Similar Accuracy but Different Topographies under Cross-Entropy and Contrastive Learning

Gerrit Sander[1,2], **Uri Hasson**[2]

[1]Institute of Cognitive Science, University of Osnabrück, Germany
[2]Center for Mind and Brain Sciences (CIMeC), University of Trento, Italy
gersander@uni-osnabrueck.de, uri.hasson@unitn.it

Abstract

The brain's topographic organization has motivated topographic deep neural networks (TDNNs) as models of perceptual and conceptual representation. However, prior TDNN studies largely paired topography with cross-entropy (CE). They have not examined whether contrastive objectives are generally compatible with topographic training, how topographic strength affects run-to-run representational consistency, or what failure modes limit the effect of the topographic constraint. We addressed these issues by training TDNNs on CIFAR-10 with a local topographic loss that minimized the average ℓ_2 distance between afferent weight vectors of neighboring units. We compared four objectives: CE, supervised contrastive, self-supervised SimCLR, and a label-aware contrastive margin loss reflecting an animacy hierarchy. Across topographic strengths, label-supervised objectives maintained high accuracy, produced smooth activation maps, and increased within-class similarity relative to CE. Two factors limited the impact of the topographic loss: 1) dropout was required to obtain smooth maps rather than sparse activations; 2) under strong penalties, networks reduced the topographic loss by shrinking weight norms rather than aligning weight directions. We also found that stronger topographic constraints reduced cross-seed representational consistency, indicating multiple comparably good topographic solutions. Nonetheless, ensembles built from sets of less-consistent models only slightly outperformed ensembles without topographic constraints. Our results indicate that contrastive objectives are a robust option for training topographic networks, producing good accuracy and high within-class similarity. The findings also identify boundary conditions for afferent-weight similarity as a topographic prior.

Code — https://github.com/gerrit-sander/ConTopo

Introduction

The modeling of processes and representations in the brain's ventral visual stream using pre-trained deep neural networks has been a subject of great interest in recent years. These models can predict neural responses and capture representational geometry across visual areas (Yamins et al. 2014; Cadieu et al. 2014; Khaligh-Razavi and Kriegeskorte 2014; Güçlü and Van Gerven 2015). Consequently, they have become potential models of brain organization, making it possible to study brain organization in silico. In this context, considerable attention has been recently devoted to the possibility of learning a spatial, that is, a topographic structure within the neural network, because in the brain, neurons in close proximity often exhibit correlated responses to similar stimuli. For example, neurons in the primary visual cortex (V1) are organized according to retinotopy and orientation selectivity (Hubel and Wiesel 1962; Sereno et al. 1995), while higher visual areas contain spatially clustered regions selective for abstract categories and features (Fujita et al. 1992; Kanwisher, McDermott, and Chun 1997; Tsao et al. 2006).

These observations have motivated recent studies examining whether DNNs can perform their tasks while at the same time producing an internal representation that maintains a topographic organization. These studies suggest that topographic DNNs seeded and trained with topographic constraints produce an organization with some similarities to cortical spatial activation patterns (Margalit et al. 2024). Most studies that used deep neural networks as models of the ventral stream have relied on supervised training with cross-entropy loss. This is natural given the long-standing view of the ventral stream as supporting object recognition. However, recent findings suggest that ventral stream representations are not limited to categorization, but instead encode semantically rich information useful for a variety of downstream tasks (Cadena et al. 2024). At the same time, supervised learning with labeled categories has been criticized as biologically implausible (Konkle and Alvarez 2022; Zhuang et al. 2021), motivating the adoption of self-supervised and contrastive learning approaches as alternative training paradigms (Konkle and Alvarez 2022; Zhuang et al. 2021). These methods have been proposed as both more neurobiologically grounded and more effective at learning flexible, transferable representations.

Building on these motivations, several approaches have been developed to induce topography in artificial networks. Examples include post-hoc projection methods such as self-organizing maps applied to feature spaces (Doshi and Konkle 2023); architectural modifications that place convolutional kernels or units on a two-dimensional sheet with lateral interactions (Qian et al. 2024); and fully topographic ar-

chitectures that impose smoothness across all layers while avoiding weight sharing entirely (Lu et al. 2025). Other approaches add regularization terms to enforce similarity among neighboring units in a predefined spatial layout (Margalit et al. 2024), or introduce a local weight-similarity constraint that arranges the weight vectors of the penultimate layer on a 2D grid and penalizes differences between neighbors (e.g., Truong and Hasson 2025).

Here we consider that supervised (cross-entropy) objectives require that class exemplars maximally activate the class logit, but do not explicitly require that class exemplars show greater similarity in embedding space than cross-class exemplars. In contrast, contrastive learning simply requires the latter constraint, which is known to produce markedly different representations in feature space (e.g., Hadsell, Chopra, and LeCun 2006).

Against this backdrop, our main objective is to determine how the *choice of task loss* shapes topographic organization and learned representations under an identical spatial prior. Prior work advances arguments for and against different objectives, yet their interaction with the same topographic constraint has not been studied. Our second objective is to evaluate how the learning constraints impact the representational consensus among within-architecture models trained from different seeds. In particular, to the extent that some architectures are more effective in producing reduced representational consensus across models, this could produce a larger mixture of experts that could be used in the context of ensemble learning.

To address this, we systematically compare several training objectives under the same local weight-similarity constraint, and probe (*i*) task performance, (*ii*) the topographic smoothness of activations, (*iii*) functional co-localization, and (*iv*) the structure of the embedding space, including within-class compactness and representational geometry. We conduct these analyses systematically varying the strength of the topographic term in a manner that is model agnostic, and also evaluate the impact of dropout in training.

