A DCT-Based Contrastive Approach for Learning Visual Representations

Daniel de Sousa Moraes
TeleMídia Lab - PUC-Rio
danielmoraes@telemidia.puc-rio.br

Pedro Cutrim dos Santos
TeleMídia Lab - PUC-Rio
thiagocutrim@telemidia.puc-rio.br

Antonio José G. Busson
TeleMídia Lab - PUC-Rio/BTG Pactual
antonio.busson@btgpactual.com

Julio Cesar Duarte
Military Institute of Engineering
duarte@ime.eb.br

Sérgio Colcher
TeleMídia Lab - PUC-Rio
colcher@inf.puc-rio.br

ABSTRACT

Recent advances in self-supervised learning have significantly improved visual representation learning, creating better alternatives to supervised methods in image understanding tasks. Motivated by the demand for robust features under low-quality image scenarios, we propose adapting the SimCLR framework to operate in the frequency domain using quantized DCT coefficients from JPEG-compressed images. We aim to investigate the integration of the Discrete Cosine Transform (DCT) with self-supervised contrastive learning to enable visual representation learning. To evaluate the learned representations, we apply linear classification and JPEG artifact removal tasks, employing ResNet-50 and Vision Transformer (ViT) encoders. Results indicate that ViT-based embeddings, especially when combined with spatial or frequency-domain inputs, mitigate performance loss degraded by compression. Furthermore, artifact removal experiments with modified UNet architectures show that incorporating ViT embeddings and attention maps can slightly improve reconstruction quality. These findings suggest that frequency-domain self-supervised learning is a promising direction for building robust and transferable visual representations. However, the observed gains remain modest and limited by computational constraints.

KEYWORDS

Discrete Cosine Transform, Visual Representation Learning, Image Classification, Image Restoration

1 INTRODUCTION

Visual representation learning is fundamental for many computer vision tasks, including image classification [29], tagging [13], object detection [27], semantic and instance segmentation [15], and image restoration [9]. The effectiveness of these tasks depends directly on the quality of the visual features and representations.

Following the success of foundation models [2] in Natural Language Processing (NLP), computer vision has also advanced significantly, developing its own suite of foundation models that introduce novel architectures and also adopting key principles from NLP. Vision Transformers (ViTs), in particular, have emerged as a powerful architecture for large-scale visual representation learning [11]. Notably, models such as DINO [5] and its successor DINOv2 [22] have demonstrated the effectiveness of ViTs in self-supervised contexts. These models learn high-quality representations without requiring manual annotations and transfer well across a wide range of tasks.

One of the central techniques in training visual foundation models in a self-supervised paradigm is contrastive learning, which optimizes the agreement between similar instances while maximizing the separation of dissimilar ones in the latent space [17, 20]. Frameworks such as SimCLR [6, 7], MoCo [14], and CLIP [25] have established new standards in self-supervised and weakly-supervised learning. These methods demonstrate the effectiveness of contrastive objectives for learning transferable visual features without manual annotation.

Traditionally, computer vision techniques, such as those briefly presented, operate in the spatial domain, where images' pixels are directly analyzed to extract features. These methods employ convolutional operations and other pixel-based transformations to understand visual patterns. However, recently, frequency domain approaches have gained increasing interest, particularly in specific domain applications such as image restoration and classification. These approaches transform images into a frequency representation, often revealing structural patterns that may not be as apparent in the spatial domain [19].

A widely used frequency-based technique in image processing is the Discrete Cosine Transform (DCT), a key component of JPEG compression. DCT decomposes image blocks into frequency components, isolating low-frequency elements (global structure) and high-frequency elements (fine details and textures). This transformation enables direct manipulation of DCT coefficients within the media bitstream, offering advantages such as improved compression efficiency, artifact removal, and feature extraction for machine learning models [23]. Busson et al. [3, 4] used DCT coefficients in deep models for artifact reduction in JPEG/MPEG media, improving decompression quality. These findings support the broader hypothesis that frequency-domain representations can be effectively used in downstream vision tasks and motivate their integration into representation learning frameworks.

Despite its utility in compression and restoration, the use of DCT for learning high-level visual representations using self-supervised techniques remains underexplored. This represents a promising opportunity to investigate whether DCT-derived representations can be used in self-supervised learning frameworks to improve

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

the robustness and transferability of visual models, particularly in scenarios involving image degradation or compression.

The DCT offers several advantages that make it an attractive choice for self-supervised representation learning, such as Frequency-Domain Insights, Quality Factor Variability, and Efficiency and Scalability.

Additionally, frequency-based approaches can offer practical advantages in resource-constrained environments. For instance, transmitting highly compressed images and performing analysis or restoration directly from the DCT coefficients on the client side can significantly reduce data transmission overhead and minimize network resource usage.

In this context, we hypothesize that the variability introduced by DCT quantization can serve as an effective form of data augmentation for contrastive learning. By training models to recognize consistent content across varying compression levels, we aim to develop robust, discriminative, and generalizable representations.

To investigate this hypothesis, we propose adapting the SimCLR framework [6, 7], replacing raw image inputs with quantized DCT coefficients. Furthermore, we evaluate and compare two different encoder architectures: the standard ResNet-50 used in SimCLR, and the Vision Transformer (ViT), which has shown strong performance in recent self-supervised learning settings. We selected ResNet-50 and ViT as representative architectures of convolutional and transformer paradigms, respectively. This allows assessing whether frequency-domain contrastive learning benefits more from local convolutional features or from the global self-attention mechanisms typical of transformers.

