Improving Image Clustering with Artifacts Attenuation via Inference-Time Attention Engineering

Kazumoto Nakamura, Yuji Nozawa, Yu-Chieh Lin,
Kengo Nakata, and Youyang Ng

Kioxia Corporation
{kazumoto1.nakamura,yuji1.nozawa,yuchieh.lin,
kengo1.nakata,youyang.ng}@kioxia.com

Abstract. The goal of this paper is to improve the performance of pretrained Vision Transformer (ViT) models, particularly DINOv2, in image clustering task without requiring re-training or fine-tuning. As model size increases, high-norm artifacts anomaly appears in the patches of multi-head attention. We observe that this anomaly leads to reduced accuracy in zero-shot image clustering. These artifacts are characterized by disproportionately large values in the attention map compared to other patch tokens. To address these artifacts, we propose an approach called Inference-Time Attention Engineering (ITAE), which manipulates attention function during inference. Specifically, we identify the artifacts by investigating one of the Query-Key-Value (QKV) patches in the multi-head attention and attenuate their corresponding attention values inside the pretrained models. ITAE shows improved clustering accuracy on multiple datasets by exhibiting more expressive features in latent space. Our findings highlight the potential of ITAE as a practical solution for reducing artifacts in pretrained ViT models and improving model performance in clustering tasks without the need for re-training or fine-tuning.

Keywords: Attention Engineering · Artifact · Image Clustering

1 Introduction

The Transformer [44] architecture revolutionized the field of natural language processing (NLP), and its success inspired the development of the Vision Transformer (ViT) [17] for computer vision tasks. It has shown promising results across various downstream tasks [28, 37, 45, 48] such as classification and retrieval. Various improvements to ViT have also been proposed [30, 40, 50]. However, ViT is known to require large amount of training data for it to achieve superior performance. In other words, large-scale training of ViT architecture potentially unlocks and scales up the ability of ViT models. DINO (self-distillation with no labels) [4], a method of self-supervised learning that allows training without labels, offers a robust pretraining strategy to ViT that yields high recognition

accuracy when fine-tuned. DINOv2 [35] further extends the concept by pretraining the ViT model on a large scale visual dataset, offering significant improvements and robustness across various downstream tasks [8, 55]. The introduction of state-of-the-art pretrained large vision models such as DINOv2 has led to the emergence of "foundation models", which challenges the longstanding practice of training small vision models for specific tasks. This shift could potentially upend the traditional approach to model development and deployment in computer vision. One critical task closely related to training strategy is unsupervised image clustering. This task involves analyzing datasets with unknown categories and labels. The training strategy plays a vital role in optimizing these unknown elements by extracting rich features for clustering but training-based image clusterings can hinder their applicability in time-sensitive scenarios. Our goal is to fully utilize the potential of pretrained ViT models for image clustering without requiring re-training or fine-tuning.

However, it was found that high-norm artifacts anomaly appears in the patches of the attention block in the multi-head attention module of ViT model trained with DINOv2 [15], limiting its potential. These artifacts are characterized by disproportionately large values in the attention map compared to other patch tokens. Such artifacts have been reported to exist in models with a large number of parameters. As the size of models increases, the issue of artifacts in ViT-based models becomes increasingly significant. In this study, we investigate these artifacts by paying our attention to the raw features in latent space generated by these ViT models pretrained by DINOv2. We consider the task of zero-shot image clustering, where unsupervised clustering algorithms such as K-Means are applied directly to the output features of a pretrained vision model. Unlike other tasks, zero-shot image clustering directly utilizes the features generated in latent-space, enabling us to directly observe the impact of artifacts on the resulting accuracy. In this paper, we define "zero-shot image clustering" as an unsupervised image clustering task by using pretrained models without in-domain training for a particular dataset. Note that a vision model pretrained on a large number of images is inevitably seeing similar images with the test datasets during its pretraining stage. Hence, this task is not a completely out-of-distribution clustering task and the definition of zero-shot in this case is limited to the aforementioned definition of clustering with a pretrained model.

In our study, we first identify the artifacts of the model by computing the L_2 norms of one of the Query-Key-Value (QKV) patches within the attention block of the final layer. Through this process, we found that artifacts exist even in the smaller models, not limited to larger models as previously thought when identifying artifacts through the norms of output tokens [15]. We infer that by introducing measures to attenuate these artifacts, richer features could be generated across various model sizes. In prior study [15], the role of artifacts was successfully moved from patch tokens to additional tokens called register tokens in the training process. This measure improved the performance of downstream tasks at the cost of model re-training. In this study, we aim to improve the image clustering performance of the pretrained model in a more efficient way.

We formulate an approach to attenuate these artifacts that does not involve any re-training step.

To address these artifacts, in this paper, we propose an approach called Inference-Time Attention Engineering (ITAE), which manipulates attention function during inference. Specifically, we identify artifacts in the attention block of the model's final transformer layer by using QKV patches and attenuate the values of their corresponding attentions (Fig. 1) during inference time. We further investigate the impact of artifacts anomaly on the zero-shot image clustering task, and employ our proposed approach in models where the artifacts are identified. We show that our proposed method improves clustering accuracy on multiple datasets. In addition, we show that our method is also effective and acting as a complimentary technique to models employing registers. This improvement is due to the more expressive nature of the model with ITAE. Our findings highlight the potential of ITAE as a practical solution for reducing artifacts in pretrained Vision Transformer models and improving model performance in clustering tasks without the need for re-training or fine-tuning.

