without needing any comparisons to the training data, making node insertion an $O(1)$ operation relative to graph construction.

2.2.2 Attention & Learning: Cross-Attention Aggregation and Combined Loss

Information flow between the batch images and the proxy set across the edges of the bipartite graph is mediated by cross-attention mechanisms within the GNN aggregation step. The purpose of attention here is crucial: given that the proxy set contains representatives for all classes, an image embedding should selectively attend more strongly to the proxy set elements

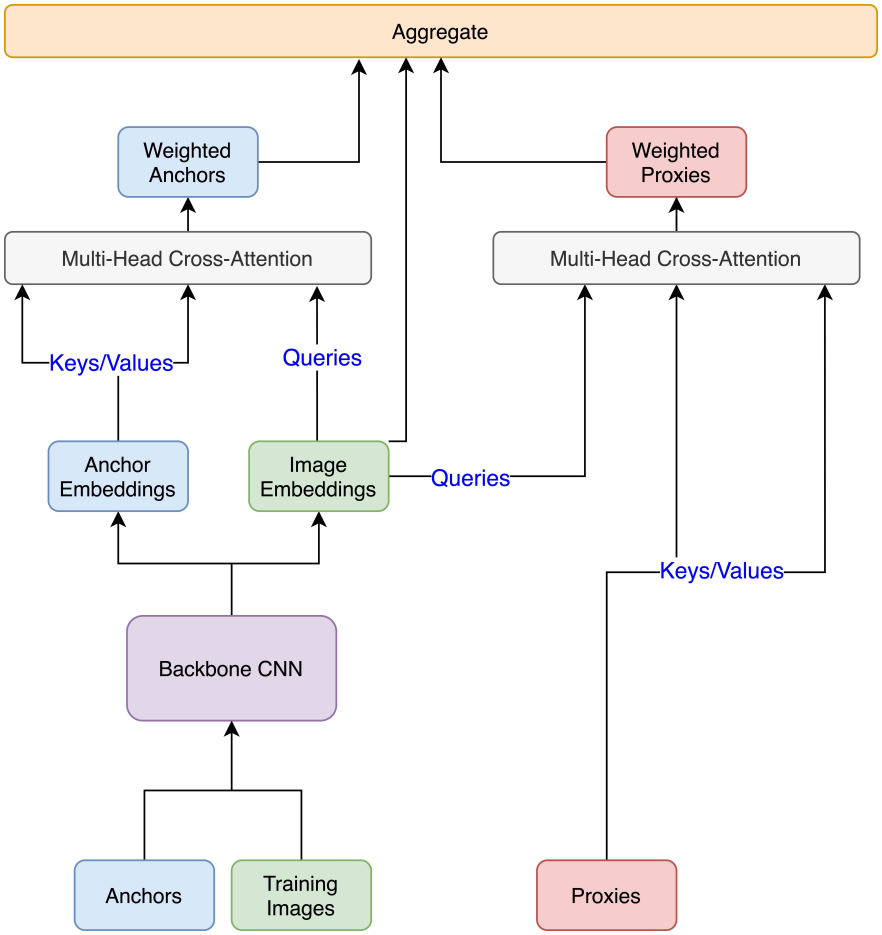

Figure 2.2: CNN2Transformer Cross-Attention Module. Shows how image embeddings (Queries) attend separately to anchor embeddings and proxy vectors (Keys/Values) before aggregation (Adapted from).

corresponding to its own class (or related classes) rather than aggregating information uniformly. This selective aggregation allows the model to refine the initial image embedding with relevant class-level context. Two specific attention mechanisms were implemented and compared:

- GAT-style Attention: Adapting the Graph Attention Network mechanism [15], the relevance $e(h_i, h_j)$ of a proxy set element h_j to an image embedding h_i is calculated via a learned projection and scoring function:

$$e(h_i, h_j) = \mathbf{a}^T \text{LeakyReLU}(\mathbf{W}[h_i || h_j]) \tag{2.2}$$

  Here, $\mathbf{W}$ projects the concatenated embeddings, and $\mathbf{a}^T$ scores the result. These scores are

normalized using softmax to produce attention weights α, which are then used to compute a weighted sum of projected proxy set element representations, yielding the updated image embedding.

- Transformer-style Cross-Attention: Leveraging the standard attention mechanism from Transformers [12], the batch image embeddings $X_{\text{emb}} = E_{\text{batch}}$ serve as the Queries (Q). The anchor embeddings $L_{\text{emb}} = E_{\text{anchors}}$ and proxy vectors P separately provide the Keys (K) and Values (V) in two distinct cross-attention modules. This computes context vectors derived from anchors (L_{mha}) and proxies (P_{mha}):

$$L_{\text{mha}} = \text{Softmax}\left(\frac{(\mathbf{W}_q X_{\text{emb}})(\mathbf{W}_{k1} L_{\text{emb}})^T}{\sqrt{d}}\right)(\mathbf{W}_{v1} L_{\text{emb}}) \tag{2.3}$$

(Eq. (2.3) in [3])

$$P_{\text{mha}} = \text{Softmax}\left(\frac{(\mathbf{W}_q X_{\text{emb}})(\mathbf{W}_{k2} P)^T}{\sqrt{d}}\right)(\mathbf{W}_{v2} P) \tag{2.4}$$

(Eq. (2.4) in [3]) The final output representation z_{out} for each image is typically formed by aggregating (e.g., concatenating or summing) its original embedding X_{emb} with these context vectors L_{mha} and P_{mha}, effectively enriching the initial representation with information gleaned from relevant class anchors and proxies.

The training of the entire system—CNN backbone, GNN attention mechanism, and learnable proxies—is driven by a carefully constructed combined loss function. This loss aims not only to achieve accurate classification but also to ensure the proxy set learns discriminative and stable class representations, preventing issues like proxy collapse where the learnable vectors become unused or redundant. The loss comprises two main parts:

- First, a standard cross-entropy classification loss ($\mathcal{L}_{\text{ce}}$) is applied to the final output embeddings (X') of the batch images after GNN aggregation. Crucially, this loss is also applied independently to the anchor embeddings (E_{anchors}) and the learnable proxy vectors

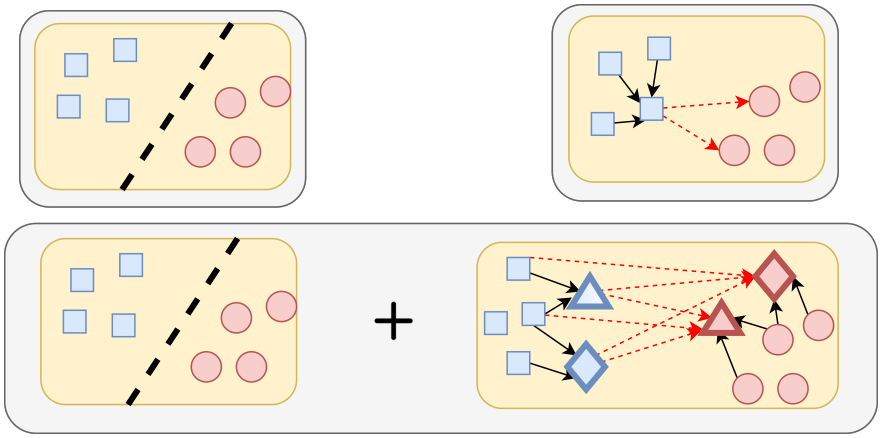

Figure 2.3: Diagram of cross-entropy loss (top-left), contrastive loss (top-right), and our combined loss (bottom). We use standard cross-entropy loss along with an adapted contrastive loss where we compute losses between training examples and a set of proxies and anchors which are uniformly distributed by class.

(P), each associated with their ground-truth class label. This direct classification objective forces the anchors and proxies to reside in regions of the embedding space suitable for linear classification according to their assigned class:

$$\mathcal{L}_{\text{classification}} = \mathcal{L}_{\text{ce}}(X') + \mathcal{L}_{\text{ce}}(L) + \mathcal{L}_{\text{ce}}(P) \tag{2.5}$$

(Eq. (2.5) in [3])

- Second, to further structure the latent space and explicitly manage the relationships between images, anchors, and proxies, several contrastive-style loss terms are incorporated. We adapt the triplet loss [17], generally defined as:

$$\mathcal{L}_{\text{triplet}}(s, g, n) = \max(||f(s) - f(g)||_2^2 - ||f(s) - f(n)||_2^2 + \alpha, 0) \tag{2.6}$$

where $f(s)$, $f(g)$, and $f(n)$ are the embeddings of a source (anchor), a positive example, and a negative example, respectively, and α is a margin. We also use the contrastive loss [16],

defined for a pair (x_1, x_2) as:

$$\mathcal{L}_{\text{contrastive}}(x_1, x_2) = (1 - Y)\frac{1}{2}D^2 + (Y)\frac{1}{2}\{\max(0, \alpha - D)\}^2 \tag{2.7}$$

where $D = ||f(x_1) - f(x_2)||_2$, $Y = 1$ if x_1, x_2 are similar (same class) and $Y = 0$ otherwise, and α is a margin.

Based on these, we define the following specific loss terms.

- $\mathcal{L}_{\text{at}} = \mathcal{L}_{\text{triplet}}(L, X, X)$: Encourages images ($X$) to be closer to anchors (L) of the same class than to anchors of different classes.
- $\mathcal{L}_{\text{pt}} = \mathcal{L}_{\text{triplet}}(P, X, X)$: Encourages images ($X$) to be closer to proxies (P) of the same class than to proxies of different classes.
- $\mathcal{L}_{\text{ap}} = \mathcal{L}_{\text{triplet}}(L, P, P)$: Encourages proxies ($P$) to be close to anchors (L) of the same class, using the stable anchors to guide the learning of the proxies.
- $\mathcal{L}_{\text{p}} = \mathcal{L}_{\text{contrastive}}(P)$: Applies a standard contrastive loss (Eq. 10 in [3]) directly between pairs of proxy vectors (P), enforcing a minimum separation between proxies assigned to different classes and explicitly counteracting collapse.

The total training objective is the sum of these components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{total\_contrastive}} + \mathcal{L}_{\text{classification}} \tag{2.8}$$

(Eq. (2.8) in [3]). This multi-faceted loss function works synergistically: the classification terms drive discriminability, while the contrastive terms organize the latent space structure and ensure the stability and relevance of the proxy set elements. (The evolution of the embedding space under these losses can be visualized via UMAP plots 2.4.

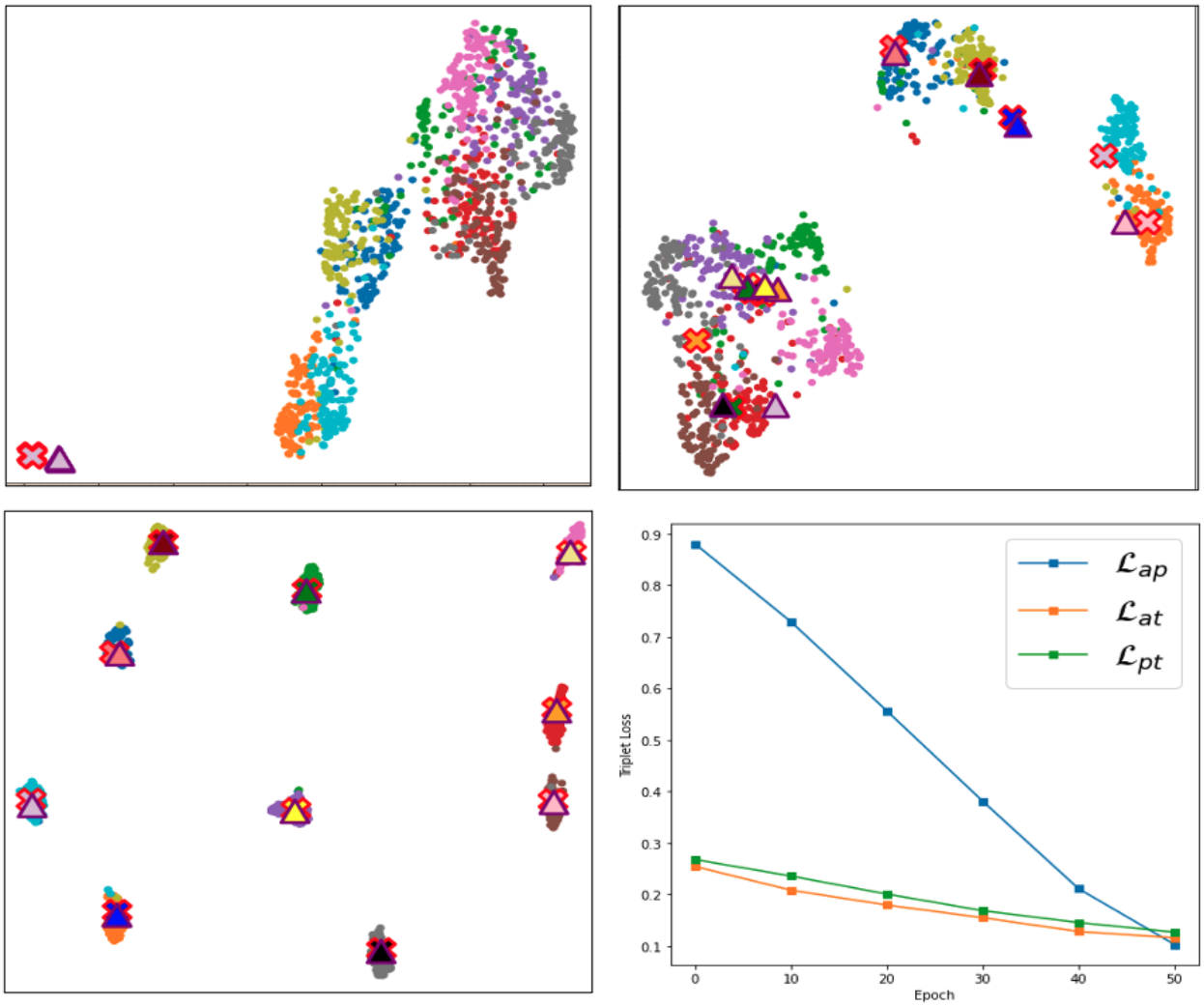

Figure 2.4: UMAP visualization of embedding space evolution during CNN2Transformer training on CIFAR-10. Shows initial state, intermediate state, and final state where anchors (X) and proxies (triangles) cluster with image embeddings according to class.

2.2.3 Key Aspects: End-to-End Learning, Inductive Inference, and Scalability

The CNN2Graph framework presents a significant step forward by enabling end-to-end differentiable learning of image representations that incorporate dataset context. Unlike methods requiring separate, potentially non-differentiable steps for graph construction or context integration, CNN2Graph allows the CNN feature extractor, the GNN attention mechanism, and the learnable proxies to be optimized jointly through standard backpropagation. This holistic

optimization is crucial for learning features that are well-suited for both initial extraction and subsequent contextual refinement.

Furthermore, the design ensures efficient inductive inference. Because the proxy set ($P \cup E_{\text{anchors}}$) is fixed after training, classifying a new, unseen test image only requires passing it through the trained CNN backbone and then performing a single GNN aggregation step involving connections to this fixed proxy set. There is no need to compare the test image to the entire training dataset or rebuild complex graph structures, making the approach scalable for deployment.

Perhaps most importantly, the framework provides an effective mechanism for capturing both fine-grained example-level features (learned by the CNN) and broader class-level context (mediated through attention over the proxy set). The empirical results confirmed the benefits of this contextual integration, showing improved classification accuracy over baseline CNNs.

