EARLY STRUCTURAL EXPERIENCE SHAPES VISUAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

During development, infants initially experience a visually degraded world, with limited spatial acuity, contrast sensitivity, chromatic sensitivity, and binocular depth perception that gradually mature over the first months of life. This developmental trajectory may bias early learning toward more stable structural cues, such as boundaries and shape, before richer appearance cues become available. Early-learned structural representations could subsequently serve as attractors that guide the interpretation of richer visual experiences, consistent with infants' and children's preference for line drawings and cartoons. Here, we investigate this developmental hypothesis in deep neural networks by using semantically meaningful line drawings as a surrogate of early structural visual experience. Line drawings preserve object boundaries, global shape, and part configurations while suppressing much of the texture and color information present in photographs. We find that early exposure to line drawings systematically reshapes subsequent learning from photographs toward more shape-centered representations. These representational changes improve performance on tasks requiring geometric and spatial understanding, including segmentation and depth estimation, in addition to enhancing recognition, detection, data efficiency, and knowledge distillation. Our findings suggest that early learning based on stable structural information creates a representational foundation to guide the learning of richer visual knowledge in later experience.

1 INTRODUCTION

Visual perception develops along a constrained developmental trajectory. Unlike adults, whose visual experience is dominated by high-resolution, richly colored, and stereoscopic information, infants initially encounter a visually limited world characterized by immature spatial resolution, contrast sensitivity, chromatic sensitivity, and binocular depth perception (Banks & Salapatek, 1981; Brown & Lindsey, 2009; Brown, 1990; Braddick & Atkinson, 2011). These early visual limitations may help focus learning on relatively stable structural information in the visual world, including object boundaries, global shape, and the spatial configuration of parts. Such information could provide a foundation for learning from later experience, when more variable appearance cues such as texture, color, and illumination become available. From this perspective, developmental constraints do not simply limit perception; they may actively scaffold the formation of efficient and robust visual representations. This raises a fundamental question about visual learning: how might early exposure to predominantly structural information shape the representations that emerge through subsequent learning?

We introduce a structure-first learning paradigm to investigate how early structural experience shapes later visual learning: models initially learn from semantically meaningful line drawings before continuing their training on photographs. Line drawings provide a useful model of such structurally biased visual input. Unlike photographs, they emphasize semantically meaningful and relatively stable object structure while suppressing fine-grained appearance information. Infants can categorize cat and dog silhouettes (Quinn et al., 2001), and some chimpanzees can generalize categories learned from photographs to drawings (Tanaka, 2007). Shape also guides children's generalization of object names, and experience learning shape-organized categories can strengthen this bias (Landau et al., 1988; Smith et al., 2002). Despite their sparse visual content, line drawings support human object recognition and serve as tools for communication, reasoning, and learning

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Fan et al., 2023). Neuroimaging further shows that natural scene categories can be decoded from visual cortical responses to line drawings as accurately as from responses to photographs (Walther et al., 2011), highlighting the information preserved in scene structure. Furthermore, drawing skill is associated with more efficient encoding of object structure and better visual memory during drawing (Perdreau & Cavanagh, 2014; 2015). Together, these observations suggest that line drawings preserve stable structural information that is highly relevant to the human visual system.

We evaluate this structure-first approach in deep networks across architectures and visual tasks, examining how early structural experience reshapes subsequent learning from photographs. Models develop stronger shape bias, more selective and sparse responses, lower-dimensional population representations, and increased shape tuning at the levels of individual neurons and neuronal populations. These representational changes improve transfer to tasks requiring spatial and geometric understanding, including segmentation and depth estimation, while also enhancing recognition, detection, data efficiency, and knowledge distillation.

2 Related Work

Developmentally Motivated Visual Learning. Computational studies have examined how the conditions of early visual experience affect learned representations. Vogelsang et al. (2018) found that initial training on blurred images promoted larger receptive fields and improved generalization across image resolutions in convolutional networks, providing a model of potential benefits from initially limited visual acuity. Zhuang et al. (2021) showed that unsupervised contrastive learning can produce representations predictive of responses in the primate ventral visual stream, including when models are trained on children's head-mounted video data. Together, these studies connect visual development to the information available during learning and the objective used to learn from it. We examine a complementary intervention: an initial phase emphasizing semantically meaningful contours and spatial structure, followed by training on photographs. We evaluate how its effects on representations and task performance persist after this transition.

Line Drawing Synthesis and Structural Representation. Line-drawing synthesis includes distinct approaches to representing and generating structure. Sketch-rnn and BézierSketch model human sketches as sequences of strokes (Ha & Eck, 2018; Das et al., 2020); reinforced adversarial learning synthesizes drawing programs without ground-truth programs (Ganin et al., 2018); and distance-field methods predict parametric shapes from images (Smirnov et al., 2020). Photo-to-drawing translation includes APDrawingGAN, trained on paired face photographs and artistic portraits (Yi et al., 2019), whereas CycleGAN learns translation between unpaired image domains (Zhu et al., 2017). Chan et al. (2022) use geometry and semantic losses to generate informative line drawings without paired photo–drawing examples. We use their structurally informative drawings as the initial experience in a learning curriculum, testing how this experience affects later learning from photographs.

Visual Pre-training and Curriculum Learning. Our curriculum changes the visual information available early in training and can complement self-supervised pre-training objectives. Methods like Masked Autoencoders (MAE) (He et al., 2022) learn by reconstructing masked patches, while DINO (Caron et al., 2021) uses self-distillation and SwAV (Caron et al., 2020) matches cluster assignments across views. These methods are typically evaluated with pre-training on photographs. This leaves open whether an initial phase of structural input can improve subsequent learning from photographs. We integrate an initial line-drawing phase into DINO and SwAV, then continue pre-training on photographs. This tests whether early structural experience remains useful when representations are learned without class-label supervision. The resulting improvements under matched or reduced pre-training epoch budgets show that the benefits of the curriculum extend beyond supervised classification to established self-supervised learning frameworks.

Texture versus Shape and Structure Bias. Shape and texture biases reveal how visual experience influences the information models use for recognition. Geirhos et al. (2019) showed that ImageNet-trained convolutional networks favor texture and that stylized training increases shape bias. Subsequent work demonstrated the influence of data augmentation (Hermann et al., 2020), including shape-focused augmentation (Lee et al., 2022), and activation sparsity (Li et al., 2023).

These findings establish that structural preferences can be shaped by training inputs and representational constraints. We ask whether such preferences can be established through early experience and retained during later learning: does emphasizing contours and spatial structure initially leave a lasting influence after training on photographs? This connects shape bias to our broader developmental question, which we examine through both recognition behavior and internal neuronal responses.

Sparsity in Neural Representations. Sparse coding has long been theorized as a mechanism for efficient visual processing. Li et al. (2023) show that Top-K activation sparsity in convolutional networks increases shape bias and promotes part-based representations. Masked autoencoders (He et al., 2022) instead sparsify the visible input by masking image patches, rather than imposing sparsity on neuronal activations. Line drawings offer another form of sparse input by emphasizing contours. Our curriculum tests whether this initial structural emphasis encourages sparse, structure-focused responses that persist after learning from photographs. The association between drawing skill and efficient encoding of object structure (Perdreau & Cavanagh, 2014) offers a perceptual parallel, although it does not establish a shared neural mechanism.

3 METHOD

We investigate the lasting effects of early structural experience using a structure-first curriculum. The pretrained model of Chan et al. (2022) generates line drawings from photographs, emphasizing contours and spatial structure while suppressing color and texture. Figure 1 summarizes this intervention and representative downstream tasks and representations.

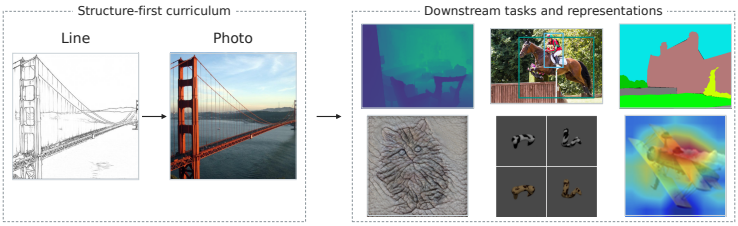


Figure 1: Models learn from structure before continuing training on photographs. The right-hand panels illustrate a subset of the downstream tasks and representations.

