Response Patterns to Rotation Angle in a Rotation Pretext Task Vary Across Datasets and Architectures

An Observation and a Negative Result

Paul Yan PAUL.YAN@MAIL.UTORONTO.CA
Amy Saranchuk AMY.SARANCHUK@MAIL.UTORONTO.CA
Michael Guerzhoy GUERZHOY@CS.TORONTO.EDU
University of Toronto

Abstract

We show that the response to rotation angle θ in a rotation-based pretext task in self-supervised pretraining (SSP) via contrastive learning interacts in systematic, dataset-dependent, and architecture-dependent ways that produce unique "signature" curves of performance versus θ. We perform a comprehensive 16×16 experiment, pre-training eight encoder architectures on 16 diverse image datasets using both SimCLR and MoCo v2, with θ swept from $0°$ to $360°$ in $0.1°$ increments. Each of the resulting 256 accuracy-versus-θ plots exhibits a distinct periodic pattern. A simple classifier trained on these curves can predict the originating dataset and encoder–method pair with high accuracy, confirming patterns specific to both datasets and architectures.

In a preliminary experiment on three medical imaging datasets (BraTS, Lung Mask, Kvasir-SEG), we measure Dice scores between ground-truth masks and saliency maps from ResNet-50, ConvNeXt-Tiny, and ViT-B/16 encoders pre-trained at fixed θ, observing clear dataset-specific oscillations. We report a negative result: Histogram-of-Gradients (HoG) features do not explain the phenomenon. We find a fascinating and previously undocumented "fingerprinting" effect linking augmentation choices to data and architecture and a negative finding about a mechanistic explanation for it.

Keywords: self-supervised pretraining, contrastive learning, rotation augmentation, periodicity, representation learning, medical imaging, learning shortcuts

1. Introduction

Self-supervised pre-training (SSP) increases a network's ability to learn low-level features, which is especially valuable when training data are scarce. In contrastive pre-training, the model is taught to tell apart different transforms of the same input — for example, distinguishing an original image from its version rotated by an angle θ versus unrelated image pairs. We show that the choice of rotation angle θ interacts in surprising ways not only with the characteristics of the dataset but also with the encoder architecture during contrastive training. We also demonstrate that a classifier trained on the curves of encoder performance versus rotation angle θ can accurately predict both the underlying dataset and the encoder architecture, indicating that each architecture and dataset have their own "signature" periodic pattern.

We conducted three complementary experiments to investigate how fixed rotation angles θ for image augmentation (without masking, shifting, or cropping) affect contrastive learning performance using MoCo and SimCLR.

We investigate the response pattern to different datasets and architectures by selecting 16 datasets (Table 2), 8 encoder architectures (Table 1), and two contrastive learning method (SimCLR and MoCo v2), training each encoder–dataset pair using both MoCo and SimCLR with rotation angles θ from 0° to 360° in 0.1° increments. We then evaluated each pre-trained encoder on a binary classification task, yielding $16 \times 16 = 256$ classification accuracy-versus-θ curves (Fig. 1 the full results). This experiment ("16×16 experiment") is described in Section 2.1.

Next, we conducted a prediction experiment to test whether the plots from the 16×16 experiment contain distinctive patterns that let a classifier identify their originating dataset and encoder–method pair. As shown in Fig. 2, classification accuracy was very high, indicating that each setting indeed has unique identifiable features. This experiment is described in Section 2.2.

In a preliminary experiment, we focus on three medical datasets — BraTS Menze et al. (2014), Lung Mask lun (2020), and Kvasir-SEG Jha et al. (2020) — and trained *ResNet-50*, *ConvNeXt-Tiny*, and *ViT-B/16* with MoCo under fixed θ augmentations. We measured performance via the Dice score between predicted and ground-truth segmentations and observed clear periodic oscillations (Fig. 3). This experiment ("segmentation experiment") is described in Section 2.3.

Finally, we test a *shortcut* hypothesis: that rotation-dependent reliance on HoG-like features Lowe (1999); Dalal and Triggs (2005) produces the observed patterns. We train an SVM on HoG descriptors to classify rotated vs. unrotated images and compare model saliency with ground-truth masks (Section 2.4). The SVM does not reproduce the characteristic periodicities (Fig. 3), so our preliminary results do not support this explanation; see Section 2.4 for details.