The comparative analysis of GAT-style versus Transformer-style attention yielded a critical insight regarding scalability. While both attention mechanisms allowed for context integration, the Transformer cross-attention demonstrated markedly superior performance and stability as the dataset complexity increased, particularly on the large-scale ImageNet-1k benchmark. The GAT-based model showed a significant drop in performance compared to its baseline in this setting, whereas the Transformer-based model provided consistent improvements. This suggests that the standard scaled dot-product attention, perhaps due to its ability to handle larger neighborhoods more effectively or its query-dependent nature, provides a more robust and scalable mechanism for aggregating information across examples in this framework. This finding has broader implications for architectural choices when designing models that need to integrate information across diverse sets of elements. (Detailed performance comparisons are provided in Table 2.1.

Method	CIFAR-10	CIFAR-100	STL-10	SVHN	ImageNet-1k
ResNet18 Baseline	94.1	77.0	95.4	95.3	69.4
CNN2GNN (ResNet18)	95.5±0.4	74.8±0.8	95.7±0.2	96.6±0.6	60.1±1.0
CNN2Tfmr (ResNet18)	95.8±0.2	77.4±0.2	95.7±0.2	96.4±0.2	71.1±0.4
ResNet34 Baseline	95.2	79.3	95.9	95.6	73.0
CNN2GNN (ResNet34)	96.4±0.4	77.9±0.9	96.9±0.3	97.0±0.3	61.0±0.8
CNN2Tfmr (ResNet34)	96.7±0.4	80.1±0.5	97.2±0.2	96.5±0.1	75.4±0.2

Table 2.1: Validation accuracy comparison of baseline ResNets, CNN2GNN, and CNN2Transformer across datasets. Demonstrates accuracy improvements and highlights the better scalability of CNN2Transformer on larger datasets like ImageNet-1k.

2.3 Conditioning Representations on Database Context (DMCAC)

Moving from the goal of improving general classification using dataset context (CNN2Graph) to the specific challenge of image retrieval, the DMCAC (Divergence Minimization with Cross-Attention Classification) framework [4] was developed. The core motivation stems from a critical observation: most methods for learning retrieval-oriented representations train the encoder using objectives (like contrastive or triplet losses on class labels, or self-supervised invariance to augmentations) that are disconnected from the actual retrieval process. These methods learn a general notion of similarity but do not explicitly optimize the representations for ranking or retrieval against the specific database that will be used in the downstream application. This potential mismatch between the training environment and the deployment scenario can lead to suboptimal retrieval performance, as the learned features might not be maximally discriminative within the specific distribution and characteristics of the target database. DMCAC proposes a paradigm shift by directly incorporating interaction with a database during the training loop, aiming to learn representations that are explicitly conditioned on and optimized for retrieval within that database context.

2.3.1 Method: Joint Query-Database Learning via Self-Supervision

DMCAC operationalizes this database conditioning through a novel training procedure that simulates retrieval within the training loop. It requires partitioning the training data into two sets: a training query set ($\mathcal{D}_Q$) and a training database set ($\mathcal{D}_D$). Importantly, the classes represented in these training sets are disjoint from those used during evaluation, ensuring that the model learns generalizable representations rather than memorizing specific database items. The central learning principle is self-supervised: the model learns by enforcing consistency among multiple augmented views of a query image, but this consistency is measured relative to their interaction with the training database.

The process unfolds as follows: First, the embeddings for all images in the training database $\mathcal{D}_D$ are computed using the current state of the encoder ϕ_{new} (typically a ViT [2]) and stored: $Z_D = \phi_{\text{new}}(\mathcal{D}_D) \in \mathbb{R}^{D \times F}$ (Eq. 3 in [4]). These database embeddings Z_D are periodically updated (e.g., every few epochs) to reflect the evolving encoder, balancing stability with responsiveness.

During a training step, a query image $q \in \mathcal{D}_Q$ is selected, and A different augmented views (X_A) are generated (including the original image). These views are embedded by the encoder: $Z_A = \phi_{\text{new}}(X_A) \in \mathbb{R}^{A \times F}$ (Eq. 5 in [4]). The embeddings Z_A (and Z_D) are typically ℓ_2-normalized.

The crucial training-time retrieval step then occurs. For each query view embedding $z_i \in Z_A$, its k nearest neighbors within the database embeddings Z_D are identified. This retrieval can be performed exhaustively (full retrieval, Eq. 6 in [4]) if Z_D is small enough to fit in GPU memory, providing complete gradient information from the database. However, for scalability to large databases, approximate retrieval using efficient libraries like FAISS [19] is employed. In the approximate case, the indices S_i of the top-k neighbors for each view i are retrieved. The union of these indices across all A views,

$$S^{\text{union}} = \bigcup_{j=1}^{A} S_j \tag{2.9}$$

(Eq. (2.9) in [4]), collects all database items considered relevant to any of the query views. The corresponding embeddings for these unique indices are fetched from Z_D to form the dynamic,

query-specific database context set $Z_S^{\text{union}} \in \mathbb{R}^{T \times F}$, where $T = |S^{\text{union}}|$. It is this set Z_S^{union} that forms the basis for the novel loss calculations in DMCAC. (The overall data flow is depicted in 2.5).

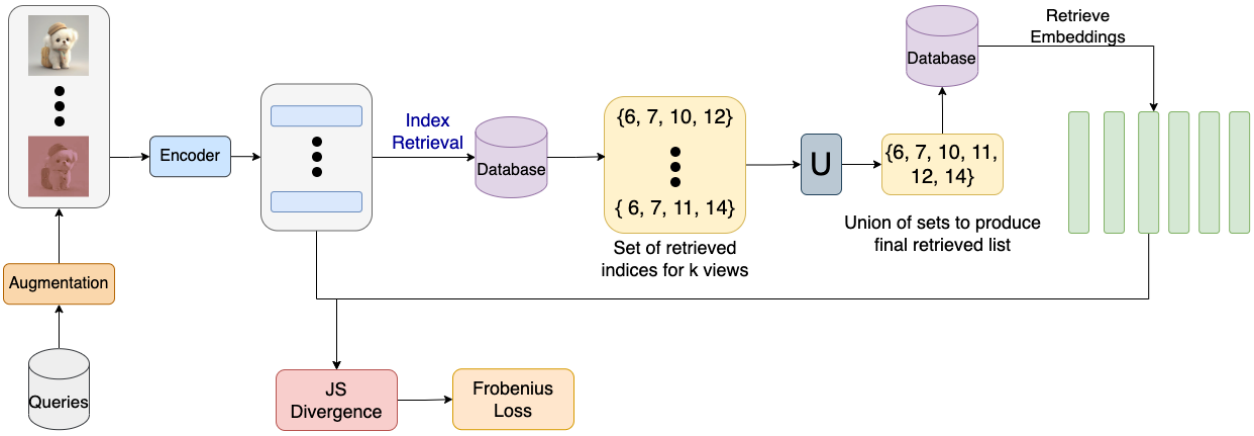

Figure 2.5: DMCAC Data Flow for Frobenius Loss Computation. Illustrates query augmentation, encoding, index retrieval from the database based on query views, union of retrieved indices, fetching corresponding database embeddings, and calculating the JS Divergence / Frobenius Loss.

2.3.2 Novel Objectives: Divergence Minimization and Cross-Attention Classification

The core innovations of DMCAC lie in its loss functions, which explicitly leverage the retrieved database context Z_S^{union} to guide representation learning:

- Frobenius Loss for Divergence Minimization ($\mathcal{L}_{\text{frob}}$): This objective replaces the standard SSL goal of simply minimizing the distance between augmented view embeddings (Z_A). Instead, it enforces consistency between the views based on how they relate to the database context. The underlying intuition is that different perspectives (augmentations) of the same underlying query image should perceive the database similarity landscape in a consistent manner. Operationally, the similarity between each query view $z_i \in Z_A$ and all retrieved database embeddings Z_S^{union} is computed, forming a similarity matrix $P = Z_A \cdot (Z_S^{\text{union}})^T$ (Eq. 9 in [4]). Applying softmax row-wise yields P', where each row P'_i is a probability

distribution reflecting the relative similarity of view i to the T database items in Z_S^{union}. DMCAC then minimizes the divergence between these similarity distributions for all pairs of views (i, j). The Jensen-Shannon (JS) divergence, a symmetric measure of distributional difference, is used:

$$L_{ij} = \text{JS}(P_i' || P_j') = \frac{1}{2}\text{KL}(P_i' || M) + \frac{1}{2}\text{KL}(P_j' || M), \quad \text{where } M = \frac{P_i' + P_j'}{2} \tag{2.10}$$

(Eq. (2.10) in [4]) The final loss term $\mathcal{L}_{\text{frob}}$ is the Frobenius norm of the matrix L containing these pairwise JS divergences:

$$\mathcal{L}_{\text{frob}} = \sqrt{\sum_{i=2}^{A} \sum_{j=1}^{i-1} L_{ij}^2} \tag{2.11}$$

(Eq. (2.11) in [4]) By minimizing this divergence, the encoder ϕ_{new} is trained to produce representations that are not only invariant to augmentations but, more importantly, exhibit consistent ranking or similarity patterns when interacting with the database. This implicitly optimizes features relevant for retrieval within that specific database context.

- Cross-Attention Classification (CAC) Loss ($\mathcal{L}_{\text{cac}}$): A significant challenge in SSL based purely on augmentation invariance is the potential for representation collapse, where the model learns trivial solutions (e.g., mapping all inputs to a single point). DMCAC introduces the CAC loss as a powerful mechanism for semantic grounding and collapse prevention, again leveraging the database context. The core idea is to test whether the retrieved database neighbors Z_S^{union} contain enough semantic information to determine the class of the original query view z. This is achieved using cross-attention: z acts as the Query, while the retrieved neighbors Z_S^{union} provide the Keys and Values.

$$Q, K, V = \mathbf{W}_q z, \mathbf{W}_k Z_S^{\text{union}}, \mathbf{W}_v Z_S^{\text{union}} \tag{2.12}$$

(Eq. (2.12) in [4]) The output of the cross-attention module is a new representation z' for the query:

$$z' = \text{Softmax}\left(\frac{QK^T}{\sqrt{F}}\right)V \tag{2.13}$$

(Eq. (2.13) in [4]) This z' can be interpreted as a database-conditioned representation of the query, effectively a projection of z onto a basis dynamically defined by its nearest neighbors in the database Z_S^{union}. This context-aware representation z' is then fed through a linear classifier and trained using a standard cross-entropy loss $\mathcal{L}_{\text{cac}}$ based on the ground-truth class label of the original query q (Eqs. 18–19 in [4]). The significance of CAC is twofold: First, it forces the encoder ϕ_{new} to learn representations such that the retrieved neighbors Z_S^{union} are semantically informative about the query's class. If the neighbors are irrelevant, classifying z' correctly becomes impossible. This implicitly pushes the encoder towards retrieving semantically relevant items, directly benefiting the retrieval task. Second, it provides a strong supervisory signal based on ground-truth class labels, effectively preventing the representational collapse that can plague pure invariance-based SSL. (The CAC mechanism is illustrated in 2.6.

Optionally, a standard cross-entropy loss $\mathcal{L}_{\text{ce}}$ (Eqs. 14–15 in [4]) can be applied directly to the original view embeddings $z \in Z_A$. The final objective combines these terms:

$$\mathcal{L}_{\text{total}} = \beta_{\text{frob}}\mathcal{L}_{\text{frob}} + \beta_{\text{ce}}\mathcal{L}_{\text{ce}} + \beta_{\text{cac}}\mathcal{L}_{\text{cac}} \tag{2.14}$$

(Eq. (2.14) in [4]) While equal weighting ($\beta = 1$) proved effective, ablation studies confirmed that both the divergence minimization ($\mathcal{L}_{\text{frob}}$) and the cross-attention classification ($\mathcal{L}_{\text{cac}}$) components are essential contributors to the final performance, highlighting the synergistic benefit of ensuring both retrieval consistency and semantic grounding through database interaction. (Ablation results are presented in Table 2.2).

Method	Betas $[\beta_{frob}, \beta_{ce}, \beta_{cac}]$	Arch	In-Shop R@k 1	10	20	30
ProxyAnchor	-	Inc-BN	91.5	98.1	98.8	99.1
Hyp-DINO	-	ViT	92.4	98.4	98.9	99.1
DMCAC-DeiT	[1,1,1]	DeiT-S	91.1	98.5	98.8	99.1
DMCAC-ViT	[1,1,1]	ViT(IN21k)	92.7	98.2	98.9	99.3
DMCAC-DeiT	[1,1,0]	DeiT-S	91.0	98.3	98.5	98.9
DMCAC-ViT	[1,1,0]	ViT(IN21k)	92.4	98.3	98.7	99.3
DMCAC-DeiT	[0,1,1]	DeiT-S	90.2	97.9	98.2	98.4
DMCAC-ViT	[0,1,1]	ViT(IN21k)	91.9	96.9	97.2	97.8

Table 2.2: Ablation study on DMCAC loss components using the In-Shop dataset. Comparing different weightings for β_{frob}, β_{ce}, β_{cac}. Shows that removing either $\mathcal{L}_{frob}$ ($\beta_{frob} = 0$) or $\mathcal{L}_{cac}$ ($\beta_{cac} = 0$) significantly degrades performance compared to the full loss or removing only $\mathcal{L}_{ce}$.

2.3.3 Key Aspects: Retrieval Alignment and Database Conditioning Effectiveness

DMCAC's primary contribution lies in its novel approach to aligning the representation learning process explicitly with the downstream task of image retrieval. By simulating retrieval within the training loop and conditioning the learning objectives on the interaction between queries and the database, DMCAC moves beyond context-free representation learning. The learned features are inherently tuned for discriminability and ranking within the context of the target database distribution, addressing the potential mismatch present in traditional methods.

The experimental results robustly validate the effectiveness of this database conditioning strategy. The combination of minimizing distributional divergence across views (relative to the database) and classifying queries based on their database neighbors (CAC) proves highly effective. DMCAC achieves state-of-the-art performance on multiple standard image retrieval benchmarks, including CUB-200, Cars-196, In-Shop Clothes, and Stanford Online Products, often outperforming prior methods based on contrastive learning or sophisticated metric learning losses, especially when employing ViT architectures.

