# Adaptive Contrastive Masked Autoencoders for Structured Representation Learning

Karan Singh
Stanford University
karanps@stanford.edu

## Abstract

*Masked Autoencoders (MAE) have emerged as a workhorse of self-supervised representation learning for Vision Transformers. However, their pixel-reconstruction objective leaves the global layout of the latent space largely unconstrained. In practice, this can entangle task-relevant and spurious factors, lead to slow convergence on downstream tasks, and diminish the separability of semantic categories. We propose* ***Adaptive Contrastive Masked Autoencoding (AC-MAE)****, a simple yet effective extension that blends standard MAE reconstruction with a supervised contrastive loss applied to visible patch tokens. A ramp-up schedule gradually increases the contrastive weight as features mature, mitigating early collapse and enabling semantically meaningful structure to emerge. Crucially, our method introduces only coarse class labels during pretraining, reducing the need for fine-grained annotation while still improving downstream performance. On CIFAR-100, AC-MAE yields substantial gains in classification accuracy—up to 12.2% in low-data regimes—by embedding superclass information during pretraining. We further demonstrate that this approach is well-suited to neuroimaging, where demographic attributes like sex can provide additional supervision during pretraining. Applying AC-MAE to resting-state fMRI data, we observe that incorporating sex labels during pretraining leads to robust clustering and strong downstream linear-probing performance.*

## 1. Introduction

Self-supervised learning (SSL) has emerged as a powerful paradigm for visual representation learning, reducing dependence on labeled data by exploiting structure inherent in the input itself. Recent advances have shown that carefully designed pretext tasks, such as predicting missing patches, solving jigsaw puzzles, or contrasting augmented views, can produce representations that rival or even surpass those learned via full supervision [9]. These approaches have been especially impactful in computer vision, where annotations are expensive and high-dimensional input structure makes supervision bottlenecks acute.

A common family of SSL methods leverages masked prediction objectives in Vision Transformers (ViTs). Inspired by the success of BERT in natural language processing, Masked Autoencoders (MAEs) [7] train models to reconstruct masked image patches from the remaining visible context. MAEs are conceptually simple, scalable, and highly effective across standard benchmarks. Their modular encoder–decoder structure enables efficient training, and their ability to learn from raw pixels without labels makes them well-suited to a variety of real-world settings where annotations are limited or unavailable.

At the same time, a separate line of work in contrastive learning has demonstrated that global semantic structure can emerge when models are trained to bring similar samples closer and push others apart in latent space. Contrastive methods like SimCLR [2], MoCo [8], and SupCon [10] learn discriminative embeddings by encouraging consistency across views or labels, often leading to more separable and task-aligned features. While masked autoencoding emphasizes fine-grained local reconstruction, contrastive learning imposes a strong global constraint on feature geometry.

In this work, we explore whether these two core objectives in self-supervised learning—masked reconstruction and contrastive alignment—can be effectively combined to improve the structure of learned representations. While masked autoencoding (MAE) enables models to learn from raw inputs without labels, it provides no explicit incentive for the latent space to reflect task-relevant or semantically meaningful organization. Our central hypothesis is that lightly injecting supervised contrastive signals during MAE pretraining can guide the latent geometry without disrupting the benefits of reconstruction-based learning. Our motivation is twofold: (1) in general vision tasks, superclass labels represent a coarse but meaningful signal that is easier to obtain than fine-grained annotations and may help impose global structure early on; and (2) in medical and neuroimag-