2. Experiments

2.1. 16×16 Experiment

This experiment quantifies how a fixed pretext rotation angle θ affects downstream classification accuracy across datasets, architectures, and contrastive methods. We evaluate 16 datasets (Section E) and 8 encoder architectures (Section D) with SimCLR and MoCo v2, using their default hyperparameters and ImageNet-pretrained weights for all encoders. The architectures and datasets, chosen to span diverse image domains, are summarized in Tables 1 and 2.

For each encoder–dataset–method triple, we run contrastive pretraining at 3,600 angles, $\theta \in \{0, 0.1, \ldots, 359.9\}^\circ$, resetting encoder weights before every run. Positives are $(x, \text{rotate}_\theta(x))$; no other augmentations are used.

After contrastive pre-training, we freeze the encoder and train a binary classifier on concatenated features from image pairs. Positives use $(f(x), f(\text{rotate}_\theta(x)))$; negatives use $(f(x), f(x'))$, where x' denotes a different image randomly sampled from the dataset. The classifier is a six-layer ReLU MLP with a final sigmoid, trained with binary cross-entropy and identical hyperparameters for all $256 = 8 \times 16 \times 2$ encoder–dataset–method combinations. We construct a balanced dataset by encoding each image and its rotated version to form positives, and by pairing each x with a randomly sampled x' to form negatives.