Furthermore, the framework demonstrates practical viability by performing competitively in both the full retrieval setting (offering complete gradient information) and the more scalable

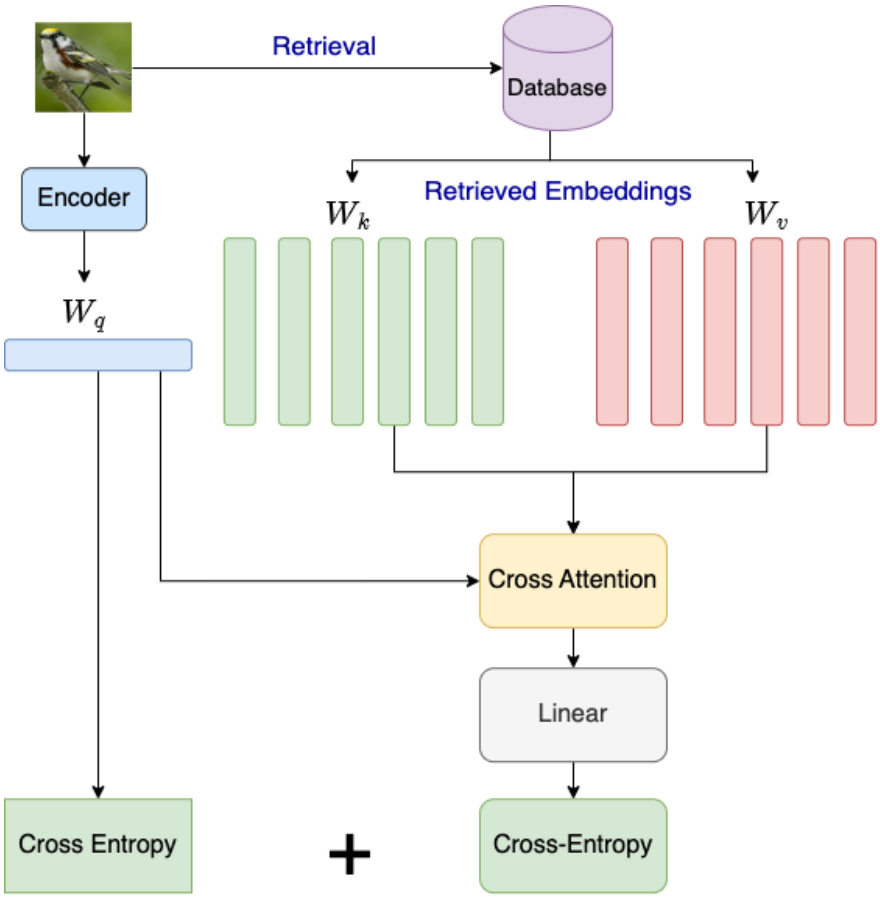

Figure 2.6: DMCAC Cross-Attention Classification (CAC) Mechanism. Shows how a query embedding attends to its retrieved database neighbors (Keys/Values) to produce a database-conditioned representation z', which is then classified using cross-entropy loss.

approximate retrieval setting using FAISS [19]. The performance difference between the two was found to be relatively minor, suggesting that the core benefits of database conditioning can be achieved even when using approximate nearest neighbor search during training, making the approach applicable to very large databases as shown in Table 2.3

2.4 Synthesis: Value of Attending Across Examples

The two frameworks detailed in this section, CNN2Graph and DMCAC, provide compelling evidence for the value of attending across examples in visual representation learning. By explicitly designing mechanisms that allow information to flow between individual data points

Dataset	Method	R@1	R@2	R@4	R@8
CUB-200	DMCAC-DeiT	78.4	87.0	92.3	95.0
	DMCAC-ViT	86.2	92.0	94.7	96.7
	DMCAC-DeiT-FR	78.6	87.2	93.0	95.5
	DMCAC-ViT-FR	**86.8**	**92.3**	**94.9**	96.7
Cars-196	DMCAC-DeiT	84.4	89.2	94.9	97.5
	DMCAC-ViT	88.5	93.9	96.7	98.1
	DMCAC-DeiT-FR	**84.8**	89.2	94.9	97.5
	DMCAC-ViT-FR	**89.2**	**94.0**	**97.0**	**97.9**

Table 2.3: Comparison of approximate retrieval (default) vs. full retrieval (FR) during training on CUB-200 and Cars-196. Full retrieval offers slightly better performance but approximate retrieval remains highly competitive and scalable.

and a broader context, both approaches successfully overcome limitations associated with processing images in isolation. They demonstrate that incorporating inter-example relationships leads to representations that are demonstrably richer, more robust, and better aligned with downstream tasks.

While both leverage cross-attention as a core mechanism for information exchange, they differ significantly in the nature of the context utilized and their primary application focus. CNN2Graph employs a static, curated dataset context embodied by learnable proxies and fixed anchors, primarily aiming to improve image classification by infusing representations with class-level structural information derived from the entire dataset. Its strengths lie in its end-to-end differentiability, efficient inductive inference, and its structured approach to modeling class relationships via the bipartite graph.

In contrast, DMCAC utilizes a dynamic, query-specific database context obtained through retrieval during training, specifically targeting the enhancement of image retrieval performance. By conditioning self-supervised learning objectives (divergence minimization) and a novel classification scheme (CAC) on the retrieved database neighbors, it directly aligns representation learning with the nuances and distribution of the target database. Its innovation lies in bridging the gap between training and retrieval deployment.

Together, these studies highlight the versatility and power of cross-example attention.

Method	Dim	Architecture	CUB-200				In-Shop				Cars-196				Stanford Online Products			
			1	2	4	8	1	10	20	30	1	2	4	8	1	2	4	8
NSoftmax [20]	512	R50	61.3	73.9	83.5	90	86.6	97.5	98.4	98.8	84.2	90.4	94.4	96.9	78.2	90.6	96.2	-
ProxyNCA++ [21]	512	R50	69.0	79.8	87.3	92.7	90.4	98.1	98.8	99.0	86.5	92.5	95.7	97.7	80.7	92.0	96.7	98.9
A-BIER [22]	512	GoogleNet	57.5	68.7	78.3	86.2	93.1	95.1	96.9	97.5	82.0	89.0	93.2	96.1	74.2	86.9	94.0	97.8
ABE [23]	512	GoogleNet	60.6	71.5	79.8	87.4	87.3	96.7	97.9	98.2	85.2	90.5	94.0	96.1	76.3	88.4	94.8	98.2
SM [24]	512	GoogleNet	56.0	68.3	78.2	86.3	90.7	97.8	98.5	98.8	83.4	89.9	93.9	96.5	75.3	87.5	93.7	97.4
Proxy-Anchor [25]	512	Inception-BN	68.4	79.2	86.8	91.6	91.5	98.1	98.8	99.1	86.1	91.7	95.0	97.3	79.1	90.8	96.2	98.7
SoftTriple [26]	512	Inception-BN	65.4	76.4	84.5	90.4	-	-	-	-	84.5	90.7	94.5	96.9	78.6	86.6	91.8	95.4
HORDE [27]	512	Inception-BN	66.8	77.4	85.1	91.0	90.4	97.8	98.4	98.7	86.2	91.9	95.1	97.2	80.1	91.3	96.2	98.7
XBM [28]	512	Inception-BN	65.8	75.9	84.0	89.9	89.9	97.6	98.4	98.6	82.0	88.7	93.1	96.1	79.5	90.8	96.1	98.7
MS [29]	512	Inception-BN	65.7	77.0	86.3	91.2	89.7	97.9	98.5	98.8	84.1	90.4	94.0	96.5	78.2	90.5	96.0	98.7
HTL [30]	512	Inception-BN	57.1	68.8	78.7	86.5	80.9	94.3	95.8	97.2	81.4	88.0	92.7	95.7	74.8	88.3	94.8	98.4
IRT_R [31]	384	DeiT-S	76.6	85.0	91.1	94.3	91.9	98.1	98.7	98.9	-	-	-	-	84.2	93.7	97.3	99.1
Sph-DeiT [32]	384	DeiT-S	76.2	84.5	90.2	94.3	89.6	97.2	98.0	98.4	81.7	88.6	93.4	96.2	82.5	92.9	97.1	99.1
Sph-DINO [32]	384	ViT	78.7	86.7	91.4	94.9	90.1	97.1	98.0	98.4	86.6	91.8	95.2	97.4	82.2	92.1	96.8	98.9
Sph-ViT [32]	384	ViT(IN21k)	85.1	90.7	94.3	96.4	90.4	97.4	98.2	98.6	81.7	89.0	93.0	95.8	82.1	92.5	97.1	99.1
Hyp-DeiT [32]	384	DeiT-S	77.8	86.6	91.9	95.1	90.5	97.8	98.5	98.9	86.4	92.2	95.5	97.5	83.3	93.5	97.4	99.1
Hyp-DINO [32]	384	ViT	80.9	87.6	92.4	95.6	92.4	98.4	**98.9**	99.1	**89.2**	**94.1**	**96.7**	**98.1**	85.1	94.4	97.8	99.3
Hyp-ViT [32]	384	ViT(IN21k)	85.6	91.4	**94.8**	**96.7**	92.5	98.3	98.8	99.1	86.5	92.1	95.3	97.3	85.9	94.9	**98.1**	**99.5**
DMCAC-DeIT	384	DeiT-S	78.4	87.0	92.3	95.0	91.1	**98.5**	98.8	99.1	84.4	89.2	94.9	97.5	84.2	93.6	97.4	99.1
DMCAC-ViT	384	ViT (IN21k)	**86.2**	**92.0**	94.7	**96.7**	**92.7**	98.2	**98.9**	**99.3**	88.5	93.9	**96.7**	98.0	**86.3**	**95.2**	97.5	**99.5**

Table 2.4: Recall@k metrics comparing across state-of-the-art methods on the CUB-200, In-Shop, Cars-196, and Stanford Online Products datasets. DMCAC (ours) performs competitively across architectures and outperforms all previous methods in several settings.

Whether the context is a curated set of class representatives or dynamically retrieved database items, allowing representations to be shaped by their relationship to other relevant examples provides a potent inductive bias. It enables models to learn features that capture not only the intrinsic content of an image but also its relative position and significance within a larger semantic landscape, ultimately leading to more effective and task-aware visual understanding. The construction of interaction pathways (graphs, retrieval) and the application of attention mechanisms (cross-attention) emerge as key enablers for unlocking the benefits of looking beyond individual data points.

Chapter 3

Attention Within Images: Discovering and Utilizing Informative Regions and Tokens

3.1 Motivation: Enhancing Representations with Localized Details

The previous chapter highlighted the benefits of incorporating context from across different images. We now turn our attention inward, focusing on the challenge of effectively capturing the rich information contained within a single image. As established earlier, the common practice of representing an entire image using a single global feature vector, such as the `[CLS]` token from a Vision Transformer (ViT), imposes a severe information bottleneck [5]. This compression is particularly detrimental for tasks that hinge on fine-grained visual distinctions. While a global vector might adequately capture the general category of an image (e.g., ”bird,” ”car”), it often fails to preserve the subtle, localized details necessary to differentiate between closely related subcategories (e.g., ”sparrow” vs. ”finch”) or specific object instances, which is the core challenge in fine-grained image retrieval. The averaging or pooling inherent in creating a single descriptor tends to wash out the very features—unique patterns, specific parts, local textures—that are critical for making these fine distinctions.

Recognizing this limitation, various approaches have explored multi-vector representations. Early computer vision techniques utilized local invariant features detected around keypoints (e.g., SIFT [33]) or features extracted from salient image regions identified by specialized algorithms [34, 35]. While demonstrating the value of local information, these methods often relied on hand-crafted features or separate detection modules that were not integrated into end-to-end deep learning pipelines. More contemporary deep learning approaches have involved using features from multiple layers of a CNN or employing region proposal networks to identify and represent distinct image parts. Within the ViT framework, the most direct multi-vector approach is to utilize the embeddings associated with all the input image patches. This dense representation theoretically captures the maximum amount of local information available to the model. However, its practical application, especially in large-scale retrieval systems containing millions or billions of images, is severely constrained by the prohibitive computational and storage costs [5]. Storing hundreds of high-dimensional patch vectors per image results in massive index sizes, and the complexity of comparing all query patch vectors against all database patch vectors during search becomes computationally infeasible.

This creates a critical gap: there is a need for representation-learning techniques that can capture essential localized details for fine-grained tasks but do so efficiently, avoiding both the information loss of single global vectors and the intractability of dense patch representations. Furthermore, an ideal solution would achieve this without resorting to complex, external modules for region detection or saliency prediction, instead leveraging the inherent capabilities of the representation learning model itself. The work presented in this section [5] directly addresses this challenge by proposing a novel framework for constructing efficient, yet powerful, multi-vector representations by augmenting the standard `[CLS]` token with a small, carefully selected set of informative tokens derived entirely from the internal structure and learned patterns of the ViT itself.

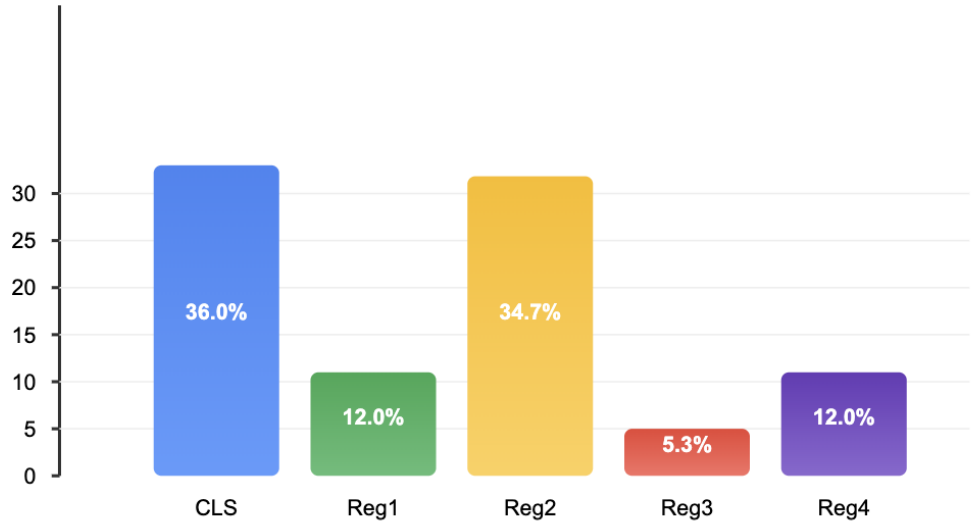

Figure 3.1: Histogram of most similar token types in database images for query [CLS] tokens (COCO dog class). Shows that query [CLS] tokens are often most similar to database register tokens, motivating their use in retrieval.

3.2 Multi-Vector Representations via Internal Discovery (Augmenting the CLS Token)

The core proposition of the work detailed in [5] is that the limitations of the single `[CLS]` token can be overcome not by incorporating all available patch tokens, but by strategically augmenting it with a compact set of additional, specialized tokens. The selection and generation of these tokens are designed to capture diverse and complementary information—global context, salient parts, and localized details—relevant for fine-grained discrimination. A key innovation is that this token discovery process is performed internally, harnessing the emergent properties and attention patterns within a specific ViT architecture, DINOv2-reg [36], thereby obviating the need for external detection or segmentation modules often employed in prior part-based or region-based approaches.

3.2.1 Leveraging DINOv2-reg and Register Tokens

The architectural foundation for this method is the DINOv2-reg ViT model [36]. A notable feature of this architecture is the inclusion of a small number (R) of extra learnable tokens, termed register tokens (t_r), processed alongside the standard `[CLS]` token (t_{cls}) and the image patch tokens (t_p). In the original DINOv2-reg work, these registers were introduced primarily as "attention sinks"—their purpose was theorized to be attracting and isolating attention patterns associated with image artifacts or potentially irrelevant background information, thus yielding cleaner, more object-focused attention maps for the `[CLS]` token. Consequently, the recommendation was to discard these register tokens for any downstream tasks.

However, a crucial insight motivating the work in [5] was the empirical observation that these register tokens, far from merely capturing noise, emergently learn to represent semantically meaningful concepts, often focusing on distinct objects or salient object parts within the image. Instead of being discarded, they represent a potentially valuable source of complementary information. Analysis on retrieval tasks revealed that the similarity between a query's `[CLS]` token and a relevant database image's register token was frequently higher than the similarity to the database image's own `[CLS]` token, particularly for images within the same fine-grained category as shown in Figure 3.1.

This discovery suggested that the registers implicitly learned discriminative features overlooked by the global `[CLS]` token and that retaining them could significantly enrich the image representation for retrieval. This repurposing of register tokens, leveraging an emergent property contrary to their original design intent, forms the first component of the augmented representation.

3.2.2 Token Selection and ROI Discovery Mechanism

Building upon the established value of the global `[CLS]` token and the newly recognized potential of the register tokens, the proposed multi-vector representation begins with the set of "cue" tokens: $\{t_{cls}, t_r^1, \ldots, t_r^R\}$ (where typically $R = 4$). To further enhance representational

granularity by incorporating highly localized information, the framework introduces a novel, fully internal mechanism for deriving an additional set of Region-of-Interest (ROI) tokens. One ROI token is generated corresponding to each of the $R + 1$ cue tokens, based entirely on the similarity patterns computed between the cue tokens and the image patch tokens during the ViT's standard forward pass.