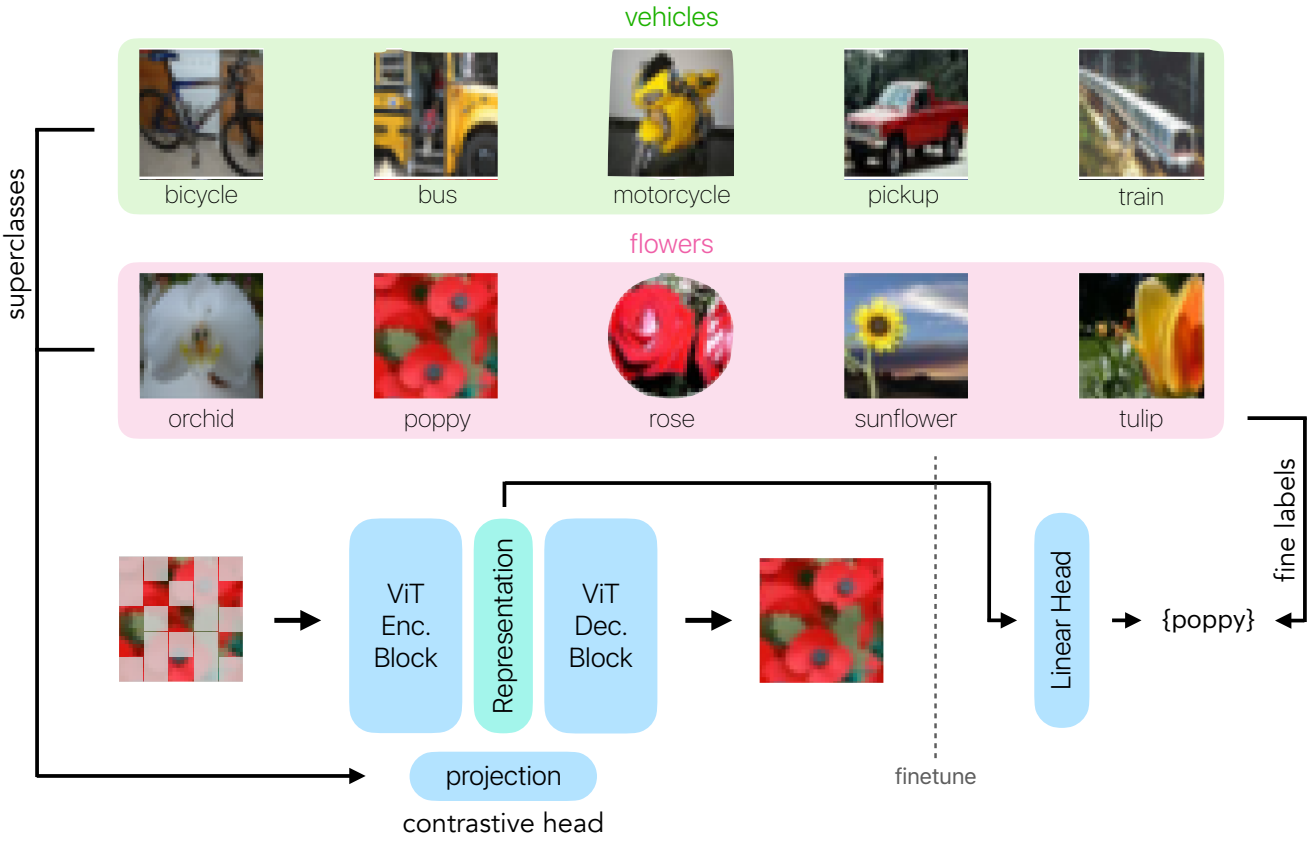


Figure 1. **AC-MAE** introduces coarse label supervision during masked autoencoder (MAE) pretraining via a contrastive head applied to the encoder representations. Positive pairs are formed using shared superclass labels (e.g., "flowers" or "vehicles"), guiding global structure in the learned representations. Following convention, in the separate fine-tuning stage, a linear classifier is added on top of the learned representations, and the encoder is finetuned to predict fine-grained labels (e.g., "poppy"). We find that adding our contrastive component improves downstream classification in low-data regimes, reducing the reliance on expensive fine-grained annotation during pretraining.

ing domains, where downstream tasks often benefit from demographic or clinical information, leveraging sparse auxiliary labels (e.g., age or sex) during pretraining may help bridge the gap between representation learning and deployment. Our main contributions can be summarized as follows:

- We propose **Adaptive Contrastive Masked Autoencoder (AC-MAE)**, a lightweight extension to MAE that incorporates a supervised contrastive loss on visible patch tokens using shared coarse labels (e.g., superclass, sex), with a ramp-up schedule for the contrastive weight.
- On CIFAR-100, we demonstrate that incorporating superclass labels during pretraining yields consistent gains in classification accuracy, particularly in low-data regimes (up to 12.2%).
- We extend AC-MAE to resting-state fMRI data and show that incorporating demographic labels (i.e. sex) during pretraining improves downstream classification of sex and Parkinson's disease.

## 2. Related Work

**Masked Autoencoders.** Masked image modeling has emerged as a powerful self-supervised paradigm for vision Transformers. In particular, Masked Autoencoders (MAE) train a ViT encoder-decoder by randomly masking a high fraction of input patches and reconstructing the missing pixels, using an asymmetric architecture that enables efficient learning of transferable representations [7]. Numerous extensions build on this idea – for example, CrossMAE [5] reduces compute by replacing the full decoder with a cross-attention mechanism from masked to visible tokens.. Another variant, MixMAE [12], replaces masked patches with patches from a different image to form a mixed input and then decodes both original images, greatly accel-

erating pre-training (by removing mask tokens) while maintaining strong performance on downstream tasks.