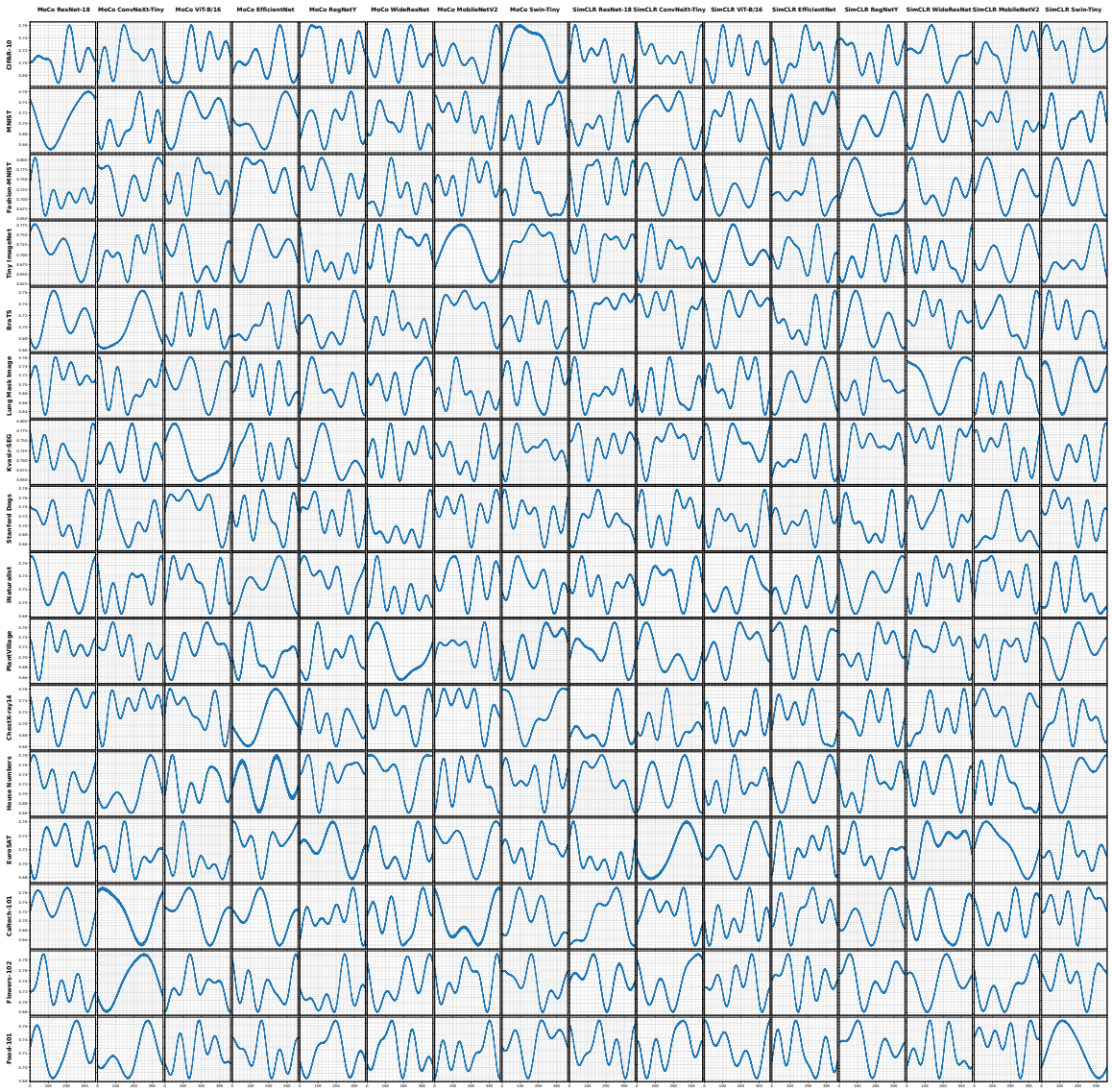

Figure 1: Each row is a dataset; each column is a contrastive-method/encoder combo. Each subplot shows accuracy vs. rotation angle (0–360°). Note the distinct periodic patterns unique to each dataset-architecture pair.

Results (Fig. 1) show dataset- and architecture-dependent periodicity; implementation details are in Section B.

2.2. Prediction Experiment

The goal of this experiment is to train a classifier that, given an accuracy-versus-θ curve from the 16×16 experiment, can identify its originating dataset (row) and architecture (column)

with high accuracy, for a held-out curve. Strong performance indicates that each dataset and architecture produces a distinct, recognizable pattern. We find that it is possible to predict with high accuracy which architecture and dataset held-out signature curves were generated from. The results are summarized in Fig. 2. More details on the prediction experiment are given in Section C.

2.3. Segmentation Experiment

This experiment revisits the effect of a fixed rotation angle θ on downstream performance using a *standard U-Net* segmentation pipeline (rather than saliency maps). We evaluate three medical datasets—BraTS Menze et al. (2014), Lung Mask Image Dataset lun (2020), and Kvasir-SEG Jha et al. (2020)—and three encoders: ResNet-50, ConvNeXt-Tiny, and ViT-B/16. For each encoder and each angle $\theta \in \{0°, 1°, \ldots, 360°\}$, we run MoCo v2 contrastive pretraining with fixed-angle positives: a positive pair is the original image and the same image rotated by θ (no other augmentations). For each $(\text{dataset}, \text{encoder}, \theta)$ triple, we instantiate a U-Net in which the encoder path is initialized from the corresponding MoCo(θ)-pretrained backbone (ResNet-50 / ConvNeXt-Tiny / ViT-B/16). The U-Net decoder is a conventional upsampling stack with skip connections from the encoder stages. We then train the U-Net supervised on the dataset's ground-truth masks, fine-tuning end-to-end with a soft dice loss function, Adam optimizer, and early stopping on a validation split. For each fixed θ, we measure the mean Dice score on a held-out set and record this as one point on the Dice-vs-θ curve. Repeating this over $\theta = 0° \ldots 360°$ yields one curve per dataset–encoder pair. We summarize these results in a 3×3 grid (rows: datasets; columns: encoders), which exhibits clear, dataset-specific oscillations (Fig. 3).

2.4. Shortcut Hypothesis Experiment

We hypothesized that MoCo may use HoG-like features as "shortcuts." To test this, we extracted HoG descriptors from each image both before and after applying a rotation of θ. A support vector machine (SVM) was then trained (using cross-validation to select its hyperparameters) to distinguish rotated from unrotated HoG feature vectors at each θ. We plotted the SVM's classification accuracy as a function of the rotation angle. A close correspondence between the SVM's accuracy curve and MoCo's orientation-dependent performance would support the idea that MoCo similarly relies on orientation-specific HoG-like cues. Fig. 3 plots SVM accuracy against rotation angle θ for all three datasets. We compare these SVM accuracy curves to the MoCo model's Dice score plots (also in Fig. 3). They do not match very well. More details are in Section F.

3. Conclusions

Encoder performance under rotation-based self-supervised pretraining varies with angle θ in a periodic, non-monotonic manner, with patterns that depend on both dataset and architecture. Each dataset–encoder pair shows a distinctive signature in our 16×16 grid. We did not find evidence that edge-based shortcuts explain the phenomenon, and a mechanistic explanation remains an open question.

References

Lung mask image dataset, 2020. Dataset available from the corresponding source or institution.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, pages 446–461, 2014.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 2020a. doi: https://doi.org/10.48550/arXiv.2002.0570.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020b. URL https://arxiv.org/abs/2002.05709.