We evaluate the learned representations using two complementary approaches. First, we adopt the standard linear evaluation protocol [1], where a linear classifier is trained on top of the frozen pre-trained encoders. Second, we apply the representations in a practical downstream task, JPEG artifact removal, using the learned features to guide the reconstruction of compressed images. The goal is to determine whether the learned representations can be effectively applied to classic image tasks, particularly in scenarios involving low-quality or compressed images.

By addressing this, our research aims to explore the underutilized synergy between frequency-domain transformations and self-supervised learning. It contributes a novel perspective to developing efficient and robust visual representations, particularly for applications involving limited bandwidth or degraded visual input.

The remainder of this work is organized as follows: section 2 reviews the related works, highlighting advancements in visual representation learning, the use of DCT with downstream tasks in the frequency domain. section 3 outlines the proposed framework, including the adaptation of SimCLR for DCT-based augmentations using varying quality factors, the definition of the ResNet-50 and ViT encoders, and the pre-training process of both versions. In section 4, we describe the experiments that evaluate the learned representations and analyze their respective results. Finally, section 5 presents the final considerations, limitations, and future works.

2 RELATED WORKS

Numerous self-supervised models for visual representation learning have successfully employed contrastive learning [6–8, 14] as well as other training paradigms [12, 26] to generate robust, generalizable representations.

Furthermore, similarly to developments observed in NLP, Vision Transformers have also been integrated into self-supervised visual learning methods, further enhancing their representational capabilities.

Chen et al. [10] investigated ViTs in self-supervised representation learning. Their empirical evaluations demonstrate that self-supervised ViTs can achieve competitive results compared to convolutional neural networks, particularly when scaled to larger architectures. This study underscores the potential of ViTs in self-supervised learning contexts and provides practical insights into effective training methodologies.

Ling et al. [21] proposed a novel deep clustering method that effectively integrates Vision Transformers (ViTs) with contrastive learning for image clustering tasks. Experiments conducted on eight benchmark datasets demonstrate VTCC's (Vision Transformers for Contrastive Clustering) superior clustering performance and improved training stability compared to existing state-of-the-art approaches. The findings highlight the efficacy of ViTs in capturing global dependencies when combined with contrastive learning principles, positioning VTCC as a significant advancement in self-supervised deep clustering research.

These studies have successfully integrated Vision Transformers with contrastive learning for visual representation learning. They operate exclusively in the conventional spatial domain, applying their methods directly to image pixels to learn spatial features and properties. Our work, on the other hand, aims at exploring the frequency domain to enable learning visual representation with the DCT information.

The DCT has been applied in domain-specific machine learning tasks, particularly in image restoration and artifact removal. For instance, Jiang et al. [18] proposed the Flexible Blind Convolutional Neural Network (FBCNN), designed to address JPEG compression artifacts across varying quality factors (QFs). FBCNN predicts the QF of a compressed image and embeds this information into the reconstruction process, allowing users to adjust the balance between artifact reduction and detail retention according to their preferences. Extensive experiments demonstrate that FBCNN outperforms existing state-of-the-art methods in both quantitative metrics and visual quality, offering a versatile and effective solution for real-world JPEG artifact removal.

Ouyang and Chen [23] introduced the DCTransformer, an approach for recovering quantized Discrete Cosine Transform (DCT) coefficients in JPEG-compressed images. While existing recovery methods often operate in the pixel domain, DCTransformer works directly in the DCT domain, effectively capturing both spatial and frequential correlations through a dual-branch architecture. They introduce a luminance-chrominance alignment head to unify features across different color components. Extensive experiments demonstrate that DCTransformer outperforms state-of-the-art techniques in mitigating JPEG artifacts and restoring image quality.

Yang et al. [31] also introduced D2LNet, a method that integrates spatial and frequency domain information to reduce artifacts in JPEG-compressed images. D2LNet addresses this by first transforming spatial domain images to the frequency domain using the Fast Fourier Transform (FFT). It then employs two core modules: the Amplitude Correction Module (ACM) and the Phase Correction Module (PCM), which collaboratively facilitate interactive learning between spatial and frequency domain information. Extensive experiments on both color and grayscale images demonstrate that D2LNet outperforms previous state-of-the-art methods in mitigating JPEG artifacts, highlighting the effectiveness of incorporating dual-domain learning in image restoration tasks.

As shown, several works have successfully applied Vision Transformers (ViTs) in self-supervised visual representation learning [10, 21], but all operate exclusively in the spatial domain. A substantial portion of the literature uses frequency-domain methods, particularly DCT, in downstream image processing applications such as artifact removal [18, 23, 31]. These studies underscored the benefits of frequency-domain information but do not address representation learning.

In contrast, our work is unique in combining self-supervised visual representation learning with frequency-domain, applying the DCT. While previous works either focus on DCT for task-specific solutions or on representation learning in the spatial domain, our method bridges these perspectives. Aiming at learning generalizable visual representations directly from DCT-transformed inputs, we hypothesize that our approach is able to support multiple downstream tasks, such as classification and artifact removal.

3 DCT-BASED CONTRASTIVE REPRESENTATION LEARNING

In this work, we aim to explore the potential of DCT for creating general visual representations using self-supervised contrastive learning. By varying the quality factors in the quantization table, we propose to generate augmented views of the same image that emphasize different frequency components, thereby creating a rich and diverse set of training samples for contrastive learning.