**Contrastive self-supervised learning.** Contrastive learning approaches have driven major advances in unsupervised visual representation learning. SimCLR pioneered an instance discrimination framework that pulls together two augmented views of the same image while pushing apart other images, requiring very large batch training but no memory bank or specialized architectures [2]. Momentum Contrast (MoCo) tackled the memory/batch size issue by maintaining a moving-averaged encoder and a queue of negative samples to serve as a consistent dictionary for contrastive learning [8]. More recent methods eliminate explicit negatives: BYOL showed that a student network can learn from a momentum teacher's representation of the same image under a different augmentation, matching state-of-the-art results without any negative pairs [6]. Likewise, Barlow Twins avoids collapse by making the cross-correlation matrix between two distorted view embeddings as close to identity as possible, thus maximizing agreement on individual features while minimizing redundancy between features [18]. In addition, a Supervised Contrastive Learning (SupCon) objective extends these ideas to labeled data by pulling together embeddings of samples with the same class label and pushing away different classes, yielding improvements over standard cross-entropy on ImageNet classification [10].

A number of works have also combined masked modeling with contrastive or alignment losses to capture both low-level reconstruction cues and high-level semantic structure. For example, CAN fuses Contrastive Learning with Masked Autoencoding (and even integrates a diffusion-model noise prediction term) in a single framework, masking 50% of patches in both views; this hybrid approach outperforms its separate MAE and SimCLR components on transfer learning tasks while also being more efficient than pure contrastive pre-training [13]. Another notable approach, iBOT, performs masked image prediction via self-distillation: a ViT encoder is trained to predict the representations of an online teacher network for masked patches (and class tokens), effectively unifying a generative masked loss with a feature alignment objective and delivering state-of-the-art results on ImageNet and dense prediction benchmarks [19]. These efforts demonstrate that combining reconstruction-based and contrastive objectives can yield representations with improved diversity and discrimination compared to either alone.

Prior work has also explored incorporating class labels or auxiliary information into masked autoencoder training in supervised or semi-supervised settings. Multimodality-guided Visual Pre-training (MVP) showed that using a pre-trained multimodal encoder to provide semantic targets for masked image modeling can significantly boost representation quality – specifically, MVP replaces the BERT-style tokenizer in a ViT MAE with the vision branch of CLIP (pre-trained on image–text data), injecting high-level semantic guidance into the reconstruction task and greatly improving downstream accuracy (e.g. +6.8 mIoU on ADE20K over prior MIM) [16]. In a semi-supervised context, Semi-MAE introduced a parallel masked-autoencoder branch into a standard ViT training pipeline: the model learns from unlabeled images by reconstructing masked patches with a high mask ratio, alongside a supervised loss on the small labeled subset – an approach that achieved state-of-the-art ImageNet results using only 10% of labels [17]. These techniques illustrate the benefit of leveraging external signals (whether from pretrained models or limited labels) to enrich the representations learned by masked autoencoders.

**Self-Supervised Learning for fMRI.** Functional neuroimaging data (fMRI) has recently become a testbed for self-supervised learning methods inspired by computer vision. Some works have applied contrastive learning to fMRI representations, finding that maximizing agreement between different views or augmentations of brain data can reduce overfitting and stabilize features in low-data regimes [15]. On a larger scale, BrainLM was proposed as a foundation model for fMRI, using a masked prediction objective on 6,700 hours of recordings to learn general-purpose brain representations; notably, BrainLM's embeddings can be fine-tuned to predict subject-specific clinical variables (like age or mental health scores) and even used in a zero-shot manner to detect intrinsic functional networks from raw fMRI data [1]. Most recently, Brain-JEPA brought the joint-embedding predictive architecture to fMRI: instead of reconstructing input signals, it trains a vision-transformer model to predict latent representations of masked-out spatiotemporal brain patches, aided by brain-specific innovations (a functional coordinate positional encoding and tailored spatiotemporal masking) for better alignment with neuroanatomy [4]. This JEPA-style approach achieves state-of-the-art results on downstream brain activity prediction tasks (covering demographics and clinical diagnoses) and shows strong cross-cohort generalization, surpassing earlier large-scale fMRI models.

**Our contributions.** While previous works have explored combining contrastive and masked modeling objectives or incorporating supervision in semi-supervised contexts, our approach is distinct in several key ways. First, we introduce a supervised contrastive loss directly into the MAE pretraining stage, but apply it only to the **visible patch tokens**—preserving the integrity of the masked reconstruction task while softly guiding the global structure of the representation space. However, unlike methods that rely

on full supervision or pretraining with auxiliary modalities (e.g., MVP or Semi-MAE), our approach leverages coarse labels (e.g., superclasses in image datasets or demographic variables in fMRI) that are significantly easier and cheaper to obtain than fine-grained annotations. We show that our method, Adaptive Contrastive Masked Autoencoding (AC-MAE), can inject minimal but meaningful supervision to yield better-structured embeddings, especially in low-data regimes, and improve downstream performance across both vision and neuroimaging domains.