The ROI token discovery process unfolds in three steps:

1. **Identify "Buddy" Patches:** For each cue token $t_{cue} \in \{t_{cls}, t_r^1, \ldots, t_r^R\}$, its affinity to every image patch token t_p^j ($j = 1, \ldots, P$, where P is the number of patches) is measured using the dot-product similarity:

$$s_j = t_{cue}^T t_p^j$$

   The patch token $t_p^{j^*}$ that exhibits the highest similarity is designated as the "buddy" patch for that specific cue token: $j^* = \arg\max_j s_j$. The intuition here is powerful: the model's own learned representations (the cue tokens) are used to identify the image region (the buddy patch) they are most strongly associated with or focused on. This provides a data-driven, attention-guided mechanism for locating potentially salient points within the image, directly reflecting the model's internal understanding, without external guidance. Different cue tokens (global `[CLS]` vs. part-focused registers) naturally identify different buddy patches, capturing diverse points of interest. Figure 3.2 visually demonstrates how buddy patches for `[CLS]` and different registers often correspond to distinct semantic parts like a dog's head, paw, or a nearby object.

2. **Define Region of Interest (Ω):** Relying on a single buddy patch token might be sensitive to noise or overly specific. To capture a more robust representation of the local area, a small spatial neighborhood around the buddy patch is defined. An $N \times N$ grid of patch tokens (e.g., $N = 3$, capturing the buddy patch and its immediate neighbors) centered on $t_p^{j^*}$ constitutes the Region of Interest, Ω.

3. **Compute ROI Token:** The final ROI token b_{cue} associated with the cue token t_{cue} is obtained

by averaging the embeddings of all patch tokens t_p falling within the defined region Ω:

$$b_{cue} = \frac{1}{|\Omega|} \sum_{t_p \in \Omega} t_p \tag{3.1}$$

(Eq. (3.1) in [5]) This local averaging acts as a spatial smoothing mechanism, yielding a descriptor b_{cue} that represents the salient local region identified by the cue token in a more stable and contextually informed manner than a single patch token could provide.

This internal discovery process generates $R+1$ ROI tokens $(b_{cls}, b_r^1, \ldots, b_r^R)$, one derived from each of the $R+1$ cue tokens. The final proposed multi-vector representation $E(I)$ for an image I is the union of the original cue tokens and these newly derived ROI tokens:

$$E(I) = \{t_{cls}, t_r^1, \ldots, t_r^R\} \cup \{b_{cls}, b_r^1, \ldots, b_r^R\} \tag{3.2}$$

(Eq. (3.2) in [5]) For a standard DINOv2-reg model ($R = 4$), this yields a highly compact set of only $2(R+1) = 10$ tokens per image, drastically smaller than the full set of patch tokens.

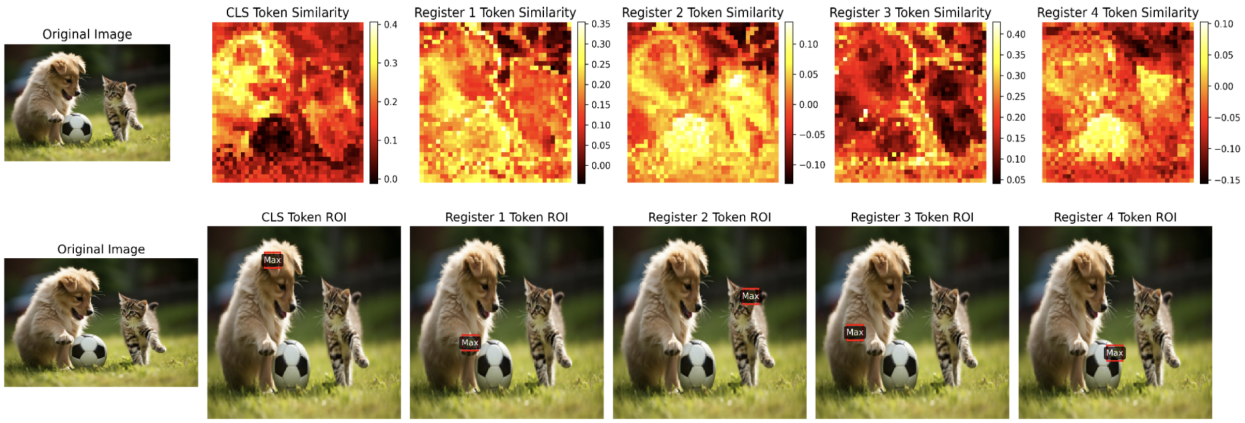

Figure 3.2: Visualization of DINOv2-reg token similarities and identified buddy patches. Top row shows similarity maps for [CLS] and register tokens. Bottom row highlights the corresponding buddy patch (black box) and the 3×3 ROI (red box), showing how different cue tokens focus on different semantic parts (dog head, paw, cat ear, ball).

3.2.3 Retrieval Framework and Multi-Vector Training

This compact yet diverse set of $E(I)$ tokens enables an efficient multi-vector retrieval framework, adopting the late-interaction strategy popularized by ColBERT [37] in the text domain. This paradigm shifts away from comparing single pre-aggregated global vectors and instead performs comparisons at the level of individual tokens, allowing for finer-grained matching.

Given the token sets for a query Q, $E(Q) = \{q_1, \ldots, q_m\}$, and for a database item D, $E(D) = \{d_1, \ldots, d_n\}$ (where $m = n = 10$ in the typical case), the late-interaction matching score $S(Q, D)$ is computed by summing the maximum similarity achieved by each query token across all database tokens:

$$S(Q, D) = \sum_{i=1}^{m} \max_{1 \le j \le n} (q_i^T d_j) \tag{3.3}$$

(Eq. (3.3) in [5]) All tokens (q_i, d_j) are L2-normalized prior to the dot product calculation. The intuition behind this scoring function is its flexibility; it allows different types of query tokens (global, part-focused, localized ROI) to independently find their best counterpart in the database representation. A strong match might occur between the global `[CLS]` tokens, or between specific register tokens capturing the same object part, or between ROI tokens representing similar local details. Aggregating these "best local matches" provides a robust similarity score that captures partial or fine-grained correspondences, which are often missed by single-vector comparisons that enforce a single global alignment.

To ensure the learned token embeddings are optimized for this late-interaction scoring, a multi-vector triplet training objective is employed. For each training triplet (Q, D^+, D^-), where D^+ is semantically similar to Q and D^- is dissimilar, the loss function aims to maximize the score $S(Q, D^+)$ relative to $S(Q, D^-)$ by at least a margin α:

$$\mathcal{L} = \sum_{\text{triplet}} [\max(0, \alpha + S(Q, D^-) - S(Q, D^+))] \tag{3.4}$$

This triplet loss directly optimizes the entire set of $2(R + 1)$ token embeddings jointly, pushing the encoder to produce representations where the aggregate late-interaction score

accurately reflects semantic similarity, thereby leveraging the fine-grained matching capability enabled by the multi-vector format. Figure 3.3 provides a visual schematic of the complete training pipeline, including ROI generation and the multi-vector triplet loss computation.

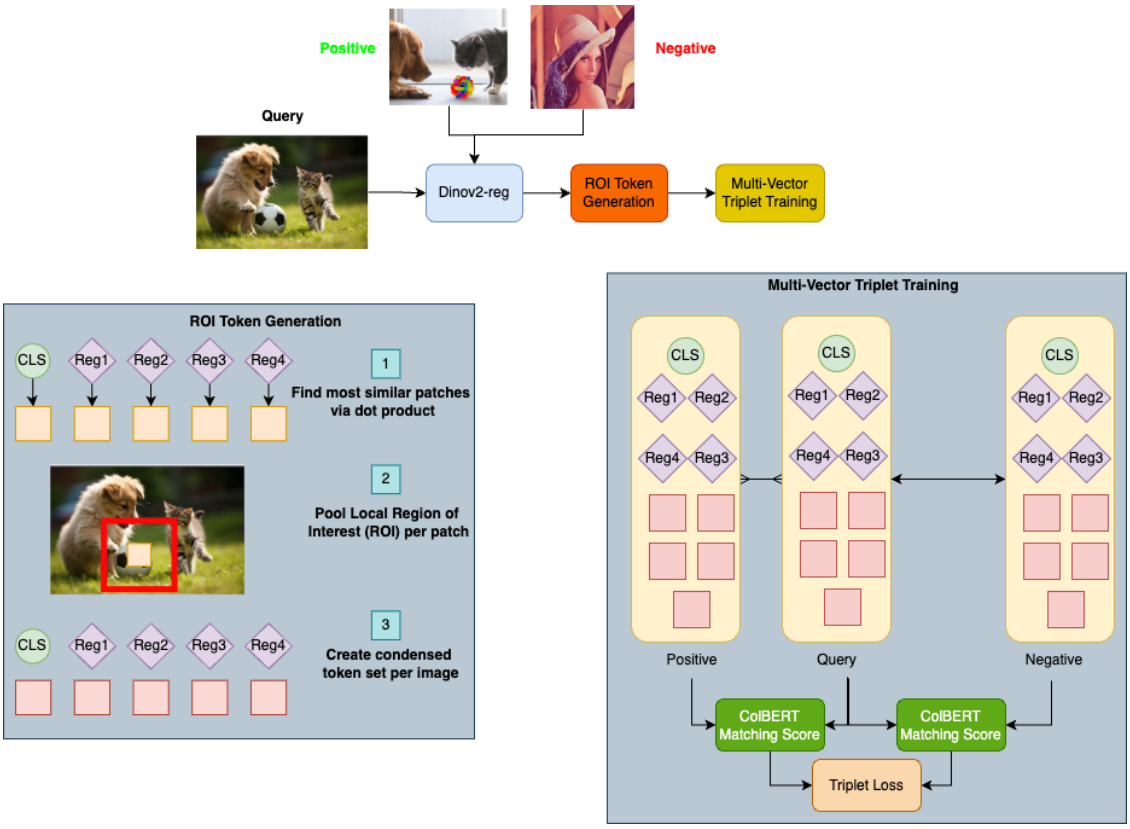

Figure 3.3: Training pipeline for the Augmenting CLS method. Top: Overall flow showing query/positive/negative images passing through DINOv2-reg, ROI token generation, and multi-vector triplet training. Bottom Left: Detail of ROI token generation (buddy patch identification, region pooling). Bottom Right: Detail of multi-vector triplet loss using ColBERT-style matching scores.

3.3 Enhanced Retrieval via Focused Information

The effectiveness of this strategy—augmenting the `[CLS]` token with internally discovered register and ROI tokens, coupled with the late-interaction framework—was rigorously validated on several standard image retrieval benchmarks, demonstrating clear advantages over baseline approaches.

Method	Dim	Architecture	CUB-200				In-Shop				Cars-196				Stanford Online Products			
			1	2	4	8	1	10	20	30	1	2	4	8	1	2	4	8
NSoftmax [20]	512	R50	61.3	73.9	83.5	90	86.6	97.5	98.4	98.8	84.2	90.4	94.4	96.9	78.2	90.6	96.2	-
ProxyNCA++ [21]	512	R50	69.0	79.8	87.3	92.7	90.4	98.1	98.8	99.0	86.5	92.5	95.7	97.7	80.7	92.0	96.7	98.9
A-BIER [22]	512	GoogleNet	57.5	68.7	78.3	86.2	93.1	95.1	96.9	97.5	82.0	89.0	93.2	96.1	74.2	86.9	94.0	97.8
ABE [23]	512	GoogleNet	60.6	71.5	79.8	87.4	87.3	96.7	97.9	98.2	85.2	90.5	94.0	96.1	76.3	88.4	94.8	98.2
SM [24]	512	GoogleNet	56.0	68.3	78.2	86.3	90.7	97.8	98.5	98.8	83.4	89.9	93.9	96.5	75.3	87.5	93.7	97.4
Proxy-Anchor [25]	512	Inception-BN	68.4	79.2	86.8	91.6	91.5	98.1	98.8	99.1	86.1	91.7	95.0	97.3	79.1	90.8	96.2	98.7
SoftTriple [26]	512	Inception-BN	65.4	76.4	84.5	90.4	-	-	-	-	84.5	90.7	94.5	96.9	78.6	86.6	91.8	95.4
HORDE [27]	512	Inception-BN	66.8	77.4	85.1	91.0	90.4	97.8	98.4	98.7	86.2	91.9	95.1	97.2	80.1	91.3	96.2	98.7
XBM [28]	512	Inception-BN	65.8	75.9	84.0	89.9	89.9	97.6	98.4	98.6	82.0	88.7	93.1	96.1	79.5	90.8	96.1	98.7
MS [29]	512	Inception-BN	65.7	77.0	86.3	91.2	89.7	97.9	98.5	98.8	84.1	90.4	94.0	96.5	78.2	90.5	96.0	98.7
HTL [30]	512	Inception-BN	57.1	68.8	78.7	86.5	80.9	94.3	95.8	97.2	81.4	88.0	92.7	95.7	74.8	88.3	94.8	98.4
IRT_R [31]	384	DeiT-S	76.6	85.0	91.1	94.3	91.9	98.1	98.7	98.9	-	-	-	-	84.2	93.7	97.3	99.1
Sph-DeiT [32]	384	DeiT-S	76.2	84.5	90.2	94.3	89.6	97.2	98.0	98.4	81.7	88.6	93.4	96.2	82.5	92.9	97.1	99.1
Sph-DINO [32]	384	ViT	78.7	86.7	91.4	94.9	90.1	97.1	98.0	98.4	86.6	91.8	95.2	97.4	82.2	92.1	96.8	98.9
Sph-ViT [32]	384	ViT(IN21k)	85.1	90.7	94.3	96.4	90.4	97.4	98.2	98.6	81.7	89.0	93.0	95.8	82.1	92.5	97.1	99.1
Hyp-DeiT [32]	384	DeiT-S	77.8	86.6	91.9	95.1	90.5	97.8	98.5	98.9	86.4	92.2	95.5	97.5	83.3	93.5	97.4	99.1
Hyp-DINO [32]	384	ViT	80.9	87.6	92.4	95.6	92.4	98.4	**98.9**	99.1	**89.2**	**94.1**	96.7	**98.1**	85.1	94.4	97.8	99.3
Hyp-ViT [32]	384	ViT(IN21k)	85.6	91.4	**94.8**	96.7	92.5	98.3	98.8	99.1	86.5	92.1	95.3	97.3	85.9	94.9	**98.1**	**99.5**
DINOv2-reg (ours)	384	ViT	**87.1**	**92.3**	94.7	**96.9**	**92.9**	98.4	**98.9**	**99.4**	89.1	93.8	**96.9**	98.0	**86.2**	**95.4**	98.0	**99.5**

Table 3.1: Recall@k metrics comparing across state-of-the-art methods on the CUB-200, In-Shop, Cars-196, and Stanford Online Products datasets. We perform competitively across architectures and outperform all previous methods in several settings.

3.3.1 Performance Gains

The empirical evaluations reported in [5] revealed substantial improvements in retrieval accuracy, with the most pronounced gains observed on datasets characterized by fine-grained visual distinctions, such as CUB-200 (birds) and Cars-196. When compared against using only the single `[CLS]` token from the identical DINOv2-reg backbone, the proposed 10-token multi-vector representation consistently delivered significantly higher recall metrics, particularly the crucial Recall@1 metric which reflects the ability to find the correct match as the top-ranked result.

Ablation studies carefully dissected the contributions of each component. Adding just the register tokens to the `[CLS]` token already provided a solid performance boost, confirming their value in capturing complementary information. However, the subsequent addition of the automatically discovered ROI tokens resulted in further, often larger, improvements, highlighting the significant benefit of incorporating these targeted, localized region descriptors. (Table 3.2 in [5] presents these key ablation results).

