Connecting Joint-Embedding Predictive Architecture with Contrastive Self-supervised Learning

Shentong Mo[1]*, **Shengbang Tong**[2]
[1]CMU, [2]NYU

Abstract

In recent advancements in unsupervised visual representation learning, the Joint-Embedding Predictive Architecture (JEPA) has emerged as a significant method for extracting visual features from unlabeled imagery through an innovative masking strategy. Despite its success, two primary limitations have been identified: the inefficacy of Exponential Moving Average (EMA) from I-JEPA in preventing entire collapse and the inadequacy of I-JEPA prediction in accurately learning the mean of patch representations. Addressing these challenges, this study introduces a novel framework, namely C-JEPA (Contrastive-JEPA), which integrates the Image-based Joint-Embedding Predictive Architecture with the Variance-Invariance-Covariance Regularization (VICReg) strategy. This integration is designed to effectively learn the variance/covariance for preventing entire collapse and ensuring invariance in the mean of augmented views, thereby overcoming the identified limitations. Through empirical and theoretical evaluations, our work demonstrates that C-JEPA significantly enhances the stability and quality of visual representation learning. When pre-trained on the ImageNet-1K dataset, C-JEPA exhibits rapid and improved convergence in both linear probing and fine-tuning performance metrics.

1 Introduction

Unsupervised learning of visual representations has recently seen remarkable progress, primarily due to the development of innovative architectures and strategies that exploit unlabeled imagery. Among these advancements, the Joint-Embedding Predictive Architecture (JEPA) [1, 2, 3] has distinguished itself as a powerful approach. I-JEPA [2] leverages a masking strategy to extract visual features, facilitating significant strides in understanding and utilizing unlabeled visual data.

Figure 1: Our C-JEPA achieves faster and better convergence than I-JEPA.

However, despite its successes, certain limitations within the JEPA framework have become apparent, particularly concerning its components I-JEPA Exponential Moving Average (EMA) and I-JEPA prediction capabilities. Specifically, I-JEPA EMA has been found to be inadequate in preventing the issue of entire collapse [4, 5], while the I-JEPA prediction mechanism struggles to accurately learn the mean of patch representations. These challenges not only hinder the performance of JEPA but also limit its applicability in broader contexts.
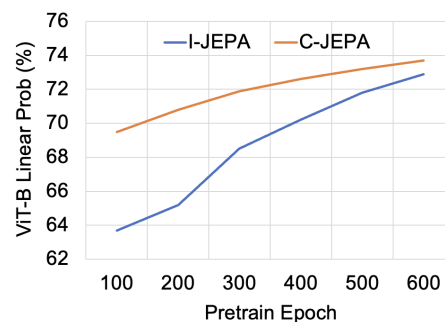
---

*Corresponding author: shentongmo@gmail.com.

To address these limitations, we introduce a novel contrastive self-supervised learning framework based on JEPA, namely C-JEPA, which aims to address the aforementioned challenges by incorporating the principles of Variance-Invariance-Covariance Regularization (VICReg) [6]. VICReg's methodology is adept at learning variance and covariance to avert entire collapse and ensure invariance for the mean of augmented views. By integrating VICReg with the Image-based JEPA, C-JEPA is designed to achieve faster and better convergence, as shown in Figure 1. In this paper, we aim to detail the theoretical underpinnings and empirical validations that substantiate the superiority of C-JEPA over previous self-supervised learning methods.

Our contributions are manifold and significant. Firstly, we identify and articulate the limitations inherent in the I-JEPA framework, specifically its EMA and prediction mechanisms. Secondly, we propose the C-JEPA framework as a novel solution that synergizes JEPA with VICReg to address these limitations effectively. Thirdly, through rigorous empirical and theoretical evaluations, we demonstrate that C-JEPA not only mitigates the issues identified but also achieves superior performance metrics when compared to existing frameworks. Particularly notable is C-JEPA's performance when pre-trained on the ImageNet-1K dataset, where it shows fast and improved convergence in linear probing and fine-tuning scenarios. These results highlight C-JEPA's potential to set a new benchmark in unsupervised visual representation learning, thereby contributing significantly to the field's advancement.

2 Related Work

Self-supervised Learning. In the self-supervised literature, researchers aim to exploit the internal characteristics of data and leverage pretext tasks to train a model. Recently, an unsupervised framework that learns effective views with data augmentation was proposed by Tian *et al.* [7] to reduce the mutual information between views. CMC [8] introduced a multi-view contrastive learning framework with any number of views to learn view-agnostic representations. Another pretext task of solving jigsaw puzzles was developed in PIRL [9] to improve the semantic quality of learned image representations, achieving better object detection results than supervised pre-training. More recently, Masked image modeling (MIM) has been explored in many previous works [10, 11, 12, 13, 14] to reconstruct the masked image patch given the unmasked counterpart as clues. Some MIM approaches [10, 11, 12, 15, 16] designed customized masking strategies (*i.e.*, random, block-wise) as pre-text tasks during pre-training. For example, block-wise masking was introduced in BEiT [10] to learn transferrable visual representations by recovering discrete tokens of masked image patches. Given features extracted from the 25% unmasked patches, the seminal work, MAE [12] directly reconstructed missing pixels of 75% masked patches. SimMIM [14] randomly masked the input image with a large square patch size (*i.e.*, 32) and used a one-layer prediction layer after the encoder to predict RGB values of raw pixels. Other researchers [15, 16, 17, 18] started to leverage a teacher network like CLIP [19] or adversarial learning to generate the mask and supervision target.

Contrastive Learning. In the past years, contrastive learning has shown its effectiveness in self-supervised learning, where various instance-wise contrastive learning frameworks [20, 21, 22, 23, 24, 25, 26, 27, 4] and prototype-level contrastive methods [28, 29, 30, 31] were proposed. The general idea of the instance-wise contrastive learning is to close the distance of the embedding of different views from the same instance while pushing embeddings of views from different instances away. One common way is to use a large batch size to accumulate positive and negative pairs in the same batch. For instance, Chen *et al.* [20] proposed a simple framework with a learnable nonlinear projection head and a large batch size to improve the quality of the pre-trained representations. To make the best use of a large amount of unlabelled data, they present a bigger unsupervised pre-training network and introduce distillation with unlabeled data in SimCLR v2 [21] to improve the performance in downstream tasks. Without involving negative instances, BOYL [22] trains the online network from an augmented view of an image to predict the target network representation of the same image under a different augmented view (positive instance). Another broadly-used approach [23] in the ICL literature is to apply a momentum encoder to update negative instances from a large and consistent dictionary on the fly. The dynamic dictionary was used with a moving-averaged encoder in MoCo series [24, 23] to build a dynamic dictionary to update negative instances in a queue of considerable size. The Variance-Invariance-Covariance (VICReg) [6] regularization strategy has been proposed to address shortcomings in self-supervised learning by enforcing stability through variance and covariance constraints.
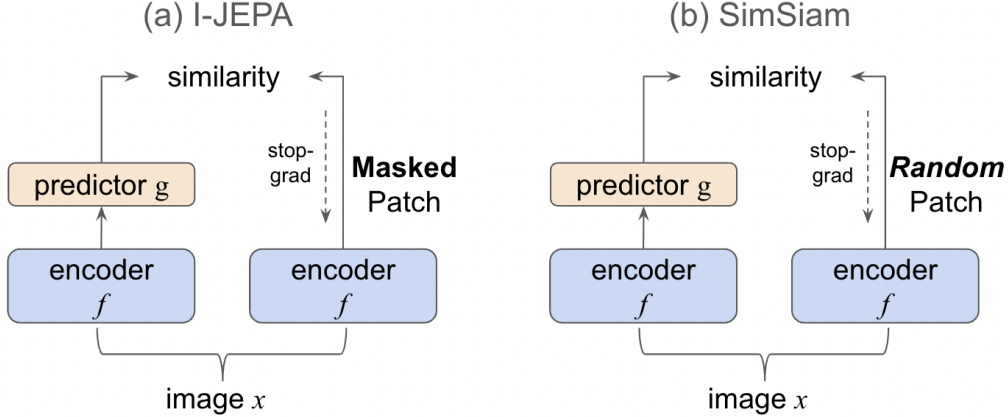
Figure 2: Illustration of I-JEPA (a) and SimSiam (b).

Joint-Embedding Predictive Architectures. The Joint-Embedding Predictive Architecture (JEPA) framework has shown considerable promise by enabling the prediction of embeddings from a compatible signal [1]. Building on this, the Image-based Joint-Embedding Predictive Architecture (I-JEPA) introduced by Assran et al. [2] utilizes a context encoder and a target encoder, together with a predictor, to manage representations under a masking regime. This approach focuses on learning embedding spaces directly, contrasting with other methods that operate on pixel or token spaces. While I-JEPA has been successful, it faces challenges such as the inefficacy of Exponential Moving Average (EMA) strategies in preventing model collapse and inaccuracies in predicting the mean of patch representations. These limitations hinder the stability and efficacy of the learned representations, prompting a need for enhanced methodologies. By maintaining sufficient variance in the embeddings and minimizing redundancy among features, VICReg supports robust feature learning. To address the limitations of I-JEPA, our work integrates VICReg into the JEPA framework, thereby aiming to prevent total model collapse and ensure invariance across different views of the same image.

3 Methodology

In this section, we present a novel masked modeling framework designed for the joint-embedding predictive architecture to avoid entire collapsing and improve the mean of patch representations. Our key idea is to integrate VICReg into the JEPA framework for alleviating entire model collapse and improving the invariance across different views of the same image. We first provide preliminaries in Section 3.1, then provide the theoretical and empirical connection between I-JEPA and SimSiam in Section 3.2, and finally connect I-JEPA with VICReg in Section 3.3 to show the benefit of reducing entire collapsing and learning the mean of patch representations.