Alexander Chowdhury, Jacob Rosenthal, Jonathan Waring, and Renato Umeton. Applying self-supervised learning to medicine: Review of the state of the art and medical implementations. *Informatics*, 2021. doi: https://doi.org/10.3390/informatics8030059.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. 2020. doi: https://doi.org/10.48550/arXiv.1911.05722.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 1–4, 2019.

Shih-Cheng Huang, Malte Jensen Anuj Pareek, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6(74), 2023. doi: https://doi.org/10.1038/s41746-023-00811-0.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 2021. doi: https://doi.org/10.3390/technologies9010002.

Debadatta Jha, Hassaan Akbari, Pål Halvorsen, Marte Lindset, and Ebrahim Abir. Kvasir-seg: A segmented polyp dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 449–458, 2020.

Aditya Khosla, Ning Jia, and Fei-Fei Sun. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4707–4714, 2011.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Available at https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Minh Le et al. Tiny imagenet visual recognition challenge, 2015. Available at https://tiny-imagenet.herokuapp.com/.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhuang Liu, Han Hu, Stephen Lin, Guoqing Hua, Xiatian Liang, Hong Qin, Yixuan Zhang, Yue Wang, Zhibo Li, and Tao Mei. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

Bjoern H Menze, Andras Jakab, Steffen Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, John Kirby, Yvonne Burren, Nicole Porz, Jan Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. In *Frontiers in Plant Science*, volume 7, page 1419, 2016.

Yuval Netzer, Tao Wang, Adam Coates, Anna Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Maureen-Eve Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *European Conference on Computer Vision (ECCV)*, pages 190–203, 2008.

R. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Veenu Rani, Syed Tufael Nabi, Munish Kumar, Ajay Mittal, and Krishan Kumar. Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, 30:2761–2775, 2023. doi: https://doi.org/10.1007/s11831-023-09884-2.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Deval Shah and Abhishek Jha. Self-supervised learning and its applications. URL https://neptune.ai/blog/self-supervised-learning.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Garrett Van Horn, Steve Branson, R. Farrell, S. Haber, C. Barry, and D. Held. The inaturalist challenge 2018, 2018. Available at https://www.inaturalist.org/pages/2018_inaturalist_challenge.

Xiaosong Wang, Yifan Peng, Lu Lu, Ziyan Lu, Mohammadhossein Bagheri, and Ronald M. Summers. Chestx-ray14: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Han Xiao, Kashif Rasul, and Robert Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jiashu Xu. A review of self-supervised learning methods in the field of medical image analysis. *International Journal of Image, Graphics and Signal Processing*, 4:33–46, 2021. doi: http://dx.doi.org/10.5815/ijigsp.2021.04.03.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

Appendix A. Background: Self-Supervised Pretraining

Self-supervised pre-training (SSP) lets the model create its own training signals from unlabeled data, rather than relying on human labels Rani et al. (2023); Chowdhury et al. (2021). SSP usually involves two stages:

- **Pretext task:** The model is trained on an artificial task (e.g., predicting image rotations or solving a jigsaw puzzle) to learn useful features.
- **Downstream task:** The pre-trained model is fine-tuned on the actual task of interest (e.g., classification or segmentation) Shah and Jha.

Pretext tasks generate pseudo-labels that force the network to discover patterns like edges, colors, and shapes in the data. Common pretext examples include image recoloring, rotation prediction, and jigsaw solving Xu (2021). Once the model learns these basic representations, it transfers them to help solve the downstream task.

Contrastive learning trains a model by bringing similar examples closer in feature space and pushing dissimilar ones apart. Two leading frameworks are SimCLR Chen et al. (2020b) and Momentum Contrast (MoCo) He et al. (2020). Both build on the idea that different augmentations of the same image share the same semantic content.

SimCLR is a simple, end-to-end approach that:

- Creates a *positive pair* by applying two random augmentations to the same image. In this study, positive pairs were created by applying a constant rotation angle θ to every image in the dataset.
- Treats every other image in the batch as a *negative example.*
- Uses the NT-Xent loss to pull positive pairs together and push negatives apart, all within each batch.

MoCo extends this idea by:

- Maintaining a dynamic queue (dictionary) of encoded "keys" from previous batches as extra negatives.
- Using a momentum-updated encoder for the keys, which stabilizes training when batch sizes are smaller.