These findings strongly validate the core hypotheses: specialized tokens like registers and internally derived ROIs capture critical fine-grained information missed by the global `[CLS]`

token, and the multi-vector late-interaction framework effectively leverages this enriched representation for superior retrieval. When compared against a wide range of existing state-of-the-art methods (many using different backbones or more complex architectures), the proposed approach demonstrated highly competitive, and in several cases superior, performance, underscoring its effectiveness. Refer to Table 3.1 for detailed benchmark comparisons. Figure 3.4 offers qualitative visualizations suggesting how different cue tokens attend to distinct patterns and how this facilitates matching across images).

Method	CUB (R@1)	Cars-196 (R@1)
DINOv2-reg (CLS only)	82.2	87.8
DINOv2-reg (CLS+Registers)	85.1	88.2
DINOv2-reg (CLS+Register+ROI)	**87.1**	**89.1**

Table 3.2: Ablation study showing the impact of adding register tokens and ROI tokens to the base [CLS] token representation on Recall@1 performance. Both additions provide significant improvements (Adapted from).

Method	#Tokens	Total Dim	Memory (1M imgs)	CUB R@1
Single-Vector	1	384	1.5 GB	83.5
Ours (CLS+Reg+ROI)	10	3,840	15 GB	87.1
All Patch+Reg+CLS	≈201	77,184	309 GB	87.3

Table 3.3: Theoretical index size comparison for 1 million images (384-dim float32 embeddings). Shows the proposed 10-token method offers a significant performance boost over single-vector retrieval with much lower memory cost than using all tokens (Adapted from).

3.3.2 Efficiency

Beyond accuracy, a defining characteristic of this framework is its efficiency, particularly concerning computational and storage resources, when contrasted with the alternative of using dense, all-patch representations. While using all P patch tokens (where P can be several hundred for standard ViT configurations) provides the most exhaustive local information, the associated costs render it impractical for most large-scale retrieval deployments. Indexing hundreds of

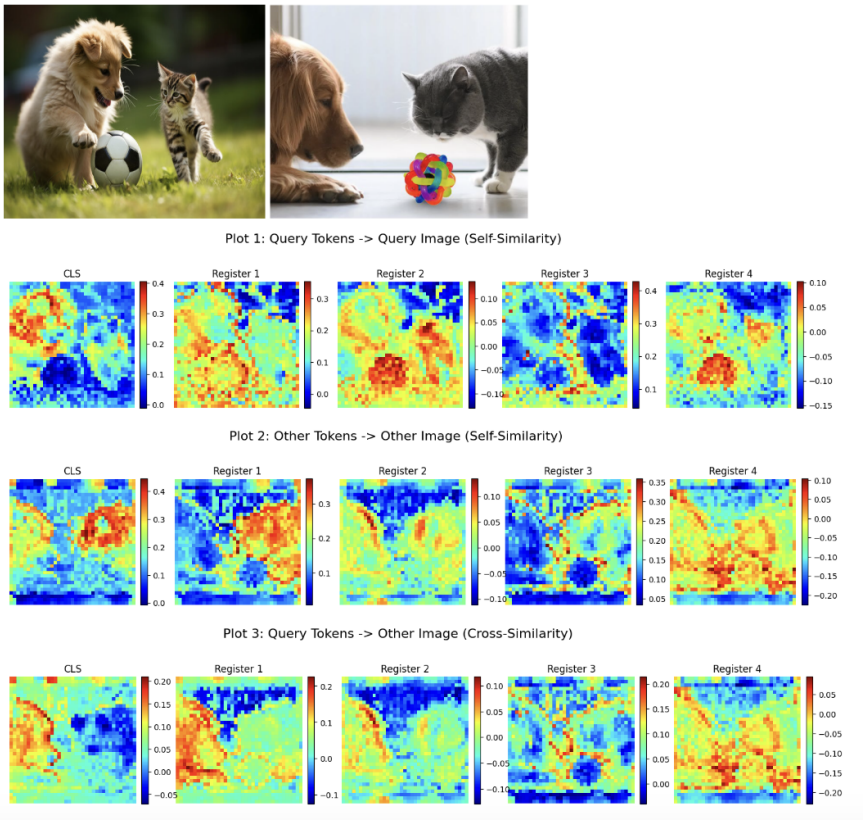

Figure 3.4: We show how the tokens from a query image search for patterns in another image by showing the heat map from a given token to all image patch tokens. Row 1 shows the heatmap from query (left) cue tokens to its image patches. Row 2 shows the same for the other (right) image. Row 3 shows the heat map computed by using query cue tokens across the other images image patch tokens. For example, register 1 in the query focuses strongly on the head shape when searching across the other image (row 3).

vectors per image leads to enormous storage requirements, and the $O(m \times n)$ complexity of late-interaction matching becomes prohibitive when m and n are large.

The proposed method achieves its significant performance improvements using only a small, fixed set of tokens (e.g., 10). This represents a highly practical trade-off between representational richness and efficiency. The memory footprint is only marginally larger than single-vector methods but remains orders of magnitude smaller than storing all patch embeddings, making large-scale indexing and deployment feasible. Similarly, the computational cost of the late-interaction matching remains manageable due to the small number of tokens involved. (Table 3.3 provides a compelling theoretical comparison of index sizes and retrieval performance for

single-vector, the proposed 10-token method, and an all-patch approach, clearly illustrating the efficiency benefits).

Furthermore, the approach demonstrated practical robustness. Ablation studies on the size ($N \times N$) of the region Ω used for pooling ROI tokens indicated that performance was relatively stable across different reasonable sizes (e.g., $N = 3, 5, 7$), with $N = 3$ often providing near-optimal results. This suggests the method is not overly sensitive to this hyperparameter. Importantly, using a pooled region ($N \geq 3$) consistently outperformed using just the single buddy patch ($N = 1$), confirming the value of incorporating local spatial context in the ROI token. See Table **??** for results varying ROI size).

The combination of the fully internal ROI discovery mechanism, the compactness of the final token set, and the robustness to hyperparameters makes this approach a practical and effective solution for enhancing fine-grained retrieval.

ROI Size	**CUB (R@1)**	**Cars-196 (R@1)**
Single Patch	85.6	88.9
3×3	87.1	89.1
5×5	87.1	89.0
7×7	87.3	89.2
9×9	87.1	89.0

Table 3.4: Ablation on region size for ROI tokens. We report Recall@1 on CUB and Cars-196 with single patch vs. $N \times N$ mean pooling for N=3, 5, 7, 9. Our default setting is N=3.

3.4 Synthesis: Leveraging Internal Attention for Richer, Efficient Representations

The research presented in this section introduces a novel and effective strategy for overcoming the inherent limitations of single global vector representations in vision transformers, particularly for challenging fine-grained image retrieval tasks. The core achievement lies in demonstrating how to construct richer, multi-vector representations by exploiting latent information within specialized

tokens (registers) and leveraging the internal similarity structures revealed by the ViT's own self-attention mechanism, all accomplished in an efficient and entirely self-contained manner.

The key contribution is the method for augmenting the global `[CLS]` token with a small, intelligently curated set of additional tokens: the part-aware register tokens (repurposed from their original design [36]) and the localized ROI tokens derived automatically via the novel "buddy patch" mechanism. This ROI discovery process, driven by cue-token-to-patch similarities, offers a unique way to identify and extract features from salient image regions without necessitating external detection, segmentation, or saliency modules, relying instead on the model's learned internal representations.

This work strongly embodies the principle of attending to focused information within the image to build better representations. It moves decisively beyond holistic descriptors by acknowledging that different learned tokens can specialize—capturing global context (`[CLS]`), distinct object parts (registers), or specific local details (ROIs). The adoption of the late-interaction matching framework [37] provides the necessary mechanism to effectively harness this diverse set of features during the retrieval comparison process, allowing for nuanced, part-to-part or region-to-region matching. Crucially, this significant enhancement in representational capability and retrieval performance is achieved while maintaining computational tractability. By using only a small number of tokens, the method strikes a practical and highly effective balance between retrieval accuracy and the resource constraints of large-scale systems. This research underscores the largely untapped potential residing within the internal workings of Vision Transformers, suggesting that further exploration of their attention patterns and emergent token behaviors can lead to more powerful and efficient approaches to visual understanding.

Chapter 4

Controlling Attention: Aligning with Object Focus for Robustness

4.1 Motivation: Mitigating Shortcut Learning in ViTs via Explicit Attention Control

Having explored methods to leverage context across examples (Chapter 2) and to select informative features within images (Chapter 3), we now turn to a more direct form of intervention: explicitly controlling the internal attention mechanism of Vision Transformers (ViTs). This focus is motivated by a well-recognized vulnerability of ViTs and other powerful deep learning models: their propensity for shortcut learning [6]. Shortcut learning occurs when models exploit superficial correlations or biases within the training data to achieve high accuracy on that data, without necessarily learning the underlying semantic concepts intended by the task. ViTs, with their ability to form global dependencies from the earliest layers and their relative lack of strong spatial inductive biases compared to CNNs, can be particularly adept at discovering and exploiting such shortcuts. Instead of grounding their predictions in the intrinsic properties of objects, such as their shape and structure, they might learn to associate classes with co-occurring background textures, specific dataset artifacts, or simple local patterns that happen to be

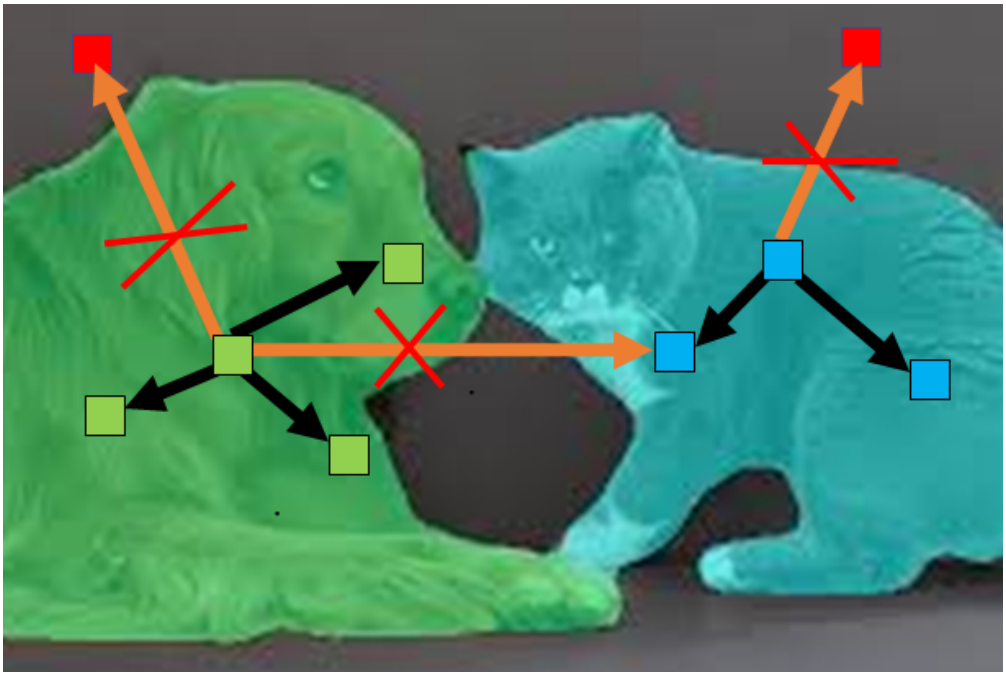

Figure 4.1: We restrict learning attention to objects of the same class.

discriminative only within the training distribution [7, 8, 9, 10]. For instance, a model trained on ImageNet might learn that ”snow” is a strong predictor for ”husky,” failing when presented with a husky indoors.

The consequences of relying on these shortcuts are severe. Models exhibit poor generalization performance when faced with out-of-distribution (OOD) data, where the spurious correlations learned during training are no longer present. This brittleness makes them unreliable for real-world applications where encountering novel contexts or variations is inevitable. Furthermore, this reliance on superficial cues often manifests as a strong bias towards texture over shape [9, 10]. While humans rely heavily on configural shape information for object recognition [7, 8], models biased towards texture fail to capture this crucial aspect of visual understanding, limiting their semantic grounding. While techniques like large-scale pre-training and sophisticated data augmentation strategies aim to mitigate shortcut learning by exposing the model to greater data diversity, they provide no explicit guarantee against it. Architectural modifications that re-introduce CNN-like locality biases (e.g., Swin Transformer [38], ConViT [39]) offer another direction, but may also limit the unique strengths of the original Transformer

architecture.

This landscape reveals a compelling need for methods that directly address the mechanism of shortcut learning within ViTs by explicitly guiding the model's focus towards semantically relevant information. This motivates the Object-Focused Attention (OFA) framework presented in [11]. OFA proposes a novel training strategy that directly intervenes in the ViT's self-attention computation. The central hypothesis is that by actively penalizing attention paid to non-object regions during training, using semantic segmentation masks as a readily available form of weak supervision, we can compel the model to prioritize learning from intra-object interactions. This explicit guidance aims to instill a stronger inductive bias towards object-centric processing, thereby reducing reliance on spurious background cues, fostering a better understanding of object shape, enhancing robustness to OOD scenarios, and ultimately leading to more semantically grounded representations.

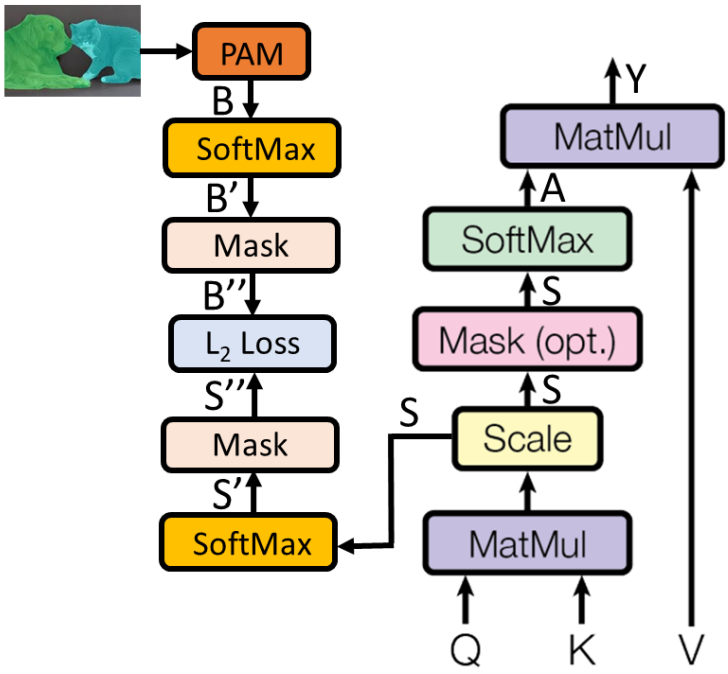

Figure 4.2: Object Focused Attention (OFA) Module. Right: Standard self-attention calculation producing output Y. Left: Parallel OFA branch calculating the L_2 loss between the model's foreground attention distribution (S'') and the target object-centric distribution (B'') derived from the Patch Attention Matrix (PAM).

4.2 Object-Focused Attention (OFA) Framework

The OFA framework introduces a modification to the standard ViT training regime, designed to encourage the self-attention mechanism to concentrate its focus within the boundaries of semantic objects. This is achieved by incorporating an auxiliary loss term computed directly from the attention weights, without altering the core ViT architecture for inference.