The DCT has long been a core image processing component, particularly in compression algorithms such as JPEG. By transforming image data from the spatial domain to the frequency domain, the DCT separates an image into its constituent frequency components, enabling efficient representation and manipulation of visual information. This property makes the DCT particularly well-suited for tasks such as image compression, artifact removal, and feature extraction [24, 30].

The Simple Framework for Contrastive Learning of Representations (SimCLR) [6] has proven highly effective for self-supervised representation learning in computer vision. At its core, SimCLR learns representations by contrasting augmented views of the same image, encouraging the model to focus on semantically meaningful features while remaining invariant to superficial transformations. In this work, we adapt SimCLR, illustrated in Figure 1, to operate in the frequency domain by applying DCT and quantization with two different quality factors in the augmented image inputs. With this adaptation, we aim to enable the model to learn representations invariant to variations in image quality, making it particularly well-suited for tasks involving compressed or low-quality images.

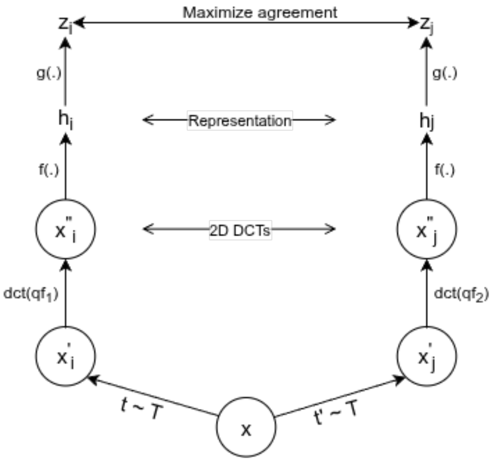

Figure 1: Adaptation of SimCLR framework [6] with DCT extraction augmentation. Both $x_i^{''}$ and $x_j^{''}$ are the DCTs extracted from the augmented images and quantized using quality factors qf_1 and qf_2, respectively.

1. **Input Image Augmentation**: For each input image, with size 224 x 224, two distinct but identically sized augmented views, $x_i^{'}$ and $x_j^{'}$, are generated using standard spatial-domain augmentations, including random flipping, zooming, color jittering, and brightness adjustments. These augmentations guarantee that the model is exposed to a diverse set of transformations, improving its ability to generalize.
2. **DCT Transformation and Quantization**: Each augmented image $x_i^{'}$ and $x_j^{'}$ is partitioned into non-overlapping 8 x 8 blocks, and a 2D DCT is applied to transform these blocks into the frequency domain. The resulting DCT coefficients are quantized using two different quality factors:
   a. For $x_i^{''}$, the DCT coefficients are quantized with quality factor $qf_1 = 10$.
   b. For $x_j^{''}$, the DCT coefficients are quantized with quality factor $qf_2 = 50$.
3. **Base Encoder Network**: The quantized DCT coefficients, $x_i^{''}$ and $x_j^{''}$, are fed into a shared encoder network $f(.)$. The encoder outputs feature representations h_i and h_j for the two augmented views. As in Chen et al. [6, 7], we use the popular ResNet, specifically defining a ResNet-50 architecture as our base encoder. We also propose using a Vision Transformer as an alternative encoder. The ViT base encoder is defined as:
   a. Given that the DCT blocks are 8 x 8, we also define this as the input patch size.
   b. The number of patches is defined as $N = HW/P^2$, where H is the image height, W width, and P is the patch size. Thus, the resulting number of patches is 784, which also is the input sequence length for the Transformer.

c. The Transformer uses a constant 384 as the length of the latent vector through its layers.

The ViT base encoder is implemented as defined by Dosovitskiy et al. [11], but uses a reduced configuration consisting of 12 transformer blocks with 6 attention heads, as described in DINO [5]. The ViT processes the quantized DCT coefficients as input tokens, using self-attention mechanisms to capture global relationships between frequency components.

4. **Projection Head**: The representations produced by the base encoder are passed through a projection head composed of three fully connected layers, following the findings of Chen et al. [7], which show that deeper projection heads can improve the results. We also added batch normalization layers between the fully connected layers to stabilize and accelerate training. The projection head maps the representations to a lower-dimensional space where the contrastive loss is applied, with $z_i = \text{MLP}(h_i)$ and $z_j = \text{MLP}(h_j)$.
5. **Contrastive Loss**: The Normalized Temperature-Scaled Cross Entropy (NT-Xent)[6], loss function is used to maximize the similarity between z_i and z_j (positive pair) while minimizing the similarity with representations of other images in the batch (negative pairs). The function is defined as:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

where $\text{sim}(z_i, z_j) = z_i^T z_j / (\|z_i\| \|z_j\|)$ is the cosine similarity between z_i and z_j. τ is the temperature parameter. The denominator sums over all negative pairs (i.e., all views k that are not i or j). And $\mathbb{1}_{[k\neq i]}$ is an indicator function that ensures $k \neq i$. The final loss is computed across all positive pairs, both (i, j) and (j, i), in a mini-batch.

6. **Memory Mechanism**: As introduced by Chen et al. [7], we also adopt a memory bank to increase the number of views used as negative examples, attempting to reduce the large batch sizes demanded by SimCLR.