We use Line and Photo as shorthand for line drawings and photographs in condition labels, figures, and tables. In Line→Photo, a model first learns from line-drawing versions of the training images, then continues from the same parameters on their photographic counterparts. This transition lets us examine which effects of early structural experience persist after photographic adaptation. Photo-only training provides the primary comparison. Our experiments span convolutional and transformer architectures and include supervised classification and self-supervised pretraining. In supervised experiments, the line drawings and photographs retain the same class labels; in self-supervised experiments, the curriculum is incorporated into DINO and SwAV.

The central prediction is that early structural experience changes which information remains accessible after photographic learning. Shape–texture cue conflicts test recognition decisions, while crossed structure–texture stimuli test neuronal sensitivity as these factors vary independently. Agreement across these measurements would connect a behavioral preference to the information encoded internally. Transfer evaluations then test whether the resulting representations support other visual tasks. Together, these tests examine both the persistence of early structural learning and its consequences for later visual competence.

Our evaluation traces these effects through three stages: recognition, data efficiency, and self-supervised learning establish benefits for photographic learning; structural preferences, neuronal responses, dimensionality, and layer alignment and truncation characterize the resulting representations; depth estimation, segmentation, and distillation test their transfer. Additional controls, spatial attribution, feature dispersity, and detection results appear in the appendix. Datasets, architectures, training budgets, and evaluation procedures accompany each experiment, with further implementation details in the appendix.

4 RESULTS

We first establish the effects of early structural experience on photographic learning, then examine the structural preferences and representations retained after adaptation and their utility for depth estimation, segmentation, and distillation.

4.1 EARLY STRUCTURAL EXPERIENCE IMPROVES PHOTOGRAPHIC LEARNING

Photograph recognition. Beginning with line drawings improves subsequent recognition of photographs (Table 1). On STL10, Line→Photo increases ResNet-18 accuracy from 68.79% to 73.58%, a gain of 4.79 percentage points. ViT-Tiny gains 5.93 points on STL10, and a YOLO-v8 backbone gains 0.95 points on ImageNet-1K, extending the benefit across the tested convolutional and transformer architectures. After photographic training, Line→Photo also improves ResNet-18 recognition of line drawings from 16.67% (Photo-only) to 29.53%. The full comparison of training and evaluation modalities is provided in Appendix A.10.

Table 1: Photograph classification accuracy (%) after Photo-only or Line→Photo training. Differences are reported in percentage points (pp).

Architecture	Dataset	Photo-only	Line→Photo	Δ (pp)
ResNet-18	STL10	68.79	**73.58**	+4.79
ViT-Tiny	STL10	55.04	**60.97**	+5.93
YOLO-v8 backbone	ImageNet-1K	67.87	**68.82**	+0.95

Learning from fewer examples. We next vary the STL10 training subset from 5% to 100% for ResNet-18 and from 10% to 100% for ViT-Tiny, using stratified sampling (Appendix A.8). Line→Photo improves photograph recognition at every tested data fraction for both architectures (Figure 2). With 70% of the training data, ResNet-18 reaches 69.15% accuracy, exceeding the Photo-only model's 68.79% with the full dataset. ViT-Tiny reaches 54.75% with 60% of the data, approaching the full-data Photo-only accuracy of 55.04%. The curriculum therefore reduces the number of examples needed to approach or exceed full-data baseline performance. Comparisons with Canny edges and artistic-style transformations are reported in Appendix A.1.



(a) ResNet-18 **(b) ViT-Tiny**

Figure 2: Photograph recognition on STL10 as a function of the training subset size. The gray dashed line in each panel marks the corresponding Photo-only accuracy with 100% of the training data.

Self-supervised learning. We next ask whether the benefit extends to learning without class-label supervision during pretraining. We incorporate the structure-first curriculum into DINO and SwAV, using ResNet-18 on ImageNet-100. For each method, Photo-only pretraining uses 400 epochs of photographs. The Line→Photo curricula begin with 150 epochs of line drawings, followed by either 250 epochs of photographs, matching the baseline's 400-epoch budget, or 100 epochs of photographs, for a reduced total of 250 epochs (Table 2).

With the matched 400-epoch budget, the structure-first curriculum increases top-1 accuracy by 2.30 percentage points for DINO and 3.34 for SwAV. Both 250-epoch curricula also achieve comparable or higher top-1 accuracy than their 400-epoch Photo-only baselines, using 37.5% fewer pretraining epochs. These comparisons show that the curriculum's benefit extends to self-supervised learning and persists under matched or reduced epoch budgets.

Table 2: Self-supervised pretraining on ImageNet-100 with ResNet-18. Matched (400-epoch) and reduced (250-epoch) curricula are compared with 400-epoch Photo-only baselines. Bold marks the highest accuracy within each method.

		Pretraining epochs			Accuracy (%)	
Method	Curriculum	Line	Photo	Total	Top-1	Top-5
DINO	Photo-only	0	400	400	74.92	92.78
DINO	Line→Photo	150	250	400	**77.22**	**94.18**
DINO	Line→Photo	150	100	250	75.54	93.18
SwAV	Photo-only	0	400	400	74.28	92.84
SwAV	Line→Photo	150	250	400	**77.62**	**94.34**
SwAV	Line→Photo	150	100	250	74.40	92.96

At ImageNet-21K scale, a ViT-B/16 initialized with two epochs of line-drawing training subsequently attains lower photograph training loss under matched total optimizer-step budgets (Appendix A.11). This experiment extends the evidence for improved photographic optimization to a larger training dataset.

Having established these learning benefits, we next ask whether the representations learned after adaptation to photographs retain a greater emphasis on object structure.

4.2 Structural preferences persist after adaptation to photographs

We examine structural preferences at two levels: the information used to resolve ambiguous recognition decisions and the sensitivity of neuronal responses when object structure and texture are varied independently. Both analyses compare models after their photographic training stage.

Shape-based recognition decisions. We evaluate shape versus texture preferences using the cue-conflict images in the model-vs-human benchmark (Geirhos et al., 2019; 2021). These images combine shape and texture cues associated with different categories, allowing us to measure which cue determines the model's classification. We analyze the seven categories shared by the benchmark and STL10.

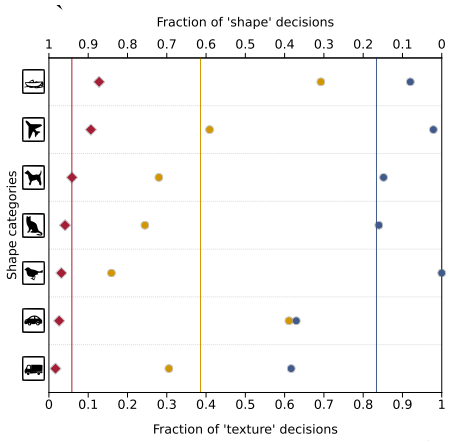


Figure 3: Shape versus texture decisions on STL10 cue-conflict categories. Yellow circles: Line→Photo; blue circles: Photo-only; red diamonds: humans. Models use ResNet-18.

Line→Photo shifts decisions toward shape relative to Photo-only training (Figure 3), bringing the pattern of choices closer to the human comparison data. The persistence of this preference after photographic adaptation shows that early structural experience continues to influence how models resolve conflicting visual cues.

Neuronal sensitivity to structure across textures. To examine whether this behavioral preference is accompanied by a change in internal sensitivity, we use ten rendered 3D object structures, each paired with seven surface textures, from the stimulus family illustrated in Figure 1G of Bian et al. (2025) (Figure 4).