# 3. Method

## 3.1. Datasets

**CIFAR-100 with Hierarchical Superclasses.** We utilize CIFAR-100 [11] as our primary dataset, leveraging its built-in hierarchical structure of 20 semantically meaningful superclasses (e.g., "aquatic mammals," "large carnivores," "vehicles"). Each superclass contains 5 fine-grained classes, providing natural groupings — this hierarchical structure enables us to apply contrastive learning at the superclass level while maintaining fine-grained reconstruction targets. We apply standard data augmentation including random horizontal flips, rotation (±15°), and color jittering (brightness/contrast/saturation ±0.2, hue ±0.1) during training.

**fMRI with Demographic Labels.** We additionally evaluate our method on resting-state functional magnetic resonance imaging (rs-fMRI) data collected from the public Human Connectome Project Young-Adults (HCP-YA) [14] dataset. rs-fMRI measures spontaneous fluctuations in brain oxygenation over time, producing 4D data volumes (3D spatial brain scans over time) that serve as a proxy for brain activity. These recordings are high-dimensional and noisy, making them challenging for direct training with deep models. To reduce this complexity, we follow standard practice and apply the DiFuMo atlas [3], which parcellates the brain into 1024 distinct functional regions of interest (ROIs). We then extract low-resolution time series for each ROI, resulting in 2D matrices (ROI × time) for each sample. For pretraining, we sample 64-timestep segments from 1,004 training samples in HCP-YA, reserving 10% each for validation and held-out testing. Importantly, each sample is associated with auxiliary demographic labels (e.g., sex), which we use here for both contrastive alignment and downstream evaluation.

## 3.2. Architecture

**CrossMAE Architecture.** Our model extends the Cross-Attention Masked Autoencoder (CrossMAE) architecture [5], which differs from standard MAE in its decoder design. While MAE concatenates learnable mask tokens and processes them through heavy self-attention layers, CrossMAE uses a single shared mask token that attends to frozen encoder features via lightweight cross-attention. This reduces decoder parameters and computational cost while maintaining reconstruction quality. We implement the optional weighted feature map aggregation as recommended in the original implementation, where decoder layers can attend to weighted combinations of encoder features from different depths, providing richer cross-attention interactions.

**Contrastive Projection Head.** We augment the CrossMAE encoder with a projection head for contrastive learning, consisting of two linear layers (embed_dim $\rightarrow$ embed_dim $\rightarrow$ 256) with ReLU activation. This head processes the CLS token embeddings from the encoder to produce normalized 256-dimensional representations for contrastive learning, following established practices in self-supervised learning. On top of these contrastive embeddings, we implement a temperature-scaled supervised contrastive loss that leverages superclass labels. For a batch of normalized embeddings $\{\mathbf{z}_i\}_{i=1}^{N}$ with superclass labels $\{c_i\}_{i=1}^{K}$ where $K$ is the number of superclasses, we define positives for sample $i$ as $P(i) = \{j \neq i \mid c_j = c_i\}$. The contrastive loss is:

$$\mathcal{L}_{\text{SCL}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p/\tau)}{\sum_{a\neq i}\exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}$$

where $\tau = 0.2$ is the temperature parameter. This formulation encourages embeddings from the same superclass to be similar while pushing apart embeddings from different superclasses.

**Contrastive Loss Delay and Warmup.** We implement a ramped training schedule that accounts for the quality of learned representations over time. The contrastive loss is delayed for the first $D$ epochs (typically $D = E/4$ where $E$ is the total number of epochs) to allow basic feature learning through the masked autoencoding objective, then gradually ramped up over $W$ warmup epochs using a linear schedule:

$$\lambda_e = \begin{cases} 0 & \text{if } e < D \\ \lambda_{\max}\frac{e-D}{W} & \text{if } D \leq e < D+W \\ \lambda_{\max} & \text{if } e \geq D+W \end{cases}$$

where $\lambda_{\max}$ is the final contrastive weight, $D$ is the delay period, and $W$ epochs is the warmup duration. This schedule prevents applying strong alignment pressure to early, noisy embeddings while allowing robust representation space development.

| Data Fraction | Baseline (%) | Contrastive (%) | Gain (%) |
|---|---|---|---|
| 0.01 | $27.39 \pm 0.93$ | $\mathbf{39.62 \pm 0.75}$ | $+12.23$ |
| 0.02 | $41.84 \pm 0.69$ | $\mathbf{50.54 \pm 0.93}$ | $+8.70$ |
| 0.05 | $54.80 \pm 0.27$ | $\mathbf{60.60 \pm 0.62}$ | $+5.80$ |
| 0.10 | $63.00 \pm 0.25$ | $\mathbf{66.55 \pm 0.37}$ | $+3.55$ |
| 0.20 | $70.44 \pm 0.15$ | $\mathbf{72.02 \pm 0.29}$ | $+1.58$ |
| 0.50 | $78.50 \pm 0.17$ | $78.41 \pm 0.15$ | $-0.09$ |
| 1.00 | $\mathbf{83.81 \pm 0.13}$ | $82.83 \pm 0.25$ | $-0.98$ |

Table 1. Classification accuracy (%) on CIFAR-100 across varying fractions of labeled training data, comparing CrossMAE (baseline) and our contrastive variant (AC-MAE). Each value reflects the best top-1 accuracy achieved after fine-tuning, averaged over 3 seeds with standard deviation. AC-MAE provides significant gains in low-data regimes—most notably a +12.2% absolute improvement at 1% data—demonstrating that contrastive alignment with coarse supervision improves sample efficiency. At higher data fractions, gains diminish and eventually reverse, suggesting that the benefits of contrastive pretraining are most prominent when labeled data is limited.