The use of two different quality factors, $qf_1 = 10$ and $qf_2 = 50$, is central to this adaptation. By quantizing the DCT coefficients with varying quality factors, we generate two distinct views of the same image that emphasize different frequency components. We chose $qf_1 = 10$ because the lower quality factor results in stronger quantization, discarding more high-frequency information. This produces a compressed representation with visible artifacts, challenging the model to focus on the underlying structure of the image. On the other hand, a higher quality factor preserves more high-frequency details, resulting in a less compressed representation. We chose $qf_2 = 50$, mainly because studies have claimed that for quality factors above 50, metrics typically indicate minimal perceptual degradation, suggesting that the human eye cannot easily distinguish between the original and compressed images. This provides the model with richer information about fine textures and edges.

Contrasting these two views encourages the model to learn representations that are invariant to compression artifacts while retaining sensitivity to semantically meaningful features. This approach aligns with the goals of contrastive learning, where the model is trained to recognize similarities between different views of the same image.

Adapting the framework to use DCT coefficients offers several key advantages. First, it provides frequency-domain insights, as operating in the frequency domain allows the model to access structural information that may not be apparent in the spatial domain. This is particularly beneficial for tasks involving compressed or low-quality images, where frequency-domain representations can reveal critical details obscured by compression artifacts or noise. Second, the framework benefits from invariance to quality variations. By contrasting images quantized with two different quality factors, the model learns to recognize the underlying content of the image regardless of compression quality, making it robust to artifacts and noise. Finally, the approach maintains scalability, as DCT-based preprocessing is computationally efficient and can be seamlessly integrated into large-scale self-supervised learning pipelines. Together, these advantages make the adapted framework well-suited for learning robust and generalizable visual representations.

To validate our proposal, we experimented with training the adapted SimCLR with the two versions of encoders: ResNet-50 and the ViT encoder. Next, we detail the training procedure of the two versions.

3.1 Contrastive Training Setup

In our experiments, we used Google Colab with a TPU v2-8 (8-core) accelerator, which represented the optimal available computational resource. We used the ImageNet-1K ILSVRC-2012 dataset [28] for training our contrastive visual representation models. While the complete dataset comprises 1,281,167 training images and 50,000 validation images across 1,000 object classes, hardware constraints necessitated working with a subset of 200 classes randomly selected only from the training set, resulting in 255,674 training images.

For model training, we implemented distinct batch size configurations to accommodate architectural differences: a 512 batch size was used for the ResNet-50 encoder, while the larger ViT architecture required a reduced batch size of 96 to maintain computational feasibility. This adaptation ensured stable training while optimizing resource utilization, given our hardware limitations.

The training protocol for our Visual Representation Contrastive Models was developed through systematic experimentation and hyperparameter optimization. Both model variants employed a 100-epoch training process with early stopping based on contrastive accuracy monitoring. We employed the AdamW optimizer with a weight decay of 0.0001, coupled with a cosine decay learning rate schedule incorporating a 20-epoch warm-up period. This configuration was empirically determined to provide optimal convergence while preventing overfitting.

3.1.1 *ResNet-50 Encoder Training.*

The ResNet-50 encoder architecture follows the standard implementation described in He et al. [16], generating 2048-dimensional embeddings. To gain an initial understanding of the quality of learned representations during contrastive training, we conducted an exploratory evaluation using dimensionality reduction and visualization.

Through Principal Component Analysis (PCA), we projected the high-dimensional embeddings onto a 2D plane, as shown in Figure 2. The visualization suggests that paired DCT representations, quantized with quality factors 10 (circles) and 50 (diamonds), respectively, tend to maintain proximity in the embedding space despite their differing compression levels, while clearly distancing from unrelated instances. While this pattern appears consistent with the contrastive learning objective of clustering similar instances, further analysis would be needed to fully validate these observations and characterize the quality and effectiveness of the learned representations.

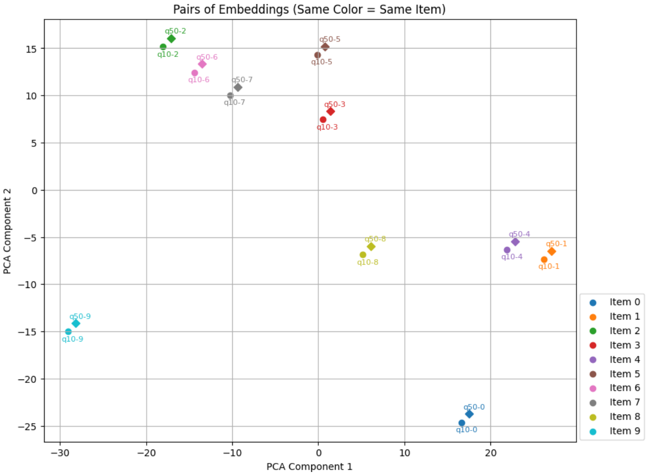

Figure 2: Visualization of validation embeddings generated with the ResNet-50 encoder. Circles are the embeddings from DCTs quantized with qf=10. Diamonds are the embeddings from DCTs quantized with qf=50.

3.1.2 ViT Encoder Training. Following the same training protocol described previously, we additionally trained a Vision Transformer (ViT) encoder that generates 384-dimensional embeddings. When applying the same PCA-based visualization approach, preliminary observation from Figure 3 suggests that the ViT embeddings may be learning some degree of the contrastive relationships between DCT pairs. The visualization reveals that representations of the same image quantized with different quality factors (10 and 50) tend to maintain closer proximity in the embedding space when compared to unrelated samples, which could indicate that the ViT encoder is capturing certain aspects of the semantic similarity between augmented views while distinguishing dissimilar instances. However, further analysis is needed to characterize the effectiveness of the embeddings and their full applications.