3.1 Preliminaries

In this section, we first describe the problem setup and notations, and then revisit I-JEPA [2], SimSiam [4], and VICReg [6] for self-supervised image modeling.

Problem Setup and Notations. Given a dataset $\mathcal{X} = \{x_i\}_{i=1}^N$ with images $x_i \in \mathbb{R}^{c \times h \times w}$, our goal is to learn a neural network $f_\theta(\cdot)$ to extract unsupervised representations from these visual samples.

Revisit I-JEPA [2]. Taking the self-supervised modeling as a joint-embedding predictive architecture, I-JEPA [2] utilized $f_\theta(\cdot)$ as a context encoder, a pair of neural network $f'_\theta(\cdot)$ as a target encoder. A predictor $g_\theta(\cdot)$ is applied to predict the target representations from M masked block patches $\mathbf{b}_y(1), ..., \mathbf{b}_y(M)$. For a target block $\mathbf{b}_{y_i}$ corresponding to a target mask $\mathcal{B}_i$, the predictor $g_\theta(\cdot, \cdot)$ takes as input the output of the context encoder $\mathbf{b}_x$ and a mask token for each patch to predict $\{\mathbf{m}_j\}_{j \in \mathcal{B}_i}$, and outputs the patch-level prediction $\{\hat{\mathbf{b}}_{y_j}\}_{j \in \mathcal{B}_i}$, that is, $\{\hat{\mathbf{b}}_{y_j}\}_{j \in \mathcal{B}_i} = g_\theta(\{\mathbf{m}_j\}_{j \in \mathcal{B}_i})$.

The masking objective is optimized by the average ℓ_2 distance between the predicted patch-level representations $\hat{\mathbf{b}}_{y_j}$ and the target patch-level representation $\mathbf{b}_{y_j}$, which is formulated as:

$$\mathcal{L}_{\text{I-JEPA}} = \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||\hat{\mathbf{b}}_{y_j} - \mathbf{b}_{y_j}||_2^2, \tag{1}$$

where $|M|$ denotes the total number of target blocks M, and $\mathcal{B}_i$ is the generated mask corresponding to the i-th target block during training.

Revisit SimSiam [4]. For contrastive self-supervised learning, SimSiam [4] tried to obtain two augmented views x_i and x_i'. Then they fed two views into a pair of neural networks $f_\theta(\cdot)$ and $f_\theta'(\cdot)$. To learn the invariance of two different views from the same image, a prediction MLP head, denoted as g, transforms the output $\mathbf{z}$ of one view and matches it to representations $\mathbf{p}$ of the other view. The overall objective of SimSiam is to minimize the distance between $\mathbf{z}$ and $\mathbf{p}$ for all random patches r_j in the same image, which is formulated as:

$$\mathcal{L}_{\text{SimSiam}} = \frac{1}{|V|} \sum_{i=1}^{V} \sum_{j \in \mathcal{P}_i} ||\mathbf{z}_{r_j} - \mathbf{p}_{r_j}||_2^2, \tag{2}$$

where $|V|$ denotes the total number of augmented views V, and $\mathcal{P}_i$ is the all random patches corresponding to the i-th view during training.

Revisit VICReg [6]. To prevent a collapse in which the encoders produce constant or non-informative vectors in Siamese net architecture, VICReg [6] introduced variance and covariance regularization terms based on the representations space in the invariance terms. Firstly, the variance regularization term v as a hinge function on the standard deviation of the embeddings $\mathbf{z}$ along the batch dimension n to prevent collapse with all the embeddings mapped on the same vector, which is denoted as:

$$v(\mathbf{z}) = \frac{1}{d} \sum_{j=1}^{d} \max(0, \gamma - \sqrt{\text{Var}(\mathbf{z}_j) + \epsilon}), \tag{3}$$

where $\gamma = 1$ denotes a constant value for the standard deviation in their experiments, and ϵ denotes a small numerical scalar to stabilize training. Secondly, the covariance regularization term c minimizes the sum of the squared off-diagonal coefficients to encourage the off-diagonal coefficients of correlation matrix C along the batch dimension to be close to 0, which is formulated as:

$$c(\mathbf{z}) = \frac{1}{d} \sum_{i \neq j} [\frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T)]_{i,j}^2, \tag{4}$$

where $\bar{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i$. This term is helpful to prevent different dimensions of the embeddings from encoding trivial information. Finally, the invariance term is defined to minimize the mean-squared Euclidean distance between each pair of vectors $\mathbf{z}$ and $\mathbf{z}_{r_j}$ from different views.

3.2 Connecting I-JEPA with SimSiam

In the realm of self-supervised learning, both I-JEPA [2] and SimSiam [4] frameworks aim to extract robust representations from images, yet they approach the problem with distinct architectures and objectives. Here, we draw theoretical and empirical parallels between these methodologies, leveraging their unique approaches to enhance the understanding of joint-embedding architectures.

Regarding theoretical connections, I-JEPA [2] uses a predictive model where the encoder and predictor work together to forecast masked parts of the input, relying on partial views. Conversely, SimSiam [4] operates without explicit masking, utilizing dual augmentations of the same image to enforce consistency between the independently processed views. Both models share the underlying principle of minimizing the distance between certain representations. As shown in Figure 2, I-JEPA focuses on the distance between predicted and actual masked patch representations, whereas SimSiam minimizes the distance between the two augmented views of an image, thereby encouraging consistency across different transformations of the same data.

Meanwhile, empirical studies such as Tian et al. [5] have shown that linear predictors in frameworks like BYOL [22], which are closely related to SimSiam, tend to learn alignment with the correlation

matrix of representations. This suggests that SimSiam's approach without a separate predictor could inherently align with the mean of augmented views, a concept central to I-JEPA's strategy of predicting masked patches. A study [32] in joint-embedding self-supervised learning also indicates that these methods primarily focus on capturing the co-occurrence and distribution of image patches, which aligns with I-JEPA's objective of learning from masked representations.

3.3 Connecting I-JEPA with VICReg

VICReg [6] introduces variance and covariance regularization to prevent the collapse of representations in Siamese networks, ensuring that the model learns informative and diverse features across different dimensions. By integrating VICReg into I-JEPA, our objective is to tackle the common challenges of model collapse and enhance the mean representation learning of image patches.

VICReg's variance regularization ensures that all dimensions of the embedding space contain meaningful variance, which is crucial for preventing the model from collapsing to trivial solutions. I-JEPA, which aims to learn diverse patch representations, can benefit from such a mechanism to ensure that each masked patch contributes informatively to the overall representation. By minimizing the off-diagonal elements of the covariance matrix, VICReg encourages the features to be uncorrelated, enhancing the diversity of the learned features. This aspect can be particularly beneficial for I-JEPA, where diverse patch predictions are essential for effective representation learning. Literature on non-contrastive self-supervised learning [33] suggests that implicit variance regularization, as seen in VICReg, can facilitate the learning dynamics in joint-embedding architectures. Such regularization helps maintain a balance between similarity and diversity in learned representations, which is crucial for effective self-supervised learning.

In the following, we consider linear predictor $W_P \in \mathbb{R}^{M \times M}$ in I-JPEA, the masking objective with the average ℓ_2 distance between the predicted patch-level representations $\mathbf{z}_{y_j}$ and the target patch-level representations $\mathbf{z}_{y_j}^a$ as

$$\mathcal{L} = \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||W_P \mathbf{z}_{y_j} - \text{SG}(\mathbf{z}_{y_j}^a)||_2^2, \tag{5}$$

where $|M|$ denotes the total number of target blocks M, and $\mathcal{B}_i$ is the generated mask corresponding to the i-th target block during training. By connecting I-JEPA with SimSiam in Section 3.2, we can can diagonalize correlation matrix of $\mathbf{z}_{y_j}$ over $\mathbb{R}$: $C_{\mathbf{z}_{y_j}} = U D_C U^T$, where U is an orthogonal matrix whose columns are the eigenvectors of $C_{\mathbf{z}_{y_j}}$ and D_C is the real-valued diagonal matrix of the eigenvalues $s_k, k \in [1, K]$. Given this eigendecomposition, the predictor is directly set to $W_P = U D_C^{\alpha} U^T$, where α is a positive constant exponent applied element-wise to D_C. The eigenvalues λ_k of the predictor matrix W_P are then $\lambda_k = s_k^{\alpha}$. Assume $\hat{\mathbf{z}}_{y_j}$ the representations expressed in the predictor's eigenbasis, the asymmetric loss $\mathcal{L}$ can be formulated as:

$$\mathcal{L} = \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} \sum_{k}^{K} ||\lambda_k \hat{\mathbf{z}}_{y_j} - \text{SG}(\hat{\mathbf{z}}_{y_j}^a)||_2^2, \tag{6}$$

Following non-contrastive self-supervised learning [33], we can use NTK [34, 35] to characterize the learning gradient dynamics of neural networks as $\nabla_{\mathbf{z}_{y_j}} \mathcal{L} = (D\mathbf{z}_{y_j} - \mathbf{z}_{y_j}^t)D$, and the representational dynamics of each mode k independently follow gradient of the loss $-\nabla_{\hat{\mathbf{z}}_{y_j}} \mathcal{L}$, and decouple as K independent differential equations:

$$\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = -\eta \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{z}}_{y_j,k}}(t) = \eta \lambda_k \left(\hat{\mathbf{z}}_{y_j,k}^a - \lambda_k \hat{\mathbf{z}}_{y_j,k} \right) \tag{7}$$

By taking the expectation over the same masking blocks, we can have the dynamics for each mask patch y_j as

$$\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = \eta \lambda_k (1 - \lambda_k) \hat{\mathbf{z}}_{y_j,k} \tag{8}$$

Note that when $\lambda_k < 1$, $d\hat{\mathbf{z}}_{y_j,k}$ has the same sign as $\hat{\mathbf{z}}_{y_j,k}$ and they have opposite sign at $\lambda_k > 1$. These convergent dynamics will push an eigenvalue λ_k of one to prevent the collapse of each mode k in representation dynamics. When removing the stop-grad operator in $\mathcal{L}$, we can have

$\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = -\eta(1-\lambda_k)^2\hat{\mathbf{z}}_{y_j,k}$, where $d\hat{\mathbf{z}}_{y_j,k}$ and $\hat{\mathbf{z}}_{y_j,k}$ always have opposite signs and the learned representations will become zero with exponentially decaying eigenvalues to collapse. Without the predictor W_P, we will have $\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = 0$, and the representations will not be updated.