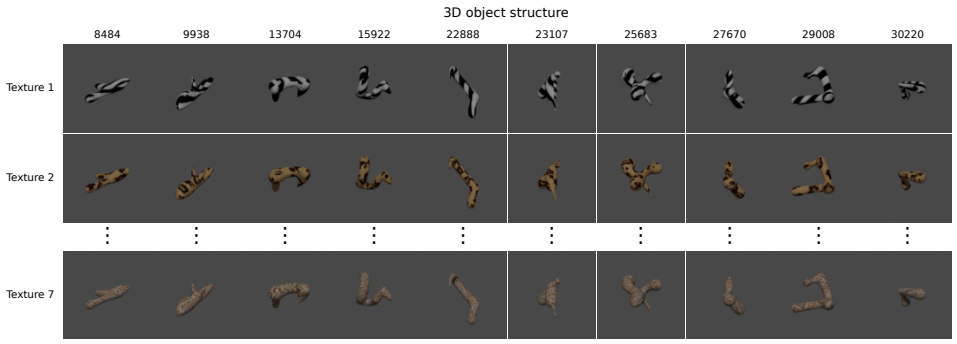


Figure 4: Controlled structure–texture stimuli. Ten 3D object structures (columns) are each rendered with seven surface textures (rows). Rows 1, 2, and 7 are shown; ellipses indicate rows 3–6. All 70 stimuli in the complete grid are used in the analysis.

We measure responses at four stages of frozen ResNet-18 models. For each sampled unit, we compare response variance across structures at a fixed texture, averaged over textures (A), with response variance across textures at a fixed structure, averaged over structures (B). The normalized structure sensitivity index is $\mathrm{SI} = A/(A + B)$; values above 0.5 indicate greater variation across the tested structures than across textures. The full measurement procedure is provided in Appendix A.12.

Line→Photo increases mean SI at all four stages (Figure 5). The proportion of sampled units with $\mathrm{SI} > 0.5$ also increases at every stage, by 10.9, 7.8, 12.1, and 15.6 percentage points at L1–L4, respectively.

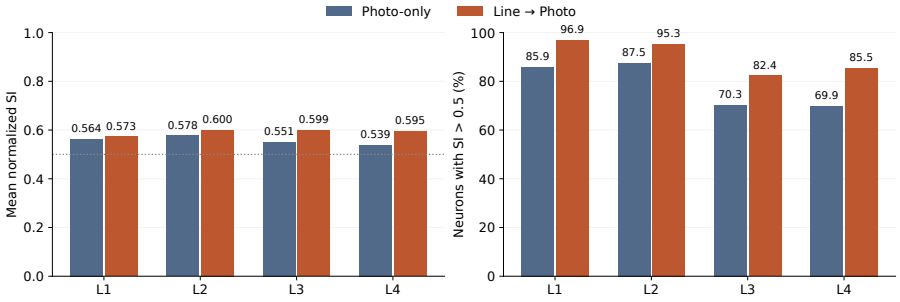


Figure 5: Structure sensitivity across the four ResNet-18 stages. Left: mean normalized SI across sampled units. Right: percentage of units with $\mathrm{SI} > 0.5$. The dotted line marks equal structure and texture sensitivity ($\mathrm{SI} = 0.5$). L1–L4 correspond to the final residual block of each stage; sampling details are provided in Appendix A.12.

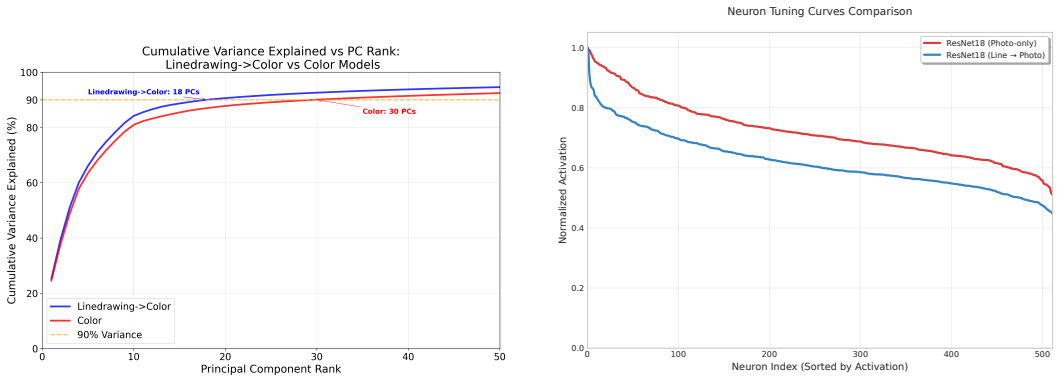
Thus, early structural experience shifts the average response balance toward structure and increases the proportion of units whose responses vary more with structure than with texture. Together with the cue-conflict results, these observations show a retained emphasis on object structure in both recognition decisions and internal responses after learning from photographs. Complementary spatial-attribution and feature-dispersity analyses are provided in Appendices A.2 and A.3.

4.3 STRUCTURE-FIRST LEARNING YIELDS COMPACT REPRESENTATIONS ACCESSIBLE AT EARLIER LAYERS

We next ask how early structural experience affects the organization of learned representations. We examine compactness across feature dimensions and channels, then test how early in the network useful information becomes accessible.

Effective representational dimensionality. We extract the 512-dimensional penultimate-layer activations of ResNet-18 for STL10 test images and apply PCA. Line→Photo reduces the number of principal components needed to explain 90% of the variance from 30 (Photo-only) to 18 (Figure 6a). Representational variation is therefore concentrated in fewer linear dimensions on the evaluated image set, indicating a more compact representation under this measure.

Concentration of channel responses. We also examine activations in the final convolutional layer over 500 test images. For each channel, we average activations over spatial positions and images, rank the resulting channel means, and normalize by the largest channel mean within each model. The ranked response profile decreases more steeply for Line→Photo than for Photo-only training (Figure 6b). This pattern indicates greater concentration of mean response strength across channels.



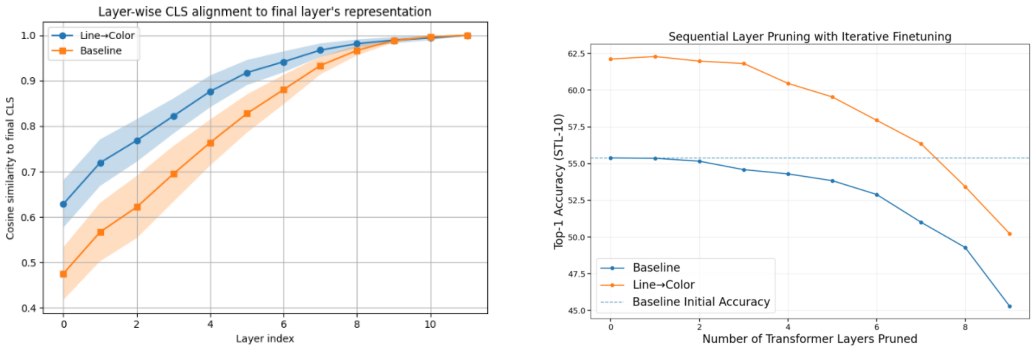
(a) Cumulative explained variance **(b) Ranked mean channel responses**

Figure 6: Population organization after photographic training. (a) PCA of 512-dimensional penultimate-layer activations on the STL10 test set. (b) Final-layer channel responses averaged over spatial positions and 500 test images, ranked and normalized by each model's largest channel mean. The curriculum produces a steeper ranked response profile.

Alignment across network depth. The preceding analyses characterize the final representation; we next ask how closely intermediate representations resemble it. In the transformer models, we measure cosine similarity between each layer's CLS representation and the same model's final-layer CLS representation. Early layers are more closely aligned with the final representation after Line→Photo training (Figure 7a), suggesting that the curriculum changes how representations develop across network depth.

Recognition after layer truncation. To test the functional implications, we sequentially remove later transformer layers with iterative fine-tuning. Line→Photo retains higher accuracy than Photo-only at every evaluated truncation level (Figure 7b). After removing seven layers, it retains approximately 56% accuracy, comparable to the unpruned Photo-only model's approximately 55%.