Beyond generating final embeddings, the ViT encoder offers additional interpretable representations through its attention mechanisms. As a transformer-based architecture, the ViT produces attention scores in each of its multi-head attention layers, which can be aggregated and visualized as attention maps. These maps indicate how strongly different regions of the input (in our case, DCT coefficient patches) contribute to the final representation at each layer.

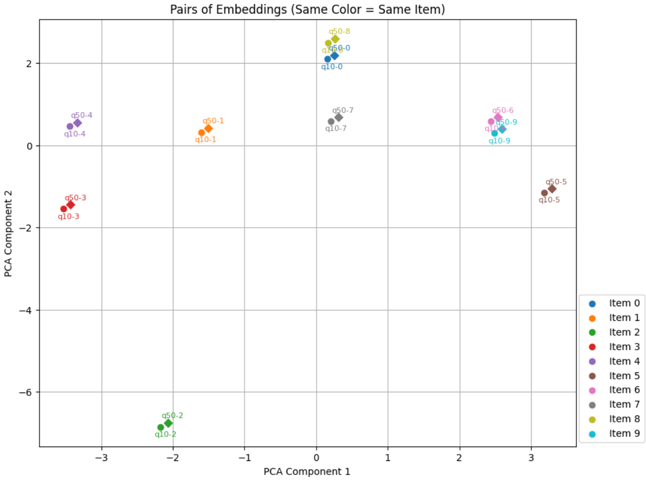

Figure 3: Visualization of validation embeddings generated with the ViT encoder. Circles are the embeddings from DCTs quantized with qf=10. Diamonds are the embeddings from DCTs quantized with qf=50.

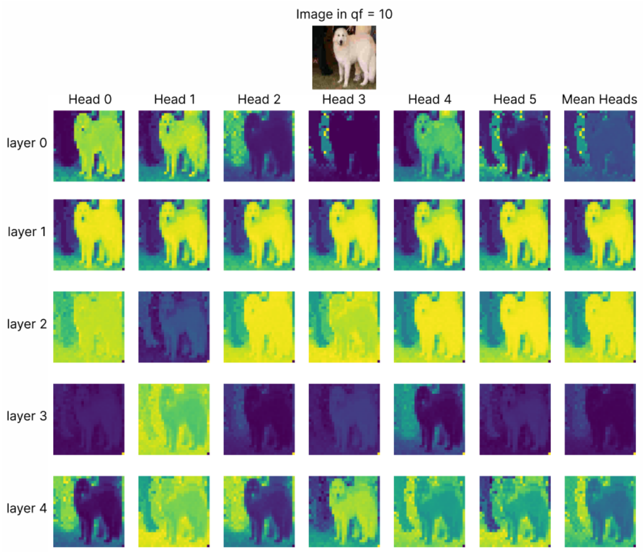

Figure 4: Attention scores visualization. The original image is quantized with $qf = 10$ and passed through the ViT encoder. The attention maps highlight regions of interest, demonstrating the potential model's ability to focus on meaningful features.

The visualizations of these attention maps, shown in Figure 4, suggest that some attention heads appear to focus on particular regions of the frequency spectrum, potentially corresponding to semantically meaningful features in the image. For instance, certain heads consistently attend to lower-frequency DCT coefficients (typically representing broader shapes and structures), while others show activation patterns around mid-frequency components (often corresponding to textures and edges). However, we emphasize that these observations require a more rigorous quantitative evaluation to fully characterize the relationship between attention patterns and semantic content.

4 DOWNSTREAM TASKS EXPERIMENTS

To assess the quality and practical utility of the learned representations, we conducted experiments using two complementary approaches: (1) the standardized linear evaluation protocol following Chen et al. [6], and (2) a practical JPEG artifact removal task. The linear evaluation protocol provides a standardized benchmark for assessing representation quality by training a simple linear classifier on frozen features, while the artifact removal task tests the representations' utility in a concrete application scenario. Detailed methodologies and results for both evaluation approaches are presented in the following sections.

4.1 Linear Classifier Evaluation

This approach measures the linear separability of the features by training a simple linear classifier on top of frozen encoder outputs, without updating the encoder weights. It provides a standardized method for comparing the discriminative power of different representations.

The evaluation process involves using both ResNet-50 and ViT encoders, pre-trained using the proposed DCT-based contrastive learning approach. During the linear evaluation phase, the encoder weights remained frozen, ensuring that classification performance directly reflected the quality of the learned embeddings.

We performed the evaluation using a subset of 200 categories from the ImageNet-1K validation set, with 10,000 images. This subset matched the classes used in the pre-training phase. To maintain consistency with the self-supervised training setup, we applied a similar set of data augmentations—random flipping, zooming, color jittering, and brightness changes—but with reduced intensity.

A linear classifier was then trained on top of the frozen representations. The classification head consisted of a single fully connected layer, followed by batch normalization and dropout for regularization. The training was performed using stochastic gradient descent (SGD) with an initial learning rate of 0.005, a cosine decay schedule, and a warmup period of 10 epochs. All models were trained from scratch, as we are using a random subset of the dataset.

4.1.1 Results for Linear Classifiers Evaluation. We conducted three evaluation setups to compare the effectiveness of different input modalities and to examine the contribution of the learned embeddings.

In the first setup, we trained a linear classifier using raw images compressed at quality factors (QFs) 50 and 10, establishing baseline results. We then evaluated the impact of combining these inputs with embeddings extracted from the ViT encoder.