Incorporating VICReg's regularization strategies into I-JEPA could prevent the entire collapsing of the model, especially when learning from a large and diverse dataset. This integration could also improve the granularity and utility of the patch-level representations by ensuring that each dimension of the embedding space remains informative and independent. Overall, the integration of VICReg and SimSiam principles into the I-JEPA framework offers promising avenues for enhancing the robustness and efficacy of self-supervised learning models, particularly in tasks requiring nuanced understanding of complex visual content.

4 Experiments

4.1 Experimental setup

Datasets. Following previous methods [12, 2], we use ImageNet-1K [36] for image classification, MS-COCO [37] for object detection and instance segmentation, and ADE20K [38, 39] for semantic segmentation. We closely follow previous work [40, 41], and adopt the Mask R-CNN [42] as the detector. The ViTs [43] backbone weights are initialized with weights pre-trained on ImageNet-1K using our C-JEPA. Following the settings in [12], we use the Semantic FPN and UPerNet approach [44] based on our ImageNet-1K pre-trained ViTs for evaluation. For a fair comparison, we fine-tune the detector with the same learning rate in [12]. For video object segmentation, we use DAVIS-2017 dataset containing 60 training, 30 validation, and 60 testing videos. For low-level tasks, we follow the previous work [2] and use Clevr/Count and Clevr/Dist on Clevr [45] dataset.

Evaluation Metrics. For image classification, we follow previous masked image modeling methods [12, 2] to report the classification accuracy of linear probing and fine-tuning. For object detection and instance segmentation on MS-COCO, we apply AP^{box} and AP^{mask} as metrics for the bounding boxes and the instance masks. mIoU results are reported to evaluate semantic segmentation on ADE20K. For video object segmentation on DAVIS-2017, we use Jabri-based $(\mathcal{J}\&\mathcal{F})_m$, $\mathcal{J}_m$, $\mathcal{F}_m$ as metrics to evaluate the quality of frozen representations of image patches by segmenting scenes with the nearest neighbor between consecutive frames. For object counting and depth prediction tasks on Clevr, we use object counting and depth prediction to evaluate the linear probing performance of our model.

Implementation. For input images, we resized the resolution to 224×224, *i.e.*, $H = W = 224$. Following prior work [12, 2], we apply a patch size of 16, *i.e.*, $P = 16$. We use the tiny, small, base, and large models of ViT [43] architecture for experiments. We set the embedding dimension of the predictor to 384, and keep the number of self-attention heads the same as the backbone context-encoder. For the ViT-T/16, ViT-S/16, and ViT-B/16 context-encoder, we set the depth of the predictor as 6. For ViT-L/16 context-encoders, we set the depth of the predictor to 12. Following I-JEPA [2], we use AdamW to optimize the context-encoder and predictor weights. We train our model using the default batch size of 2048, and the learning rate linearly increased from 1e-4 to 1e-3 during the first 15 epochs of pre-training, and decay to 1e-6 following a cosine schedule. The weight decay is linearly increased from 0.04 to 0.4, and the target-encoder weights are initialized the same as the context-encoder weights, and updated via an exponential moving average. We use a momentum value of 0.996, and linearly increase this value to 1.0. For masking, we use the same strategy and settings as I-JEPA [2] for 4 possibly overlapping target block masks.

4.2 Experimental comparisons

In this work, we propose a novel and effective framework for connecting Joint-Embedding Predictive Architecture with VICReg in non-contrastive self-supervised learning. In order to demonstrate the effectiveness of the proposed C-JEPA, we comprehensively compare it to previous baselines [41, 10, 12, 46, 47, 2] on non-contrastive self-supervised learning and mask image modeling.

ImageNet-1K image classification. For image classification on the ImageNet-1K benchmark, we report the quantitative comparison results on linear evaluation and fine-tuning results in Table 1. We can observe that our C-JEPA achieves the best results in terms of ViT-B/16 compared to previous

Table 1: **Comparison with prior work.** We perform linear evaluation, fine-tuning, COCO detection/segmentation, and ADE20K semantic segmentation on pre-trained ViT-B/16 models. We report linprob, fine-tune, AP^{box}, AP^{mask}, and mIoU to evaluate the quality of pre-trained representations. The best results are indicated in **bold**.

Method	Pretrain Epochs	linprob	fine-tune	AP^{box}	AP^{mask}	mIoU
DINO [41]	1600	78.2	82.8	50.1	43.4	46.8
BEiT [10]	800	56.7	83.4	49.8	44.4	47.1
MAE [12]	1600	68.0	83.6	50.3	44.9	48.1
iBOT [46]	1600	79.5	84.0	51.2	44.2	50.0
data2vec [47]	800	60.8	84.2	–	–	48.2
I-JEPA [2]	600	72.9	83.5	49.9	44.5	47.6
C-JEPA (ours)	600	**73.7**	**84.5**	**50.7**	**45.3**	**48.7**

Table 2: **Scaling up to Large Models.** We perform linear evaluation, fine-tuning, video object segmentation, and low-level tasks on pre-trained ViT-L/16 models. We report linprob, fine-tune, $(\mathcal{J}\&\mathcal{F})_m$, Clevr/Count and Clevr/Dist metrics to evaluate the quality of pre-trained representations. The best results are indicated in **bold**.

Method	Pretrain Epochs	linprob	fine-tune	$(\mathcal{J}\&\mathcal{F})_m$	Clevr/Count	Clevr/Dist
I-JEPA [2]	600	77.5	85.3	56.6	85.6	71.2
C-JEPA (ours)	600	**78.1**	**86.2**	**58.3**	**86.8**	**71.6**

masked image modeling approaches. Specifically, the proposed method outperforms the I-JEPA by 0.8 and 1.0 in terms of top-1 accuracy on linear evaluation and fine-tuning settings. Meanwhile, we enjoy the advantage of fewer pre-training epochs compared to other masked image model frameworks.

COCO object detection and instance segmentation. For COCO object detection & instance segmentation benchmarks, we also report the quantitative comparison results in Table 1. As can be seen, we achieve significant performance gains of 0.8@AP^{box} and 0.8@AP^{mask} on COCO object detection and instance segmentation compared to I-JEPA. We also achieve better results than DINO [41] and MAE [12] regarding both settings.

ADE20K semantic segmentation. For the ADE20K semantic segmentation, we report the quantitative comparison results in Table 1. As can be seen, the proposed C-JEPA outperforms I-JEPA [2], the current image-based joint-embedding predictive architecture by 1.1@mIoU. Also, we observed that our C-JEPA can achieve better performance than other masked image modeling baselines, including MAE [12] and data2vec [47]. These improvements suggest the importance of leveraging the VICReg's regularization to capture better semantics for dense prediction tasks.

DAVIS video object segmentation. In addition, we scale up our experiments from ViT-B to ViT-L and report the results in Table 2. The proposed C-JEPA still achieves better performance than I-JEPA [2] in terms of linear evaluation and fine-tuning settings. For DAVIS video object segmentation, our method also shows significant and consistent gains as shown in Table 2. Compared to I-JEPA, ours achieved the results gains of 1.7@$(\mathcal{J}\&\mathcal{F})_m$ on ViT-L/16.

Clevr object counting and depth prediction. We further measure the abilities of object-counting and depth prediction on Clevr benchmarks for our pre-trained models. Table 2 shows linear evaluation results of C-JEPA on the Clevr counting and depth benchmarks. Compared to I-JEPA [2], we achieve the results gains of 1.2@Clevr/Count and 0.4@Clevr/Dist using ViT-L/16. These results further indicate the benefit of the proposed method in learning better representations than our vanilla I-JEPA baseline without VICReg regularization.

4.3 Experimental analysis

In this section, we performed ablation studies to demonstrate the benefit of the proposed Variance/Covariance and Invariance terms from VICReg. Here, we conducted extensive experiments on ImageNet-1k pre-trained ViT-B/16 to explore the impact of representation mean and collapse, and learned meaningful qualitative patch-level representations.

Table 3: **Ablation studies on component analysis for faster convergence.** We perform ablation studies on Variance/Covariance and Invariance modules in VICReg using ViT-B/16 model. The best results are indicated in **bold**.

I-JEPA	Variance/Covariance	Invariance	Backbone	Pretrain Epoch	linprob	fine-tune	(J & F)_m
✓	✗	✗	ViT-B/16	100	63.7	82.5	52.3
✓(mean)	✓(collapse)	✗	ViT-B/16	100	68.3	83.2	54.6
✓	✓	✓	ViT-B/16	100	**69.5**	**83.6**	**55.2**
✓(EMA for collapse)	✗	✓(mean)	ViT-B/16	100	67.6	82.8	53.9

Table 4: **Ablation studies on component analysis for better convergence.** We perform ablation studies on Variance/Covariance and Invariance modules in VICReg using ViT-B/16 model. The best results are indicated in **bold**.