Together, these analyses suggest that early structural experience leaves representations more concentrated across feature dimensions and channels, with useful information accessible at shallower network depths. The alignment and truncation results link similarity across layers to the ability to preserve recognition performance in a shallower, fine-tuned network. This combination motivates testing whether the resulting representations also support transfer to new tasks and models.



(a) Layer alignment.

(b) Sequential layer truncation.

Figure 7: Representational alignment and recognition across network depth. (a) Cosine similarity between each layer's CLS representation and the same model's final-layer representation; shading denotes standard deviation across self-attention heads. (b) STL10 accuracy after sequential removal of later transformer layers with iterative fine-tuning. The dashed line marks the unpruned Photo-only accuracy. In the plots, "Line→Color" denotes Line→Photo and "Baseline" denotes Photo-only.

4.4 Structure-first representations support transfer across tasks and models

We test three forms of transfer: depth estimation from frozen features, adaptation to semantic segmentation, and knowledge distillation from trained classifiers to smaller models.

Depth estimation from frozen features. Inspired by the depth-transfer evaluation of Zhao et al. (2025), we train a feature-pyramid depth head for ten epochs on NYU Depth V2 (Silberman et al., 2012), keeping the pretrained ResNet-18 backbone frozen. We report test absolute relative error (REL) and root mean squared error (RMSE, in metres), with means and sample standard deviations across five separately trained heads per backbone.

Line→Photo reduces mean REL by 2.63% and RMSE by 4.04% (Table 3). The frozen features therefore support more accurate depth readout under the same head-training budget.

Table 3: Depth estimation from frozen ResNet-18 features on NYU Depth V2. Values are mean ± sample standard deviation across five separately trained depth heads per backbone. Bold indicates the better mean.

Backbone training	REL ↓	RMSE (m) ↓
Photo-only	0.3317 ± 0.0140	0.9923 ± 0.0145
Line→Photo	$\mathbf{0.3230} \pm 0.0101$	$\mathbf{0.9521} \pm 0.0064$

Adaptation to semantic segmentation. We next use the two ResNet-18 models for semantic segmentation on ADE20K over 20 epochs. Both decoders are trained; the encoder learning rate is 0.02 for Photo-only and zero for Line→Photo (Appendix A.14). Under these settings, Line→Photo improves mean intersection over union (mIoU) by 1.95 percentage points and pixel accuracy by 1.51 points (Table 4).

These experiments probe complementary aspects of transfer. With the backbone frozen, improved depth estimation indicates that geometric information is more accessible in the learned features. The segmentation results show that the Line→Photo representation supports spatially organized predictions while retaining its encoder parameters and training a decoder. Together, these findings connect the representational differences observed above to their utility on new visual tasks under the reported adaptation settings.

Table 4: Semantic segmentation on ADE20K after 20 epochs of adaptation. Encoder learning rates differ by condition as described in the text; both decoders are trained. Metrics are evaluated at the final epoch.

Initialization	mIoU (%) ↑	Pixel accuracy (%) ↑
Photo-only	23.32	70.23
Line→Photo	**25.27**	**71.74**

Knowledge transfer to smaller models. We distill MobileNetV1, ResNet-8, and VGG-8 students from ResNet-18 teachers. To compare teachers with similar recognition accuracy, we use Line→Photo trained with 70% of the data (69.01% accuracy) and Photo-only trained with 100% (68.89%). Students distilled from Line→Photo outperform those distilled from Photo-only in all three architectures (Table 5).

A Line→Photo teacher trained on the full dataset achieves 73.58% accuracy and yields further improvements in all three students. The advantage under approximately matched teacher accuracy indicates that the structural curriculum improves the usefulness of the teacher's supervision beyond what its recognition score alone captures. Teachers with similar accuracy can therefore differ in how effectively their knowledge transfers to smaller networks. This links the influence of early structural experience to the performance of models that learn from the resulting teacher.

Table 5: Knowledge distillation on STL10 from ResNet-18 teachers. Bold marks the better student accuracy for teachers with approximately matched recognition accuracy. The 70% subset is used in both curriculum stages; the full-data Line→Photo teacher provides a higher-accuracy reference.

			Student accuracy (%)		
Teacher curriculum	Data (%)	Teacher acc. (%)	MobileNetV1	ResNet-8	VGG-8
Teachers with approximately matched accuracy					
Photo-only	100	68.89	73.83	68.81	73.15
Line→Photo	70	69.01	**75.48**	**71.08**	**75.41**
Higher-accuracy teacher for reference					
Line→Photo	100	73.58	77.06	72.75	77.25

Across these evaluations, the curriculum benefits geometric readout, adaptation to dense prediction, and knowledge transfer to smaller models. Additional object-detection experiments are reported in Appendix A.4.

5 Conclusion

Motivated by the hypothesis that early visual constraints can guide later learning, we use line drawings as a computational surrogate of early structural experience. Across the tested architectures and tasks, learning from line drawings before photographs leaves a lasting structural bias after photographic adaptation. The resulting representations are more sensitive to object structure and more compact, with useful information accessible at shallower network depths. These representational differences accompany improvements in segmentation, depth estimation, recognition, data efficiency, and knowledge transfer.

Together, these findings provide computational support for the developmental hypothesis that early emphasis on stable structural cues can organize subsequent learning from richer visual input. The persistence of structural preferences after photographic adaptation suggests that early experience establishes a representational foundation that continues to guide learning as appearance information becomes available. In this view, developmental constraints can be beneficial because they shape the information around which later representations are built. Early structural experience can therefore shape both what a visual system learns next and how effectively that knowledge transfers.

AI USE STATEMENT

Generative AI tools assisted with implementing the authors' experimental designs in software, scientific figure preparation, and manuscript editing. The authors conceived the study, designed the experiments, and made the final scientific decisions. The authors take responsibility for the final content, including AI-assisted text, code, and figures.

REFERENCES

Martin S. Banks and Philip Salapatek. Infant pattern vision: A new approach based on the contrast sensitivity function. *Journal of Experimental Child Psychology*, 31(1):1–45, 1981. doi: 10.1016/0022-0965(81)90002-3.

Yingjue Bian, Tianye Wang, Shiming Tang, and Tai Sing Lee. Integrative characterization of the topography of V4 neural codes using deep learning approaches, 2025. URL https://arxiv.org/abs/2511.22050.

Oliver Braddick and Janette Atkinson. Development of human visual function. *Vision Research*, 51(13):1588–1609, 2011. doi: 10.1016/j.visres.2011.02.018.

Angela M. Brown. Development of visual sensitivity to light and color vision in human infants: A critical review. *Vision Research*, 30(8):1159–1188, 1990. doi: 10.1016/0042-6989(90)90173-I.

Angela M. Brown and Delwin T. Lindsey. Contrast insensitivity: The critical immaturity in infant visual performance. *Optometry and Vision Science*, 86(6):572–576, 2009. doi: 10.1097/OPX.0b013e3181a72980.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7915–7925, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chan_Learning_To_Generate_Line_Drawings_That_Convey_Geometry_and_Semantics_CVPR_2022_paper.html.

Adam Coates, Honglak Lee, and Andrew Y. Ng. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 215–223. PMLR, 2011. URL https://proceedings.mlr.press/v15/coates11a.html.

Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. BézierSketch: A generative model for scalable vector sketches. In *European Conference on Computer Vision*, pp. 632–647. Springer, 2020. doi: 10.1007/978-3-030-58574-7_38.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Judith E. Fan, Wilma A. Bainbridge, Rebecca Chamberlain, and Jeffrey D. Wammes. Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, 2(9):556–568, 2023. doi: 10.1038/s44159-023-00212-w.

Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, S. M. Ali Eslami, and Oriol Vinyals. Synthesizing programs for images using reinforced adversarial learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1666–1675. PMLR, 2018. URL https://proceedings.mlr.press/v80/ganin18a.html.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23885–23899, 2021. URL https://papers.nips.cc/paper/2021/hash/c8877cff22082a16395a57e97232bb6f-Abstract.html.