Overall, our contributions are summarized as follow.

(1) We identify the artifacts anomaly in DINOv2-based ViT models by computing the L_2 norms of the QKV patches in the attention block of the multi-head attention module and evaluate baseline performance on zero-shot image clustering tasks.
(2) We formulate ITAE, a method to manipulate attention function during inference of ViT models to attenuate artifacts.
(3) We perform empirical study and show that our proposed method improved the zero-shot clustering accuracy on multiple datasets by eliminating the negative impact of artifacts on the clustering task for models with artifacts identified.

2 Related works

2.1 Pretrained Large Vision Models and their Potential

The development of pretrained vision models can be dated back particularly to a series of self-supervised learning strategy, especially those contrastive learning methods [6, 7, 10, 11, 18, 19]. These methods did not pretrain their models with ViT architecture, thus limiting their potential. However, since the emergence of ViT, there has been a growing interest in applying self-supervised and weakly-supervised learning to train ViT-based models, including DINO, CLIP [36], ALIGN [22] and MoCov3 [12]. Unlike supervised learning, self-supervised learning offers the advantage of training with large amounts of unlabeled or weakly labeled data without the need for costly annotations. Notably, DINOv2 is a model pretrained on a substantial amount of data using DINO's approach. CLIP is a model pretrained on large amount of vision and language data pairs. Their robustness allows for easy adaptation to diverse downstream tasks. With only minimal fine-tuning and task-specific heads, these general pretrained models can

be transformed to tackle specific tasks. However, some form of parametric learning is still necessary in this scenario. Here, we focus on a specific use case of pretrained vision models, where no fine-tuning whatsoever is required and the pretrained models are employed as-is [34]. This concept is already widely seen in the field of NLP, where large language models (LLM) are used as-is in various NLP tasks through the introduction of prompt-engineering during inference time [2, 24, 46]. In prompt-engineering, only the input tokens are modified but not the LLM itself. For computer vision tasks such as image retrieval and k-NN, it is already possible to directly utilize the output features of pretrained model in calculating distance in latent space between data samples [34]. We focus our study on another task that similarly able to utilize the output features of a pretrained model as-is but less studied in this way: unsupervised image clustering.

2.2 Artifacts Anomaly in Pretrained ViT and Attention Engineering

DINOv2 is a pretrained model that achieved state-of-the-art performance in various computer vision tasks. However, it has been noted that artifacts [15] in the form of patches with abnormally large norms can exist in ViT models such as DINOv2, DeiT III [42], and OpenCLIP [39]. Previous study [15] addressed these artifacts by incorporating register tokens during the model training stage. In contrast, our approach applies inference-time attention engineering to the pretrained models. This allows us to mitigate the artifact issue without requiring model re-training. Several works on manipulating self-attention have been reported [9, 27, 50, 53]. However, these manipulations occurred during the learning phase, whereas our work focuses solely on the implementation during inference time. The most relevant works to our study are SATA [9] and LSA [27], which attempt to filter out unwanted attentions in ViT. However, these methods involve a learning phase and do not address the specific challenges of pretrained models and image clustering tasks. SATA used direct observation to suppress attention weights after the softmax function, while LSA introduced diagonal masking of attention and leveraged a temperature parameter to sharpen the attention values. In contrast, our approach attenuates the attention of artifact patches based on the observed artifacts in the QKV patches.

2.3 Deep Image Clustering and Application of Pretrained Models

For the image clustering task, there are two primary approaches to consider. The first approach [3, 5, 16, 21, 23, 32, 43, 49, 51, 52] focuses on developing learning methods that jointly optimize both image feature and clustering end-to-end. The second approach [20,41,47] is a two-stage method that involves feature extraction suitable for clustering, follow by applying clustering algorithms like K-Means for the actual clustering process. In the first approach, DAC [5] formulates clustering as a binary pairwise-classification problem while DeepCluster [3] repeatedly assigns pseudo-labels to learn the neural network. In the second approach, ID [47] formulates feature learning as a non-parametric instance discrimination