**Combined Loss Function.** The total training loss combines reconstruction and contrastive objectives:

$$\mathcal{L} = (1 - \lambda_e)\mathcal{L}_{\text{rec}} + \lambda_e \mathcal{L}_{\text{SCL}}$$

where $\mathcal{L}_{\text{rec}}$ is the standard MAE reconstruction loss (MSE between predicted and actual pixel values, optionally normalized per patch).

## 4. Experiments

**Experimental Setup.** We evaluate our approach on two domains: image classification with CIFAR-100 and representation learning on resting-state fMRI. For CIFAR-100, we use the standard `ViT-Base` architecture with patch size 16, an embedding dimension of 384, 12 transformer encoder blocks, and 8 decoder blocks. For our contrastive loss, we use a final contrastive weight $\lambda_{\text{max}} = 0.05$, with a contrastive delay $D = 100$ and contrastive warmup $W = 100$. For our fMRI data, we adopt a lightweight ViT variant with an embedding dimension of 128 and 4 encoder and decoder layers, which we found more appropriate given the limited dataset size. We additionally use a much higher contrastive weight of $\lambda_{\text{max}} = 0.7$, as we observed very quick saturation of the reconstruction loss and therefore required a stronger contrastive signal to align the embeddings to our coarse labels.

In both settings, we randomly mask 75% of patch tokens and train using the Adam optimizer with a learning rate of $1 \times 10^{-4}$. We use a batch size of 128, with early stopping based on validation loss. Our learning rate follows a warmup-stable-decay schedule (1000 steps of linear warmup and cosine decay), with a base learning rate of $1.5 \times 10^{-4}$ scaled by batch size, and we train for 100-400 epochs depending on the dataset size.

**Fine-Tuning.** For CIFAR-100, we fine-tune the entire pretrained model by adding a single linear classification head on top of the [CLS] token and train all parameters end-to-end. For fMRI, we adopt a linear probing setup, freezing the encoder and training a three-layer MLP (embed_dim $=$ 512) on top of the [CLS] token embedding to predict binary sex labels.

**Baselines.** We compare our proposed method (AC-MAE) against a reconstruction-only CROSSMAE baseline. Both models share the same encoder architecture, masking strategy, and training schedule in each domain to isolate the effect of our supervised contrastive component. For CIFAR-100, both methods are pretrained without access to fine class labels, using only superclasses for contrastive alignment when applicable.

**Metrics.** For CIFAR-100, we report top-1 classification accuracy for various subsets of the full dataset (1%, 2%, 5%, 10%, 20%, 50%, and 100%). For the binary fMRI sex classification task, we use balanced accuracy to account for class imbalance. In addition, we visualize the learned embeddings via t-SNE and report silhouette scores as a quantitative measure of cluster separation in embedding space.

## 5. Results

We evaluate AC-MAE against a reconstruction-only CrossMAE baseline across both visual and neuroimaging domains. On CIFAR-100, we assess transfer performance via top-1 and top-5 classification accuracy after fine-tuning. As shown in Table 1, contrastive pretraining significantly boosts accuracy in low-data regimes: for instance, at 1% of training data, AC-MAE improves top-1 accuracy by +12.2%, with diminishing returns as data increases. On the full dataset, baseline performance slightly surpasses AC-MAE (83.81% vs. 82.83%), suggesting that contrastive alignment primarily benefits sample-efficient learning. For reference, one-shot and five-shot accuracy after fine-tuning