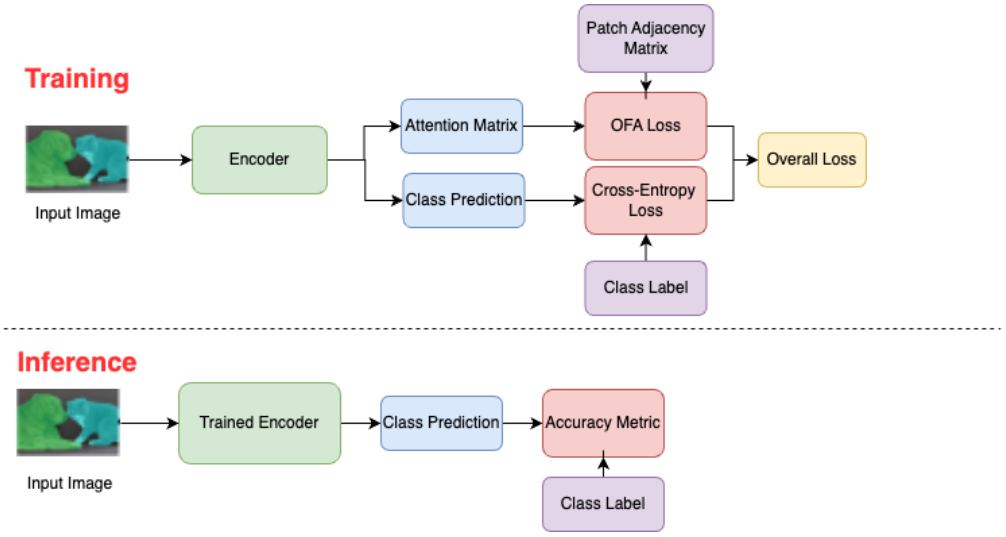

Figure 4.3: Data flow showing differences in training and inference. OFA is shown explicitly as a training time method and thus can be used without any segmentation labels during inference.

4.2.1 Method: Auxiliary OFA Loss Guided by Semantic Masks

Recall that in a standard ViT self-attention layer, input patch tokens $X \in \mathbb{R}^{N \times d}$ are projected to Query (Q), Key (K), and Value (V) matrices. The pairwise compatibility between patches is captured in the pre-softmax attention score matrix:

$$S = \frac{QK^T}{\sqrt{d}} \in \mathbb{R}^{N \times N}$$

Applying row-wise softmax yields the final attention weights:

$$A = \text{softmax}(S) \in \mathbb{R}^{N \times N} \tag{4.1}$$

(Eq. (4.1) in [11]) Row A_i dictates how patch p_i aggregates information from all patches p_k (via the Value matrix V) to form its updated representation.

OFA introduces a parallel computation during training that compares the model's learned attention patterns (S) against an ideal object-focused pattern derived from semantic segmentation masks. Given an image segmented into regions $\mathcal{R} = \{R_1, \ldots, R_r\}$ (where some regions correspond to foreground objects and potentially one or more to background), the goal is to enforce that a patch p_i primarily attends to other patches p_k that fall within the same foreground object region.

To quantify this ideal behavior, a target Patch Attention Matrix (PAM) $B \in \{0,1\}^{N \times N}$ is constructed. For a pair of patches (p_i, p_k), the entry B_{ik} is set to 1 if both patches intersect the same foreground object mask R_j, and B_{ik} is set to 0 otherwise (including cases where one or both patches are background, or they belong to different foreground objects). This matrix B encodes the desired sparse connectivity, restricting attention flow strictly within object boundaries. (Handling patches overlapping multiple regions might involve assigning them to the region with the largest overlap or other heuristics, though the core principle remains restricting attention based on shared object identity).

Directly comparing the raw scores S with the binary target B is problematic. Instead, both are processed to represent attention distributions focused on foreground objects:

- **Normalization:** Both the model's scores S and the target matrix B are normalized row-wise using the softmax function: $S' = \text{softmax}(S)$ and $B' = \text{softmax}(B)$. For a row i in B' corresponding to an object patch, this normalization creates a uniform distribution over the k patches belonging to that object ($B'_{ik} = 1/k$ if p_k is in the same object, 0 otherwise), representing the target of attending equally to all parts of the same object.

- **Background Masking:** To concentrate the learning signal on relevant foreground interactions, rows in both S' and B' corresponding to patches primarily identified as background are masked out or ignored during the loss computation. This yields the final matrices S'' and B'' representing the model's and the target's foreground attention

distributions, respectively.

The Object-Focused Attention (OFA) loss, $\mathcal{L}_{OFA}$, quantifies the discrepancy between these two distributions. The squared L2 distance was chosen in [11] as a simple and effective measure:

$$\mathcal{L}_{OFA} = ||S'' - B''||_2^2 \tag{4.2}$$

(Eq. (4.2) in [11]) Minimizing this loss directly penalizes the model whenever its attention distribution S'' for a foreground patch deviates significantly from the ideal object-centric distribution B''. It pushes the model to reduce attention weights assigned to patches outside the current object boundary and potentially encourages a more uniform spread of attention within the object. This auxiliary objective gently guides the self-attention mechanism towards learning object structures defined by the masks, complementing the primary task objective. This approach differs from methods that might enforce hard constraints or use attention supervision for specific downstream tasks (like localization), offering instead a general training adaptation aimed at improving the underlying representation quality. Figure 4.2 provides a clear diagram of this auxiliary loss computation.

4.2.2 Integration and Training

The OFA loss is seamlessly integrated into the ViT training pipeline:

- **Placement:** Rather than applying the loss only at the final layer, OFA is typically computed at multiple self-attention layers distributed throughout the ViT architecture. The rationale is that attention patterns evolve through the network depth; guiding attention at early layers might influence low-level feature grouping, while guidance at later layers can shape higher-level semantic focus. Empirical studies in [11] found that applying $\mathcal{L}_{OFA}$ at early, middle, and late layers (specifically layers 1, 7, and 14) yielded superior results compared to single-layer application or other combinations. Though this application is dataset dependent and is left as a design choice.

When using multiple OFA losses, their contributions are typically aggregated. For instance, in [11], a weighting scheme gave increasing importance to guidance at deeper layers using geometrically decaying weights. For OFA losses at layers $l_1, l_2, ..., l_k$ (ordered from shallow to deep), the total OFA loss might be computed as:

$$\mathcal{L}_{OFA} = \frac{1}{N_{ofa}} \sum_{i=1}^{N_{ofa}} w_i \cdot \mathcal{L}_{OFA}^{(l_i)} \quad \text{where } w_i = \gamma^{k-i} \tag{4.3}$$

with γ being a decay factor (e.g., 0.9) and N_{ofa} the number of layers where the loss is applied. For the specific case of layers [1, 7, 14] ($N_{ofa} = 3$), this becomes:

$$\mathcal{L}_{OFA} = \frac{1}{3}(\gamma^0 \mathcal{L}_{OFA}^{(14)} + \gamma^1 \mathcal{L}_{OFA}^{(7)} + \gamma^2 \mathcal{L}_{OFA}^{(1)}) \tag{4.4}$$

And for layers [1, 14] ($N_{ofa} = 2$):

$$\mathcal{L}_{OFA} = \frac{1}{2}(\gamma^0 \mathcal{L}_{OFA}^{(14)} + \gamma^1 \mathcal{L}_{OFA}^{(1)}) \tag{4.5}$$

- **Total Loss Function:** The overall objective function for training the network becomes a weighted combination of the standard loss for the primary downstream task ($\mathcal{L}_{task}$, e.g., binary cross-entropy for multi-label classification) and the aggregated OFA loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha \cdot \mathcal{L}_{OFA} \tag{4.6}$$

(Eq. (4.6) in [11]) The hyperparameter α controls the strength of the object-focusing regularization relative to the main task objective.

- **No Inference Cost:** A key practical advantage is that the entire OFA auxiliary loss computation, including the need for semantic segmentation masks, is confined to the training phase. During inference, the OFA branch is simply removed, and the ViT operates exactly as a standard pre-trained model, using the learned weights. Consequently, OFA

enhances the model's intrinsic properties—its robustness and semantic understanding—without imposing any additional computational cost, latency, or architectural modifications at deployment time.

While the primary focus of [11] was on using ground-truth or high-quality predicted masks, the paper also explored the potential for self-supervised OFA. It proposed integrating OFA principles with Masked Autoencoders (MAE) [18], utilizing a novel multi-scale masking strategy compatible with the Musiq transformer's spatial grid encoding [40]. Although preliminary, this suggests pathways for extending object-focused attention principles to scenarios where explicit pixel-level masks are unavailable, potentially by leveraging unsupervised segmentation methods (like SAM [41]) or other self-supervised signals related to object coherence.

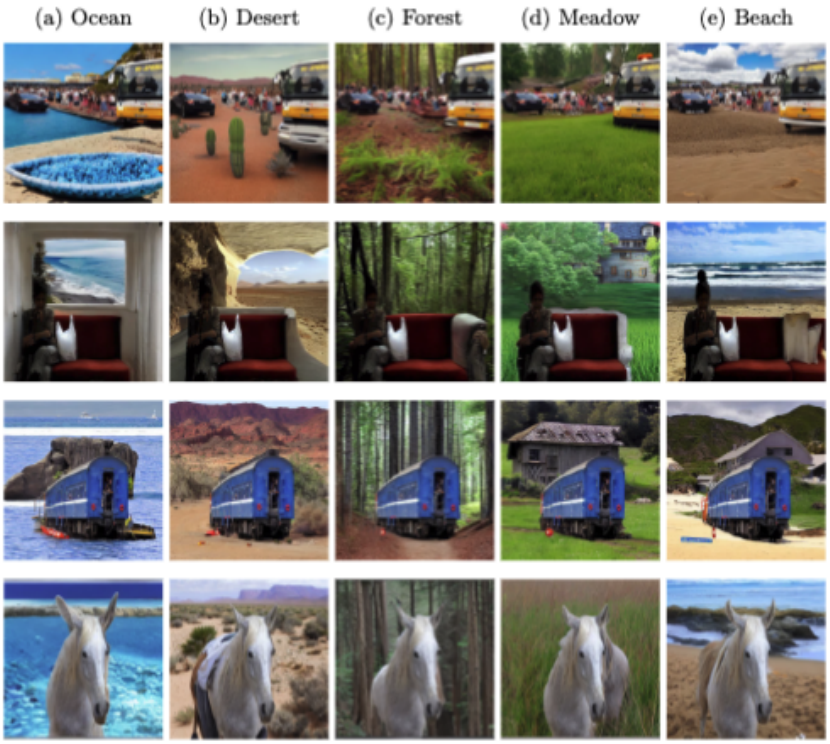

Figure 4.4: Examples from the OOD benchmark created by inpainting MS COCO validation image backgrounds using Stable Diffusion with different scene prompts (Ocean, Desert, Forest, Meadow, Beach). Foreground objects remain unchanged.

Base Model	Resolution	Baseline ViT mAP (Δ)	ViT+OFA mAP (Δ)
ViT-Base-Patch16 (1k)	224	73.9 (-7.0)	78.6 (-2.2)
ViT-Base-Patch16 (21k)	224	73.6 (-9.3)	81.7 (-2.2)
ViT-Large-Patch16 (21k)	384	79.0 (-6.9)	83.7 (-3.0)

Table 4.1: OOD robustness results on the Stable Diffusion inpainted MS COCO test set. Shows mAP on original test set and performance drop (Δ) on the inpainted set. ViT+OFA demonstrates significantly less degradation, indicating better robustness to background changes.

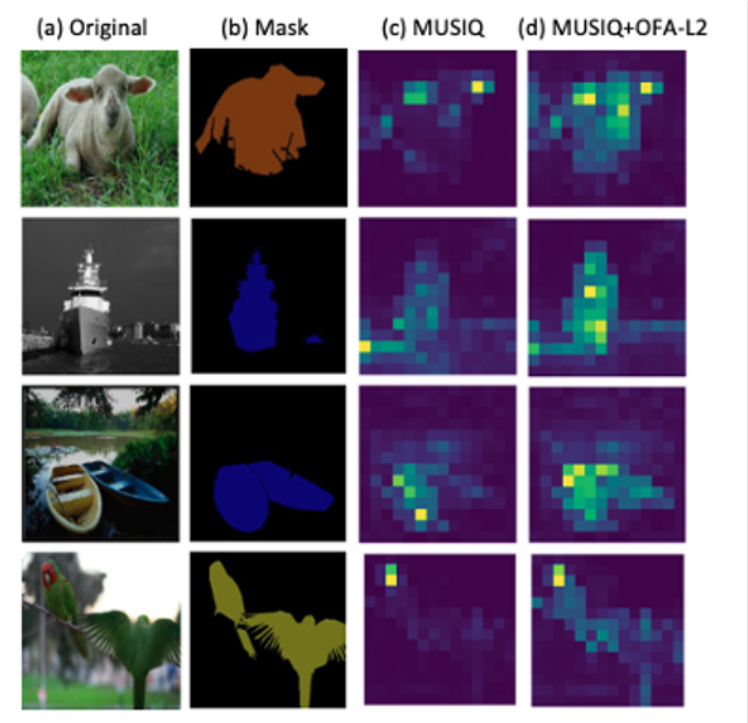

Figure 4.5: Comparison of attention maps of proposed MUSIQ + OFA and baseline MUSIQ.

4.3 Benefits – Robustness and Semantic Understanding

The integration of the OFA loss during training confers significant advantages that manifest primarily as enhanced model robustness against distributional shifts and a demonstrably improved semantic understanding, particularly regarding object shape, moving beyond superficial texture reliance.

Methods	MS COCO	zero-shot VOC2012
ViT-Base	86.6	81.7
ViT-Base + OFA	87.3	87.8
MUSIQ-single	87.5	89.7
MUSIQ-multi	88.0	90.2
MUSIQ-single + OFA	89.0	90.9
MUSIQ-single + MAE	89.7	92.3
MUSIQ-multi + OFA	89.9	93.2
MUSIQ-multi + MAE	91.6	93.6
MUSIQ-single + MAE + OFA	91.7	94.7
MUSIQ-multi + MAE + OFA	**92.1**	**95.4**

Table 4.2: mAP multilabel classification results on the MS COCO and Pascal VOC2012 datasets. All models are trained and evaluated on MS COCO. They are then applied on Pascal VOC2012 without any finetuning besides the linear head.

OFA at Different Layers (40% data)	1	2	3	4	5	6	7	8	9	10	11	12	mAP
[12]												✓	83.5
[1]	✓												83
[1,12]	✓											✓	83.6
[1,6,12]	✓					✓						✓	83.7
[1,3,7,10,12]	✓			✓			✓			✓		✓	**84.0**
[1,3,5,7,9,11]	✓		✓		✓		✓		✓		✓		83.7
[all]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	83.6

Table 4.3: Ablation of computing OFA loss on multiple attention blocks in ViT+OFA using the ViT-Base-Patch16 (21k) on a subset of MS COCO.

4.3.1 OOD Robustness: Resilience to Background Perturbations

A central claim of OFA is its ability to mitigate reliance on spurious background correlations, thereby improving generalization to OOD data. To provide rigorous evidence for this, a novel OOD evaluation benchmark was specifically created for the study in [11]. This involved taking the standard MS COCO validation images and systematically altering their backgrounds using the powerful generative capabilities of Stable Diffusion inpainting [42]. For each image, the existing semantic segmentation masks were used to precisely define the foreground object regions, which were kept unchanged. The background region was then inpainted using Stable Diffusion guided by diverse text prompts corresponding to distinct scene types (ocean, desert, forest, meadow, beach). This process yielded multiple versions of each validation image where the foreground objects were realistically placed into novel and often dramatically different background contexts,

directly challenging models that might have learned spurious object-background associations from the original training data. Figure 4.4 showcases examples from this challenging OOD dataset.

Evaluating models trained on the original MS COCO data against this benchmark revealed the effectiveness of OFA. Standard ViT baselines suffered a substantial drop in classification accuracy when faced with the inpainted images, confirming their sensitivity to the statistics of the background context they were trained on. In stark contrast, ViT models trained with the addition of the OFA loss exhibited significantly greater resilience, maintaining much higher accuracy and showing considerably less performance degradation on the OOD dataset. (Table 4.1 provides the quantitative comparison).

This result strongly supports the hypothesis that OFA successfully encourages the model to ground its predictions in the properties of the foreground objects themselves, making the learned representations more invariant to background variations and thus more robust for real-world deployment where context shifts are common.

