SELF-SUPERVISED IMAGE FRAGMENT MATCHING VIA CONTRASTIVE REPRESENTATION LEARNING

1 PROBLEM FORMULATION AND EVALUATION METRICS

We address the task of reconstructing original images from an unordered collection of fragments, without relying on spatial or positional priors. To this end, we train a self-supervised embedding function that maps fragments from the same image to nearby points in representation space, while pushing fragments from different images apart. Clustering in the learned embedding space then serves to recover the original image groupings. Our dataset comprises millions of images spanning 1,000 semantic categories. However, labels are not used during training, as the objective is fully self-supervised. Given a batch of N images, each resized to 64×64, we partition each image into a 4×4 grid of non-overlapping patches, resulting in $16N$ fragments per batch. The embedding function is defined as $f_\theta \colon \mathbb{R}^{16\times16\times3} \to \mathbb{R}^d$. To evaluate the quality of the learned embeddings and the resulting reconstructions, we adopt a two-stage evaluation protocol that balances local pairwise discriminability and global structure recovery:

(i) Pairwise Separability: We sample random pairs of fragments and compute the Cosine distance between their embeddings. Based on the ground-truth label indicating whether the fragments originate from the same source image, we compute the **Area Under the ROC Curve (AUC)** as a threshold-free measure of class separability. We also report the **Matthews Correlation Coefficient (MCC)**, where the classification threshold is selected to directly maximize the MCC on the evaluation set. Both metrics are robust to class imbalance, which is important in our setting where positive pairs (same image) are vastly outnumbered by negative pairs (different images). This pairwise evaluation provides a lightweight and intuitive assessment of the model's ability to distinguish between same-image and different-image fragment pairs.

(ii) Clustering-Based Reconstruction: We further assess the structural quality of the embedding space via clustering. Specifically, we apply standard k**-means clustering** with $k = N$, where each cluster is expected to correspond to one original image. While k-means does not enforce equal-sized clusters, it is significantly faster than constrained alternatives and scales well to large batches. We evaluate the clustering with the Adjusted Rand Index (ARI), which measures agreement between the predicted and ground-truth groupings while correcting for chance. ARI reflects the model's ability to reconstruct full images from their constituent fragments, though it can be a noisy metric due to the stochasticity in sampling and clustering.

2 MODEL ARCHITECTURE AND LOSS FUNCTION FOR EMBEDDING LEARNING

Architecture. We use a compact CNN encoder to process each 16×16 image fragment independently. Despite the small input size, convolutional layers effectively capture local color and texture patterns due to their inductive biases. Our encoder consists of two convolutional layers (with 32 and 64 filters), each followed by batch normalization and ReLU activation, and a `GlobalAveragePooling2D` layer to collapse spatial dimensions. A final dense layer projects the features into a d-dimensional embedding space. We use low-dimensional embeddings (e.g., $d = 8$ or 16) to encourage compact, discriminative representations while minimizing overfitting. Higher dimensions increased computation without significant gains. We favor a CNN over an autoencoder, as our goal is not pixel reconstruction but structured representation learning for contrastive or clustering-based objectives.

Contrastive vs. Binary Cross-Entropy Self-Supervision. We explore two training strategies: a contrastive learning objective (NT-Xent loss) and a binary classification objective (BCE). In both cases, we construct all pairwise combinations of fragments in a batch and assign a binary target indicating whether the fragments originate from the same image.

The NT-Xent loss (Chen et al., 2020) encourages positive pairs (from the same image) to have higher similarity than negative pairs (from different images). For a batch of $n = 160$ L2-normalized embeddings $\{z_1, \ldots, z_n\}$, the cosine similarity between fragments i and j is given by:

$$s_{ij} = \frac{z_i^\top z_j}{\tau}$$

where τ is a temperature parameter. Let $P = \{(i, j) \mid y_{ij} = 1, i \neq j\}$ be the set of all fragment pairs from the same image. The NT-Xent loss is defined as:

$$\mathcal{L}_{\text{NT-Xent}} = -\frac{1}{|P|} \sum_{(i,j)\in P} \log \frac{\exp(s_{ij})}{\sum_{k\neq i} \exp(s_{ik})}$$

This loss encourages each fragment to be more similar to its true partners than to all others in the batch. We omit self-similarities s_{ii} from the denominator.

As an alternative, we implement a weighted binary cross-entropy (WBCE) loss applied over all unordered fragment pairs. Using the same similarity scores s_{ij}, the model predicts whether each pair belongs to the same image. To address the strong class imbalance inherent to our setup—where roughly 1 in 10 pairs are positive—we apply a higher weight to positive pairs:

$$\mathcal{L}_{\text{WBCE}} = \frac{1}{n(n-1)} \sum_{i\neq j} [\alpha \cdot y_{ij} \log \sigma(s_{ij}) + (1 - y_{ij}) \log(1 - \sigma(s_{ij}))]$$

where $\sigma(\cdot)$ is the sigmoid function and $\alpha > 1$ is a positive weighting factor reflecting the inverse class ratio (typically $\alpha = 9$ in our setup).