Both methods learn high-quality image representations that transfer well to downstream tasks by leveraging unlabeled data Rani et al. (2023); Xu (2021); Jaiswal et al. (2021); Huang et al. (2023); Chen et al. (2020a).

Appendix B. Details: 16×16 experiment

All classifiers were trained using identical protocols and hyperparameter settings:

- Optimizer: SGD with a learning rate of 0.015.
- Batch size: 64.
- Number of epochs: 20.
- Data split: 80% for classifier training and 20% for evaluation.

After training, we evaluate each classifier on the test set to obtain its classification accuracy. Each accuracy value represents a single point on an accuracy-versus-θ curve, and since θ ranges from 0° to 360° in 0.1° increments, each of the 256 curves contains 3600 data points.

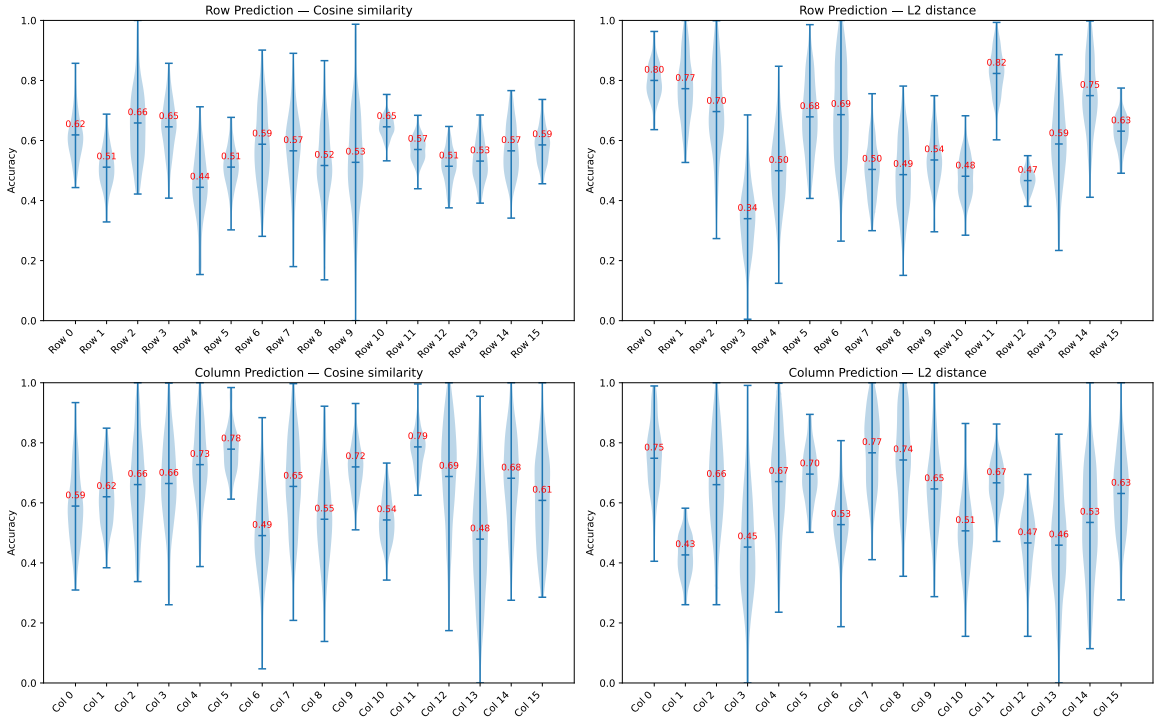

Figure 2: 2×2 violin plots of prediction accuracy: rows—row (top) vs. column (bottom) prediction; columns—cosine (left) vs. L2 (right). Each violin shows the distribution; red numbers mark means.

Appendix C. Details: Prediction experiment

We organize all $R = 16$ datasets and $C = 16$ encoder–method combinations into a tensor

$$\mathbf{X} \in \mathbb{R}^{R \times C \times T} \tag{1}$$

where $T = 3600$ is the number of measurements at 0.1° increments over a full 0°–360° rotation. Each slice $\mathbf{X}_{i,j,:}$ is the curve for dataset i and encoder–method j.