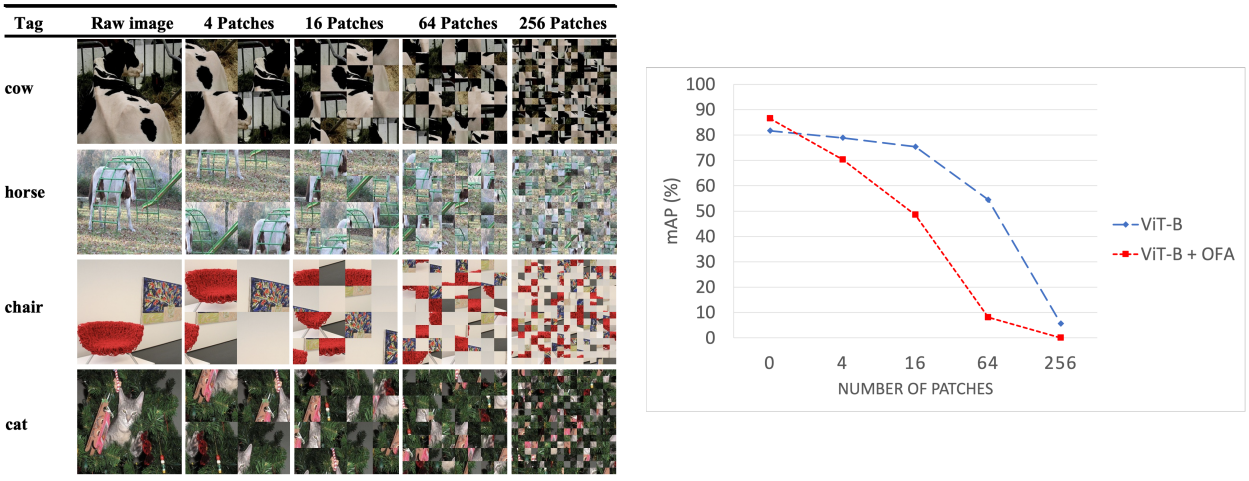

Figure 4.6: Example shuffle operation applied to a varying number of patches. For humans, the objects in a shuffled grid with 4 patches already seem unrecognizable. The mAP over 20 classes on PASCAL VOC2012 when patches are shuffled. While the classification performance of ViT + OFA drops significantly, those of ViT hardly drops.

4.3.2 Mitigating Shortcuts & Enhancing Shape Understanding

The OFA framework directly confronts the well-documented tendency of deep networks, including ViTs, to develop a strong bias towards texture cues at the expense of shape information [7, 8, 9, 10]. Shortcut learning often involves exploiting these easily discriminable texture patterns, even if they are not semantically central to the object category. By explicitly forcing the self-attention mechanism to operate within the confines of object masks via the $\mathcal{L}_{OFA}$ loss, OFA discourages attention links based merely on texture similarity between object patches and unrelated background patches. Instead, it promotes the integration of information across the spatial extent of the object itself. This process inherently encourages the model to learn features related to the object's internal structure, the spatial arrangement of its parts, and its overall configural or holistic shape.

To empirically investigate whether OFA indeed fosters a better grasp of shape, a random patch shuffling experiment was designed [11]. This experiment serves as a direct probe for sensitivity to spatial configuration. Input images were divided into regular grids of increasing granularity (e.g., 2×2, 4×4, 8×8, 16×16), and the patches within each grid were randomly permuted before being fed into the trained ViT models. This manipulation preserves all local patch information (including textures) but systematically destroys the global spatial arrangement, i.e., the object's shape. Figure 4.6 visually illustrates the effect of shuffling.

The results of this experiment were particularly illuminating. Standard ViT baselines demonstrated a surprising degree of insensitivity to this shuffling; their classification performance degraded only moderately even with significant spatial disruption (e.g., 4×4 or 8×8 grids). This suggests that their predictions relied heavily on a "bag-of-features" approach, primarily using local patch statistics (like texture) rather than the overall spatial configuration. In dramatic contrast, the performance of ViT models trained with OFA plummeted much more rapidly as the degree of shuffling increased. The graph in Figure 4.6 presents this stark difference in performance curves.

This heightened sensitivity to spatial scrambling provides compelling evidence that the OFA training regime successfully induced the learning of representations that are dependent on the

global configuration of object parts—in other words, OFA fostered a significantly better understanding of holistic object shape compared to the baseline ViT.

4.3.3 Qualitative Evidence: Focused Attention Maps

The mechanism through which OFA achieves these benefits is further corroborated by qualitative analysis of the models' internal attention patterns. Visualizing the final-layer self-attention maps often reveals clear differences between baseline ViTs and their OFA-trained counterparts. Baseline models frequently exhibit diffuse attention that spreads across the image, including significant attention paid to background regions or scattered, seemingly arbitrary high-attention patches. Conversely, the attention maps generated by OFA-trained models typically show a much stronger concentration on the foreground objects, often clearly outlining their silhouettes and resembling coarse segmentation masks. Figure 4.5 provides several illustrative examples comparing attention maps.

This direct visual evidence confirms that the auxiliary $\mathcal{L}_{OFA}$ loss effectively fulfills its intended purpose during training: guiding the self-attention mechanism to learn patterns that prioritize and align with the semantic objects present in the scene.

4.4 Synthesis: Impact of Directly Controlling Attention Learning

The Object-Focused Attention (OFA) framework, as detailed in this section, offers a targeted strategy for enhancing the robustness and semantic fidelity of Vision Transformers by exerting direct control over the self-attention learning process. Through the introduction of an auxiliary loss function guided by semantic segmentation masks, OFA actively discourages attention to non-object image regions during training, thereby promoting a focus on intra-object feature relationships. This contrasts with other approaches that might rely solely on data diversity or architectural biases to implicitly shape attention.

The primary and most significant impact of this explicit attention control is the effective mitigation of shortcut learning. By reducing the model's tendency to exploit spurious correlations involving background or texture cues, OFA cultivates representations that are more robustly grounded in the intrinsic properties of the objects depicted. This manifests in tangible benefits, including markedly improved generalization to out-of-distribution data where context may vary, and a demonstrably enhanced understanding and utilization of holistic object shape information, moving beyond superficial texture matching. A key advantage of OFA is that these substantial improvements in representation quality and model reliability are achieved through a training-only modification, leveraging existing semantic priors (masks) without incurring any additional computational overhead or architectural complexity at inference time.

Positioning OFA within the overarching theme of this thesis—the critical role of attention and information selection—it represents a powerful example of explicit attentional control. While the methods discussed in Chapter 2 focused on leveraging context across examples and Chapter 3 focused on selecting informative features within images based on emergent model properties, OFA actively intervenes to enforce a desired attentional behavior aligned with high-level semantic understanding (objectness). This highlights the significant potential of incorporating domain knowledge or semantic priors, when available and appropriate, to guide the learning dynamics of powerful but potentially under-constrained models like ViTs. Such explicit control mechanisms are crucial stepping stones towards building AI vision systems that are not only accurate on standard benchmarks but also reliable, generalizable, and semantically coherent in their understanding of the visual world. Future work could explore extending these control principles to fully self-supervised settings or applying them to regulate attention for other desirable properties beyond object focus.

Chapter 5

Synthesis, Conclusion, and Future Work

5.1 Synthesis of Contributions

This thesis has embarked on an investigation into enhancing visual representation learning by focusing on the pivotal roles of attention mechanisms and strategic information selection. Moving beyond traditional approaches that often treat images as isolated entities or rely on monolithic global descriptors, the research presented herein has explored diverse strategies for optimizing what visual information models utilize and how they process it. Through four interconnected studies, we have demonstrated that carefully guiding and structuring the flow of information—whether across examples, within images, or by directly controlling the attention mechanism itself—leads to representations that are demonstrably more robust, efficient, and effective for a range of challenging computer vision tasks, including image classification, multi-label classification, and fine-grained image retrieval.

The journey began in Chapter 2 by challenging the prevalent independent example processing paradigm. We first introduced CNN2Graph [3], a framework designed to infuse dataset-level context into image classification. By constructing a differentiable bipartite graph between mini-batch images and a fixed proxy set (comprising learnable class prototypes and fixed data anchors), and employing cross-attention for information aggregation, CNN2Graph demonstrated

the feasibility and benefit of end-to-end learning that incorporates inter-example relationships, while also providing efficient inductive inference capabilities. This work highlighted the importance of the attention mechanism's scalability in handling such interactions. Subsequently, focusing on the specific demands of image retrieval, we presented DMCAC [4]. This framework addressed the disconnect between typical representation learning objectives and the retrieval task itself by explicitly conditioning representation learning on interactions with a target database during training. Through a novel self-supervised objective based on minimizing the divergence of retrieval distributions across augmented query views, coupled with a Cross-Attention Classification (CAC) loss for semantic grounding, DMCAC demonstrated that aligning the training process with the downstream task via database conditioning yields state-of-the-art retrieval performance. Both CNN2Graph and DMCAC underscored the power of leveraging context across examples, mediated by cross-attention.

Chapter 3 shifted the focus to attention within images, tackling the information bottleneck imposed by single global vector representations, particularly for fine-grained retrieval. The "Augmenting CLS" approach [5] proposed an efficient multi-vector representation strategy. Instead of relying solely on the ViT's `[CLS]` token or resorting to computationally expensive dense patch representations, this method augments the `[CLS]` token with a small, curated set of informative tokens: specialized register tokens (repurposed from DINOv2-reg [36] based on their emergent part-representing properties) and novel Region-of-Interest (ROI) tokens. Crucially, these ROI tokens are discovered internally by leveraging the ViT's own cue-token-to-patch similarity patterns, requiring no external modules. Combined with a ColBERT-inspired [37] late-interaction matching framework, this compact multi-vector representation was shown to significantly boost fine-grained retrieval accuracy while maintaining computational tractability, effectively balancing representational richness and efficiency.

Finally, Chapter 4 addressed the critical issue of shortcut learning and the lack of robustness often observed in ViTs by exploring methods for explicitly controlling the attention mechanism. The Object-Focused Attention (OFA) framework [11] introduced an auxiliary loss term during

training, guided by semantic segmentation masks. This $\mathcal{L}_{OFA}$ loss directly penalizes self-attention weights assigned to non-object regions, effectively forcing the model to concentrate its attention within object boundaries. This intervention was shown to significantly mitigate reliance on spurious background cues, leading to enhanced robustness against out-of-distribution (OOD) data (validated on a novel dataset created using Stable Diffusion inpainting [42]) and fostering a better understanding of holistic object shape (demonstrated via patch shuffling experiments). Importantly, this semantic guidance is achieved without incurring any additional computational cost at inference time.

Collectively, these four contributions illustrate a multi-faceted approach to improving visual representations. By operating at different scopes—inter-example context, intra-image feature selection, and direct attention control—and employing diverse mechanisms—graph structures, database interaction, internal signal exploitation, and auxiliary losses—this body of work consistently demonstrates that principled strategies for managing attention and selecting information are key to unlocking more powerful and reliable visual understanding.

5.2 Overarching Principles

Across the diverse methodologies explored in this thesis, several overarching principles emerge as central to achieving advancements in visual representation learning through the lens of attention and information selection:

- **The Primacy of Context:** A recurring theme is the limitation of processing visual information in isolation. Both CNN2Graph [3] and DMCAC [4] explicitly demonstrated that incorporating broader context—whether the structural context of classes within a dataset or the specific content of a target database—leads to more informative and task-aligned representations. Mechanisms like graph-based message passing and dynamic retrieval coupled with cross-attention proved effective in integrating this contextual information, enabling models to learn relative similarities and task-specific nuances missed

by context-free approaches.

- **The Power of Guided Attention:** Unconstrained attention mechanisms, while powerful, are susceptible to latching onto superficial correlations (shortcut learning). The OFA framework [11] provided strong evidence that explicitly guiding attention using semantic priors (object masks) significantly enhances robustness and semantic grounding. Even implicit guidance, such as the structure imposed by the proxy set in CNN2Graph or the database conditioning in DMCAC, helps steer the model towards learning more meaningful features compared to purely unsupervised or weakly constrained learning paradigms. Directing where and how a model attends is crucial for reliable performance.

- **Beyond Single Vectors: The Utility of Multi-Vector Representations:** The work on augmenting the CLS token [5] clearly illustrated the inadequacy of single global descriptors for tasks demanding fine-grained detail. Moving towards multi-vector representations allows for a more nuanced encoding of visual information, capturing global context, distinct parts, and localized details simultaneously. However, the key lies in efficient construction and utilization. Leveraging internal model signals (registers, attention patterns) to select a compact, informative set of tokens, combined with late-interaction matching, provides a practical path to harness the benefits of multi-vector representations without succumbing to the intractability of dense methods.

- **Object Focus Enhances Shape Understanding and Robustness:** The OFA study [11] directly linked controlled, object-focused attention to improved robustness and a better grasp of holistic object shape. By mitigating the model's reliance on background cues and forcing it to integrate information across object regions, OFA countered the prevalent texture bias and led to representations more sensitive to configural structure. This highlights that controlling attention is not just about improving accuracy on standard benchmarks but about fostering deeper, more human-like semantic understanding and resilience to real-world variations.

- **Task Alignment Through Training:** The DMCAC framework [4] emphasized the importance of aligning the representation learning process with the specifics of the downstream task. By simulating retrieval and conditioning learning on database interactions, DMCAC achieved superior performance compared to methods trained with generic objectives, demonstrating that tailoring the learning environment to the target application yields more effective representations.

These principles are often interconnected. For instance, guided attention (OFA) naturally promotes object focus and shape understanding. Incorporating context (CNN2Graph, DMCAC) can implicitly guide attention towards more relevant inter-example relationships. Efficient multi-vector representations (Augmenting CLS) provide the richer substrate needed for fine-grained tasks where context and object focus are paramount. Ultimately, this thesis argues that progress in visual representation learning hinges on intelligently managing the vast information landscape of visual data through mechanisms that strategically select, contextualize, and guide the focus of attention.

5.3 Limitations and Future Directions

While the research presented in this thesis offers significant contributions, it is also important to acknowledge its limitations and identify promising avenues for future investigation.

5.3.1 Limitations

The CNN2Graph framework [3], while providing end-to-end learning, relies on a proxy set whose size scales linearly with the number of classes, potentially posing scalability challenges for datasets with extremely large label spaces. Its performance might also exhibit sensitivity to the initial random sampling of anchor examples. The DMCAC approach [4], although effective, introduces additional complexity during training due to the need for database embedding storage, periodic updates, and the training-time retrieval step (even if approximate). Its performance might

also depend on the quality of the query augmentations and the representativeness of the training query/database split. The Augmenting CLS method [5], while efficient, is currently tied to the specific DINOv2-reg architecture [36] and its register tokens; generalizing the principle of leveraging specialized internal tokens to other architectures requires further investigation. Furthermore, while the 10-token set proved effective, its optimality across all datasets and tasks is not guaranteed, and scaling retrieval still necessitates integration with approximate nearest neighbor (ANN) indexing techniques [5]. The OFA framework's [11] primary limitation in its supervised form is the reliance on semantic segmentation masks, which are not always available, and the generation of high-quality masks can be computationally expensive itself. The interplay between the task loss and the auxiliary OFA loss might also require careful balancing via the α hyperparameter. Generally, across all studies, the focus remained predominantly on the visual modality, with limited exploration of multimodal interactions.

5.3.2 Future Directions

These limitations and the insights gained throughout this work point towards several exciting future research directions:

- **Scaling Contextual Methods:** Developing more scalable versions of context-aware methods like CNN2Graph [3], perhaps using hierarchical proxy structures or more efficient graph sampling techniques, is crucial for application to massive datasets. Similarly, optimizing the training-time retrieval and database management in DMCAC [4] remains an important practical challenge.