Methods

Model and Dataset

For all task-loss and topographic weighting conditions, we used a modified version of the ResNet-18 architecture (He et al. 2015) trained end-to-end on CIFAR-10 (Krizhevsky, Hinton et al. 2009). The architecture consisted of an initial convolutional layer, followed by eight residual blocks with two convolutional layers each, and a final fully connected layer that created the embedding. We adapted ResNet-18 to the low-resolution CIFAR-10 dataset. In particular, the original 7×7 convolution with stride 2 followed by a 3×3 max-pooling layer was replaced with a 3×3 convolution with stride 1 and no max-pooling, following related work (Poli, Dupoux, and Riad 2023).

Furthermore, the original 1000-dimensional fully connected layer was replaced with a 256-dimensional embedding layer, such that the ResNet-18 backbone functioned as an encoder producing 256-dimensional representations. For contrastive learning settings, this encoder was followed by a lightweight projection head composed of batch normalization, ReLU activation, dropout ($p = 0.5$), and a 128-dimensional fully connected layer, yielding 128-dimensional embeddings used to compute the contrastive loss. For classification under cross-entropy loss, the encoder was followed by dropout ($p = 0.5$) and a 10-dimensional fully connected output layer.

Task Loss Functions

Models were trained in four different task loss settings: (i) a pairwise, label-aware hierarchical margin loss with a contrastive term that imposes an animacy hierarchy: within-class pairs are pulled closer and cross-class pairs are pushed apart; additionally, images from animate categories are pulled closer to each other than to inanimate categories (and vice versa), (ii) the classical cross-entropy (CE) loss, (iii) the fully self-supervised SimCLR loss (Chen et al. 2020), and (iv) its supervised counterpart, SupCon (Khosla et al. 2020).

Margin Loss. First, we applied the margin loss. For a mini-batch of size B, let $z_i \in \mathbb{R}^{128}$ be the projection-head output for sample i. Using cosine similarity, define the distance $d_{ij} = 1 - \mathrm{sim}_{\cos}(z_i, z_j)$. Let c_i denote the (fine-grained) class label and $\sigma_i \in \{\text{animate}, \text{inanimate}\}$ its animacy superclass. We use three margins (hyperparameters): (i) m_p for positive pairs ($c_i = c_j$), (ii) $m_{\tilde{n}}$ for negative pairs that share the same superclass ($c_i \neq c_j$ but $\sigma_i = \sigma_j$), and (iii) m_n for negative pairs across superclasses ($\sigma_i \neq \sigma_j$). We chose $m_p = 0.05$, $m_{\tilde{n}} = 0.3$, and $m_n = 0.5$.

Writing the loss over unordered pairs, our objective is

$$\mathcal{L}_{\text{margin}} = \frac{2}{B(B-1)} \sum_{1 \le i < j \le B} \begin{cases} \max(0, d_{ij} - m_p)^2, & A, \\ \max(0, m_{\tilde{n}} - d_{ij})^2, & B, \\ \max(0, m_n - d_{ij})^2, & C. \end{cases}$$

A: $c_i = c_j$; $\quad B$: $c_i \neq c_j, \sigma_i = \sigma_j$; $\quad C$: $\sigma_i \neq \sigma_j$.

Thus positives are penalized only when their distance exceeds m_p; negatives are penalized only when their distance falls below the appropriate negative margin ($m_{\tilde{n}}$ within superclass, m_n across superclasses).

SimCLR and SupCon. In SimCLR and SupCon, two random augmentations of each image were generated. In SimCLR, the loss considers one positive pair for each anchor image (its augmented twin) and $2 \times (B - 1)$ negatives. SupCon follows the same principle, but instead of treating only the twin augmentation as positive, it also includes all other images in the batch that share the same label. We used a standard augmentation pipeline consisting of random resized cropping, horizontal flipping, color jittering, and occasional conversion to grayscale, in line with prior work (e.g., Chen et al. 2020; Khosla et al. 2020).

Cross-Entropy. The cross-entropy loss was used as in typical applications, pushing the model to assign high probability to the correct class.

Topographic Learning

Topographic Constraint. All models were trained either with or without an additional topographic constraint. To impose a topographic structure on the model, the weights of

the 256-dimensional embedding layer were arranged as a 16×16 grid. Following prior works (e.g., Truong and Hasson 2025), the local topographic loss $\mathcal{L}_{\text{topo}}$ reflected the average ℓ_2 distance between afferent (incoming) neighboring weight vectors within this grid, emulating a push towards similar afferent excitation in spatially adjacent units. In Eq. (1), p denotes a reference unit and q indexes its neighboring units, $q \in \mathcal{N}_M(p)$, the Moore neighborhood (i.e., all adjacent locations with $\|p - q\|_\infty = 1$). The neighborhood size varies with grid location: $|\mathcal{N}_M(p)| = 3$ at corners, 5 along edges, and 8 elsewhere. The per-unit topographic loss $\ell_{\text{topo}}(p)$ averages the ℓ_2 distance between $\mathbf{w}_p$ and its neighbors' incoming weight vectors; the total loss $\mathcal{L}_{\text{topo}}$ aggregates these terms over $p \in \Omega$ (the set of grid locations):

$$\begin{aligned} \ell_{\text{topo}}(p) &= \frac{1}{|\mathcal{N}_M(p)|} \sum_{q \in \mathcal{N}_M(p)} \left\| \mathbf{w}_p - \mathbf{w}_q \right\|_2, \\ \mathcal{N}_M(p) &= \{\, q \in \Omega : \, \|p - q\|_\infty = 1 \,\}, \\ \mathcal{L}_{\text{topo}} &= \sum_{p \in \Omega} \ell_{\text{topo}}(p). \end{aligned} \tag{1}$$