For row prediction, we split each row's 16 curves into a training set of columns 1–10 and a test set of columns 11–16:

$$\mathcal{C}_{\text{train}} = \{1, \ldots, 10\} \quad \mathcal{C}_{\text{test}} = \{11, \ldots, 16\} \tag{2}$$

We then learn a weight vector $\mathbf{a} \in \mathbb{R}^R$ for each row r by maximizing intra-row similarity on its training curves:

1. **Initialization & optimization.** Set $\mathbf{a} = \mathbf{1}$. Use the Adam optimizer (learning rate 0.01) for 200 epochs.

2. **Mini-batch sampling.** In each epoch, sample 20 column indices j (with replacement) from $\mathcal{C}_{\text{train}}$.

3. **Similarity & logits.** For each sampled j, let $\mathbf{x} = \mathbf{X}_{r,j,:}$. Compute

$$S_{i,k} = \mathrm{sim}\big(\mathbf{x},\, \mathbf{X}_{i,k,:}\big) \quad i \in \{0, \ldots, R-1\},\ k \in \mathcal{C}_{\mathrm{train}}. \tag{3}$$

Then form per-row logits

$$s_i = a_i \times \begin{cases} \dfrac{1}{|\mathcal{C}_{\mathrm{train}}| - 1} \displaystyle\sum_{\substack{k \in \mathcal{C}_{\mathrm{train}} \\ k \neq j}} S_{r,k}, & i = r \\ \dfrac{1}{|\mathcal{C}_{\mathrm{train}}|} \displaystyle\sum_{k \in \mathcal{C}_{\mathrm{train}}} S_{i,k}, & i \neq r \end{cases} \tag{4}$$

Apply cross-entropy loss against the one-hot label for row r and backpropagate to update $\mathbf{a}$.

4. **Evaluation on held-out columns.** After convergence, we assess performance on the test columns as follows:

   (a) Sample 100 indices j from $\mathcal{C}_{\mathrm{test}}$ with replacement.

   (b) For each sampled j, compute

$$S_{i,k} = \mathrm{sim}\big(\mathbf{X}_{r,j,:},\, \mathbf{X}_{i,k,:}\big), \quad i \in \{1, \ldots, 16\},\ k \in \mathcal{C}_{\mathrm{test}}. \tag{5}$$

   (c) For each i, compute the mean similarity across the test columns by applying (4) with $\mathcal{C}_{\mathrm{test}}$ in place of $\mathcal{C}_{\mathrm{train}}$.

   (d) Multiply by the learned weights and predict:

$$s_i = a_i\, \bar{S}_i, \quad \hat{r} = \arg\max_i s_i. \tag{6}$$

   (e) The row-classification accuracy for dataset r is the fraction of trials where $\hat{r} = r$.

An analogous procedure applies when predicting encoder–method combinations by swapping the roles of rows and columns.

Appendix D. Architectures

Table 1: Summary of Encoder Architectures

Encoder	**Architecture Type**	**# Params**
ResNet-18He et al. (2016)	CNN with residual connections	11.7M
ConvNeXt-TinyLiu et al. (2022)	Modernized CNN (inspired by transformers)	29M
ViT-B/16Dosovitskiy et al. (2020)	Vision Transformer (Transformer-based)	86M
EfficientNet-B0Tan and Le (2019)	CNN with compound scaling	5.3M
RegNetY-400MFRadosavovic et al. (2020)	CNN (simple design space-based)	2.7M
WideResNet-50-2Zagoruyko and Komodakis (2016)	Wider variant of ResNet-50 (CNN)	68M
MobileNetV2Sandler et al. (2018)	Depthwise separable CNN	3.4M
Swin-TinyLiu et al. (2021)	Hierarchical Vision Transformer	29M