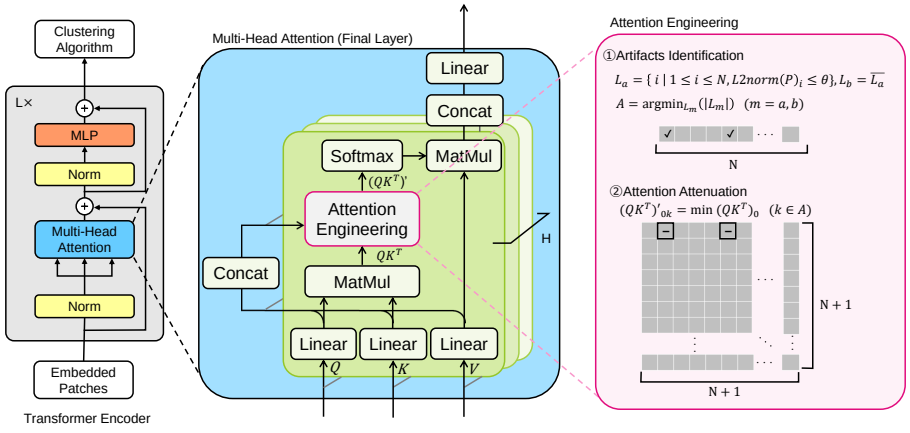


Fig. 1: Overall introduction of ITAE approach: We manipulate self-attention function of a pretrained vision model during inference. Specifically, we identify artifact in the final layer of the model's attention block during inference time by computing the L_2 norms of the QKV patches, and attenuate the values of their corresponding attentions. $N + 1$ represents the length of the QK^T matrix, including the CLS token.

problem. IDFD [41] proposes to perform instance discrimination and feature decorrelation simultaneously. In this work, we opt for the second approach that involves a two-stage setting, which affords better interpretability. Prior studies have employed in-domain training for each dataset, but this limitation can hinder their applicability in time-sensitive scenarios. In contrast, our proposed method leverages off-the-shelf pretrained vision models, offering a promising solution for image clustering tasks without the need for domain-specific training. However, the use of ViT-based pretrained models for image clustering has garnered limited attention, with only a handful of papers [1,13,31,54] being published in this area. These works focused on improving the clustering phase of the algorithms while our method focuses on improving the feature extraction phase without fine-tuning. In this paper, we seek to harness the untapped potential of pretrained models through inference-time engineering techniques. Our work aims to illuminate the underexplored territory of image clustering using state-of-the-art pretrained large vision models.

3 Approach

In this section, we first revisit the mechanism of attention block in the multi-head attention module of transformer architecture. We then describe the approach we use to identify artifacts anomaly. Next, we explain the details of our proposed method of Inference-Time Attention Engineering (ITAE). In this paper, we define "Attention Engineering" as an attention manipulation technique that modifies the attention value of patches in attention block. The overall picture of our approach is shown in Fig. 1. Finally, we describe the details of clustering

algorithm (K-Means) that we apply in this study but note that the investigation of clustering algorithm itself is out of the scope of this paper.

3.1 Self-Attention Block in Vision Transformer

In ViT, attention in the Self-Attention (SA) block is calculated from the three matrices query, key, value as follows [17].

$$\mathrm{SA}(Q, K, V) = \mathrm{Attention}(Q, K)V = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q denotes the query matrix, K denotes the key matrix, and V denotes the value matrix. Also, d_k denotes the number of dimensions of the key.

3.2 Identifying the Artifacts in Self-Attention Block

Previous study [15] identified artifacts based on the feature norms of the output tokens of the model. However, we argue that artifacts are better analyzed within the self-attention block. We utilize the L_2 norms distribution of the QKV patches, P within the self-attention block across all heads in the final layer. Using QKV is also computationally efficient as no back pass loop from the output is needed. In our study, we select query among the QKV patches to identify artifacts as we observed that norms distribution of query patches shows clearer bimodality between normal patches and artifacts. Note that it is also possible to use key or value patches to identify artifacts as they differ only by a linear projection. From this point onwards, we will use the term QKV to refer to query patches. In our work, the L_2 norms distribution is collected across a dataset. We divide the L_2 norms into two groups $L_a, L_b \in \{1, 2, ..., N\}$ with a threshold θ and identify those patches in the minority group as artifacts, A, where

$$L_a = \{i \,|1 \leq i \leq N,\ L2norm(P)_i \leq \theta\},\ L_b = \overline{L_a}, \tag{2}$$

$$A = \underset{L_m}{\mathrm{argmin}}(|L_m|) \quad (m = a, b), \tag{3}$$

and where

$$L2norm(P)_i = \left(\sum_{1\leq h\leq H}\sum_{1\leq j\leq d_p}(\frac{P_{ij}^h}{\sqrt{d_p}})^2\right)^{\frac{1}{2}}. \tag{4}$$

Here, H represents the number of heads in the multi-head attention module, N represents the number of patches and d_p denotes the number of dimensions of P. Through our observation, the optimal value of θ is roughly dataset-agnostic and model-dependent. This observation is consistent with the previous study [15] that the characteristic of artifacts depends on the scale of the models. Hence, we only need to pre-determine the value of θ. θ is not modified for different datasets in our study. Our approach of analyzing artifacts via the QKV patches in self-attention block allows us to identify artifacts present in smaller ViT models

that are difficult to recognize in the feature norms of the output tokens of the model. Specific analysis of L_2 norms of QKV patches will be discussed in detail in Sec. 4.2.