Joint Loss Term. The overall loss was defined as $\mathcal{L} = \mathcal{L}_{\text{task}} + \hat{\lambda}\, \mathcal{L}_{\text{topo}}$. Because we compare the impact of induced topography across task losses from different models, we consider that task losses can have fundamentally different scales/magnitudes, and control for this factor. We define a dynamic $\hat{\lambda}$ that matches the gradient magnitudes of the two losses and is updated at each mini-batch update step. We define ρ as a hyperparameter controlling the contribution of the topographic loss to the model's update step, $\|\cdot\|$ as the ℓ_2 norm, and $\nabla_\theta \mathcal{L}_{\text{task}}$ and $\nabla_\theta \mathcal{L}_{\text{topo}}$ as the separately computed gradients of the task and topographic losses, respectively, each with respect to the encoder's parameters θ. The gradient-matched coefficient λ^* is defined as

$$\lambda^* = \rho \cdot \frac{\|\nabla_\theta \mathcal{L}_{\text{task}}\|}{\|\nabla_\theta \mathcal{L}_{\text{topo}}\| + \varepsilon}, \tag{2}$$

where ε is a small positive constant for numerical stability, with $0 < \varepsilon \ll 1$. The exponential moving average of λ^*, denoted $\hat{\lambda}_t$, with a fixed β set to 0.1 is then given by

$$\hat{\lambda}_t = (1 - \beta)\hat{\lambda}_{t-1} + \beta\, \lambda^*. \tag{3}$$

Training and Evaluation

We evaluated 24 model types, produced by combining the four task loss settings with six magnitudes of the topographic constraint weight ($\rho \in \{0, 0.008, 0.04, 0.2, 1, 5\}$). Note that in our gradient-matched scheme (Equation 2), ρ is not a raw loss weight but a target ratio that controls the relative strength of the topographic update to the task update. Since $\hat{\lambda}_t \approx \rho \cdot \frac{\|\nabla_\theta \mathcal{L}_{\text{task}}\|}{\|\nabla_\theta \mathcal{L}_{\text{topo}}\|}$ (up to EMA smoothing and ε), the resulting update magnitudes satisfy the relation $\|\hat{\lambda}_t \nabla_\theta \mathcal{L}_{\text{topo}}\| \approx \rho \cdot \|\nabla_\theta \mathcal{L}_{\text{task}}\|$. Thus, $\rho = 1$ aims for parity between the two loss signals; $\rho < 1$ assigns greater importance to the task gradient; and $\rho > 1$ assigns greater importance to the topographic constraint. Setting $\rho = 0$ disables the topographic term. Because the scaling uses gradient norms, this interpretation holds across very different task losses whose raw magnitudes are otherwise incomparable.

Each condition was trained with five random seeds to estimate variability. We refer to these five independent runs as *trials*. We used the Adam optimizer (learning rate 0.002) with a batch size of 512 to have enough pairs in the contrastive regimes. From the training set, 5k samples (500 per class) were designated as a validation split. Training employed early stopping based on validation performance, with a patience of 25 epochs. For contrastive regimes, early stopping was determined by the validation task loss, whereas for cross-entropy classification it was based on validation accuracy.

For contrastive training, we evaluated the class accuracy by adding a linear readout, which was trained based on the same stopping rules. Throughout training, we tracked the best model based on the lowest total validation loss (task plus topographic term with the corresponding $\hat{\lambda}_t$), and this checkpoint was used for all subsequent experiments. During readout, the encoder was frozen, and a linear classifier was trained using AdamW (learning rate 0.0003, betas = (0.9, 0.999)).

Results

Test Accuracy and t-SNE Visualization

Test accuracy was generally high and comparable across models and parameter combinations, except for the SimCLR condition (Figure 1). Topographic constraints did not produce a drop in accuracy (see $\rho = 0$ condition in Figure 1). This shows that under supervised cross-entropy and supervised contrastive settings, models can be trained with varying levels of topographic regularization while achieving consistently high test performance, occasionally surpassing the baseline. This clarifies some prior results where topographic networks either surpassed or performed below non-topographic ones, and we suggest the crucial factor is precise balancing of the loss terms that have different scales. The self-supervised SimCLR regime was an exception. As a feature extractor, it achieved low accuracy after training a linear readout and showed a monotonic decline in performance as ρ increased. While SimCLR performance could likely be improved by more extensive searches for optimizers and hyperparameters, or by employing a more expressive projection head, we deliberately kept its configuration aligned with the other regimes for the sake of comparability.

t-SNE visualization was performed on 2,000 feature embeddings extracted from the trained encoder using the CIFAR-10 test set. To compare topographic and non-topographic embedding spaces, we focused on the conditions with $\rho = 0$ and $\rho = 0.2$. The embeddings were flattened and reduced to two dimensions using the scikit-learn t-SNE algorithm, employing default parameters and a fixed random seed for reproducibility. The resulting two-dimensional projections were plotted with color-coded points corresponding to each CIFAR-10 class (Figure 2). Across all supervised conditions, we observed clear clustering by class and, in many cases, a distinct separation by the animacy superclass. In contrast, SimCLR produced smoother, more entangled representations with substantial overlap between classes.