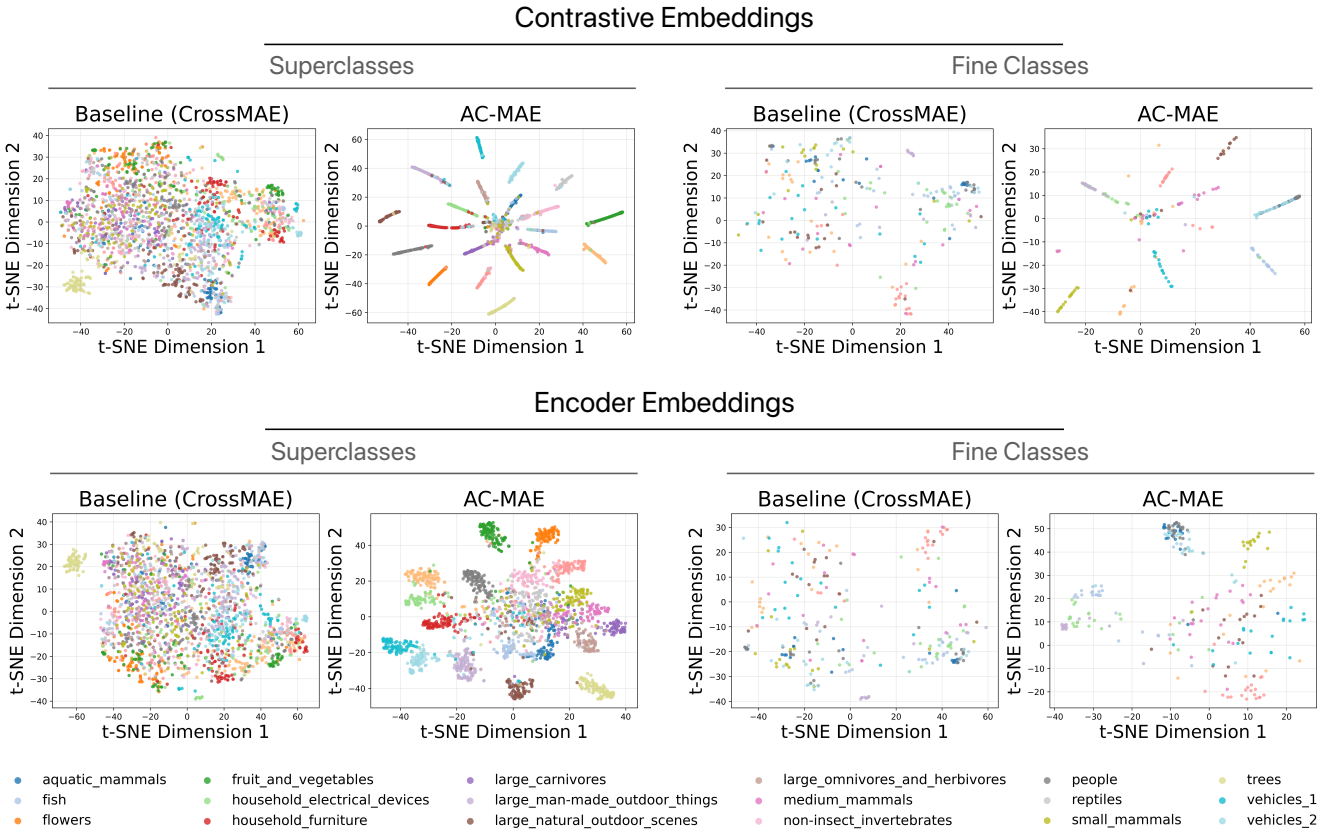


Figure 2. **AC-MAE improves semantic structure in the learned representations.** We visualize t-SNE projections of embeddings from CrossMAE (baseline) and AC-MAE on CIFAR-100, colored by either superclass (left) or fine-grained class (right). **Top:** Contrastive projections (after the projection head) show that AC-MAE produces highly separable clusters aligned with superclass labels (0.1745 silhouette score versus -0.0381 for the baseline), even without using fine-grained class supervision. This structure is preserved down to the fine-class level, suggesting improved global organization. **Bottom:** Encoder embeddings (pre-projection) show similar trends: AC-MAE yields cleaner and more distinct clusters (0.0917 silhouette score versus -0.381 for the baseline), especially for superclasses, while CrossMAE embeddings remain diffuse. These results highlight that lightly injecting coarse label information during pretraining encourages more semantically meaningful representations.

on 100% of CIFAR-100 are 83.91% and 95.67% for AC-MAE, compared to 85.34% and 96.31% for CrossMAE.

| Model | Balanced Accuracy (%) | F1 Score (%) |
|---|---|---|
| Baseline (CLS) | 59.13 | 61.55 |
| **AC-MAE (CLS)** | **96.83** | **97.14** |

Table 2. Test set performance on binary sex classification using resting-state fMRI representations from CrossMAE (baseline) and our proposed AC-MAE model. We report balanced accuracy and F1 score using CLS token embeddings with a frozen encoder and a lightweight MLP classifier (linear probing). AC-MAE achieves a substantial improvement over the reconstruction-only baseline, boosting balanced accuracy from 59.13% to 96.83% and F1 score from 61.55% to 97.14%.

Beyond classification, we assess the semantic quality of learned representations. Figure 2 visualizes t-SNE projections of embeddings for both methods. AC-MAE forms tighter clusters aligned with superclass and class labels, reflected quantitatively by improved silhouette scores (contrastive: 0.1745 vs. baseline: -0.0381 for projection head; 0.0917 vs. -0.381 for encoder output).