I-JEPA	Variance/Covariance	Invariance	Backbone	Pretrain Epoch	linprob	fine-tune	(J & F)_m
✓	✗	✗	ViT-B/16	600	72.9	83.9	56.2
✓(mean)	✓(collapse)	✗	ViT-B/16	600	73.5	84.3	56.9
✓	✓	✓	ViT-B/16	600	**73.7**	**84.5**	**57.5**
✓(EMA for collapse)	✗	✓(mean)	ViT-B/16	600	73.2	84.2	56.6

Variance/Covariance and Invariance in VICReg. We first conducted an ablation study to determine the efficacy of incorporating Variance and Covariance terms, along with the Invariance term, in preventing model collapse and improving the fidelity of representation means. These experiments are crucial for understanding how each component contributes to the overall stability and effectiveness of the learned embeddings. The results, as presented in Tables 3 and 4, provide clear insights into the performance enhancements driven by these modifications. The inclusion of the Variance and Covariance terms significantly accelerated the training process, as evidenced by the metrics in Table 3. Table 4 highlights improved accuracy and stability in the embeddings, demonstrating the beneficial impact of the Invariance term in aligning the representation means across different mask blocks of the same images.

Representation Mean and Collapse. Further analysis focused on how well the C-JEPA framework mitigates the risk of model collapse and accurately learns the mean of patch representations: The integration of VICReg's terms with C-JEPA's architecture addresses the previously noted deficiencies in I-JEPA's EMA component, which was prone to collapsing. The results (referenced in the last rows of Tables 3 and 4) underscore the robustness added by these regularization terms from VICReg.

Qualitative Attention Visualizations. To complement our quantitative findings, we also examined qualitative aspects through attention visualization techniques. Figures 3 and **??** showcase the attention maps generated by both the base I-JEPA and the enhanced C-JEPA models. These visualizations illustrate the more focused and contextually relevant attention in C-JEPA, which correlates with the theoretical improvements expected from incorporating VICReg's terms. The attention maps further validate our approach by visually demonstrating the enhanced capability of C-JEPA to maintain stable and meaningful patch-level representations across various image contexts, an improvement over the base I-JEPA model. This qualitative evidence supports the quantitative improvements and highlights the benefits of our approach in capturing invariance.

5   Conclusion

In this work, we introduced C-JEPA, a novel enhancement to the Joint-Embedding Predictive Architecture that integrates the robust features of Variance-Invariance-Covariance Regularization (VICReg) to address critical limitations in the existing I-JEPA model. By refining the EMA and prediction components of I-JEPA, C-JEPA successfully prevents entire model collapse and more effectively learns the mean of patch representations, thereby advancing the state of unsupervised visual representation learning. Our theoretical analysis and empirical results have demonstrated the effectiveness of C-JEPA, particularly when pre-trained on the ImageNet-1K dataset. The framework shows marked improvements in convergence rates and performance in both linear probing and fine-tuning scenarios, outperforming existing methods. This underscores C-JEPA's capability to not only address the weaknesses of its predecessor but also to provide substantial improvements in learning quality and stability.
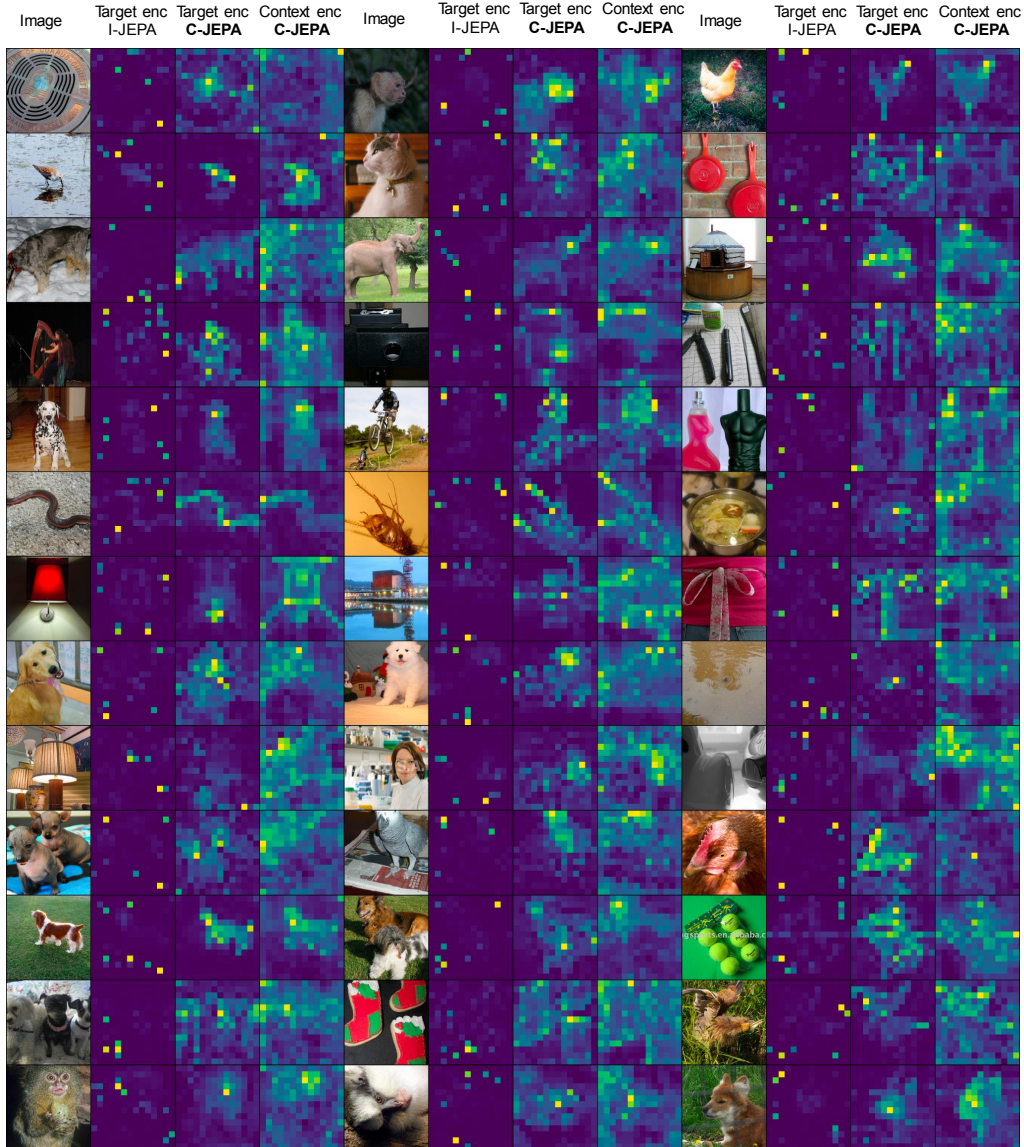
Figure 3: **Qualitative visualization of learned attention maps using ViT-B/16 model.** Columns for each sample denote the original image, attention maps from the target encoder in I-JEPA [2], attention maps from the target encoder in our C-JEPA, and attention maps from the context encoder in our C-JEPA. Our C-JEPA achieves much better attention maps.

Broader Impact

Our C-JEPA's adoption of VICReg principles enhances its versatility and applicability across various unsupervised learning contexts, making it a valuable tool for tasks requiring robust and reliable visual representations. The implications of this research are significant, offering a pathway for future studies to explore and expand upon the integration of contrastive and joint-embedding techniques. As we move forward, further refinements and explorations into the scalability of C-JEPA in more diverse and challenging datasets will be crucial. Additionally, investigating the adaptability of C-JEPA to different domains and its effectiveness in real-world applications will be essential to fully realize its potential. C-JEPA represents a significant step forward in the field of machine learning, particularly in the unsupervised learning of visual representations. It sets a new benchmark for future research and development in this area, promising enhanced performance and broader applicability in tackling complex visual understanding tasks.

References

[1] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022. 1, 3

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023. 1, 3, 4, 6, 7, 9, 19, 20, 22, 23, 24

[3] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. 1

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021. 1, 2, 3, 4

[5] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 10268–10278. PMLR, 2021. 1, 4

[6] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. 2, 3, 4, 5

[7] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 6827–6839, 2020. 2

[8] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[9] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[10] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 6, 7

[11] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021. 2

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 2, 6, 7

[13] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14668–14678, June 2022. 2

[14] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, June 2022. 2

[15] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. MST: Masked self-supervised transformer for visual representation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[16] Yuge Shi, N Siddharth, Philip HS Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *Proceedings of International Conference on Machine Learning (ICML)*, 2022. 2

[17] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. MVP: multimodality-guided visual pre-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 337–353, 2022. 2

[18] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Sem-MAE: Semantic-guided masking for learning masked autoencoders. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020. 2

[21] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[22] 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. Bootstrap your own latent - a new approach to self-supervised learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 2

[24] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[26] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 15614–15624, 2020. 2

[27] Hu Qianjiang, Wang Xiao, Hu Wei, and Qi Guo-Jun. AdCo: adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1074–1083, 2021. 2

[28] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[29] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 2

[30] Xudong Wang, Ziwei Liu, and Stella X Yu. CLD: unsupervised feature learning by cross-level instance-group discrimination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[31] Shentong Mo, Zhun Sun, and Chao Li. Siamese prototypical contrastive learning. In *BMVC*, 2021. 2

[32] Yubei Chen, Adrien Bardes, ZENGYI LI, and Yann LeCun. Bag of image patch embedding behind the success of self-supervised learning. *Transactions on Machine Learning Research*, 2023. 5

[33] Manu Srinath Halvagal, Axel Laborieux, and Friedemann Zenke. Implicit variance regularization in non-contrastive SSL. In *Proceedings of Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 5, 14

[34] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, 2018. 5, 14

[35] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5, 14

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 6, 19

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6, 19

[38] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 6, 19

[39] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision (IJCV)*, 127:302–321, 2018. 6, 19

[40] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 6

[41] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 6, 7

[42] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 6, 19

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations*, 2021. 6

[44] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 432–448, 2018. 6, 19

[45] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2016. 6, 19

[46] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 6, 7

[47] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 6, 7

[48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. 17, 18

[49] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 17, 18

[50] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 19

Appendix

In this appendix, we provide the following materials:

- proofs on the loss $\mathcal{L}$ in the predictor eigenspace and representation dynamics under $\mathcal{L}$ Section A,
- derivation of learning dynamics without the stop-grad and predictor in Section B,
- detailed loss functions and algorithm in Section C,
- additional experimental results in small-scale models and VICReg hyperparameters in Section D,
- additional visualizations on the learned attention maps in Section E.