As shown in Table 1, classification performance decreases with reduced image quality. Top-1 accuracy drops from 53.36% and 78.64% top-5 at QF 50 to 44.00% top-1 and 70.64% top-5 at QF 10. Combining ViT embeddings with image inputs did not lead to improvement in this setting. For QF 50, the hybrid input slightly underperformed the baseline, reaching 52.72% top-1 and 78.48% top-5 accuracy. A similar pattern is observed for QF 10, where the combination achieved 42.08% top-1 and 71.44% top-5 accuracy. These results suggest that the ViT embeddings, while informative, may not have provided additional discriminative power beyond the raw image features in this experimental setup. The observed differences across folds, as reflected by the standard deviations, are relatively small, indicating stable behavior of the models across data splits.

Table 1: K-fold (k=5) results for Linear Evaluation of ViT embeddings combined with images in the spatial domain.

data	QF	Encoder arch	top-1 acc (%) with s.d.	top-5 acc (%) with s.d.
Image	50	-	0.5336 ± 0.04	0.7864 ± 0.02
Image	10	-	0.4400 ± 0.03	0.7064 ± 0.03
Image + Embedding	50	ViT	0.5272 ± 0.03	0.7848 ± 0.01
Image + Embedding	10	ViT	0.4208 ± 0.02	0.7144 ± 0.02

The second setup, summarized in Table 2, uses only frequency domain data—specifically, quantized DCT coefficients with QFs 10 and 50—and evaluates both raw DCTs and DCTs combined with ViT embeddings. The accuracy is generally lower than with spatial images, with a top-1 accuracy of 30.64% at QF 50 and 23.06% at QF 10. Combining DCTs with embeddings results in a slight improvement. At QF 50, top-1 accuracy increases to 34.16%, and at QF 10, it rises to 36.88%. The top-5 accuracy also improves considerably, particularly at QF 10, reaching 62.88% compared to 43.99% without embeddings. This indicates that the embeddings contribute meaningful complementary information that supports class separability even in severely compressed conditions. Additionally, the standard deviations are within acceptable bounds, indicating consistent performance across folds.

Table 2: K-fold (k=5) results for Linear Evaluation of ViT embeddings combined with images' DCTs

data	QF	Encoder arch	top-1 acc (%) with s.d.	top-5 acc (%) with s.d.
DCT	50	-	0.3064 ± 0.03	0.6512 ± 0.17
DCT	10	-	0.2306 ± 0.02	0.4399 ± 0.03
DCT + Embedding	50	ViT	**0.3416 ± 0.04**	0.5952 ± 0.04
DCT + Embedding	10	ViT	**0.3688 ± 0.03**	**0.6288 ± 0.02**

The final setup, shown in Table 3, evaluates classifiers trained exclusively on embeddings. As anticipated, accuracy in this case is substantially lower due to the absence of visual input. Still, ViT consistently outperforms ResNet-50 across both QFs. For example, at QF 10, ViT achieves 12.40% top-1 and 22.40% top-5 accuracy, whereas ResNet-50 achieves 6.80% and 12.00%, respectively and the standard deviations remain within low ranges, indicating stability again. This may indicate that ViT-based representations retain more semantic information under compressed training conditions.

In summary, these experiments show that:

- Compression degrades classification accuracy, but learned embeddings help recover part of the lost performance;
- DCT-based inputs are viable but less discriminative than spatial images; however, embeddings can still provide a slight improvement under low-quality conditions;
- ViT encoders generally produce more informative embeddings than ResNet-50 in the tested settings.

Together, these findings are consistent with the hypothesis that DCT-based contrastive learning, particularly when paired with ViT

Table 3: k-fold (k=5) results for Linear Evaluation of ViT and ResNet-50 encoder embeddings

data	QF	Encoder arch	top-1 acc (%) with s.d.	top-5 acc (%) with s.d.
Embedding	50	ViT	0.0879 ± 0.02	0.1959 ± 0.02
Embedding	10	ViT	0.1240 ± 0.02	0.2240 ± 0.02
Embedding	50	ResNet-50	0.0480 ± 0.02	0.1400 ± 0.01
Embedding	10	ResNet-50	0.0680 ± 0.02	0.1200 ± 0.03

encoders, can support learning transferable representations under varying image quality conditions.

4.2 JPEG Artifact Removal Evaluation

The JPEG artifact removal task focuses on reducing or eliminating distortions introduced by lossy JPEG compression, which typically manifest as blocking artifacts, blurring, ringing effects, and quantization noise. The goal is to reconstruct a higher-quality image that closely resembles the original uncompressed version while preserving essential fine details and textures.

For this experiment, we randomly selected 3,100 images from the ImageNet validation set (guaranteeing no overlap with pretraining data) and divided them into 3,000 training images and 100 validation samples. Each image was processed as follows:

- Converted to DCT domain using $8x8$ blocks;
- Quantized with QF 10 as model input;
- Quantized with QF 50 as reconstruction target.

This setup challenges the model to recover higher-frequency components discarded during aggressive QF 10 compression, effectively learning an artifact-removal mapping between low- and high-quality DCT representations.

The frozen pre-trained ResNet-50 and ViT encoders, without any fine-tuning, were employed to extract the embeddings and attention maps. The ResNet-50 embeddings were extracted from the final global average pooling layer. For the ViT encoder, the embeddings were also extracted from the global average pooling layer following the transformer layers, and the attention maps were derived from the mean attention scores from the first transformer block.