3.3 Inference-Time Attention Engineering

After identifying the artifacts in self-attention block, we apply attention engineering by manipulating attention function during inference. Specifically, we attenuate the values of the identified artifact patches in the self-attention block of the model's final transformer layer. For every identified artifact, k in A, the attention value of the artifact is transformed into a minimum value independently for every head in the multi-head attention module by the equation

$$(QK^T)'_{ik} = \min_{1 \le j \le N} \left((QK^T)_{ij} \right) \quad (k \in A)\,, \tag{5}$$

where i and j have the same dimension. However, since we only focus on the final layer of the model and we use only the output CLS token when applying clustering algorithm, only the row 0 that corresponds to the CLS token needs to be calculated in our proposed approach. Consequently, we apply the following part of attention function in our study:

$$(QK^T)'_{0k} = \min_{1 \le j \le N} \left((QK^T)_{0j} \right) \quad (k \in A)\,. \tag{6}$$

The full self-attention function can then be calculated through Eq. (1) with component from Eq. (6). In our experiment, we also apply the same function in Eq. (6) to models pretrained with register tokens. The pretrained models with register tokens we utilize in this study have a fixed number of register tokens, which is 4. As mentioned above, we extract the output features of CLS tokens in latent space for the subsequent clustering tasks. Our focus on applying the proposed method to the final layer of the model is justified by the simplicity of its modification, and the observability of the effect, where the output features are directly applied to clustering algorithm.

3.4 Clustering Algorithm

Finally, we discuss the clustering algorithm that we utilize in our study. Note that in this work, our focus is on the improvement of feature extractor and the latent features it produces towards the performance of image clustering through ITAE. Tuning of clustering algorithm is out of the scope of this study. Since we do not modify the format of the latent features, any feasible clustering algorithm could be applied with our approach. Classical clustering algorithms like K-Means, spectral clustering and agglomerative clustering are among the candidates. Modern learnable dense algorithms are also applicable. In our work, we apply K-Means, one of the simplest and robust clustering algorithms available, and show in the subsequent section that our proposed method is effective in creating a richer feature space for image clustering task.

4 Experiments

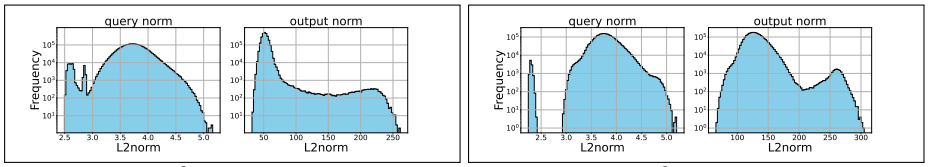
4.1 Experimental Settings

In this study, we evaluate the performance of the zero-shot image clustering task on four common clustering datasets: Tiny ImageNet [26], CIFAR-100 [25], CIFAR-10 [25] and STL-10 [14]. Tiny ImageNet contains image data of 200 classes extracted from the ImageNet dataset [38] and downsized to 64×64 colored images. Each class has 500 training images, 50 validation images, and 50 test images. The validation images were used for the experiments. CIFAR-10 is a dataset consisting of 10 classes (5,000 training images and 1,000 test images for each class) with an image size of 32×32 pixels. For the experiments, all 10,000 test images were used. CIFAR-100 is a dataset consisting of 100 classes (500 training images and 100 test images for each class) with an image size of 32×32 pixels. All 10,000 test images were used for the experiments. STL-10 is also a dataset derived from ImageNet dataset, with image size 96×96 pixels. It consists of 10 classes of data, 100,000 unlabeled data, 500 training images for each class, and 800 test images for each class. The test image set of 8,000 images was used for the experiments. For all datasets, the size of the images was standardized by resizing them to 224×224 pixels before feeding them into the pretrained model to perform the evaluation. This preprocessing step ensured uniformity of the input dimensions for all datasets.

For ViT models, we use the publicly available models pretrained by DINOv2 in our experiments (`https://github.com/facebookresearch/dinov2/tree/main`). Specifically, we use the small (ViT-S/14 distilled), base (ViT-B/14 distilled), large (ViT-L/14 distilled) and giant (ViT-g/14) variation of ViT models pretrained on DINOv2 for our experiments with models without register tokens. Similarly, we use the same small, base, large and giant variation of ViT models but with register tokens pretrained on DINOv2 for our experiments with models with register tokens. Note that the small, base and large models are distilled from the giant models.

For the clustering algorithm, we use the standard K-Means algorithm without any prior parameters fine-tuning. We extracted the feature representation of each image by normalizing the CLS token output from the final layer of the model. As the initialization of K-Means can impact its results, for each experimental setting, we conducted 20 sets of 25 clustering runs (a total of 500 runs) to account for this initial value dependence. We calculate the mean and standard error of accuracy across these multiple clustering sets and report their values in this paper. Note that we do not fine-tune our models for clustering hence we report zero-shot image clustering results. As stated in Sec. 1, we define "zero-shot image clustering" as an unsupervised image clustering task by using pretrained models without in-domain training for a particular dataset.

For quantitative comparison, we compare the clustering accuracy of our proposed method with the original DINOv2 without applying ITAE. DINOv2 by itself achieved state-of-the-art performance across various downstream tasks. Similarly, for image clustering with pretrained model, DINOv2 is one of the



without register with register

Fig. 2: The left plots show the L_2 norms distribution for model pretrained without register tokens while the right plots show the L_2 norms distribution including register tokens for model pretrained with register tokens. *Query norm* graphs show the L_2 norms distribution of QKV patches excluding CLS token. *Output norm* graphs show the L_2 norms distribution of the final output. Clear bimodality is observed in the L_2 norms distribution of QKV patches (model: ViT-B/14 distilled, dataset: CIFAR-100).

most competitive models available today and this model presents us a strong baseline for comparison. We also compare our method to previous work where the model is pretrained with register tokens [15]. We use standard clustering accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI) as the metrics to measure clustering performance.