A Proofs

A.1 Loss $\mathcal{L}$ in the predictor eigenspace

We consider linear predictor $W_P \in \mathbb{R}^{M \times M}$ in I-JPEA, the masking objective with the average ℓ_2 distance between the predicted patch-level representations $\mathbf{z}_{y_j}$ and the target patch-level representations $\mathbf{z}_{y_j}^a$ as

$$\mathcal{L} = \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||W_P \mathbf{z}_{y_j} - \text{SG}(\mathbf{z}_{y_j}^a)||_2^2, \tag{9}$$

Assume $\hat{\mathbf{z}}_{y_j}$ the representations expressed in the predictor's eigenbasis. Since $W_P \in \mathbb{R}^{M \times M}$ is a symmetric matrix with eigendecomposition $W_P = U D_C U^T$, where U is an orthogonal matrix and $UU^T = I$, we can substitute W_P in Eq. 9 as

$$\begin{aligned}
\mathcal{L} &= \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||W_P \mathbf{z}_{y_j} - \text{SG}(\mathbf{z}_{y_j}^a)||_2^2 \\
&= \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||U D_C U^T \mathbf{z}_{y_j} - \text{SG}(UU^T \mathbf{z}_{y_j}^a)||_2^2 \\
&= \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||D_C \mathbf{z}_{y_j} - \text{SG}(\mathbf{z}_{y_j}^a)||_2^2 \\
&= \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} \sum_{k}^{K} ||\lambda_k \hat{\mathbf{z}}_{y_j} - \text{SG}(\hat{\mathbf{z}}_{y_j}^a)||_2^2
\end{aligned} \tag{10}$$

A.2 Representational dynamics under $\mathcal{L}$

Following non-contrastive self-supervised learning [33], we can use NTK [34, 35] to characterize the learning gradient dynamics of neural networks as $\nabla_{\mathbf{z}_{y_j}} \mathcal{L} = (D \mathbf{z}_{y_j} - \mathbf{z}_{y_j}^t) D$, and the representational dynamics of each mode k independently follow gradient of the loss $-\nabla_{\hat{\mathbf{z}}_{y_j}} \mathcal{L}$, and decouple as K independent differential equations:

$$\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = -\eta \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{z}}_{y_j,k}}(t) = \eta \lambda_k \left(\hat{\mathbf{z}}_{y_j,k}^a - \lambda_k \hat{\mathbf{z}}_{y_j,k} \right) \tag{11}$$

By taking the expectation over the same masking blocks, we can have the dynamics for each mask patch y_j as

$$\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = \eta \lambda_k (1 - \lambda_k) \hat{\mathbf{z}}_{y_j,k} \tag{12}$$

Suppose the parameters of the neural network are denoted as θ with weights $W \in \mathbb{R}^{M \times M}$, we follow Lemma 3 proposed in [33] and have the empirical NTK $\hat{\Theta}(\mathcal{Y}, \mathcal{Y})$ in the orthogonal eigenbasis is equal

to the empirical NTK$\Theta(\mathcal{Y}, \mathcal{Y})$ in the original basis. Each block of the full empirical NTK is denoted as $\hat{\Theta}_t(\boldsymbol{y}_j, \boldsymbol{y}_q) = (\boldsymbol{y}_j^T \boldsymbol{y}_q) I_M$, where $\boldsymbol{y}_j, \boldsymbol{y}_q$ denote the high-dimensional inputs from mask patches j and q in the training samples. $I_M \in \mathbb{R}^{M \times M}$ is the identity. Based on this and the central limit theorem for an i.i.d standard Gaussian distribution, we can formulate the representational dynamics for each mask patch y_j under the $\mathcal{L}$ as

$$
\begin{aligned}
\frac{d\hat{\mathbf{z}}_{y_j}}{dt} &= -\eta \hat{\Theta}_t(\boldsymbol{y}_j, \mathcal{Y}) \nabla_{\hat{\mathcal{Z}}_y} \mathcal{L} \\
&= -\eta \hat{\Theta}_t(\boldsymbol{y}_j, \boldsymbol{y}_j) \nabla_{\hat{\mathbf{z}}_{y_j}} \mathcal{L} - \eta \sum_{q \neq j} \hat{\Theta}_t(\boldsymbol{y}_j, \boldsymbol{y}_q) \nabla_{\hat{\mathbf{z}}_{y_q}} \mathcal{L} \\
&= -\eta (\boldsymbol{y}_j^T \boldsymbol{y}_j) \nabla_{\hat{\mathbf{z}}_{y_j}} \mathcal{L} - \eta \sum_{q \neq j} (\boldsymbol{y}_j^T \boldsymbol{y}_q) \nabla_{\hat{\mathbf{z}}_{y_q}} \mathcal{L} \\
&= -\eta \nabla_{\hat{\mathbf{z}}_{y_j}} \mathcal{L} = -\eta \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{z}}_{y_j}}(t) \\
&= -\eta \left(D_t \hat{\mathbf{z}}_{y_j} - \hat{\mathbf{z}}_{y_j}^a \right) D_t
\end{aligned}
\tag{13}
$$

where D_t is a diagonal matrix composed of the eigenvalues λ_k. For kth entry of $\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt}$, we can formulate the representational dynamics under the $\mathcal{L}$ as

$$
\begin{aligned}
\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} &= -\eta \nabla_{\hat{\mathbf{z}}_{y_j,k}} \mathcal{L} = -\eta \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{z}}_{y_j,k}}(t) \\
&= -\eta \left(\lambda_k \hat{\mathbf{z}}_{y_j,k} - \hat{\mathbf{z}}_{y_j,k} \right) \lambda_k \\
&= \eta \lambda_k \left(\hat{\mathbf{z}}_{y_j,k}^a - \lambda_k \hat{\mathbf{z}}_{y_j,k} \right)
\end{aligned}
\tag{14}
$$

Since the expectation over the same masking blocks is $\hat{\mathbf{z}}_{y_j,k}$, we can have the dynamics for each mask patch y_j as

$$
\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = \eta \lambda_k (1 - \lambda_k) \hat{\mathbf{z}}_{y_j,k}
\tag{15}
$$

B   Representation learning dynamics without the stop-grad and predictor

B.1   Learning dynamics of the loss $\mathcal{L}$ without the stop-grad

Recall we consider linear predictor $W_P \in \mathbb{R}^{M \times M}$ in I-JPEA, the masking objective with the average ℓ_2 distance between the predicted patch-level representations $\mathbf{z}_{y_j}$ and the target patch-level representations $\mathbf{z}_{y_j}^a$ as

$$
\mathcal{L} = \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||W_P \mathbf{z}_{y_j} - \mathrm{SG}(\mathbf{z}_{y_j}^a)||_2^2,
\tag{16}
$$

When removing the stop-grad operator in $\mathcal{L}$, we can have the masking objective as

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||W_P \mathbf{z}_{y_j} - \mathbf{z}_{y_j}^a||_2^2 \\
&= \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} \sum_{k}^{K} ||\lambda_k \hat{\mathbf{z}}_{y_j,k} - (\hat{\mathbf{z}}_{y_j,k}^a)||_2^2
\end{aligned}
\tag{17}
$$

The representational dynamics for each mask patch y_j under the $\mathcal{L}$ is formulated as

$$
\begin{aligned}
\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} &= -\eta \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{z}}_{y_j,k}}(t) \\
&= -\eta \left(\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{z}}_{y_j,k}}(t) + \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{z}}_{y_j,k}^a}(t) \right) \\
&= -\eta \left((\lambda_k \hat{\mathbf{z}}_{y_j,k} - \hat{\mathbf{z}}_{y_j,k}^a)\lambda_k + (\lambda_k \hat{\mathbf{z}}_{y_j,k} - \hat{\mathbf{z}}_{y_j,k}^a) \right) \\
&= -\eta(\lambda_k \hat{\mathbf{z}}_{y_j,k} - \hat{\mathbf{z}}_{y_j,k}^a)(\lambda_k - 1)
\end{aligned}
\tag{18}
$$

Since the expectation over the same masking blocks is $\hat{\mathbf{z}}_{y_j,k}$, we can have the dynamics for each mask patch y_j as

$$
\frac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = -\eta(1 - \lambda_k)^2 \hat{\mathbf{z}}_{y_j,k}
\tag{19}
$$

where $d\hat{\mathbf{z}}_{y_j,k}$ and $\hat{\mathbf{z}}_{y_j,k}$ always have opposite signs and the learned representations will become zero with exponentially decaying eigenvalues to collapse.