On fMRI, we evaluate representation quality via sex classification using linear probing (Figure 3). As shown in Table 2, AC-MAE achieves a test F1 score of 97.14% and balanced accuracy of 96.83% using CLS token representations, a dramatic improvement over the baseline (61.55% F1, 59.13% balanced accuracy), and most notably, much higher than prior models like BrainLM [1] or Brain-JEPA

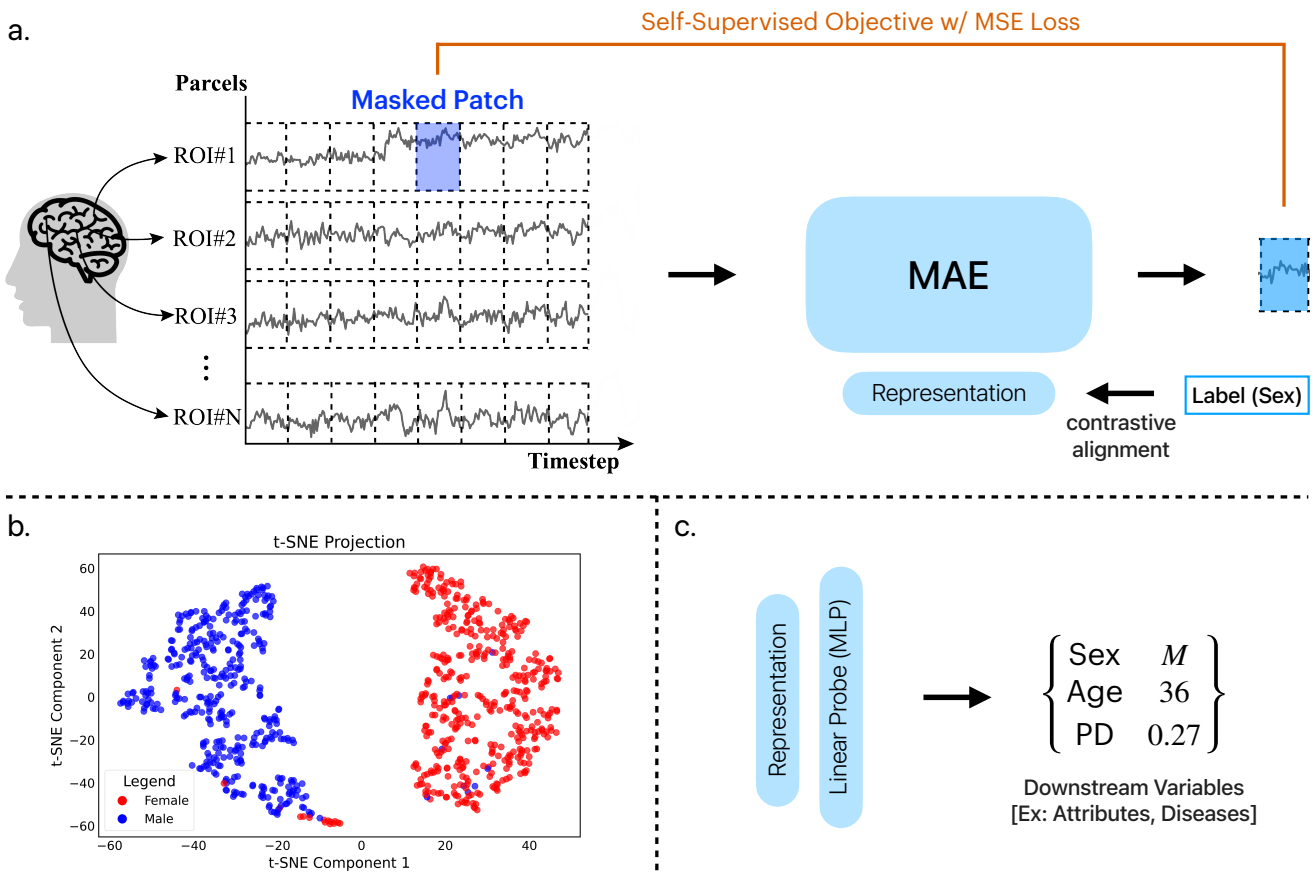


Figure 3. **AC-MAE with fMRI. (a)** We apply masked autoencoding to resting-state fMRI data by predicting randomly masked spatiotemporal patches. Simultaneously, we introduce a contrastive loss to encourage alignment of representations based on coarse demographic attributes (e.g., sex). **(b)** This supervision robustly encourages the embedding space to reflect the demographic signal, as visualized by a t-SNE projection colored by patients' sex. **(c)** While we evaluate the representations on sex classification via linear probing, the same alignment strategy could support a range of downstream tasks involving clinical or demographic variables, including cognitive traits or disease risk.