Jacob Gildenblat and contributors. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/pdf/1e7e33ffa8fe030c42dad328e4f075c614ccce55.pdf.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19000–19015, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/db5f9f42a7157abe65bb145000b5871a-Abstract.html.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html.

Dunhan Jiang, Tianye Wang, Yingjue Bian, Shiming Tang, and Tai Sing Lee. Computational constraints underlying shape and texture functional domain organization in macaque V4. *Cerebral Cortex*, 36(2):bhaf345, 2026. doi: 10.1093/cercor/bhaf345. URL https://academic.oup.com/cercor/article/36/2/bhaf345/8488897.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. Software, version 8.0.0, 2023. URL https://github.com/ultralytics/ultralytics.

Barbara Landau, Linda B. Smith, and Susan S. Jones. The importance of shape in early lexical learning. *Cognitive Development*, 3(3):299–321, 1988. doi: 10.1016/0885-2014(88)90014-7.

Sangjun Lee, Inwoo Hwang, Gi-Cheon Kang, and Byoung-Tak Zhang. Improving robustness to texture bias via shape-focused augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4323–4331, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/HCIS/html/Lee_Improving_Robustness_to_Texture_Bias_via_Shape-Focused_Augmentation_CVPRW_2022_paper.html.

Tianqin Li, Ziqi Wen, Yangfan Li, and Tai Sing Lee. Emergence of shape bias in convolutional neural networks through activation sparsity. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71755–71766, 2023. doi: 10.52202/075280-3141.

Florian Perdreau and Patrick Cavanagh. Drawing skill is related to the efficiency of encoding object structure. *i-Perception*, 5(2):101–119, 2014. doi: 10.1068/i0635.

Florian Perdreau and Patrick Cavanagh. Drawing experts have better visual memory while drawing. *Journal of Vision*, 15(5):5, 2015. doi: 10.1167/15.5.5.

Paul C. Quinn, Peter D. Eimas, and Michael J. Tarr. Perceptual categorization of cat and dog silhouettes by 3- to 4-month-old infants. *Journal of Experimental Child Psychology*, 79(1):78–94, 2001. doi: 10.1006/jecp.2000.2609.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Computer Vision – ECCV 2012*, pp. 746–760. Springer, 2012. doi: 10.1007/978-3-642-33715-4_54. URL https://cs.nyu.edu/~fergus/datasets/nyu_depth_v2.html.

Dmitriy Smirnov, Matthew Fisher, Vladimir G. Kim, Richard Zhang, and Justin Solomon. Deep parametric shape predictions using distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 561–570, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Smirnov_Deep_Parametric_Shape_Predictions_Using_Distance_Fields_CVPR_2020_paper.pdf.

Linda B. Smith, Susan S. Jones, Barbara Landau, Lisa Gershkoff-Stowe, and Larissa Samuelson. Object name learning provides on-the-job training for attention. *Psychological Science*, 13(1):13–19, 2002. doi: 10.1111/1467-9280.00403.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=4nPswr1KcP.

Masayuki Tanaka. Recognition of pictorial representations by chimpanzees (Pan troglodytes). *Animal Cognition*, 10(2):169–179, 2007. doi: 10.1007/s10071-006-0056-1.

Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23 (56):1–6, 2022. URL https://www.jmlr.org/papers/v23/21-1155.html.

Lukas Vogelsang, Sharon Gilad-Gutnick, Evan Ehrenberg, Albert Yonas, Sidney Diamond, Richard Held, and Pawan Sinha. Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, 115(44):11333–11338, 2018. doi: 10.1073/pnas.1800901115.

Dirk B. Walther, Barry Chai, Eamon Caddigan, Diane M. Beck, and Li Fei-Fei. Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, 108(23):9661–9666, 2011. doi: 10.1073/pnas.1015666108.

Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10743–10752, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Yi_APDrawingGAN_Generating_Artistic_Portrait_Drawings_From_Face_Photos_With_Hierarchical_CVPR_2019_paper.html.

Junru Zhao, Tianqin Li, Dunhan Jiang, Shenghao Wu, Alan Ramirez, and Tai Sing Lee. Perceptual inductive bias is what you need before contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9621–9630, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhao_Perceptual_Inductive_Bias_Is_What_You_Need_Before_Contrastive_Learning_CVPR_2025_paper.html.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/papers/Zhou_Scene_Parsing_Through_CVPR_2017_paper.pdf.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3): 302–321, 2019. doi: 10.1007/s11263-018-1140-0.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021. doi: 10.1073/pnas.2014196118.

A Appendix

A.1 Controls for the Initial Visual Experience

We compare line drawings with alternative initial visual representations to test whether the benefits of the curriculum also arise from other transformations of the training images. In each two-stage curriculum, ResNet-18 first learns from a transformed version of STL10 and then trains on photographs, using the same class labels in both stages. These comparisons examine how the content of early visual experience affects subsequent recognition.

For the artistic-style controls, we use the AdaIN implementation from `https://github.com/xunhuang1995/AdaIN-style` (Huang & Belongie, 2017) to generate Matisse-, Mondrian-, and Picasso-style images with stylization strength $\alpha = 0.6$. These provide alternative initial image domains for comparison with line-drawing pretraining. We also include Canny edge images as a control that removes color and emphasizes local intensity boundaries. This comparison tests whether a conventional edge representation provides the same benefit as the learned line drawings produced by Chan et al. (2022). Figure 8 illustrates the input representations.

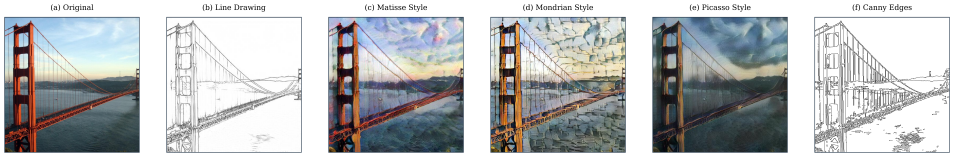


Figure 8: Input representations for the control curricula: the original photograph, a learned line drawing, three artistic-style transformations, and Canny edges. Each transformed domain is used for initial training before adaptation to photographs.

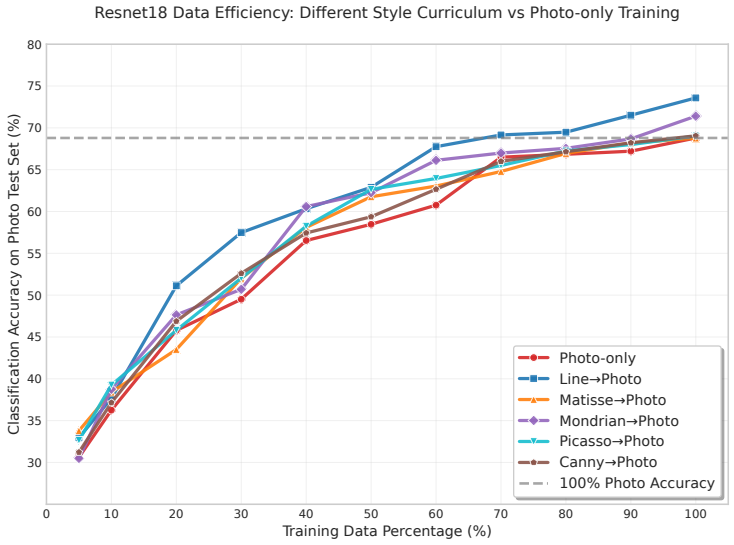


Figure 9: ResNet-18 data efficiency on STL10 across the full set of control curricula. Each transformed-image curriculum is followed by training on photographs. The gray dashed line marks photograph-only accuracy with 100% of the training data.

The full comparison in Figure 9 shows that Line→Photo exceeds the photograph-only baseline at every tested data fraction and achieves the highest accuracy among the tested curricula at eight of the eleven fractions. At 100% of the data, Line→Photo reaches 73.58%, compared with 71.40% for Mondrian→Photo, 69.05% for Canny→Photo, 69.03% for Picasso→Photo, and 68.77% for Matisse→Photo. Some artistic-style controls perform better at individual smaller data fractions, but their gains do not consistently match those from line drawings. These results support line drawings

as a particularly effective initial visual representation among the transformations tested, and indicate that the choice of early visual experience matters for subsequent learning.