B.2 Learning dynamics of the loss $\mathcal{L}$ without the predictor

Without the predictor W_P, we can have the masking objective as

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} ||\mathbf{z}_{y_j} - \text{SG}(\mathbf{z}_{y_j}^a)||_2^2 \\
&= \frac{1}{|M|} \sum_{i=1}^{M} \sum_{j \in \mathcal{B}_i} \sum_{k}^{K} ||\hat{\mathbf{z}}_{y_j,k} - \text{SG}(\hat{\mathbf{z}}_{y_j,k}^a)||_2^2
\end{aligned}
\tag{20}
$$

In this case, we have $\lambda_k = 1$ for all independent eigenvalues in the eigenbasis. Based on the representation dynamics for each mask patch y_j as $\dfrac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = \eta \lambda_k (1 - \lambda_k) \hat{\mathbf{z}}_{y_j,k}$, we will have $\dfrac{d\hat{\mathbf{z}}_{y_j,k}}{dt} = 0$, and thus the representations will not be updated in the end.

C Algorithm

In this section, we provide detailed algorithmic and implementation aspects of the Contrastive-JEPA (C-JEPA) framework, elaborating on the loss functions, pseudo-code, and experimental setup used throughout our studies. These details aim to enable reproducibility and deeper understanding of the modifications introduced in the C-JEPA compared to its predecessors.

C.1 Loss

The loss function for C-JEPA is a combination of the original JEPA loss and the VICReg-inspired regularization terms, structured as follows:

$$
\mathcal{L} = \mathcal{L}_{JEPA} + \beta_{vicreg} * \mathcal{L}_{VICReg}
\tag{21}
$$

where $\beta vicreg = 0.001$ provides a scaling factor to balance the original JEPA loss with the new regularization terms introduced from VICReg. The VICReg loss component is defined as:

$$
\mathcal{L}_{VICReg} = \beta_{sim} * \mathcal{L}_{sim} + \beta_{std} * \mathcal{L}_{std} + \beta_{cov} * \mathcal{L}_{cov}
\tag{22}
$$

with $\beta_{sim} = 25$, $\beta_{std} = 25$, and $\beta_{cov} = 1$, which scale the similarity, standard deviation, and covariance regularization terms respectively, integral to maintaining the diversity and stability of the learned representations.

16

C.2 Pseudo Code

The following pseudo-code outlines the batch-wise application of VICReg within the C-JEPA framework, providing clarity on the computational steps involved during training:

```
# Pseudo-code for batch-wise VICReg application in C-JEPA
for each batch in dataset:
    # Forward pass through the context and target encoders
    context_embeddings, target_embeddings = encode(batch)

    # Apply VICReg terms
    sim_loss = compute_similarity_loss(context_embeddings, target_embeddings)
    std_loss = compute_standard_deviation_loss(context_embeddings)
    cov_loss = compute_covariance_loss(context_embeddings)

    # Combine losses
    vicreg_loss = beta_sim * sim_loss + beta_std * std_loss + beta_cov * cov_loss
    total_loss = jepa_loss(batch) + beta_vicreg * vicreg_loss

    # Backward pass and update weights
    optimizer.zero_grad()
    total_loss.backward()
    optimizer.step()
```

Algorithm 1 outlines the procedure for applying VICReg in a batch-wise manner within the C-JEPA framework, focusing on multiple views in the context of contrastive learning.

Table 5: **Pretraining setting for ViT-T**. All models trained for 100 and 600 epochs.

config	value
optimizer	AdamW [48]
epochs	100 or 600
learning rate	$1e^{-3}$
weight decay	$(0.04, 0.4)$
batch size	2048
learning rate schedule	cosine decay [49]
warmup epochs	15
encoder arch.	ViT-T
predicted targets	4
predictor depth	6
predictor attention heads	12
predictor embedding dim.	192

Table 6: **Pretraining setting for ViT-S**. All models trained for 100 epochs.

config	value
optimizer	AdamW [48]
epochs	100
learning rate	$8e^{-4}$
weight decay	$(0.04, 0.4)$
batch size	2048
learning rate schedule	cosine decay [49]
warmup epochs	15
encoder arch.	ViT-S
predicted targets	4
predictor depth	6
predictor attention heads	12
predictor embedding dim.	384

Algorithm 1 VICReg loss for our framework

Require: zc: Context embeddings tensor of size [batch_size, num_block, num_mask_patch, dimension]
Ensure: $loss$: Computed loss for the batch

1: **function** VICREGLOSS(zc, args)
2:     $zc \leftarrow$ MEAN(zc, dim $= 2$)   ▷ Average over the mask patches
3:     $zc \leftarrow$ PROJECTOR(zc)   ▷ Apply projection to the embeddings
4:     **if** 'vicreg' in args **then**
5:         $criterion \leftarrow$ vicreg_criterion
6:     **else**
7:         $criterion \leftarrow$ CosineSimilarity(dim=1)
8:     **end if**
9:     $loss \leftarrow 0$
10:    $num\_masks \leftarrow$ SHAPE(zc)[1]
11:    **for** $i \leftarrow 0$ **to** $num\_masks - 1$ **do**
12:       **for** $j \leftarrow 0$ **to** $num\_masks - 1$ **do**
13:          **if** $i \neq j$ **then**
14:             $zc_i \leftarrow zc[:, i, :]$
15:             $zc_j \leftarrow zc[:, j, :]$
16:             **if** 'vicreg' in args **then**
17:                 $zc\_gather \leftarrow$ FULLGATHERLAYER(zc)
18:                 $zc\_gather_i \leftarrow zc\_gather[:, i, :]$
19:                 $zc\_gather_j \leftarrow zc\_gather[:, j, :]$
20:                 $loss \leftarrow loss +$ CRITERION($zc_i, zc_j, zc\_gather_i, zc\_gather_j$)
21:             **else**
22:                 $loss \leftarrow loss -$ CRITERION(zc_i, zc_j)
23:             **end if**
24:          **end if**
25:       **end for**
26:    **end for**
27:    $loss \leftarrow loss/(\text{num\_masks} \times (\text{num\_masks} - 1))$
28:    **return** $loss$
29: **end function**

Table 7: **Pretraining setting for ViT-B**. All models trained for 600 epochs.

config	value
optimizer	AdamW [48]
epochs	600
learning rate	$8e^{-4}$
weight decay	$(0.04, 0.4)$
batch size	2048
learning rate schedule	cosine decay [49]
warmup epochs	15
encoder arch.	ViT-B
predicted targets	4
predictor depth	6
predictor attention heads	12
predictor embedding dim.	384

Table 8: **Pretraining setting for ViT-L**. Trained for 600 epochs.

config	value
optimizer	AdamW [48]
epochs	600
learning rate	$8e^{-4}$
weight decay	$(0.04, 0.4)$
batch size	2048
learning rate schedule	cosine decay [49]
warmup epochs	15
encoder arch.	ViT-L
predicted targets	4
predictor depth	12
predictor attention heads	16
predictor embedding dim.	384

C.3 Implementation Details

For a comprehensive implementation, we detail the specific settings used during our experiments:

- **Image Preprocessing:** All input images were resized to a resolution of 224×224 pixels.
- **Patch Size:** Following prior work, a patch size of 16×16 pixels was used.

Table 9: **ImageNet-1K image classification.** We perform a linear evaluation on pre-trained ViT-T/16 and ViT-S/16 models for image classification on ImageNet-1K benchmark. We report the Top-1 accuracies using knn, linprob and fine-tune settings to evaluate the quality of pre-trained representations. The best results are indicated in **bold**.

Method	Backbone	Epochs	knn	linprob	fine-tune
I-JEPA [2]	ViT-T/16	100	15.16	22.13	64.46
	ViT-S/16	100	24.36	29.61	76.87
C-JEPA (ours)	ViT-T/16	100	**19.18**	**23.31**	**66.77**
	ViT-S/16	100	**26.85**	**32.35**	**78.17**

Table 10: **COCO object detection and instance segmentation.** We fine-tuned pre-trained ViT-T/16 and ViT-S/16 models to perform COCO object detection and instance segmentation. The AP^{box} and AP^{mask} metrics denote the results of COCO detection, and COCO segmentation, respectively. The best results are indicated in **bold**.

Method	Backbone	Pretrain Epoch	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
I-JEPA [2]	ViT-T/16	100	30.5	52.6	32.8	29.1	49.5	30.6
	ViT-S/16	100	36.9	58.5	39.2	35.2	56.3	37.3
C-JEPA (ours)	ViT-T/16	100	**32.8**	**54.5**	**34.6**	**30.9**	**51.4**	**32.2**
	ViT-S/16	100	**38.5**	**59.7**	**40.3**	**37.1**	**57.5**	**39.2**

- **Vision Transformer Configurations:** We utilized tiny, small, base, and large models of the ViT architecture, with specific settings for each model scale provided in Tables 5, 6, 7, and 8.
- **Optimizer:** The AdamW optimizer was used for training, with specific learning rate schedules and weight decay settings described.
- **Batch Size and Training Schedule:** A default batch size of 2048 was used, with a learning rate and momentum adjustment schedule to optimize convergence.

To validate the efficacy of C-JEPA, we employed several benchmark datasets and tasks:

- **Image Classification on ImageNet-1K [36]**
- **Object Detection and Instance Segmentation on MS-COCO [37]**
- **Semantic Segmentation on ADE20K [38, 39]**
- **Video Object Segmentation on DAVIS-2017 [50]**
- **Low-Level Tasks on Clevr Dataset [45]**

For each task, we adopted established methodologies and frameworks, such as Mask R-CNN [42] for object detection and Semantic FPN and UPerNet [44] for semantic segmentation, ensuring that our findings were comparable with state-of-the-art results.