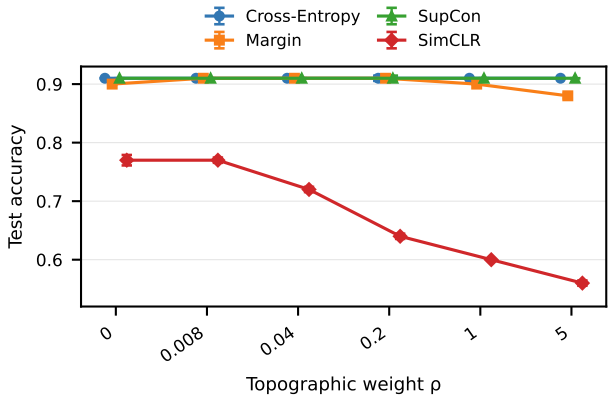

Figure 1: Test set accuracies for CIFAR-10 at each topographic weight ρ. Under gradient-matched topographic regularization, topographic regularization matched or exceeded performance of non-topographic networks ($\rho = 0$) for all supervised models (Margin, Cross-Entropy, SupCon). Values indicate mean $\pm$ sd across five seeds. Small horizontal jitter was applied to the x-positions for visualization only.

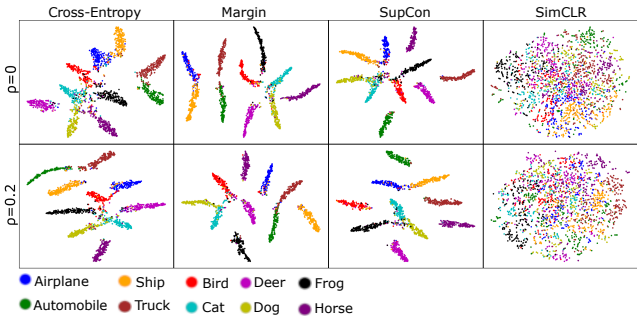

Figure 2: Two-dimensional t-SNE visualizations of feature embeddings extracted from the trained encoders on the CIFAR-10 validation set. Each plot corresponds to the first trial of a given model condition, comparing non-topographic ($\rho = 0$) and topographic ($\rho = 0.2$) settings.

Smoothness

Figure 3 demonstrates the smoothness induced by different models, presenting activation maps on the topographic grid for a single image across the different training conditions. To show the effect of dropout ablation, we also present results for the cross-entropy models trained without dropout (for all ρ values). It can be observed that in the no-dropout variants, the network could satisfy the topographic constraint by producing highly sparse maps: a few edge units show strong activation (carrying the post-ReLU relevant information), while many central units remain redundant and weakly active, leading to less coherent spatial patterns and implying that central weight vectors become very small and thus close in Euclidean space. This effect appears across task losses, but is most pronounced in the cross-entropy setting. In such cases, the topographic constraint is still met because these

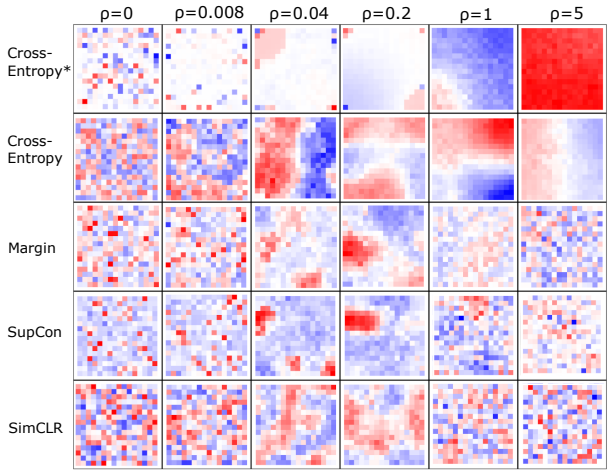

Figure 3: Smoothness across task loss and ρ conditions (activation maps). Activation maps (from the same CIFAR-10 test image of a truck) arranged as a 16×16 grid for each model condition from the first trial. In the no-dropout variants, the network often satisfies the topographic constraint by producing highly sparse maps: a few edge units show strong activation, while many central units remain redundant and weakly active, leading to less coherent spatial patterns. By contrast, with dropout, activations are more evenly distributed across the grid, resulting in qualitatively smoother maps. (*Conditions marked with an asterisk denote models trained without dropout.)

central weight vectors are highly similar, so that the difference in weight vectors of the edge units does not significantly affect the overall topographic loss. By contrast, with dropout, activations are more evenly distributed across the grid, resulting in qualitatively smoother maps. In these cases the network can no longer rely on sparse codes and fulfills the topographic constraint by creating smooth maps.

Following prior work (e.g., Rathi et al. 2025), we quantified spatial smoothness in these 16×16 activation maps using Moran's I (Moran 1950). Higher values indicate stronger positive spatial autocorrelation (smoother maps), values near zero correspond to spatial independence, and negative values reflect checkerboard-like organization. We evaluated smoothness scores for each model configuration, also including a single additional run per condition trained without dropout (i.e., following the procedure as described in the Methods section but omitting dropout) (dotted lines) (Fig. 4a). We observe that for the Cross-Entropy and SimCLR settings, models trained without dropout consistently yielded lower smoothness scores across ρ values. In the supervised contrastive setting, however, this gap was relatively small. As a result, Cross-Entropy was associated with the least-smooth topographies without dropout and the smoothest ones with dropout. Overall, these results highlight the importance of including dropout for preventing trivial solutions where few grid units carry most information.