[4], though these results are not directly comparable due to differences in cohort demographics (e.g., age). Though we omit these results for brevity, this improvement holds even when using average pooled embeddings. Figure 3b further highlights this effect: the contrastive model's embeddings are clearly separable by sex in t-SNE space, demonstrating successful alignment with demographic attributes through our pretraining objective.

These results collectively demonstrate that injecting coarse supervision via contrastive loss during masked autoencoding enhances both the structure and downstream utility of learned representations, especially under limited data.

## 6. Discussion and Future Work

**Conclusion.** We introduce AC-MAE, a simple yet effective extension of masked autoencoders that incorporates weak supervision through a supervised contrastive loss during pretraining. By aligning learned representations with coarse labels such as superclasses (in CIFAR-100) or demographic attributes (in fMRI), our method improves downstream performance in low-data regimes and enhances the semantic structure of the embedding space. On fMRI, contrastive alignment dramatically boosts linear probing accuracy for sex classification—achieving over 96% balanced accuracy, compared to under 60% for the baseline. On CIFAR-100, contrastive pretraining improves top-1 accuracy by over 12% when training on 1% of the data, highlighting its effectiveness for sample-efficient learning.

**Limitations.** Despite strong performance improvements in our experiments, we could not evaluate our method on richer datasets like ImageNet due to computational and time constraints. Additionally, in our applications to fMRI, our method currently relies on a single attribute (sex) for supervision and was only tested on relatively small datasets (especially compared to prior work in the field). A larger set of both datasets and downstream tasks would allow us to evaluate whether contrastive alignment generalizes beyond sex classification, particularly to cognitive phenotypes and neurological disease risk. Finally, while we demonstrate generality across vision and neuroimaging domains, the full potential of contrastive alignment for large-scale pretraining remains underexplored.

**Future Work.** Future extensions may focus on refining the contrastive objective to more directly shape embedding geometry and semantic disentanglement. This includes exploring alternative contrastive formulations such as class-conditional variants, introducing harder or curriculum-based negative sampling strategies, and incorporating multi-head projections for disentangling different attribute axes.

## References


[1] J. O. Caro, A. H. d. O. Fonseca, C. Averill, S. A. Rizvi, M. Rosati, J. L. Cross, P. Mittal, E. Zappala, D. Levine, R. M. Dhodapkar, I. Han, A. Karbasi, C. G. Abdallah, and D. van Dijk. BrainLM: A foundation model for brain activity recordings, 9 2023. 3, 6

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. 2 2020. 1, 3

[3] K. Dadi, G. Varoquaux, A. Machlouzarides-Shalit, K. J. Gorgolewski, D. Wassermann, B. Thirion, and A. Mensch. Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage*, 221:117126, 11 2020. 4

[4] Z. Dong, R. Li, Y. Wu, T. T. Nguyen, J. S. X. Chong, F. Ji, N. R. J. Tong, C. L. H. Chen, and J. H. Zhou. Brain-JEPA: Brain Dynamics Foundation Model with Gradient Positioning and Spatiotemporal Masking. 9 2024. 3, 7

[5] L. Fu, L. Lian, R. Wang, B. Shi, X. Wang, A. Yala, T. Darrell, A. A. Efros, and K. Goldberg. Rethinking Patch Dependence for Masked Autoencoders. 1 2024. 2, 4

[6] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised Learning. 6 2020. 3

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. 11 2021. 1, 2

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. 11 2019. 1, 3

[9] L. Jing and Y. Tian. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. 2 2019. 1

[10] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised Contrastive Learning. 4 2020. 1, 3

[11] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009. 4

[12] J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li. MixMAE: Mixed and Masked Autoencoder for Efficient Pretraining of Hierarchical Vision Transformers. 5 2022. 2

[13] S. Mishra, J. Robinson, H. Chang, D. Jacobs, A. Sarna, A. Maschinot, and D. Krishnan. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. 10 2022. 3

[14] D. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. Curtiss, S. Della Penna, D. Feinberg, M. Glasser, N. Harel, A. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. Petersen, F. Prior, B. Schlaggar, S. Smith, A. Snyder, J. Xu, and E. Yacoub. The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231, 10 2012. 4

[15] X. Wang, L. Yao, I. Rekik, and Y. Zhang. Contrastive Graph Learning for Population-based fMRI Classification. 3 2022. 3

[16] L. Wei, L. Xie, W. Zhou, H. Li, and Q. Tian. MVP: Multimodality-guided Visual Pre-training. 3 2022. 3

[17] H. Yu, K. Zhao, and X. Xu. Semi-MAE: Masked Autoencoders for Semi-supervised Vision Transformers. 1 2023. 3

[18] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. 3 2021. 3

[19] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. 11 2021. 3