A.2 Spatial Attribution with GradCAM

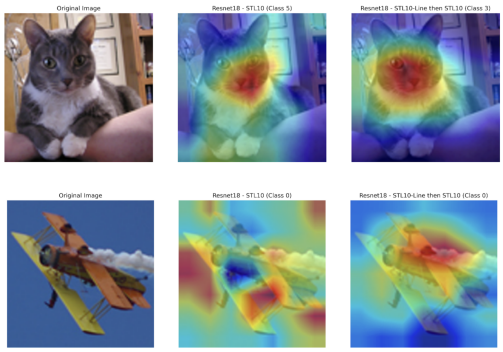


Figure 10: GradCAM attribution maps for Photo-only and Line→Photo models after photographic training.

To complement the representational analyses in Section 4.2, we analyze spatial attribution after photographic training. Figure 10 presents GradCAM (Selvaraju et al., 2017) visualizations for Photo-only and Line→Photo ResNet-18 models, computed using `pytorch-grad-cam` (Gildenblat & contributors, 2021). We compute the maps at the final residual block of `layer4`, targeting each model's own predicted class.

The Line→Photo model shows more focused attribution, with heatmaps concentrated on the primary subject. In the cat example, attribution emphasizes facial features and body contours, while the Photo-only map is more diffuse. In the airplane example, Line→Photo similarly emphasizes the aircraft's structural elements, whereas the Photo-only map is more scattered. These examples illustrate differences in the spatial distribution of attribution after photographic training. We quantify that distribution across test images in the analysis below.

Distribution analysis. We analyze the spatial distribution of GradCAM attribution by counting distinct high-activation regions. This measurement characterizes whether attribution is concentrated or distributed across multiple regions.

The single-image examples use an 85th-percentile threshold; the full-dataset region-count analysis uses the 70th percentile, as set in the released dataset runner. For each heatmap, pixels at or above its threshold are retained. We remove regions smaller than 10 pixels, fill holes smaller than 5 pixels, and count connected components using 8-connectivity.

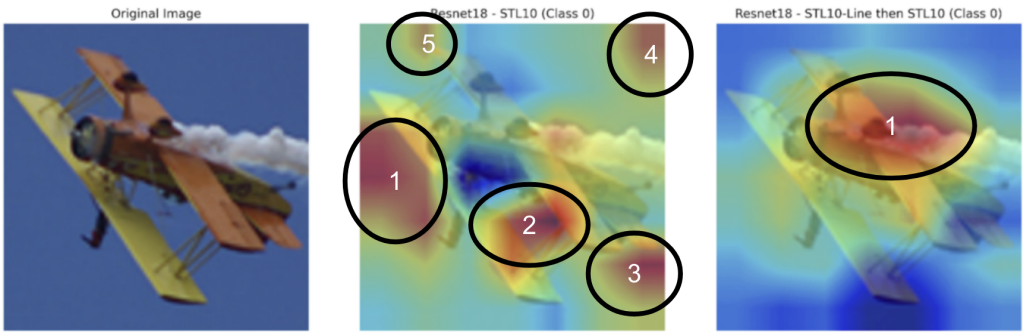


Figure 11: Counting high-activation regions in GradCAM maps. For this image, Photo-only ResNet-18 has five distinct regions, whereas Line→Photo has one concentrated region.

We define spatial scatter as the number of retained regions per image and plot its distribution across test images.

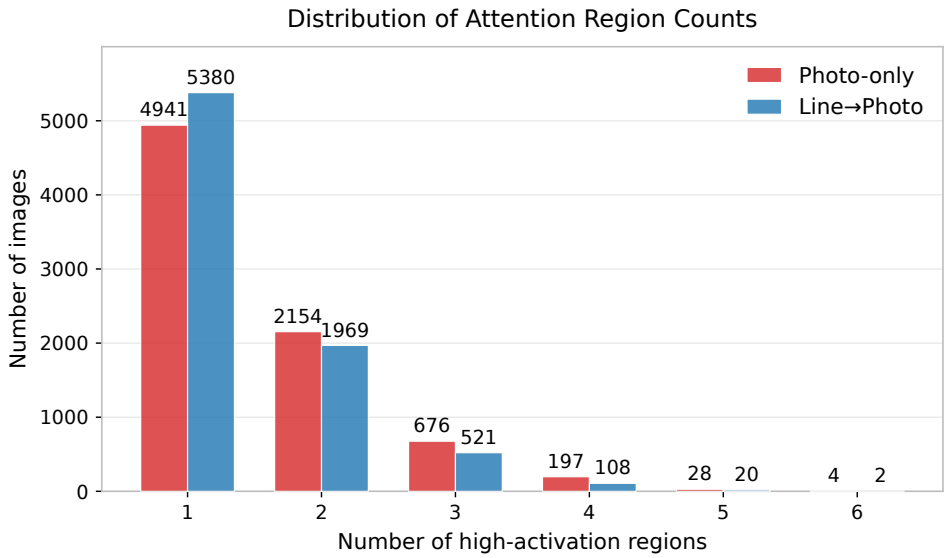


Figure 12: Distribution of high-activation region counts across STL10 test images. Regions are connected components retained after thresholding each GradCAM map at its 70th percentile and removing small regions.

The region-count distribution shows more images with a single attribution region for Line→Photo, whereas Photo-only produces more fragmented attribution (Figure 12). The implementation is provided in `Evaluation/gradcam_activations.py` in the anonymous code repository.

A.3 Population Feature Dispersity

We compare the population distributions of neuronal feature dispersity in the Photo-only and Line→Photo ResNet-18 models, following the analysis in Figure 4 of Jiang et al. (2026). We examine `layer3.1` and `layer4.0`, matching the layers in that analysis. Each neuron's dispersity is its critical image area divided by its estimated receptive-field area. The critical area is the image area at which preserving or removing the most influential pixels produces equal responses. Lower dispersity indicates that the features driving the response are more spatially localized.

The implementation in `Evaluation/Dispersity` ranks a shared pool of 50,000 natural images for each neuron, estimates receptive fields from attribution maps of its top 1,000 images, and computes preservation/removal curves on its top 25. Responses are normalized per image before averaging; the denominator is the fitted receptive field's half-maximum area. The accompanying configuration specifies image aperture, attribution noise, and mask settings.

Line→Photo shifts the population distribution toward lower dispersity at both layers (Figure 13). This indicates more spatially concentrated feature dependence across the neuronal population. Consistent with the relationship between dispersity and feature preference reported by Jiang et al. (2026), this pattern suggests greater emphasis on localized structural features and reduced reliance on dispersed texture features, complementing the structure-sensitivity analysis in Section 4.2.

A.4 Object Detection in PASCAL VOC

We extend the data-efficiency analysis in Figure 2 to object detection on PASCAL VOC. Line→Photo outperforms Photo-only at every evaluated training-data fraction (Figure 14). With 90% of the data, Line→Photo achieves 56.74 mAP, exceeding the Photo-only model's 56.14 mAP with the full 100% dataset. The advantage is particularly pronounced when data are scarce: at 10% of the training data, Line→Photo reaches 22.16 mAP, compared with 18.17 for Photo-only. The

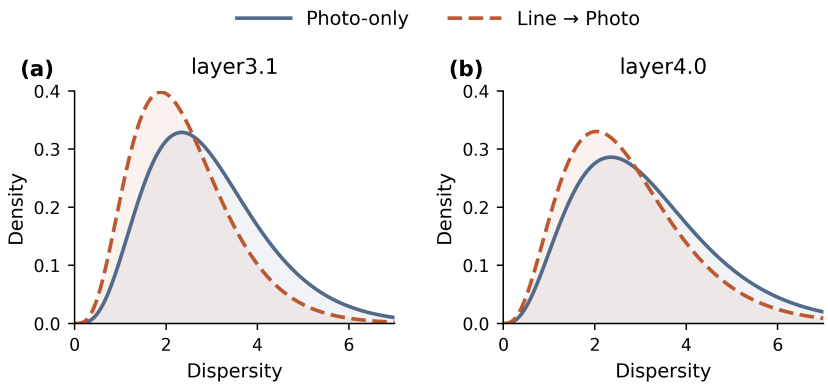


Figure 13: Population distributions of neuronal feature dispersity at (a) `layer3.1` and (b) `layer4.0`. Line→Photo (dashed orange) shows lower dispersity than Photo-only (solid blue) at both layers, indicating more spatially localized features driving neuronal responses.

benefit therefore extends from photograph classification to a task that also requires locating objects. These findings suggest that an initial phase of learning from structure-focused line drawings develops features useful for subsequent detection, improving generalization to test images and the efficiency with which training examples are used.