Appendix E. Datasets

Table 2: Summary of Datasets

Dataset	**# Examples**	**Image Dimensions**	**# Channels**	**Domain**
CIFAR-10Krizhevsky (2009)	60,000	32×32	3	Natural objects
MNISTLeCun et al. (1998)	70,000	28×28	1	Handwritten digits
Fashion-MNISTXiao et al. (2017)	70,000	28×28	1	Clothing items
Tiny ImageNetLe et al. (2015)	100,000	64×64	3	Natural images
BraTSMenze et al. (2014)	~ 300 (cases)	$\sim 240 \times 240$ (slice)	4	Medical (MRI)
Lung Mask Image Datasetlun (2020)	$\sim 1{,}000$	$\sim 512 \times 512$	1	Medical (X-ray)
Kvasir-SEGJha et al. (2020)	$\sim 1{,}000$	Variable (e.g., $\sim 512 \times 512$)	3	Medical (Endoscopy)
Stanford DogsKhosla et al. (2011)	20,580	Variable (resized to $\sim 224 \times 224$)	3	Natural (Animals)
iNaturalistVan Horn et al. (2018)	$\geq 100{,}000$	Variable (resized to $\sim 224 \times 224$)	3	Natural (Biodiversity)
PlantVillageMohanty et al. (2016)	$\sim 54{,}000$	Variable (resized to $\sim 256 \times 256$)	3	Agricultural
ChestX-ray14Wang et al. (2017)	112,120	$\sim 1024 \times 1024$ (often resized)	1	Medical (X-ray)
Street View House NumbersNetzer et al. (2011)	$\sim 100{,}000$	32×32	3	Natural (Scene text)
EuroSATHelber et al. (2019)	27,000	64×64	3	Remote sensing
Caltech-101Fei-Fei et al. (2004)	9,146	Variable (e.g., $\sim 227 \times 227$)	3	Natural (Objects)
Flowers-102Nilsback and Zisserman (2008)	8,189	Variable (e.g., $\sim 224 \times 224$)	3	Natural (Floral)
Food-101Bossard et al. (2014)	101,000	Variable (e.g., $\sim 224 \times 224$)	3	Natural (Food)

Appendix F. Shortcut hypothesis details

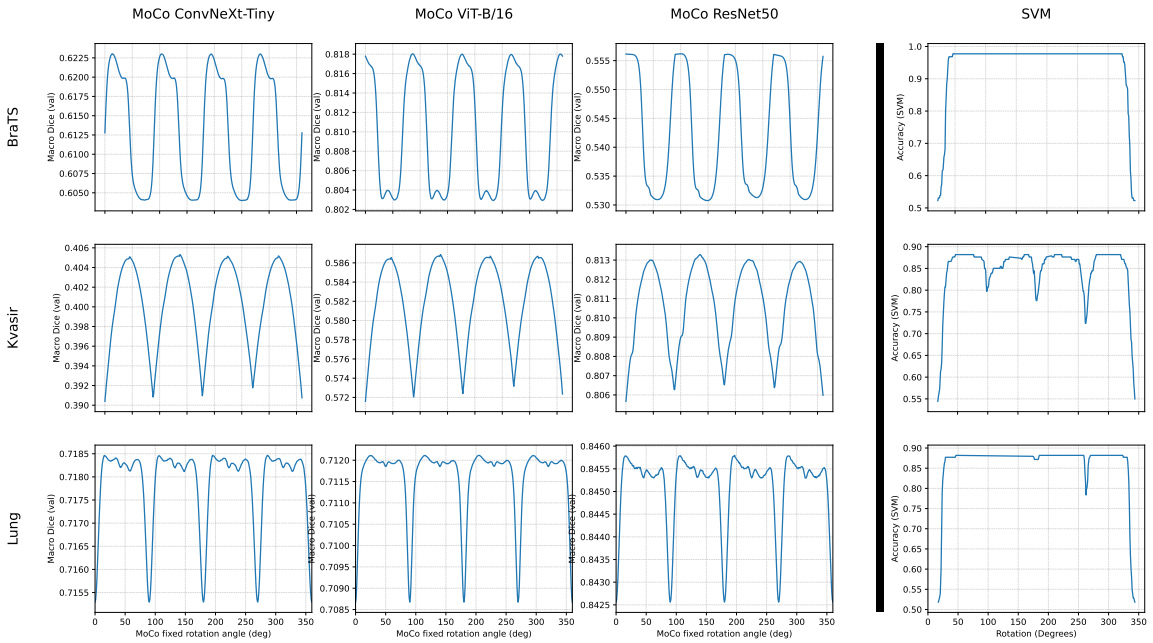

Figure 3: Average correspondence (Dice score) between ground-truth segmentations and saliency maps (left three columns) and SVM classification accuracy for original vs. rotated images (right column), each plotted against rotation angle (0°−360°). Rows correspond to datasets (BraTS, Kvasir, Lung); left columns are MoCo encoders (ConvNeXt-Tiny, ViT-B/16, ResNet50). Dice panels report performance vs. the rotation angle used during pre-training—higher is better—and reveal dataset-specific periodic patterns.