D   More Experiments

In this part, we provide additional experimental analyses to validate the effectiveness of the proposed C-JEPA framework further. These experiments encompass various downstream tasks, ablation studies, and the impact of different components and coefficients in the model, specifically focusing on the Variance, Covariance, and Invariance terms from VICReg across different configurations of Vision Transformers.

Downstream Results on ViT-T and ViT-S. We evaluated the performance of pre-trained ViT-T/16 and ViT-S/16 models across a series of downstream tasks to assess the quality of the representations learned by the C-JEPA framework. For image classification, we performed linear evaluations using knn, linear probing (linprob), and fine-tuning methods: The results of top-1 accuracies, presented in Table 9, indicate the robustness and quality of the features extracted by our pre-trained models

Table 11: **ADE20K semantic segmentation.** We fine-tuned pre-trained ViT-T/16 and ViT-S/16 models to perform ADE20K semantic segmentation. The mIoU, aAcc, and mAcc metrics denote the results of ADE20K segmentation. The best results are indicated in **bold**.

Method	Backbone	Pretrain Epoch	Semantic FPN			UPerNet		
			mIoU	aAcc	mAcc	mIoU	aAcc	mAcc
I-JEPA [2]	ViT-T/16	100	21.56	72.28	30.26	33.15	75.36	43.05
	ViT-S/16	100	27.17	73.64	37.11	36.98	78.30	46.66
C-JEPA (ours)	ViT-T/16	100	**24.40**	**73.56**	**32.88**	**34.44**	**76.96**	**44.56**
	ViT-S/16	100	**31.36**	**76.27**	**41.34**	**38.68**	**79.07**	**48.86**

Table 12: **Ablation studies on component analysis for faster convergence.** We perform ablation studies on Variance/Covariance and Invariance modules in VICReg using ViT-T/16 model. The best results are indicated in **bold**.

I-JEPA	Variance/Covariance	Invariance	Backbone	Pretrain Epoch	knn	linprob	fine-tune
✓	✗	✗	ViT-T/16	100	15.16	22.15	64.46
✓(mean)	✓(collapse)	✗	ViT-T/16	100	18.75	22.92	66.15
✓	✓	✓	ViT-T/16	100	**19.18**	**23.39**	**66.77**
✓(EMA for collapse)	✗	✓(mean)	ViT-T/16	100	18.57	22.57	65.92

under different evaluation settings. For object detection and instance segmentation, we fine-tuned the pre-trained models to perform tasks on the COCO dataset. The average precision for bounding box detection (AP^{box}) and mask segmentation (AP^{mask}) are reported in Table 10, demonstrating the applicability of C-JEPA pre-trained models in complex vision tasks. The models were also fine-tuned for semantic segmentation tasks, and metrics such as mean Intersection over Union (mIoU), average class accuracy (aAcc), and mean accuracy (mAcc) are detailed in Table 11, highlighting the effectiveness of our approach in handling pixel-level classification tasks.

Variance/Covariance and Invariance in VICReg. To dissect the contributions of individual VICReg components within the C-JEPA, we conducted focused ablation studies using the ViT-T/16 model. For faster convergence, initial ablations in Table 12 analyze the impact of the Variance/Covariance and Invariance modules on speeding up the model training. Further studies on better convergence, shown in Table 13, investigate how these modules contribute to achieving better training dynamics and model performance.

VICReg and Invariance Coefficients. For VICReg coefficients β_{vicreg}, we varied the coefficients to understand their influence on model training and final performance, with results compiled in Table 14. The impact of varying the invariance coefficient β_{sim} is explored to optimize the balance between similarity and variance, with findings in Table 15.

These experiments underline the nuanced contributions of the VICReg components to the overall efficacy of the C-JEPA framework. By systematically varying key parameters and assessing their impacts across a broad spectrum of tasks, we establish a comprehensive understanding of the model's behavior and performance characteristics. This rigorous empirical foundation supports the theoretical enhancements proposed in C-JEPA, affirming its potential to advance the state-of-the-art in unsupervised visual representation learning.

E More Visualizations

In order to provide a comprehensive understanding of the advancements offered by the C-JEPA framework, we include a detailed visual analysis alongside our quantitative evaluations. This section of the appendix is dedicated to showcasing attention visualizations that illustrate the practical impacts of the theoretical improvements achieved through the integration of VICReg components within the C-JEPA framework.

To better understand how the C-JEPA model processes visual information, we utilize attention visualization techniques. These techniques highlight the model's ability to focus on relevant features within an image, which is crucial for tasks involving fine-grained recognition and localization. The following figures 4, 5, 6, 7, and 8 provide a visual comparison between the base I-JEPA and the

Table 13: **Ablation studies on component analysis for better convergence.** We perform ablation studies on Variance/Covariance and Invariance modules in VICReg using ViT-T/16 model. The best results are indicated in **bold**.

I-JEPA	Variance/Covariance	Invariance	Backbone	Pretrain Epoch	knn	linprob	fine-tune
✓	✗	✗	ViT-T/16	600	20.53	26.83	68.68
✓(mean)	✓(collapse)	✗	ViT-T/16	600	25.87	29.65	71.06
✓	✓	✓	ViT-T/16	600	**27.61**	**31.86**	**71.52**
✓(EMA for collapse)	✗	✓(mean)	ViT-T/16	600	24.95	27.92	70.58

Table 14: **Ablation studies on component analysis for VICReg coefficients.** We perform ablation studies on VICReg coefficients $(\cdot)$ using ViT-T/16 model. The best results are indicated in **bold**.

VICReg β_{vicreg}	Backbone	Pretrain Epoch	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.0001	ViT-T/16	100	43.6	43.0	44.3
0.0002	ViT-T/16	100	45.7	45.3	46.1
0.0005	ViT-T/16	100	47.3	46.8	47.8
0.001	ViT-T/16	100	**50.2**	**49.8**	**50.6**
0.005	ViT-T/16	100	46.0	45.4	46.6
0.01	ViT-T/16	100	31.9	33.1	30.7
0.02	ViT-T/16	100	29.5	30.6	28.4
0.05	ViT-T/16	100	24.4	25.8	23.1
0.1	ViT-T/16	100	Collapsing		

enhanced C-JEPA models. For each sample, columns denote the original image, attention maps from the target encoder in I-JEPA, attention maps from the target encoder in our C-JEPA, and attention maps from the context encoder in our C-JEPA.

These attention visualizations demonstrate the qualitative improvements in the C-JEPA's ability to focus and contextualize features compared to the base I-JEPA model. The C-JEPA model produces more focused and contextually relevant attention maps, which is a direct result of the enhanced model architecture that incorporates VICReg's variance, covariance, and invariance regularization strategies. The enhanced attention capabilities of C-JEPA facilitate better feature extraction from complex visual inputs, which is critical for improving performance on downstream tasks. The more detailed and precise attention maps correlate strongly with the theoretical enhancements integrated into the C-JEPA framework, such as improved handling of variance and invariance across different visual domains.

The attention visualizations provided in this section validate the theoretical improvements posited in the earlier sections of this paper. They offer qualitative evidence that supports the quantitative results, underscoring the benefits of the C-JEPA model in capturing more nuanced and robust representations. Overall, these visualizations highlight the C-JEPA's ability to maintain stable and meaningful patch-level representations across a variety of image contexts, marking a significant advancement over its predecessor.

Table 15: **Ablation studies on component analysis for invariance coefficients.** We perform ablation studies on invariance coefficients (·) using ViT-T/16 model. The best results are indicated in **bold**.

I-JEPA	Variance/Covariance	Invariance	Backbone	Pretrain Epoch	knn	linprob	fine-tune
✓	✗	✓(25)	ViT-T/16	100	18.03	21.72	65.68
✓	✗	✓(20)	ViT-T/16	100	18.35	22.25	65.73
✓	✗	✓(15)	ViT-T/16	100	**18.57**	**22.53**	**65.92**
✓	✗	✓(10)	ViT-T/16	100	18.45	22.36	65.81