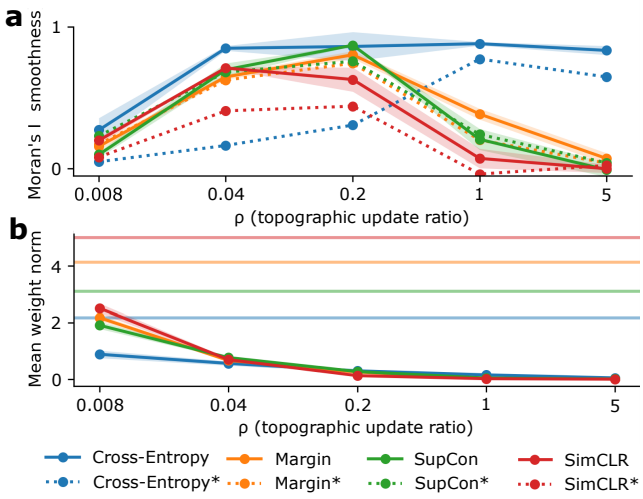

Figure 4: Smoothness metrics across task loss and ρ conditions. a) Moran's I smoothness score (averaged across the CIFAR-10 test set) for all model conditions. Shaded bands indicate standard deviation across trials, and dotted lines indicate scores of the corresponding no-dropout models. b) Relation between the average weight vector ℓ_2 norm of the embedding layer and ρ values for each task loss condition. Horizontal lines indicate the control conditions with $\rho = 0$. (*Conditions marked with an asterisk denote models trained without dropout.)

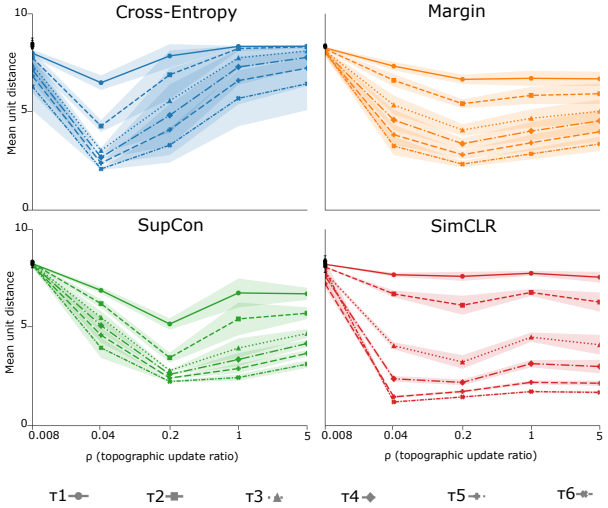

Figure 5: Functional co-localization. Mean grid distance between unit pairs with correlation $r \geq \tau_k$ as a function of ρ for each loss. Curves were shown for $\tau_k \in \{0.1, 0.3, 0.5, 0.6, 0.7, 0.8\}$; the $\rho = 0$ baseline (expected ≈ 8.3) was annotated for reference and was shown as black dots on the y-axis.

For dropout-regularized models, smoothness, particularly in contrastive regimes, showed a non-monotonic relationship with ρ: it increased from $\rho = 0.008$ to $\rho = 0.2$ but declined at higher values ($\rho \in \{1, 5\}$). At first glance, this may seem counterintuitive, since larger ρ places greater emphasis on the topographic objective, which minimizes the ℓ_2 norm of differences between neighboring *weights*. However, inspection of the topographic loss during training indicated the following: while the loss decreased monotonically with increasing ρ, at higher values of ρ this was achieved by shrinking weight magnitudes toward zero, thereby minimizing neighbor differences without necessarily preserving smooth activation-level organization. This is due to the fact that the weight-similarity loss is scale sensitive. To verify this behavior, we computed the average embedding-layer weight norm, and found that it monotonically decreases as ρ increases (Fig. 4b). Notably, in the absence of the topographic constraint ($\rho = 0$), weight norms are substantially larger than in models with the constraint, even for small ρ values such as 0.008 or 0.04.

Functional Co-Localization

We tested the extent to which similarly tuned units were spatially proximate, which speaks to the localization of function in the topographic grid. For each model, we passed the CIFAR-10 test set through the network, recorded each unit's activation vector across images, and computed the Pearson correlation r for every pair of units. We then set correlation thresholds $(\tau_1, \ldots, \tau_6) = (0.1, 0.3, 0.5, 0.6, 0.7, 0.8)$. For each threshold τ_k, we kept all pairs with $r \geq \tau_k$ and measured their mean Euclidean distance on the 16×16 grid. Units were indexed by their sheet coordinates, and distance was the standard Euclidean distance on this grid. We summarized results in Fig. 5; for comparison, Moran's I was reported in Fig. 4a.

As a baseline, with no topography ($\rho = 0$), the mean pairwise grid distance was about 8.3 and was essentially unaffected by the choice of τ_k. Introducing the topographic prior produced clear structure: higher thresholds (i.e., keeping only pairs with larger r) yielded shorter mean distances, so strongly co-activating units clustered more tightly. Distances were relatively large at $\rho = 0.008$ and dropped sharply by $\rho = 0.04$, especially for Cross-Entropy and SimCLR. The tightest neighborhoods were typically observed for pairs with $r \geq 0.8$, followed by progressively weaker correlation bands. The minima of these high-r curves usually aligned with the ρ values that maximized Moran's I, linking smoother activation maps to tighter spatial clustering of functionally similar units.