4.2 Artifact Identification

A distribution analysis of the L_2 norms of the QKV patches in self-attention block and the L_2 norms of the output of the final layer of the Base model is presented in Fig. 2. Previous study has identified anomalous tokens based on the L_2 norms distribution of the final output. For the model pretrained without register tokens, the L_2 norms distribution of the patch tokens in the final output does not exhibit clear bimodality. Interestingly, the L_2 norms distribution of the QKV patches does show bimodality, suggesting that the presence of artifacts can be determined using a threshold value in the L_2 norms distribution of the QKV patches. On the other hand, for the model pretrained with the register tokens, the L_2 norms distribution of the patch tokens alone does not display bimodality [15]. By including both the patch tokens and the register tokens in the distribution analysis, it is confirmed that a portion of the register tokens functions as an artifact, as depicted in Fig. 2. Additionally, a comparison between models pretrained with and without register tokens reveals that the model incorporating the register tokens exhibits more pronounced bimodality. Through visually inspecting the bimodality in the L_2 norms distribution of the QKV patches, we can easily set the value of θ for each model and identify the minority group as artifacts. Note that the value of θ is predicted to be roughly dataset-agnostic so we fix the value of θ throughout our experiments unless stated otherwise. Specifically, We set θ to 2.0 for small model pretrained with register tokens through visual inspection. For other models pretrained with and without register tokens, we set θ to a common value of 3.0.

Table 1: Image clustering results. *original* denotes the original DINOv2 model while *registers* denotes DINOv2 model pretrained with register tokens as implemented in previous work [15]. Our proposed method outperformed the baseline in all cases and previous work in most cases (model: ViT-B/14 distilled).

Method	Metric	CIFAR-10	CIFAR-100	STL-10	Tiny ImageNet
original	ACC	83.63 ± 1.13	64.26 ± 0.30	75.65 ± 1.04	67.81 ± 0.24
	ARI	77.75 ± 1.08	48.94 ± 0.18	61.49 ± 1.57	50.68 ± 0.25
	NMI	86.04 ± 0.46	76.58 ± 0.09	81.61 ± 0.72	81.28 ± 0.06
registers	ACC	82.12 ± 1.35	$\mathbf{66.79 \pm 0.23}$	72.70 ± 1.13	$\mathbf{68.88 \pm 0.21}$
	ARI	73.50 ± 1.25	50.30 ± 0.21	56.92 ± 1.37	50.13 ± 0.23
	NMI	84.70 ± 0.41	$\mathbf{78.42 \pm 0.08}$	79.47 ± 0.54	81.69 ± 0.05
ours	ACC	$\mathbf{84.49 \pm 1.19}$	65.02 ± 0.14	$\mathbf{82.76 \pm 1.27}$	68.23 ± 0.25
	ARI	$\mathbf{79.46 \pm 1.14}$	$\mathbf{50.53 \pm 0.18}$	$\mathbf{75.94 \pm 1.60}$	$\mathbf{52.27 \pm 0.22}$
	NMI	$\mathbf{86.82 \pm 0.51}$	77.10 ± 0.08	$\mathbf{88.18 \pm 0.62}$	$\mathbf{81.78 \pm 0.07}$

4.3 Image Clustering Result

The results of the proposed method, previous work and the baseline are presented in Tab. 1. Notably, by applying our method, consistent improvements in performance are observed. The most substantial enhancement in accuracy is observed in the STL-10 dataset, with a remarkable increase of 7.11 for ACC, 14.46 for ARI, and 6.57 for NMI. Overall, an average accuracy improvement of 2.24 was observed for our method compared to baseline. We also observe that accuracy of baseline is lower than previous work in 2 datasets, indicating the negative impact brought by the artifacts. However, interestingly, model pretrained with register tokens are not always performing better when comparing to the original baseline model. In fact, the accuracy on the datasets is reduced by -0.27 in average. On the other hand, our method, without any re-training, outperformed previous work in most cases. We speculate that models pretrained with register tokens managed to reduce the negative effect brought by the artifacts during pretraining but sacrificed global optimization of the model to a certain extent.

4.4 Analysis with Different Model Sizes

Table 2 summarizes the findings regarding the relationship between model size and accuracy. It is observed that the optimal model size for achieving the highest accuracy varies depending on the dataset. Specifically, for the Tiny ImageNet dataset, accuracy consistently improves as the model size increases. Conversely, for the STL-10 dataset, accuracy tends to decrease as the model size increases. These results show that the optimal model size for image clustering is dataset-dependent. In addition to exploring the impact of model size, we also evaluated the effectiveness of our method on both original models and models pretrained

Table 2: Image clustering result across various model sizes (*small*: ViT-S/14 distilled, *base*: ViT-B/14 distilled, *large*: ViT-L/14 distilled, *giant*: ViT-g/14) reported in ACC. For model size *small*, we did not apply our method due to the lack of bimodality confirmation. Our proposed method outperformed the strong baseline in all cases across model sizes. Interestingly, models pretrained with register tokens are not always performing better when comparing to the original models.