Building on Busson et al. [3, 4], which demonstrated the efficacy of UNet architectures for DCT-based artifact removal, we use the UNet network as our baseline model. To enable the integration of the embeddings and/or the attention maps extracted from the attention scores with the UNet decoding process, we also modify the UNet architecture with two dedicated fusion mechanisms: (1) a feature concatenation for combining DCT coefficients with encoder embeddings, and (2) an attention gating module for incorporating spatial attention maps. This dual adaptation allows flexible experimentation with three input configurations: DCT coefficients alone, DCT-enhanced embeddings, or their combination with attention guidance.

In the first mechanism, we project the embedding to match the UNet bottleneck feature space. Then, the projected embedding is reshaped and tiled across spatial dimensions, aligning with the bottleneck feature map. The tiled embeddings are then concatenated channel-wise with the UNet features, creating a composite representation that merges global semantic information from the encoder with local frequency patterns from the DCT coefficients.

The attention maps integration follows a more dynamic approach. During decoding, we resize the attention maps to match the current feature resolution. Then, a 1x1 convolution with a sigmoid activation is used to process the attention maps, transforming them into spatial importance weights that highlight regions requiring special focus. Finally, the attention map is applied through element-wise multiplication to the decoder feature maps to focus on important regions.

Thus, we trained 5 distinct UNet versions to isolate the contribution of each component:

#1 **DCT Baseline Model**: Uses only the quantized DCT coefficients as input;
#2 **DCT + ResNet-50 Embedding**: Combines the quantized DCT coefficients with frozen ResNet-50 embeddings;
#3 **DCT + ViT Embedding**: Combines the quantized DCT coefficients with frozen ViT embeddings;
#4 **DCT + Attention map**: Combines the quantized DCT coefficients with the mean attention maps from the ViT encoder's first transformer layer;
#5 **DCT + Attention map + ViT Embedding**: Full integration of the quantized DCT coefficients, ViT embedding, and the mean attention maps from the ViT encoder's first transformer layer.

All models were trained for 1,000 epochs using the Adam optimizer ($lr = 10^{-4}$) with Mean Squared Error (MSE) loss, minimizing the difference between predicted and target DCT coefficients. To measure the performance of the model at artifact removal, we employed three standard image quality metrics:

- Peak Signal-to-Noise Ratio (PSNR): to provide a quantitative measure of how well the reconstructed image approximates the original image in terms of pixel-level accuracy;
- Structural Similarity Index (SSIM): to evaluate the perceived quality of an image by comparing luminance, contrast, and structure between the original and reconstructed images. And;
- Normalized Root Mean Square Error (NRMSE): to measure the normalized difference between the original and reconstructed images.

This multi-metric approach enables a comprehensive analysis of both quantitative reconstruction accuracy and perceptual artifact removal performance. To apply these metrics, we convert the model's outputs to the spatial domain that, as the inputs, are in the frequency domain.

4.2.1 Results for Artifact Removal Evaluation. The experimental results presented in Table 4 reveal the notable potential for the application of the learned representations with the proposed integration architectures for JPEG artifact removal.

The marginally superior performance of model #3 (DCT + ViT embedding), with SSIM: 0.64, and PSNR: 23.86 dB, compared to the baseline, with ΔSSIM +0.01, and ΔPSNR +0.06 dB, and model

Table 4: Results for a linear evaluation on the artifact removal task. SSIM and PSNR are expected to increase while NRMSE is expected to decrease.

#	Unet Version	SSIM↑	PSNR↑	NRMSE↓
1	DCT baseline	0.63	23.80	0.14
2	DCT + Resnet-50 embedding	0.62	23.76	0.14
3	DCT + ViT embedding	0.64	23.86	0.14
4	DCT + attention map	0.64	23.90	0.14
5	DCT + attention map + ViT embedding	0.63	23.74	0.15

#2 (DCT + ResNet-50 embedding) variant may indicate that transformer architectures could possess certain advantages for frequency-domain processing. This observation is consistent with the theoretical framework, suggesting that ViTs' patch-based self-attention mechanism might better preserve the block-wise relationships in DCT coefficients, though the relatively small performance differences require cautious interpretation.

The variant #4 (DCT + attention map) model's performance, with PSNR: 23.90 dB, appears to demonstrate the potential benefits of spatially adaptive processing for artifact removal. These results may support the hypothesis that dynamic feature modulation may help address localized compression artifacts, though the identical SSIM scores (0.64) across multiple variants suggest the perceptual improvements might be less pronounced than the pixel-level gains.

The reduced performance of the variant #5 (combined ViT attention map + ViT embedding model), with PSNR: 23.74 dB, and NRMSE: 0.15, could potentially indicate challenges in feature integration strategies. Some plausible explanations might account for this observation, such as:

- *Information Redundancy*: ViT embeddings and attention maps may encode overlapping frequency-space information;
- *Scale Conflict*: Global embeddings could interfere with local attention guidance;
- *Projection Effects*: The linear embedding projection may distort pre-trained feature relationships.

Moreover, the metric-specific patterns offer additional insights worth considering:

- The narrow SSIM range (0.62-0.64) may suggest that structural similarity is largely maintained across variants;
- The wider PSNR variation (23.74-23.90 dB) could indicate differential sensitivity to architectural modifications;
- The stable NRMSE values (0.14), except for the #5 variant (0.15), might reflect generally robust feature integration.