At larger ρ, mean distances sometimes rose again even when Moran's I remained flat, most clearly seen for Cross-Entropy. One possible explanation could be that effective spatial degrees of freedom were reduced as clusters broadened, which could increase within-cluster pair distances despite similar global smoothness. Overall, moderate ρ yielded compact, localized functional clusters, whereas very strong regularization tended to produce broader, less compact groupings.

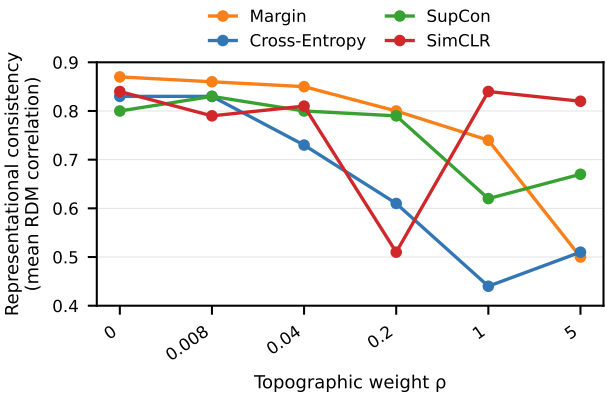

Figure 6: Representational consistency over random weight initializations decreases with topographic constraint strength. Model representations were operationalized as representational dissimilarity matrices (RDMs). For each condition, the consistency score is the mean of all pairwise correlations between models' RDMs across the five re-initializations.

Representational Geometry

Consistency of Representational Geometry. Considering that all models reached comparable accuracy, we aimed to determine whether some models produced less consistent representations than others. A model's representation was defined as the representational-dissimilarity matrix (RDM) produced at the penultimate layer of the model (whether trained as topographic or not). RDMs were computed by generating embeddings from 100 images of each of the 10 classes. Then, the similarity of any two representations was quantified as the Pearson linear correlation between the upper triangles of the two RDMs. Note that, for each model, we had 5 iterations from random initializations, at 6 levels of the topographic constraint (including 0). This produced 30 RDMs for each of the 4 models. Figure 6 shows, for each of the four models, the average pairwise consensus across RDMs. Consistency generally declined with increasing topographic regularization relative to the $\rho = 0$ controls, with the drop becoming larger at higher ρ. The main exception was SimCLR, which showed comparatively high consistency, particularly at $\rho = 1$.

Overall, these observations suggest that, in the supervised settings and especially at higher ρ, models converge on different—but equally effective—representations: despite decreasing consistency, test accuracy remains high on this 10-way classification task.

Within vs. Across-Class Similarity. To verify whether the contrastive loss produced greater within-class similarity in the topographic layer, we computed cosine-similarity between image embeddings in that layer, summarizing the distributions separately for within-class and across-class pairs (Table 1). As would be expected, across models, the within-class mean exceeded the across-class mean. Speaking to the effectiveness of contrastive learning, within-class means were higher than for training under the CE objective, and under low topographic constraints, across-class similarity was also kept low. Taken together with prior data it appears that contrastive topographic learning with a moderate $\rho = 0.2$ smoothing constraint successfully produces several interesting behaviors: a smooth topography, strong functional localization, and strong within-category similarity.

The magnitude of the differentiation between within- and across-class similarity varied with both the objective and ρ. Cross-entropy maintained a clear separation even at large ρ (e.g., at ρ=5: μ_W=0.89 vs. μ_A= − 0.03), aligning with its stable test accuracy (Figure 1). Label-aware contrastive objectives (SupCon, margin) also showed strong separation at small–moderate ρ, but at very large ρ the across-class mean approached the within-class mean (e.g., SupCon at ρ=5: $\mu_W \approx \mu_A \approx 1.00$), effectively compressing cosine differences. SimCLR exhibited the smallest gap overall, and the gap narrowed with ρ as both within- and across-class means moved toward ceiling for $\rho \geq 1$. Consistent with Figure 1, this narrowing did *not* harm test accuracy for the supervised objectives (CE, SupCon, margin), which remained high across ρ, whereas SimCLR accuracy declined monotonically with increasing ρ (while remaining above chance).

Ensemble Methods and Noise Robustness

The representational consistency analysis showed that models trained with stronger topographic regularization tend to converge on less similar representational geometries across random seeds, while still achieving comparable test accuracy at the individual model level. This suggested that ensembles of such models might benefit from complementary representations, as diverging representations could mean that the models from different seeds might learn different features.

To test this, we focused on the cross-entropy models. For each ρ setting, we trained a total of 10 independent trials (instead of 5 as in the main experiments). Performance was evaluated in two ways: first, each of the 10 models was tested individually on the CIFAR-10 test set, and results were averaged across runs. These results replicate test accuracies in Figure 1, with standard deviations around 0.004, confirming that accuracy is highly stable across seeds. Second, for each test image we collected the logits from all 10 models, averaged them across runs, and used the argmax of the averaged logits as the ensemble prediction (i.e., a soft-voting scheme).

As expected, ensembles consistently improved test accuracy across all ρ conditions (Fig. 7), whether or not the ensemble was produced from topographic or non-topographic models had a minor impact. For example, for non-noisy images, the ensemble boosted prediction by 2.9% over the average individual model performance when constructed from non-topographic models, whereas the boost was between 3.0-3.15% when constructed from topographic models. This suggests that while ensembles reliably reduce model-specific errors, representational diversity induced by higher ρ does not directly translate into larger ensemble gains.