Dataset	Model Size	original	registers	ours	registers + ours
CIFAR-10	small	70.92 ± 1.54	77.11 ± 1.60	-	$\mathbf{81.57 \pm 1.04}$
	base	83.63 ± 1.13	82.12 ± 1.35	$\mathbf{84.49 \pm 1.19}$	82.61 ± 1.40
	large	82.16 ± 1.48	78.67 ± 1.50	$\mathbf{82.49 \pm 1.55}$	79.92 ± 1.47
	giant	78.09 ± 1.25	76.38 ± 1.44	$\mathbf{78.59 \pm 1.91}$	77.42 ± 1.69
CIFAR-100	small	50.84 ± 0.24	57.33 ± 0.20	-	$\mathbf{60.80 \pm 0.27}$
	base	64.26 ± 0.30	66.79 ± 0.23	65.02 ± 0.14	$\mathbf{68.85 \pm 0.28}$
	large	68.69 ± 0.34	68.01 ± 0.38	$\mathbf{69.04 \pm 0.22}$	68.98 ± 0.31
	giant	68.99 ± 0.39	69.22 ± 0.27	69.50 ± 0.28	$\mathbf{69.94 \pm 0.25}$
STL-10	small	83.33 ± 1.69	85.20 ± 2.04	-	$\mathbf{85.89 \pm 1.45}$
	base	75.65 ± 1.04	72.70 ± 1.13	$\mathbf{82.76 \pm 1.27}$	78.71 ± 1.14
	large	65.78 ± 1.22	56.84 ± 1.21	$\mathbf{70.51 \pm 1.42}$	59.62 ± 1.49
	giant	55.91 ± 1.14	53.73 ± 1.75	$\mathbf{56.01 \pm 0.93}$	55.06 ± 1.41
Tiny ImageNet	small	55.49 ± 0.19	57.78 ± 0.15	-	$\mathbf{59.56 \pm 0.14}$
	base	67.81 ± 0.24	68.88 ± 0.21	68.23 ± 0.25	$\mathbf{69.62 \pm 0.18}$
	large	71.98 ± 0.15	71.53 ± 0.19	$\mathbf{73.19 \pm 0.21}$	72.53 ± 0.17
	giant	73.25 ± 0.16	73.44 ± 0.17	73.54 ± 0.17	$\mathbf{73.80 \pm 0.20}$

with register tokens. In this case, we apply our method to models pretrained with register tokens, effectively attenuating those attentions absorbed by register tokens. We show this result in the "registers + ours" column of Tab. 2. Our experiments demonstrated similar improvement in accuracy for both types of models. Specifically, with the exception of the ViT-S/14 distilled, our method improved the accuracy by an average of 1.43 for the original model and 1.56 for the model pretrained with register tokens. This indicates that our method is robust and can enhance the clustering performance regardless of the presence or absence of register components. When observing across model sizes, similar to results in Tab. 1, models pretrained with register tokens are not always performing better when comparing to models pretrained without register tokens.

5 Discussion

5.1 Attention Values and Maps Visualization

A histogram analysis of the attention values associated with the identified artifacts in our proposed method is shown in Fig. 3. These attention values are collected from the $\text{Attention}(Q, K)_{0j}$ for $1 \leq j \leq N$ in Eq. (1). Yellow colored distribution shows the artifact patches identified while blue colored distribution shows normal patches. The distribution shows that the identified artifact's attention values are substantially larger than normal. This observation suggests

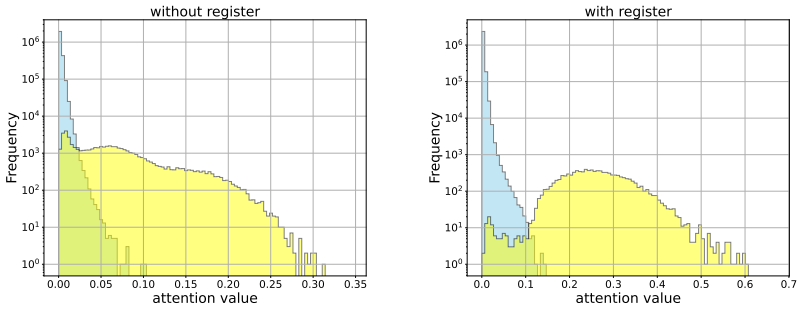


Fig. 3: Histogram of attention values: Yellow colored distribution shows the artifact patches identified while blue colored distribution shows normal patches. The left and right plots are for model pretrained without register tokens and model pretrained with register tokens, respectively (model: ViT-B/14 distilled, dataset: CIFAR-100). Best viewed in color.