- **Generalizing Internal Token Discovery:** Extending the core idea from Chapter 3—leveraging internal model signals to discover informative tokens—beyond the specific DINOv2-reg architecture [36] holds significant promise. Can similar principles be applied to identify salient patches, object parts, or other specialized features within standard ViTs or even CNNs by analyzing attention maps, activation patterns, or gradient flows?

Exploring adaptive or learned token selection mechanisms is a rich area for future work.

- **Self-Supervised Object Focus:** Advancing the preliminary explorations into self-supervised OFA [11] is a key direction. Can models learn to focus attention on objects without explicit masks? This could involve leveraging unsupervised segmentation techniques (e.g., using foundation models like SAM [41] to generate pseudo-masks), exploiting motion cues in videos to distinguish foreground from background, or developing novel self-supervised objectives based on object coherence or Gestalt principles. Successfully decoupling object-focused learning from mask supervision would vastly broaden its applicability.

- **Advanced Attention Control:** Moving beyond penalizing background attention (OFA) [11], future work could explore more sophisticated forms of attention control. Can we guide attention towards specific attributes, functional parts, or regions relevant for fine-grained reasoning tasks? Can attention patterns be regularized for improved interpretability or fairness?

- **Multimodal Representation Learning and Reasoning (Vision-Language Models):** A particularly compelling avenue is extending the principles of attention analysis and control explored in this thesis to the rapidly evolving domain of Vision-Language Models (VLMs). Current large VLMs, such as Google's Gemini, OpenAI's GPT-4V, or open models like LLaVA, while demonstrating impressive capabilities, still exhibit significant failure modes, particularly in tasks requiring complex visual reasoning, understanding fine-grained details, spatial relationships, or handling compositionality. It is hypothesized that some of these failures stem from challenges in effectively fusing information across modalities—specifically, how visual attention interacts with textual attention and concepts. Misalignments or suboptimal integration between visual grounding (what the model "sees") and textual processing (what the model "understands" or is asked) can lead to errors in visual question answering (VQA), image captioning, and instruction following. Ongoing

research, extending the work of this thesis, focuses on analyzing the evolution and behavior of attention mechanisms within VLMs. This involves dissecting how attention patterns shift when processing combined visual and textual inputs and identifying specific failure modes where the interplay between visual and language attention leads to incorrect reasoning (e.g., failing to correctly bind attributes to objects, misinterpreting spatial prepositions relative to the visual scene, hallucinating objects or relationships not present). Understanding these cross-modal attention dynamics is crucial for diagnosing VLM limitations and developing new methods—potentially inspired by the attention control techniques explored here—to improve the grounding, reasoning capabilities, and reliability of these powerful multimodal systems.

- **Efficiency and Scalability:** Continued research into optimizing the computational efficiency of attention mechanisms (e.g., sparse attention, linear attention variants) and developing scalable indexing solutions for multi-vector representations [5] remains essential for deploying advanced representation learning techniques in real-world, resource-constrained environments.

5.4 Concluding Remarks

This thesis has systematically investigated the critical role of attention mechanisms and strategic information selection in advancing the state of visual representation learning. By moving beyond simplistic assumptions of independent processing and global feature compression, we have developed and validated novel frameworks that leverage context across examples, exploit informative features within images, and exert explicit control over the attention process itself. The presented contributions—CNN2Graph [3], DMCAC [4], Augmenting CLS [5], and OFA [11]—demonstrate tangible improvements in classification accuracy, retrieval performance, fine-grained understanding, and out-of-distribution robustness.

The overarching message is clear: how a model attends to and selects information is as crucial

as the architecture itself. By thoughtfully designing mechanisms that guide attention towards semantically relevant signals, whether derived from dataset structure, database context, internal model patterns, or explicit semantic priors, we can build visual representations that are not only more accurate but also more efficient, robust, and semantically grounded. As AI systems, particularly large foundation models and multi-modal architectures, continue to grow in complexity and capability, the principles of understanding, analyzing, and controlling attention and information flow will remain paramount. The research presented here offers several concrete steps in this direction, providing both effective methodologies and valuable insights to fuel continued progress towards machines that can truly see and understand the visual world.

Bibliography

[1] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[2] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[3] Vivek Trivedy and Longin Jan Latecki. "CNN2Graph: Building Graphs for Image Classification". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 1–11. DOI: 10.1109/WACV56688.2023.00009.

[4] Vivek Trivedy and Longin Jan Latecki. "Image Retrieval with Self-Supervised Divergence Minimization and Cross-Attention Classification". In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*. 2024, pp. 1344–1352.

[5] Anonymous. "Augmenting the CLS Token with Region of Interest Tokens for Efficient Multi-Vector Image Retrieval". In: *ICCV 2025 Submission*. Paper ID 2302. 2025.

[6] Yehui Tang et al. *Augmented Shortcuts for Vision Transformers*. 2021. arXiv: 2106.15941 [cs.CV].

[7] N. Baker and J.H. Elder. "Deep Learning Models Fail to Capture the Configural Nature of Human Shape Perception". In: *iScience* 25 (9 2022), p. 104913. DOI: 10.1016/j.isci.2022.104913.

[8] Nicholas Baker et al. “Deep convolutional networks do not classify based on global object shape”. In: *PLoS Computational Biology* 14.12 (2018), e1006613. DOI: 10.1371/journal.pcbi.1006613.

[9] Robert Geirhos et al. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. In: *ICLR* (2019).

[10] Robert Geirhos et al. “Generalisation in humans and deep neural networks”. In: *NeurIPS* (2018).

[11] Vivek Trivedy, Amani Almalki, and Longin Jan Latecki. “Learning Object Focused Attention”. In: *Pattern Recognition: 27th International Conference, ICPR 2024, Proceedings, Part IX*. Berlin, Heidelberg: Springer-Verlag, 2024, pp. 291–306.

[12] Ashish Vaswani et al. “Attention is All You Need”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17)*. Curran Associates Inc., 2017, pp. 6000–6010.

[13] Thomas N. Kipf and Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017. arXiv: 1609.02907 [cs.LG].

[14] William L. Hamilton, Rex Ying, and Jure Leskovec. *Inductive Representation Learning on Large Graphs*. 2018. arXiv: 1706.02216 [cs.SI].

[15] Petar Veličković et al. *Graph Attention Networks*. 2018. arXiv: 1710 . 10903 [stat.ML].

[16] R. Hadsell, S. Chopra, and Y. LeCun. “Dimensionality Reduction by Learning an Invariant Mapping”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*. Vol. 2. 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015). DOI: 10.1109/cvpr.2015.7298682.

[18] Kaiming He et al. *Masked Autoencoders Are Scalable Vision Learners*. June 2021. arXiv: 2111.06377 [cs.CV].

[19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs". In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547.

[20] Andrew Zhai and Hao-Yu Wu. *Classification is a Strong Baseline for Deep Metric Learning*. 2019. arXiv: 1811.12649 [cs.CV].

[21] Eu Wern Teh, Terrance DeVries, and Graham W. Taylor. *ProxyNCA++: Revisiting and Revitalizing Proxy Neighborhood Component Analysis*. 2020. arXiv: 2004.01113 [cs.CV].

[22] Michael Opitz et al. "Deep Metric Learning with BIER: Boosting Independent Embeddings Robustly". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 276–290. DOI: 10.1109/TPAMI.2018.2848925.

[23] Wonsik Kim et al. *Attention-based Ensemble for Deep Metric Learning*. 2018. arXiv: 1804.00382 [cs.CV].

[24] Yumin Suh et al. "Stochastic Class-Based Hard Example Mining for Deep Metric Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[25] Sungyeon Kim et al. "Proxy Anchor Loss for Deep Metric Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[26] Qi Qian et al. "SoftTriple Loss: Deep Metric Learning Without Triplet Sampling". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[27] Pierre Jacob et al. *Metric Learning With HORDE: High-Order Regularizer for Deep Embeddings*. 2019. arXiv: 1908.02735 [cs.CV].

[28] Xun Wang et al. *Cross-Batch Memory for Embedding Learning*. 2020. arXiv: 1912.06798 [cs.LG].

[29] Xun Wang et al. "Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[30] Weifeng Ge et al. *Deep Metric Learning with Hierarchical Triplet Loss*. 2018. arXiv: 1810.06951 [cs.CV].

[31] Alaaeldin El-Nouby et al. *Training Vision Transformers for Image Retrieval*. 2021. arXiv: 2102.05644 [cs.CV].

[32] Aleksandr Ermolov et al. "Hyperbolic Vision Transformers: Combining Improvements in Metric Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 7409–7419.

[33] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110. URL: https://api.semanticscholar.org/CorpusID:174065.

[34] Artem Babenko et al. *Neural Codes for Image Retrieval*. 2014. arXiv: 1404.1777 [cs.CV]. URL: https://arxiv.org/abs/1404.1777.

[35] Yunchao Gong et al. *Multi-scale Orderless Pooling of Deep Convolutional Activation Features*. 2014. arXiv: 1403.1840 [cs.CV]. URL: https://arxiv.org/abs/1403.1840.

[36] Timothée Darcet et al. *Vision Transformers Need Registers*. 2024. arXiv: 2309.16588 [cs.CV]. URL: https://arxiv.org/abs/2309.16588.

[37] Omar Khattab and Matei Zaharia. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT*. 2020. arXiv: 2004.12832 [cs.IR]. URL: https://arxiv.org/abs/2004.12832.

[38] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 10012–10022.

[39] Stéphane d’Ascoli et al. *ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases*. 2021. DOI: 10.48550/ARXIV.2103.10697. URL: https://arxiv.org/abs/2103.10697.

[40] Junjie Ke et al. “Musiq: Multi-scale image quality transformer”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 5148–5157.

[41] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV].

[42] Robin Rombach et al. “High-Resolution Image Synthesis With Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 10684–10695.

[43] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV]. URL: https://arxiv.org/abs/2304.07193.

[44] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[45] Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. arXiv: 2002.05709 [cs.LG].

[46] C. Wah et al. *Caltech-UCSD Birds-200-2011*. Tech. rep. CNS-TR-2011-001. California Institute of Technology, 2011.

[47] Ziwei Liu et al. “DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1096–1104. DOI: 10.1109/CVPR.2016.124.

[48] Hyun Oh Song et al. “Deep Metric Learning via Lifted Structured Feature Embedding”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2016.

[49] Mathilde Caron et al. *Unsupervised Learning of Visual Features by Contrasting Cluster Assignments*. 2021. arXiv: 2006.09882 [cs.CV].

[50] Jonathan Krause et al. "3D Object Representations for Fine-Grained Categorization". In: *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*. Sydney, Australia, 2013.

[51] Hugo Touvron et al. *Training data-efficient image transformers distillation through attention*. 2021. arXiv: 2012.12877 [cs.CV].

[52] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[53] Mathilde Caron et al. *Emerging Properties in Self-Supervised Vision Transformers*. 2021. arXiv: 2104.14294 [cs.CV].

[54] Shiv Ram Dubey. *A Decade Survey of Content Based Image Retrieval using Deep Learning*. Nov. 2020. DOI: 10.48550/arXiv.2012.00641.

[55] Zhong-Yu Li et al. *Multi-Token Enhancing for Vision Representation Learning*. 2024. arXiv: 2411.15787 [cs.CV]. URL: https://arxiv.org/abs/2411.15787.

[56] Hyeonwoo Noh et al. *Large-Scale Image Retrieval with Attentive Deep Local Features*. 2018. arXiv: 1612.06321 [cs.CV]. URL: https://arxiv.org/abs/1612.06321.

[57] Bingyi Cao, Andre Araujo, and Jack Sim. *Unifying Deep Local and Global Features for Image Search*. 2020. arXiv: 2001.05027 [cs.CV]. URL: https://arxiv.org/abs/2001.05027.

[58] Philip Sun et al. *SOAR: Improved Indexing for Approximate Nearest Neighbor Search*. 2024. arXiv: 2404.00774 [cs.LG]. URL: https://arxiv.org/abs/2404.00774.

[59] Yongming Rao et al. *DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification*. 2021. arXiv: 2106.02034 [cs.CV]. URL: https://arxiv.org/abs/2106.02034.

[60] Michael S. Ryoo et al. *TokenLearner: What Can 8 Learned Tokens Do for Images and Videos?* 2022. arXiv: 2106.11297 [cs.CV]. URL: https://arxiv.org/abs/2106.11297.

[61] Hyun Oh Song et al. *Deep Metric Learning via Lifted Structured Feature Embedding*. 2015. arXiv: 1511.06452 [cs.CV]. URL: https://arxiv.org/abs/1511.06452.

[62] Ziwei Liu et al. "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1096–1104. DOI: 10.1109/CVPR.2016.124.

[63] Yair Movshovitz-Attias et al. *No Fuss Distance Metric Learning using Proxies*. 2017. arXiv: 1703.07464 [cs.CV]. URL: https://arxiv.org/abs/1703.07464.

[64] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV]. URL: https://arxiv.org/abs/1405.0312.

[65] Shichao Xu et al. "A Dual Modality Approach For (Zero-Shot) Multi-Label Classification". In: (2022).

[66] Vignesh Ramanathan et al. *PACO: Parts and Attributes of Common Objects*. 2023.

[67] Tal Ridnik et al. "Ml-decoder: Scalable and versatile classification head". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 32–41.

[68] Vladislav Sovrasov. "Combining Metric Learning and Attention Heads For Accurate and Efficient Multilabel Image Classification". In: *arXiv preprint arXiv:2209.06585* (2022).

[69] Ruyang Liu et al. "Causality Compensated Attention for Contextual Biased Visual Recognition". In: *The Eleventh International Conference on Learning Representations*.

[70] Tal Ridnik et al. “Imagenet-21k pretraining for the masses”. In: *arXiv preprint arXiv:2104.10972* (2021).

[71] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *International Conference on Learning Representations (ICLR)* (2015).

[72] Bin-Bin Gao and Hong-Yu Zhou. “Learning to discover multi-class attentional regions for multi-label image recognition”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 5920–5932.

[73] Bin-Bin Gao et al. “Deep label distribution learning with label ambiguity”. In: *IEEE Transactions on Image Processing* 26.6 (2017), pp. 2825–2838.

[74] Tal Ridnik et al. “Asymmetric loss for multi-label classification”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 82–91.

[75] Xing Cheng et al. “MlTr: Multi-label Classification with Transformer”. In: *IEEE Int. Conf. on Multimedia and Expo (ICME)* (2022).

[76] Youwei Liang et al. “Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations”. In: *ICLR* (2022).

[77] Ashish Vaswani et al. “Scaling Local Self-Attention for Parameter Efficient Visual Backbones”. In: *CVPR* (2021).

[78] Lingchen Meng et al. “AdaViT: Adaptive Vision Transformers for Efficient Image Recognition”. In: *CVPR* (2022).

[79] Tongzhou Wang and Phillip Isola. “Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere”. In: *ICML* (2020).

[80] Xiangxiang Chu et al. “Twins: Revisiting Spatial Attention Design in Vision Transformers”. In: *NeurIPS* (2021).

[81] Jianwei Yang et al. “Focal Modulation Networks”. In: 2022.

[82] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. “Crossvit: Cross-attention multi-scale vision transformer for image classification”. In: 2021.

[83] Li Yuan et al. “Tokens-to-token vit: Training vision transformers from scratch on imagenet”. In: 2021.

[84] Amani Almalki and Longin Jan Latecki. “Self-Supervised Learning with Masked Image Modeling for Teeth Numbering, Detection of Dental Restorations, and Instance Segmentation in Dental Panoramic Radiographs”. In: 2023.

[85] Kan Wu et al. “Rethinking and improving relative position encoding for vision transformer”. In: 2021, pp. 10033–10041.