Figure 4: **Qualitative visualization of learned attention maps using ViT-B/16 model.** Columns for each sample denote the original image, attention maps from the target encoder in I-JEPA [2], attention maps from the target encoder in our C-JEPA, and attention maps from the context encoder in our C-JEPA. Our C-JEPA achieves much better attention maps.
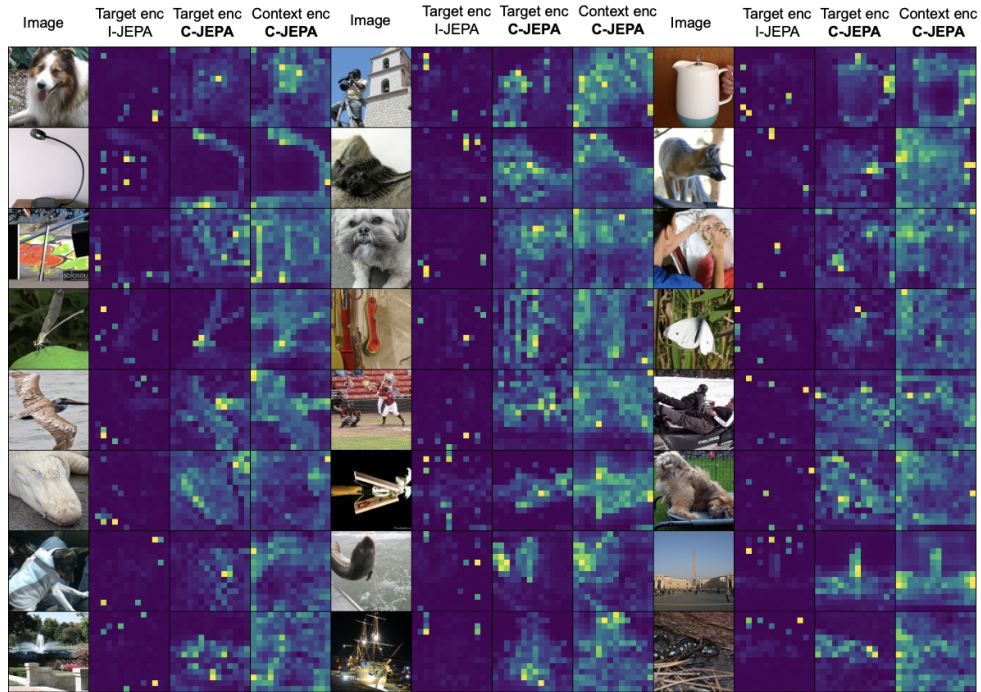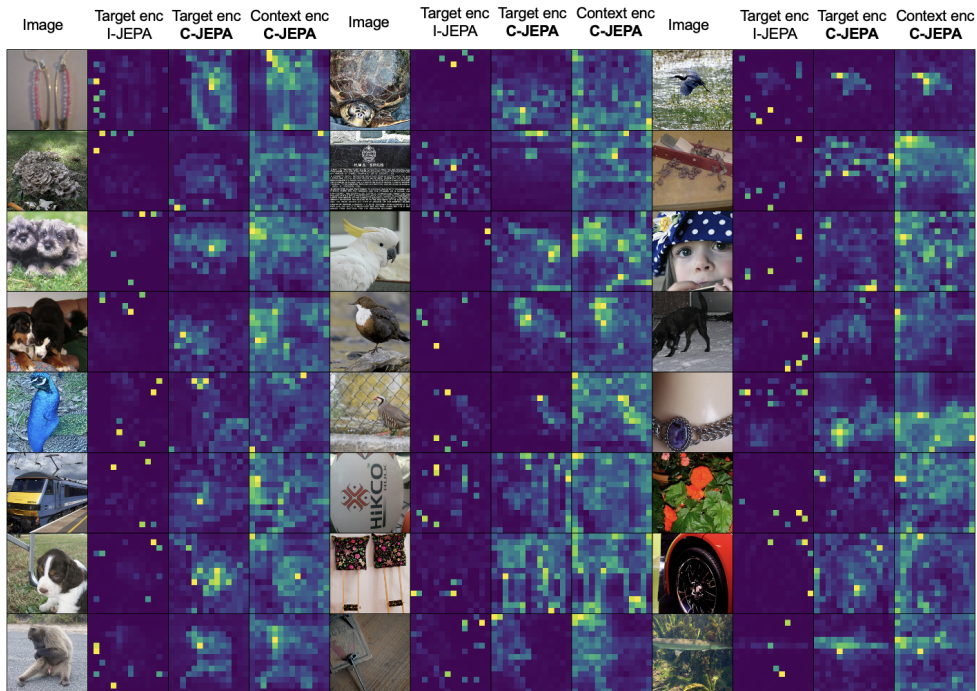
Figure 5: **Qualitative visualization of learned attention maps using ViT-B/16 model.** Columns for each sample denote the original image, attention maps from the target encoder in I-JEPA [2], attention maps from the target encoder in our C-JEPA, and attention maps from the context encoder in our C-JEPA. Our C-JEPA achieves much better attention maps.

Figure 6: **Qualitative visualization of learned attention maps using ViT-B/16 model.** Columns for each sample denote the original image, attention maps from the target encoder in I-JEPA [2], attention maps from the target encoder in our C-JEPA, and attention maps from the context encoder in our C-JEPA. Our C-JEPA achieves much better attention maps.
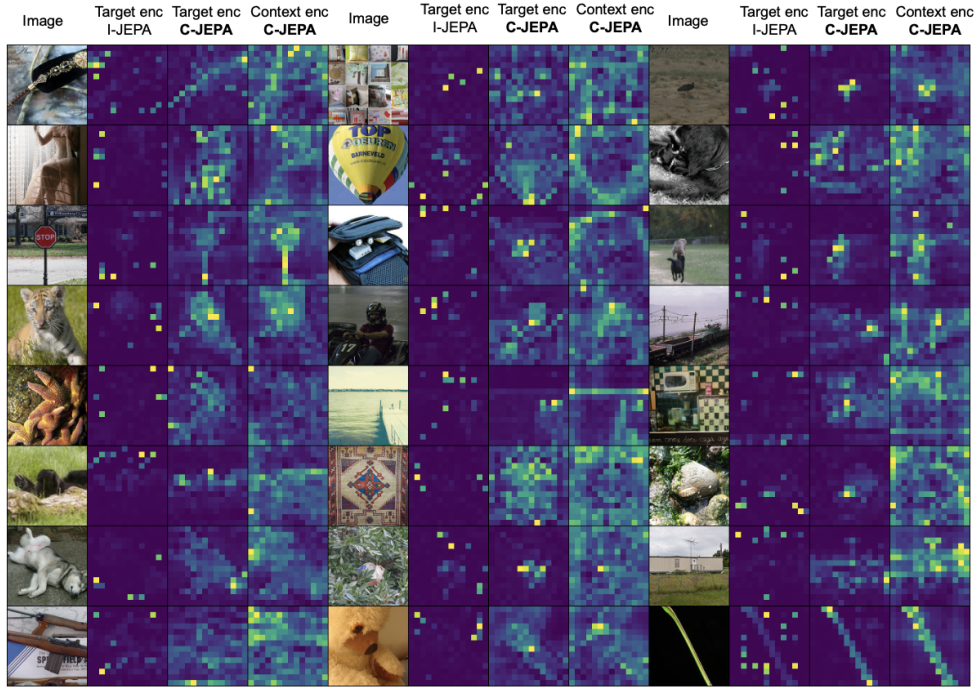
Figure 7: **Qualitative visualization of learned attention maps using ViT-B/16 model.** Columns for each sample denote the original image, attention maps from the target encoder in I-JEPA [2], attention maps from the target encoder in our C-JEPA, and attention maps from the context encoder in our C-JEPA. Our C-JEPA achieves much better attention maps.
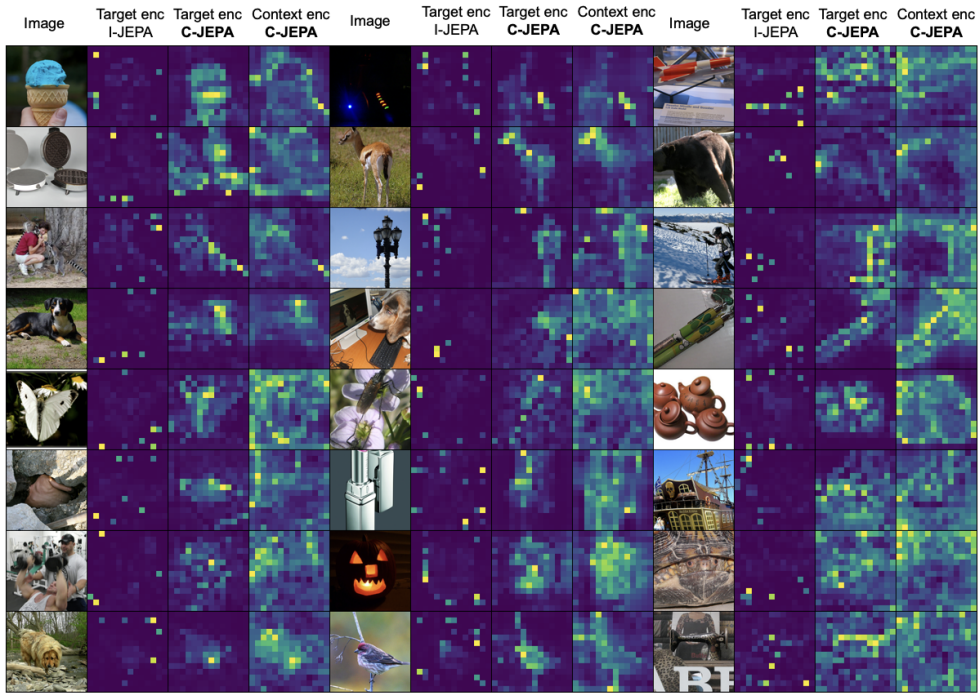

Figure 8: **Qualitative visualization of learned attention maps using ViT-B/16 model.** Columns for each sample denote the original image, attention maps from the target encoder in I-JEPA [2], attention maps from the target encoder in our C-JEPA, and attention maps from the context encoder in our C-JEPA. Our C-JEPA achieves much better attention maps.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: See Section 1

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See our discussion in the limitation section.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [NA]

Justification: We do not have theory assumptions and proofs.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: See Section 4.1

Guidelines:

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

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

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

Answer: [Yes]

Justification: See the supplemental material.

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: See Section 4.1

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [Yes]

Justification: See Section 4.2

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

Answer: [Yes]

Justification: See Section 4.1

Guidelines:

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

9. **Code Of Ethics**

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

Answer: [Yes]

Justification: We have conformed in every respect with the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader Impacts**

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

Answer: [Yes]

Justification: See our discussions in the broader impact section

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: We do not use data or models that have a high risk for misuse.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: See Section 4.1

Guidelines:

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

13. **New Assets**

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

Answer: [NA]

Justification: We do not have new assets.

Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

Answer: [NA]

Justification: We do not have experiments with human subjects.

Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

Answer: [NA]

Justification: We do not have potential risks incurred by study participants.

Guidelines:

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