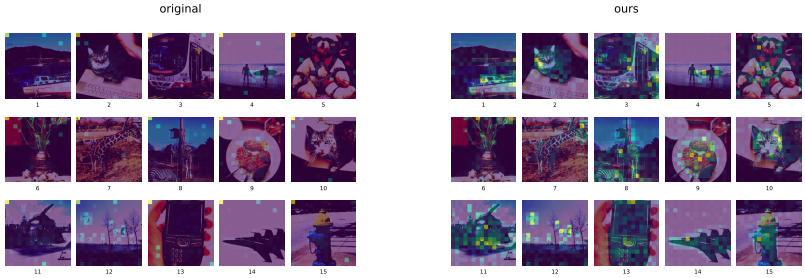


Fig. 4: Attention map: The left maps are the attention maps of original model and the right maps are the attention maps of the model incorporating our proposed method. The attention map is the averaged map across each attention head in the multi-head attention module. Best viewed in color. Details and licenses for the images are provided in supplementary material (model: ViT-B/14 distilled, dataset: MS COCO).

that the artifact's attention values contribute significantly to the output features, effectively masking the function of normal token features. By attenuating these identified artifacts during inference time, our proposed method enhances the model's expressiveness by redistributing attentions from the artifact to other tokens. In addition, the impact of our proposed method on the attention map is shown in Fig. 4. The left maps are the attention maps of original model and the right maps are the attention maps of the model incorporating our proposed method. We show images from the dataset of MS COCO [29] for these visualization due to its clear license info. We observe that our proposed method succeeds in expanding the attentions that were previously concentrated in certain areas of the attention map without any re-training or fine-tuning. These richer attentions

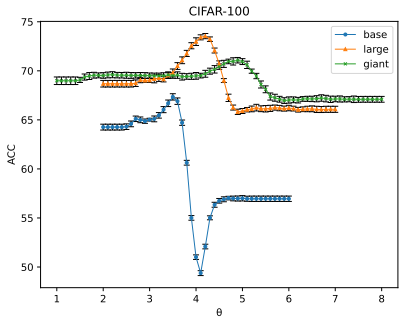


Fig. 5: θ-accuracy graph: Change in clustering accuracy (ACC) when θ is varied from 1.0 to 8.0 for each model (base: ViT-B/14 distilled, large: ViT-L/14 distilled, giant: ViT-g/14, dataset: CIFAR-100). Best viewed in color.

are beneficial to tasks that directly employing output feature representations of the model, particular the zero-shot image clustering task studied in this paper.

5.2 Ablation Study of Threshold Values for Identifying Artifacts

In this session, we present an ablation study to investigate the impact of the threshold θ on clustering accuracy. To conduct this study, we adjust the threshold within the range of QKV patches. We remove Eq. (3) and treat L_a as artifacts A directly to perform a parameter scan. The results are illustrated in Fig. 5, which displays the change in clustering accuracy corresponding to different threshold values for each model. In the following discussion, we will mainly consider the base model but note that similar observations can be made for other models.

Initially, at the lowest threshold, the clustering accuracy is similar to that of the original model. This is because tokens are not ignored at this stage. As we gradually increase the threshold, specifically around $\theta = 2.5$, the accuracy starts to improve. This improvement can be attributed to the exclusion of artifacts, which are now taken into consideration for accuracy calculation. Continuing our analysis, we observed that the accuracy continues to increase as the threshold reaches approximately $\theta = 3.5$. This is due to the removal of artifacts from the clustering process. However, beyond this threshold, we noticed a decline in accuracy. We hypothesize that this decline is a result of the attenuation of tokens required for clustering and the increased presence of non-artifact tokens that are irrelevant for clustering. As the threshold value θ increases, so does the value of $\text{Attention}(Q, K)_{00}$. This leads to the attention mechanism utilizing the CLS tokens as they are. In the base model, around the threshold of $\theta = 4.5$, most tokens other than the CLS tokens are attenuated, resulting in convergence of accuracy. At this point, there is a significant improvement in accuracy compared to the lowest accuracy value in the base model. However, it is worth noting that the degree of improvement varies across different models.

Table 3: k-NN classification results across various model sizes (*base*: ViT-B/14 distilled, *large*: ViT-L/14 distilled, *giant*: ViT-g/14) reported in accuracy.

Dataset	Model Size	original	registers	ours	registers + ours
ImageNet-1k	base	82.04	82.02	82.07	**82.35**
	large	83.50	83.84	83.62	**83.87**
	giant	83.51	83.65	83.54	**83.72**
CIFAR-100	base	87.31	87.60	87.58	**88.09**
	large	91.12	90.88	**91.39**	91.01
	giant	91.79	91.59	**91.99**	91.81

An intriguing finding from our study is that the accuracy peaks in the range greater than $\theta = 3$, where artifacts have already been ignored. This suggests that, in addition to artifacts, there are tokens that were previously considered normal but are not necessary for clustering. By appropriately processing these tokens, it is possible to further enhance the clustering accuracy.

5.3 Generalization of the Proposed Method to Other Tasks

To evaluate the tasks generalizability of our proposed method, we conduct experiments with k-NN classification task [33]. In our experiment, we follow standard k-NN implementation and evaluation metric in Ref. [35] and set the value of k to 10. For datasets, since large scale datasets are applicable to k-NN classification, we use the full set of ImageNet-1k [38], as well as CIFAR-100. The results are shown in Tab. 3. It is observed that our method consistently improves over original model by a slight margin for different model sizes. Previous work of *registers* [15] managed to improve the accuracy by a bigger margin for *large* & *giant (ImageNet-1k)*, and *base (CIFAR-100)* but performance degradation occurred for others. The complementary aspect of *registers* and our method is extensively visible here as *registers + ours* achieves highest accuracy in more categories. We report other ablation studies in the supplementary material.