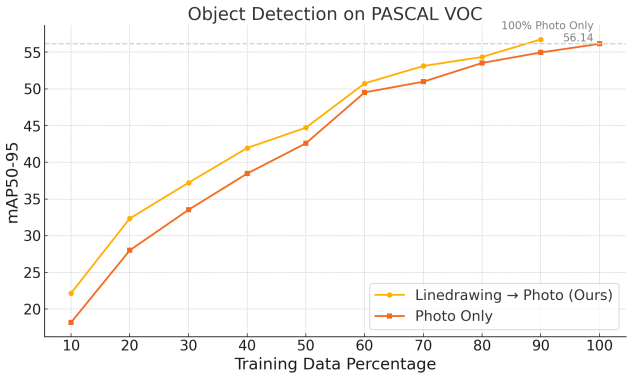


Figure 14: Improved data efficiency with Line→Photo training on PASCAL VOC for object detection.

Training setup. We use YOLO-v8-n and the Ultralytics training setup (Jocher et al., 2023) on PASCAL VOC. Line→Photo uses 200 epochs on line drawings followed by 200 epochs on photographs; Photo-only uses 200 epochs on photographs. Thus, the photographic training budgets are equal, while the curriculum includes an additional initial stage. The implementation is provided in `Training/ObjectDetection`.

A.5 Supervised Learning Statistical Analysis

Table 6 reports ResNet-18 photograph classification on STL10 at each training-data fraction. Results are means and standard deviations across five training runs per configuration, using the protocol in Appendix A.8.

Comparisons across data fractions. We use one-sided Wilcoxon signed-rank tests to compare each initial-experience curriculum with Photo-only, pairing observations at the same training-data fraction.

Table 6: STL10 photograph classification accuracy (%, mean ± standard deviation across five runs) for Photo-only and alternative initial-experience curricula.

Data (%)	Photo-only	Line→Photo	Matisse→Photo	Mondrian→Photo	Picasso→Photo
5	$30.62 \pm 0.77\%$	$32.95 \pm 0.59\%$	$33.79 \pm 0.92\%$	$30.49 \pm 1.87\%$	$32.64 \pm 0.68\%$
10	$36.25 \pm 1.66\%$	$37.68 \pm 0.30\%$	$38.40 \pm 2.34\%$	$38.76 \pm 0.94\%$	$39.25 \pm 1.56\%$
20	$45.78 \pm 3.22\%$	$51.12 \pm 2.96\%$	$43.50 \pm 1.73\%$	$47.64 \pm 2.48\%$	$45.80 \pm 3.15\%$
30	$49.51 \pm 4.10\%$	$57.48 \pm 1.08\%$	$52.01 \pm 3.67\%$	$50.70 \pm 1.42\%$	$52.04 \pm 2.19\%$
40	$56.53 \pm 0.75\%$	$60.33 \pm 1.58\%$	$58.10 \pm 0.84\%$	$60.59 \pm 2.73\%$	$58.24 \pm 1.29\%$
50	$58.47 \pm 1.27\%$	$62.89 \pm 1.60\%$	$61.76 \pm 2.05\%$	$62.24 \pm 0.76\%$	$62.64 \pm 2.81\%$
60	$60.76 \pm 1.87\%$	$67.75 \pm 1.15\%$	$63.04 \pm 1.38\%$	$66.12 \pm 2.94\%$	$63.95 \pm 0.89\%$
70	$66.49 \pm 0.71\%$	$69.15 \pm 4.03\%$	$64.78 \pm 2.67\%$	$66.99 \pm 1.03\%$	$65.50 \pm 1.85\%$
80	$66.84 \pm 2.09\%$	$69.48 \pm 2.13\%$	$66.94 \pm 0.95\%$	$67.55 \pm 1.74\%$	$67.24 \pm 2.36\%$
90	$67.21 \pm 2.34\%$	$71.51 \pm 0.46\%$	$68.19 \pm 3.08\%$	$68.68 \pm 0.62\%$	$68.00 \pm 1.47\%$
100	$68.79 \pm 0.91\%$	$73.58 \pm 0.42\%$	$68.77 \pm 1.53\%$	$71.40 \pm 0.56\%$	$69.03 \pm 0.77\%$

For each comparison, we tested the null hypothesis H_0: median difference = 0 against the alternative hypothesis H_1: method performance $>$ Photo-only baseline (one-tailed test). The test was applied to 11 paired observations corresponding to the performance differences at each data percentage level (5%, 10%, 20%, ..., 100%).

Line→Photo outperforms Photo-only at all 11 fractions ($W = 66$, $p = 0.002$). The corresponding tests for Mondrian→Photo, Picasso→Photo, and Matisse→Photo give $p = 0.037$, $p = 0.195$, and $p = 0.500$, respectively. At the nominal 0.05 level, the Line and Mondrian comparisons are significant; the Picasso and Matisse comparisons are not.

Comparison at the full training-data fraction. At 100% of the training data, the mean accuracy differences from Photo-only are +4.79 percentage points for Line→Photo, +2.61 for Mondrian→Photo, +0.24 for Picasso→Photo, and -0.02 for Matisse→Photo. These are differences at a single data fraction; the tests above summarize comparisons across all 11 fractions.

A.6 Computing Infrastructure

Experiments run on a computing cluster equipped with dual NVIDIA L40 GPUs and AMD EPYC 7443 24-core processors running at 2.84 GHz. Each experimental job uses 8 CPU cores and 32 GB of system RAM. The cluster runs Springdale Open Enterprise Linux 8.6 (Modena) on x86-64 architecture.

A.7 Dataset Preparation

We generate line-drawing versions of STL10 (Coates et al., 2011) and ImageNet-1K (Russakovsky et al., 2015) using the pretrained "Anime style" model of Chan et al. (2022), available at `https://github.com/carolineec/informative-drawings`.

We use ADE20K (Zhou et al., 2017) for semantic segmentation, PASCAL VOC (Everingham et al., 2010) for object detection, and NYU Depth V2 (Silberman et al., 2012) for depth estimation. Task-specific training and evaluation procedures are described in the corresponding sections.

A.8 Supervised Learning Experiments

We compare Photo-only training with the Line→Photo curriculum on STL10's ten classes. Alternative initial representations are described in Appendix A.1. For each data fraction, we first select a

class-balanced subset, then randomly shuffle its indices and split it into 80% training and 20% validation. The same indices are used for photographs and their line drawings. ResNet-18 is evaluated at 5%, 10%, and 20% through 100% in increments of 10%.

We train ResNet-18 from scratch with a final linear layer ($512 \rightarrow 10$) for ten-class classification. The model has approximately 11.2M parameters.

During hyperparameter development, we explored learning rates from 0.0001 to 0.01, batch sizes from 4 to 64, early stopping patience from 5 to 15 epochs, and weight decay from 10^{-5} to 10^{-3}. In the released ResNet-18 trainer, the learning rate is 0.001 for fewer than 400 training samples, 0.003 for 400–1199, 0.005 for 1200–2399, and 0.01 for at least 2400. Batch size is selected separately: 4 for at most 300 samples, 8 for 301–1000, 16 for 1001–2000, and 32 above 2000.