In Fig. 3, angles of 45°, 90°, 135°, ... consistently coincide with the local minima and maxima in the correspondence between saliency maps and the segmentation ground truth. We hypothesize that, when good features are learned, the saliency maps correspond to the segmentation ground truth better. We hypothesize that worse features are learned when the network can take "shortcuts" in figuring out the angle θ. For example, the network could rely on Histogram-of-Gradients (HoG) Lowe (1999)-like features. (Although those features are famously good, they are not specific to our dataset; we do not expect that learning HoG features would be a part of a successful fine-tuning of a network that was already pretrained on ImageNet.)

To explore this hypothesis, we train SVMs to classify the HoG features of images rotated by θ vs unrotated images. We compute and concatenate HoG descriptors for every 64×64 cell. We use cross-validation to select the best parameters for a Guassian-kernel SVM.

The results are in Fig. 3. We observe that the HoG classification accuracy is low for θ close to 0° (and 360°), reflecting the increased difficulty of the task.

For the Kvasir-SEG dataset, we observe minima in the accuracy of the HoG classifier that correspond to minima in the correspondence between the saliency map and the ground truth segmentation. This seems to be evidence against our theory: when it is more difficult to classify based on HoGs and there are no shortcuts (at least via HoGs), it seems that the correspondence between the saliency map and the ground truth segmentation is lower.

Appendix G. Segmentation experiments: illustrations

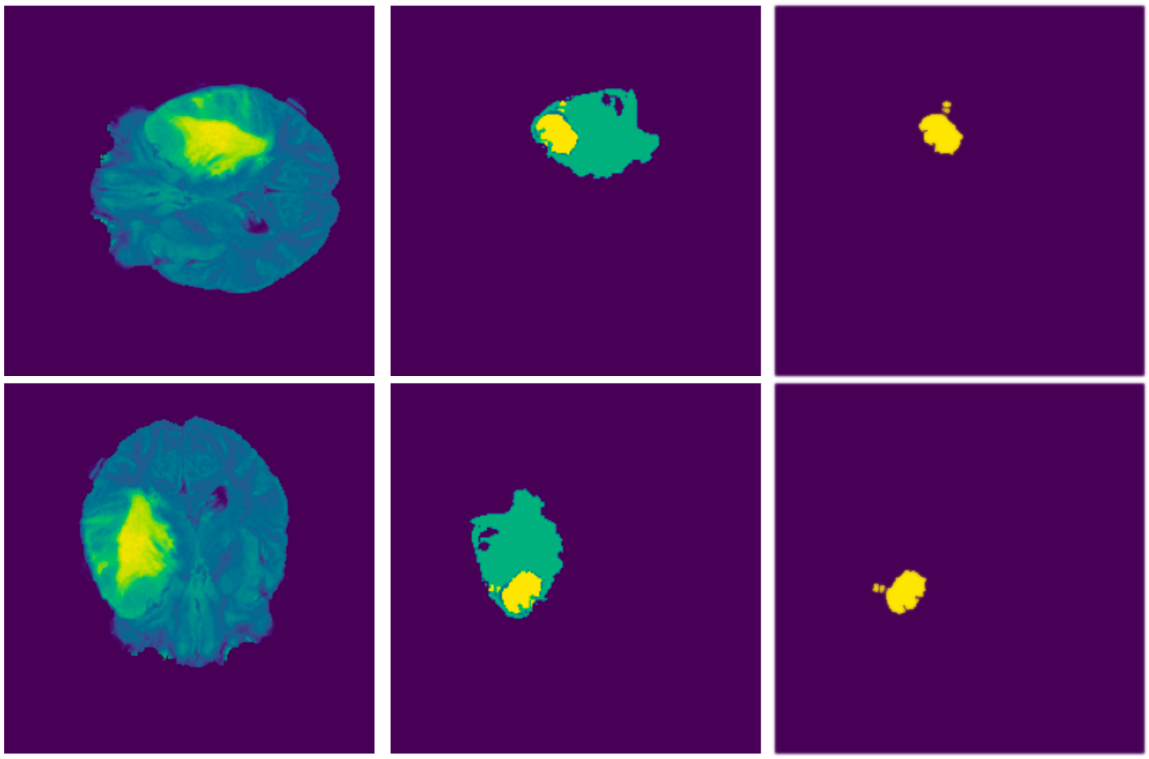

Figure 4: Comparison of original images from BraTS2020 dataset, true segmentation masks, and predicted segmentation. Top row: Original orientation. Bottom row: Rotated by 95 °.