3 TRAINING

Training Setup. At each training step, we sample 10 images from the training set. Each image is divided into a 4×4 grid of non-overlapping patches, yielding 16 fragments per image and 160 fragments in total. These fragments are randomly shuffled to remove spatial and sequential cues. We consider all possible fragment pairs within the batch and assign a binary label indicating whether the two fragments originate from the same image. For each anchor fragment, there are 15 positive pairs and 144 negative pairs. Each fragment is processed independently by the encoder, which outputs a low-dimensional embedding.

We compute cosine similarity between all pairs and experiment with two loss functions: (1) the NT-Xent loss and (2) binary cross-entropy applied to pairwise similarities. The model is trained using the Adam optimizer with a learning rate of 10^{-3}. Since both the training and validation sets are much larger than the model's observed rate of learning, we do not use traditional epochs. Instead, we treat ongoing training batches as fresh validation, and apply early stopping when the 100-step rolling average of the training loss does not improve for 500 steps. This enables efficient and fully self-supervised training using only instance-level co-occurrence.

4 RESULTS AND OBSERVATIONS

Training loss plateaued after approximately 1,000 steps, suggesting limited additional benefit from extended optimization (Figure 1A). Using embeddings trained with the NT-Xent contrastive loss, we achieve an ARI of 0.32, MCC of 0.33, and AUC of 0.81 on held-out validation batches (see AUC and distance matrix visualizations in Figure 1B,C). In contrast, training with the BCE loss yields an ARI of 0.22, MCC of 0.29, and AUC of 0.76. These results suggest that contrastive learning promotes more structured and separable embeddings. While BCE supervision provides direct pairwise feedback, it lacks the batch-level consistency imposed by contrastive objectives, which may contribute to its relatively lower performance.

The contrastive approach tends to produce tighter and more distinct clusters, especially when fragment appearance is ambiguous or lacks strong semantic cues. We also explored variations such as using Euclidean distance instead of cosine similarity and altering the encoder architecture, but

these adjustments did not lead to substantial improvements. Overall, the results support the potential benefits of contrastive objectives in capturing fragment-level structure under self-supervised constraints.

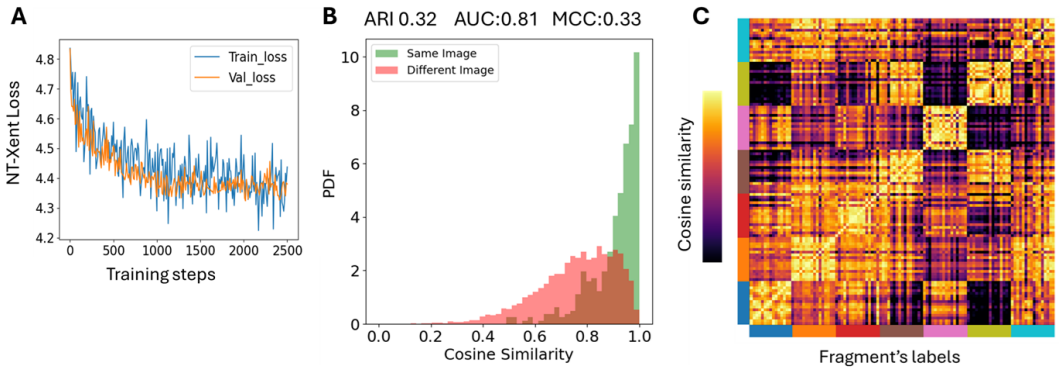

Figure 1: (A) Training and validation loss curves for the contrastive model, indicating convergence and generalization performance. (B) Distribution of pairwise distances between fragments originating from the same source image (positive pairs) versus different images (negative pairs), computed on held-out validation batches and demonstrating clear separation in the learned embedding space. (C) Representative distance matrices for seven validation images, illustrating the model's ability to cluster fragments from the same image into coherent blocks.

5 Outlook

While our model achieves moderate performance on the 4×4 fragmentation task, our analysis highlights several avenues for improvement, both in model robustness and task design. Contrastive training shows promise for self-supervised fragment grouping, but the achieved ARI of 0.3 remains modest—especially considering evaluation was limited to reconstructing just 10 images per batch. As the number of images grows, which is realistic given the scale of the training data, the task becomes significantly harder due to increased ambiguity and inter-image similarity.

Several aspects of the current setup merit further exploration. The use of fixed-size fragments presents a trade-off: smaller patches reduce semantic content and increase difficulty, while larger ones risk trivializing the task. Uniform sampling across images could be improved by emphasizing ambiguous or visually similar fragments, which may encourage more discriminative embeddings. Robustness to low-level visual cues like texture or brightness could be enhanced with data augmentations such as rotation or color jitter. Improvements in clustering could come from relaxing our reliance on balanced k-means, potentially using constrained clustering or learning assignments jointly with embeddings (Ji et al., 2019).

On the architectural side, hybrid models combining convolutional networks with transformers could better capture long-range dependencies and global context. Finally, while we currently treat each training batch as independent, revisiting a fixed subset of data in inner epochs could allow the model to refine its representations over time through repeated exposure to challenging configurations.

References

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

Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9865–9874, 2019.