6 Conclusion

In conclusion, our study has first re-identified the presence of artifacts in models of smaller size which were previously believed to be artifact-free. We then successfully improved the zero-shot image clustering accuracy by addressing the negative impact of these artifacts. Specifically, we achieved this by introducing ITAE, which manipulates attention function during inference. Our findings highlight the potential of ITAE as a practical solution for reducing artifacts in pretrained Vision Transformer models and improving model performance in clustering tasks without the need for re-training or fine-tuning. Potential future works include optimizing attention allocation to further improve model performance.

References

1. Adaloglou, N., Michels, F., Kalisch, H., Kollmann, M.: Exploring the limits of deep image clustering using pretrained models. arXiv preprint arXiv:2303.17896 (2023)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (2020)
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
4. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9630–9640. IEEE Computer Society (2021)
5. Chang, J., Wang, L., Meng, G., Xiang, S., Pan, C.: Deep adaptive image clustering. In: Proceedings of the IEEE international conference on computer vision. pp. 5879–5887 (2017)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)
8. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6593–6602 (June 2024)
9. Chen, X., Hu, Q., Li, K., Zhong, C., Wang, G.: Accumulated trivial attention matters in vision transformers on small datasets. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3973–3981. IEEE (2023)
10. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
11. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
12. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9620–9629. IEEE (2021)
13. Chu, T., Tong, S., Ding, T., Dai, X., Haeffele, B.D., Vidal, R., Ma, Y.: Image clustering via the principle of rate reduction in the age of pretrained models. arXiv preprint arXiv:2306.05272 (2023)
14. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223. JMLR Workshop and Conference Proceedings (2011)
15. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024)

16. Dilokthanakul, N., Mediano, P.A., Garnelo, M., Lee, M.C., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648 (2016)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9726–9735. IEEE Computer Society (2020)
20. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
21. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9865–9874 (2019)
22. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
23. Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H.: Variational deep embedding: An unsupervised and generative approach to clustering. arXiv preprint arXiv:1611.05148 (2016)
24. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (2024)
25. Krizhevsky, A., Hinton., G.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
26. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
27. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021)
28. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision. pp. 280–296. Springer (2022)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
31. Lowe, S., Haurum, J., Oore, S., Moeslund, T., Taylor, G.: Zero-shot clustering of embeddings with pretrained and self-supervised learnt encoders. In: Workshop on robustness of zero/few-shot learning in foundation models (NeurIPS 2023) (2023)

32. Mukherjee, S., Asnani, H., Lin, E., Kannan, S.: Clustergan: Latent space clustering in generative adversarial networks. In: Proceedings of the AAAI conference on artificial intelligence. pp. 4610–4617 (2019)
33. Nakata, K., Ng, Y., Miyashita, D., Maki, A., Lin, Y.C., Deguchi, J.: Revisiting a knn-based image classification system with high-capacity storage. In: European Conference on Computer Vision (2022)
34. Nara, R., Lin, Y.C., Nozawa, Y., Ng, Y., Itoh, G., Torii, O., Matsui, Y.: Revisiting relevance feedback for clip-based interactive image retrieval. arXiv preprint arXiv:2404.16398 (2024)
35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (2021)
37. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (2021)
38. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)
39. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
40. Shen, K., Guo, J., Tan, X., Tang, S., Wang, R., Bian, J.: A study on relu and softmax in transformer. arXiv preprint arXiv:2302.06461 (2023)
41. Tao, Y., Takagi, K., Nakata, K.: Clustering-friendly representation learning via instance discrimination and feature decorrelation. In: International Conference on Learning Representations (2020)
42. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
43. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: Scan: Learning to classify images without labels. In: European Conference on Computer Vision. pp. 268–285 (2020)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8741–8750 (2021)
46. Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems (2022)

47. Wu, Z., Xiong, Y., Stella, X.Y., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
48. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
49. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International conference on machine learning. pp. 478–487. PMLR (2016)
50. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641 (2021)
51. Yang, L., Cheung, N.M., Li, J., Fang, J.: Deep clustering by gaussian mixture variational autoencoders with graph embedding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6440–6449 (2019)
52. Yang, X., Deng, C., Zheng, F., Yan, J., Liu, W.: Deep spectral clustering using dual autoencoder network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4066–4075 (2019)
53. Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., Susskind, J.M.: Stabilizing transformer training by preventing attention entropy collapse. In: International Conference on Machine Learning. pp. 40770–40803. PMLR (2023)
54. Zhou, X., Zhang, N.L.: Deep clustering with features from self-supervised pretraining. arXiv preprint arXiv:2207.13364 (2022)
55. Zou, Z.X., Yu, Z., Guo, Y.C., Li, Y., Liang, D., Cao, Y.P., Zhang, S.H.: Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10324–10335 (June 2024)