We next evaluated the effect of the ensemble method for noisy datasets. We applied three common noise types—pink,

Table 1: Topographic regularization tends to increase within-class similarity and reduce across-class similarity. Within vs. across-class cosine similarity at each topographic weight ρ. For each objective, the first row *(W)* reports *within*-class similarity and the second row *(A)* reports *across*-class similarity. Entries are μ (SD), rounded to two decimals. Boldface marks topographic conditions for which, relative to the baseline model, within-class similarity is matched or higher and across-class similarity is matched or lower.

Objective	$\rho = 0$	$\rho = 0.008$	$\rho = 0.04$	$\rho = 0.2$	$\rho = 1$	$\rho = 5$
Margin (W)	.87 (.20)	.85 (.21)	.85 (.24)	**.87 (.24)**	.93 (.11)	.99 (.02)
Margin (A)	.18 (.19)	.15 (.19)	.02 (.21)	**.00 (.29)**	.51 (.21)	.93 (.04)
Cross-Entropy (W)	.77 (.23)	.79 (.23)	**.80 (.24)**	**.82 (.26)**	**.85 (.27)**	**.89 (.25)**
Cross-Entropy (A)	-.01 (.21)	.00 (.21)	**-.03 (.22)**	**-.05 (.32)**	**-.05 (.44)**	**-.03 (.56)**
SimCLR (W)	.26 (.29)	.26 (.29)	.28 (.29)	.78 (.16)	.98 (.02)	1.00 (.00)
SimCLR (A)	.05 (.29)	.05 (.29)	.07 (.30)	.71 (.18)	.97 (.02)	1.00 (.00)
SupCon (W)	.83 (.24)	**.83 (.26)**	**.84 (.25)**	**.84 (.27)**	.95 (.08)	1.00 (.00)
SupCon (A)	.10 (.18)	**.04 (.17)**	**.06 (.18)**	**-.03 (.21)**	.70 (.13)	1.00 (.00)

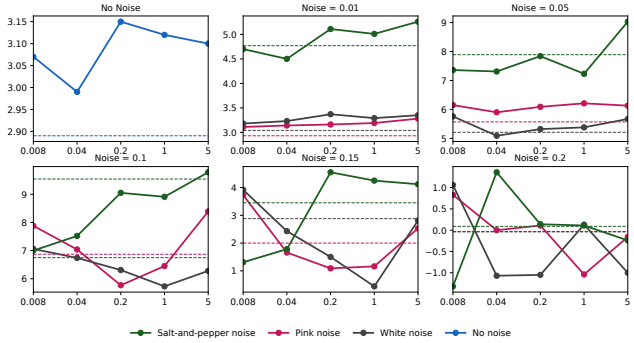

Figure 7: Ensemble boost under noisy test conditions. Each subplot shows the increase in test accuracy when averaging predictions across 10 cross-entropy models compared to single-model performance. The x-axis denotes the ρ value, and the y-axis the accuracy gain. Subplot 1 shows results on clean CIFAR-10, while subplots 2–6 show results under pink, white, and salt-and-pepper noise at varying strengths. Dotted lines indicate the non-topographic control condition.

white, and salt-and-pepper—at five different intensity levels. Here, ensemble effects were more variable across ρ values: sometimes boosts were larger for topographic models, sometimes not. Independent of ρ, however, the largest ensemble gains were consistently observed at intermediate noise levels (e.g., noise strengths of 0.05–0.15), where ensembles improved accuracy by up to 10% compared to single models (Fig. 7). At very strong noise levels (e.g., 0.2), the ensemble effect diminished toward zero, with absolute accuracies falling close to chance level ($\approx 10\%$), indicating that models were no longer able to extract meaningful features from the corrupted inputs.

Discussion

We controlled the relative weights of the task and topographic losses to enable matched comparisons across objectives. For contrastive objectives, the backbone was trained as a feature extractor and a linear classification head was added only for evaluation. Under these matched settings, accuracy remained consistently high and similar for CE, SupCon, and margin-based contrastive models, across all tested topographic strengths.

Topographic strength impacted spatial smoothness in a non-monotonic way. Increasing the topographic weight to $\rho \in \{0.008, 0.04, 0.20\}$ resulted in visibly smoother activation maps (Fig. 4a) and higher values of Moran's I. However, further increases of the topographic constraint to $\rho = 1$ and $\rho = 5$ reduced smoothness. In this high-penalty regime, the network satisfied the topographic objective by shrinking weight norms instead of producing similar weight vectors, which weakened the intended effect of the prior. We consider that cosine distance would avoid norm shrinkage. However, it breaks the intended biological analogue that Euclidean distance captures, which is pushing for similar afferent weights for nearby units.

Dropout was necessary for the topographic prior to induce spatial smoothness in the topographic grid. Without dropout, we observed that a small subset of units carried most task-relevant information while many units mainly carried the topographic constraint and often remained below the ReLU threshold. In contrast, including dropout spread activity across units producing smooth maps (Fig. 4).