These findings suggest that the ViT embeddings may maintain the global frequency relationships, whilst the attention maps seem to mitigate local artifacts better due to their spatial awareness. The ResNet-50 embeddings appear compromised by the DCT block misalignment, suggesting that convolutional operations may disrupt block-wise frequency relationships, even though, theoretically, they should benefit from spatial invariance. These interpretations remain provisional pending further ablation studies, but are consistent with established principles of frequency-domain processing and the observed metric variations across architectures.

5 CONCLUSIONS

This work investigated the integration of the Discrete Cosine Transform (DCT) with contrastive learning to enable visual representation learning under varying image quality conditions. We proposed a frequency-domain adaptation of the SimCLR framework, using quantized DCT coefficients as input and training encoders based on ResNet-50 and Vision Transformer (ViT) architectures. This design aimed to exploit DCT's structural properties, particularly its ability to separate low- and high-frequency image components and simulate compression-induced distortions.

To assess the quality of the learned representations, we performed linear classification under different conditions. Results indicate that the classification accuracy degrades as image quality decreases, and that combining spatial inputs with learned embeddings does not improve linear classification performance. While DCT-only inputs performed worse than spatial images, their combination with embeddings led to measurable gains, especially in more degraded scenarios. ViT consistently outperformed ResNet-50 across conditions, reinforcing its ability to learn robust features from frequency-domain signals.

While the performance gap between frequency-domain and spatial-domain baselines suggests areas for further optimization, this work highlights the potential of compressed-domain learning and contributes to ongoing efforts toward efficient, flexible, and label-free visual understanding.

We further evaluated the learned representations in a JPEG artifact removal task using a UNet-based decoder. The best-performing configuration used ViT embeddings alongside DCT coefficients, achieving marginal improvements in PSNR and SSIM. The integration of ViT attention maps also contributed positively, although combining embeddings and attention maps showed diminishing returns, likely due to redundancy or integration conflicts. In contrast, combining both the embeddings and the attention maps did not improve performance, possibly due to redundancy or feature integration conflicts.

These findings demonstrate that frequency-domain contrastive learning is a viable strategy, particularly when paired with transformer-based architectures. Although performance gains were modest, the results support further exploration of compressed-domain learning as a complementary pathway to traditional spatial-domain approaches.

This study presents encouraging preliminary evidence but remains limited by computational resources and dataset scale. Broader evaluations, including additional quality factors, datasets, and downstream tasks such as segmentation or retrieval, are left for future work. Moreover, quantitative analyses of ViT attention patterns and systematic ablation studies could clarify the roles of frequency components and architecture design. Despite these constraints, the proposed DCT-based contrastive approach opens new possibilities for efficient compressed-domain learning and multimedia processing.

REFERENCES

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning Representations by Maximizing Mutual Information Across Views. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs.LG] https://arxiv.org/abs/2108.07258

[3] Antonio J.G. Busson, Paulo R.C. Mendes, Daniel de S. Moraes, Álvaro M. da Veiga, Álan L. V. Guedes, and Sérgio Colcher. 2020. Video Quality Enhancement Using Deep Learning-Based Prediction Models for Quantized DCT Coefficients in MPEG I-frames. In *2020 IEEE International Symposium on Multimedia (ISM)*. 29–32. https://doi.org/10.1109/ISM.2020.00012

[4] Antonio José G. Busson, Paulo Renato C. Mendes, Daniel de S. Moraes, Álvaro Mário G. da Veiga, Sérgio Colcher, and Álan Lívio V. Guedes. 2020. Decoder-Side Quality Enhancement of JPEG Images Using Deep Learning-Based Prediction Models for Quantized DCT Coefficients. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*. 129–136.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big Self-Supervised Models are Strong Semi-Supervised Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 22243–22255. https://proceedings.neurips.cc/paper_files/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf

[8] Xinlei Chen and Kaiming He. 2021. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15750–15758.

[9] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. 2023. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22367–22377.

[10] Xinlei Chen, Saining Xie, and Kaiming He. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9640–9649.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 21271–21284. https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf

[13] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. 2019. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 729–739.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[17] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A Survey on Contrastive Self-Supervised Learning. *Technologies* 9, 1 (2021). https://doi.org/10.3390/technologies9010002

[18] Jiaxi Jiang, Kai Zhang, and Radu Timofte. 2021. Towards Flexible Blind JPEG Artifacts Removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4997–5006.

[19] Joint Photographic Experts Group JPEG et al. 2004. JPEG standards: ISO/IEC IS 10918-1, ITU-T Recommendation T.81.

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

[21] Hua-Bao Ling, Bowen Zhu, Dong Huang, Ding-Hua Chen, Chang-Dong Wang, and Jian-Huang Lai. 2022. Vision Transformer for Contrastive Clustering. arXiv:2206.12925 [cs.CV] https://arxiv.org/abs/2206.12925

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=a68SUt6zFt Featured Certification.

[23] Mingyu Ouyang and Zhenzhong Chen. 2023. JPEG Quantized Coefficient Recovery via DCT Domain Spatial-Frequential Transformer. *arXiv preprint arXiv:2308.09110* (2023).

[24] William B Pennebaker and Joan L Mitchell. 1993. *JPEG: Still image data compression standard.* Van Nostrand Reinhold.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

[27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[29] K Simonyan and A Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society.

[30] Gregory K Wallace. 1992. The JPEG still picture compression standard. *IEEE transactions on consumer electronics* 38, 1 (1992), xviii–xxxiv.

[31] Guang Yang, Lu Lin, Chen Wu, and Feng Wang. 2023. Dual-Domain Learning for JPEG Artifacts Removal. In *International Conference on Neural Information Processing*. Springer, 556–568.