With respect to functional localization on the grid, stronger topographic constraints in the $\rho \in \{0.008, 0.04, 0.20\}$ range increased the tendency of highly correlated unit pairs to be within a close distance on the grid. This is a signature of functional localization. However, different objectives produced slightly different localization profiles. CE positioned units with correlations above ≈ 0.5 very close together with limited further differentiation as correlation increased (Fig. 5). The margin-based contrastive objective produced inter-unit distances that scaled more continuously with correlation. This scaling was most evident for SimCLR, which showed the tightest relationship between pairwise correlation and typical grid distance, although SimCLR was less sensitive to the absolute ρ value. These findings effectively indicate

that topography produces a spatial kernel whose properties depend on the hyperparameters of the learning objective.

In the embedding space, supervised contrastive objectives (margin and SupCon) produced greater within-class similarity relative to CE. Furthermore, CE kept between-class similarity near zero, indicating the expected hard separation across categories. The contrastive objectives allowed modest between-class similarity, which is consistent with characteristics of CIFAR-10 where items from different classes can be visually similar.

Topographic constraints reduced cross-seed representational consistency. Without a topographic constraint, cross-seed RDM alignment within a model family was high, approximately $0.80-0.87$. At $\rho = 0.20$, alignment decreased to approximately $0.60-0.80$ for the label-supervised models (Figure 6). Thus, while CE, margin, and SupCon achieved similar accuracy with and without topography, adding the topographic loss pushed different training runs to produce different representations. Given this, we tested whether soft-voting ensembles of topographic models would produce larger gains than ensembles of non-topographic models. The effect we found was in the predicted direction, but weak. On average, topographic soft-voting ensembles outperformed non-topographic ensembles by only about 0.1%. We also evaluated out-of-distribution robustness by testing the same ensembles on images to which different types of noise were added, at different levels. For some noise levels (e.g., 0.05, 0.15, 0.20), topographic ensembles performed slightly better on pink and white noise (solid purple/black lines above dotted purple/black in Fig. 7), not under salt-and-pepper noise. These out-of-distribution effects therefore require further study.

The study has several implications for future work. The fact that topographic priors are equally compatible with contrastive objectives allows choosing training losses that provide some control over the desired feature space for downstream goals (e.g., improved transfer) or to pair topography with additional constraints (e.g., hierarchical representations as shown here, or neurosymbolic structure). The study suggests that distances between weight vectors can be effectively used, but should be monitored to avoid norm-collapse. Adding norm penalties may help address this issue. The dependence on dropout indicates that achieving smooth, cortex-like activity requires coordination of multiple mechanisms that promote distributed activation. Balancing activation sparsity (often found in cortical systems) with the strength of the topographic prior is therefore a key design trade-off and an important direction for future work.

References

Cadena, S. A.; Willeke, K. F.; Restivo, K.; Denfield, G.; Sinz, F. H.; Bethge, M.; Tolias, A. S.; and Ecker, A. S. 2024. Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks. *PLOS Computational Biology*, 20(5): e1012056.

Cadieu, C. F.; Hong, H.; Yamins, D. L.; Pinto, N.; Ardila, D.; Solomon, E. A.; Majaj, N. J.; and DiCarlo, J. J. 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS computational biology*, 10(12): e1003963.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Doshi, F. R.; and Konkle, T. 2023. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25): eade8187.

Fujita, I.; Tanaka, K.; Ito, M.; and Cheng, K. 1992. Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360(6402): 343–346.

Güçlü, U.; and Van Gerven, M. A. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, 1735–1742. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Hubel, D. H.; and Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1): 106.

Kanwisher, N.; McDermott, J.; and Chun, M. M. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311.

Khaligh-Razavi, S.-M.; and Kriegeskorte, N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11): e1003915.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Konkle, T.; and Alvarez, G. A. 2022. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1): 491.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lu, Z.; Doerig, A.; Bosch, V.; Krahmer, B.; Kaiser, D.; Cichy, R. M.; and Kietzmann, T. C. 2025. End-to-end topographic networks as models of cortical map formation and human visual behaviour. *Nature Human Behaviour*, 1–17.

Margalit, E.; Lee, H.; Finzi, D.; DiCarlo, J. J.; Grill-Spector, K.; and Yamins, D. L. 2024. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14): 2435–2451.

Moran, P. A. 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2): 17–23.

Poli, M.; Dupoux, E.; and Riad, R. 2023. Introducing topography in convolutional neural networks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Qian, X.; Dehghani, A. O.; Farahani, A. B.; and Bashivan, P. 2024. Local lateral connectivity is sufficient for replicating cortex-like topographical organization in deep neural networks. *bioRxiv*, 2024–08.

Rathi, N.; Mehrer, J.; AlKhamissi, B.; Binhuraib, T.; Blauch, N. M.; and Schrimpf, M. 2025. TopoLM: brain-like spatio-functional organization in a topographic language model. arXiv:2410.11516.

Sereno, M. I.; Dale, A. M.; Reppas, J. B.; Kwong, K. K.; Belliveau, J. W.; Brady, T. J.; Rosen, B. R.; and Tootell, R. B. 1995. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212): 889–893.

Truong, N.; and Hasson, U. 2025. Improved Robustness and Functional Localization in Topographic CNNs Through Weight Similarity. *arXiv preprint arXiv:2508.00043*.

Tsao, D. Y.; Freiwald, W. A.; Tootell, R. B.; and Livingstone, M. S. 2006. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761): 670–674.

Yamins, D. L.; Hong, H.; Cadieu, C. F.; Solomon, E. A.; Seibert, D.; and DiCarlo, J. J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23): 8619–8624.

Zhuang, C.; Yan, S.; Nayebi, A.; Schrimpf, M.; Frank, M. C.; DiCarlo, J. J.; and Yamins, D. L. 2021. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3): e2014196118.