Training uses SGD with momentum 0.9 and weight decay 10^{-4}, a ReduceLROnPlateau scheduler with patience 3, early stopping on validation loss with patience 10, and a maximum of 100 epochs per stage. Both curriculum stages use random resized cropping, horizontal flipping, and rotation; test images are center-cropped. The primary evaluation metric is held-out test accuracy.

For these supervised recognition experiments, we use five runs per configuration with independently generated random seeds. The training code is provided in `Training/train_resnet18.py` in the anonymous code repository.

For comparison with ResNet-18, we also train Vision Transformer Tiny (ViT-Tiny) from scratch on STL10. We use the `vit_tiny_patch16_224` configuration from the timm library, with approximately 5.7M parameters, 16×16 patches, and 224×224 input resolution. The architecture has 12 transformer layers, 3 attention heads, a hidden dimension of 192, learnable position embeddings, and a classification head ($192 \rightarrow 10$) for ten-class classification.

Motivated by the data, augmentation, and regularization trade-offs studied by Steiner et al. (2022), we use the following STL10-specific configuration: AdamW with learning rate 0.0005, weight decay 0.01, and betas $(0.9, 0.999)$. Each stage uses cosine decay with a 10% warmup. Batch sizes are 8, 16, 24, and 32 for at most 300, 301–1000, 1001–2000, and more than 2000 training samples, respectively. Light augmentation uses RandomResizedCrop (scale 0.85–1.0), RandomHorizontalFlip ($p = 0.5$), and mild ColorJitter. Label smoothing is 0.05 and gradient clipping uses a maximum norm of 0.5. Mixup ($\alpha = 0.1$) is applied to 30% of batches in the photographic stages and disabled during line-drawing pretraining. The released trainer uses maximum budgets of 150 epochs for Photo-only, 120 for the line stage, and 100 for the subsequent photo stage, with early stopping patience 20 in each stage.

The configuration addresses training on small datasets through reduced regularization, light augmentation, smaller batch sizes, and extended patience. ViT-Tiny follows the same Line→Photo curriculum as ResNet-18, with Xavier uniform initialization for linear layers and standard LayerNorm initialization. Training code is provided in `Training/train_vit_tiny.py`.

A.9 Ranked Mean Channel Responses

To measure the concentration of responses across channels (Figure 6), we extract activations from the final convolutional stage, `layer4`, of each ResNet-18 model. Each response corresponds to one of 512 channels, averaged across spatial positions in its feature map. The structure-sensitivity analysis in Appendix A.12 instead samples the center position of each channel.

We pass the test images (up to 500) through both ResNet-18 models: Photo-only, trained on photographs, and Line→Photo, trained first on line drawings and then on photographs. Forward hooks on `layer4` record the feature maps. Spatial averaging gives a 512-dimensional channel-response vector for each image, providing the inputs to the population summary below.

For each model, we average each channel's activation across the processed images, sort the 512 channel means in descending order, and divide by the largest channel mean. Channels are ranked independently within each model, so the horizontal axis denotes response rank rather than a shared channel identity.

The steeper profile for Line→Photo indicates that mean response strength is concentrated in fewer channels, consistent with the population analysis in Section 4.3. The implementation is provided in `Evaluation/neuron_tuning_curves.py`.

A.10 Recognition across training and evaluation modalities

Table 7 reports ResNet-18 recognition of photographs and line drawings after four training strategies on STL10 and STL10-Line. Line→Photo improves photograph accuracy from 68.79% to 73.58% relative to Photo-only training, while line-drawing accuracy increases from 16.67% to 29.53%. The reverse curriculum, Photo→Line, reaches 66.92% on line drawings and 13.43% on photographs. These measurements characterize recognition in both domains after each curriculum's final training stage. The primary comparison in the main text is between Photo-only and Line→Photo, both of which finish with photographic training.

Table 7: Classification accuracy on photographs and line drawings for ResNet-18 under four training strategies.

Training strategy	Photograph accuracy (%)	Line-drawing accuracy (%)
Photo-only	68.79	16.67
Line-only	10.67	62.62
Line→Photo	73.58	29.53
Photo→Line	13.43	66.92

A.11 Optimization on ImageNet-21K

To evaluate subsequent photographic optimization at a larger scale, we train a ViT-B/16 on ImageNet-21K under matched total optimizer-step budgets. Photo-only begins with two epochs on photographs, whereas Line→Photo begins with two epochs on line drawings. Both models then train on photographs for the same number of additional steps. After a brief adaptation to photographs, Line→Photo attains lower photograph training loss and maintains this advantage through the end of the schedule (Figure 15). This comparison assesses optimization on the photographic training data.

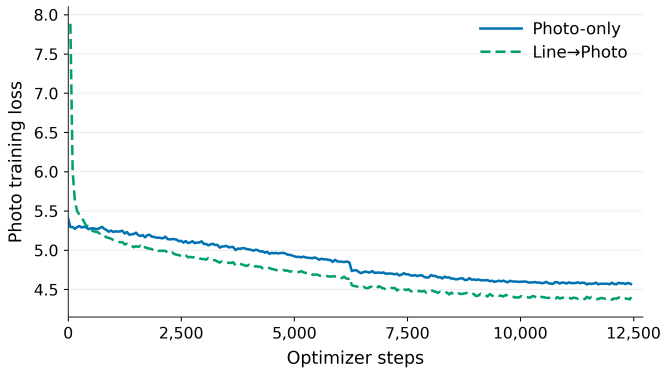


Figure 15: Photograph training loss for ViT-B/16 on ImageNet-21K after two initial epochs on photographs (Photo-only, solid blue) or line drawings (Line→Photo, dashed green). The horizontal axis starts at zero when both models begin the displayed photograph-training phase. Initial and subsequent optimizer-step budgets are matched, so equal plotted steps correspond to equal total optimizer-step counts.

A.12 Measurement of structure sensitivity

We use ten rendered 3D object structures, each with seven surface textures, from the stimulus family illustrated in Figure 1G of Bian et al. (2025). All 70 stimuli in the crossed grid contribute to the

analysis. We evaluate one frozen ResNet-18 checkpoint per training condition, with no additional training for this measurement.

L1–L4 denote the final residual blocks of the four ResNet-18 stages: `layer1.1`, `layer2.1`, `layer3.1`, and `layer4.1`, respectively. For each image, we record the response at the center of each channel's feature map, yielding 64, 128, 256, and 512 sampled units at L1–L4. Each sampled unit is treated as a neuron for the structure-sensitivity analysis.

For each unit, its responses form a 7×10 matrix R, where $R_{t,s}$ is the response to structure s rendered with texture t. We define

$$A = \frac{1}{7} \sum_{t=1}^{7} \mathrm{Var}(R_{t,1}, \ldots, R_{t,10}), \tag{1}$$

$$B = \frac{1}{10} \sum_{s=1}^{10} \mathrm{Var}(R_{1,s}, \ldots, R_{7,s}), \tag{2}$$

$$\mathrm{SI} = \frac{A}{A + B}. \tag{3}$$

Here A measures variation across structures while holding texture fixed, averaged over textures; B measures variation across textures while holding structure fixed, averaged over structures. Thus, A first takes the variance across the ten entries in each row, whereas B first takes the variance across the seven entries in each column. Variances use division by the number of entries (10 or 7), without a sample correction. Each unit therefore receives one normalized structure sensitivity index summarizing the full stimulus grid. An SI above 0.5 indicates greater response variation across the tested structures than across textures; 0.5 indicates equal variation.

At each stage, we report the mean of the individual units' SI values and the percentage of units with $\mathrm{SI} > 0.5$. Line→Photo increases both measures at all four stages. The percentage above 0.5 rises by 10.9, 7.8, 12.1, and 15.6 percentage points at L1–L4, respectively, as shown in Figure 5 in the main text.

The implementation and stimulus-manifest preparation are provided in `Evaluation/StructureSensitivity`. Its configuration uses the four stage endpoints listed above and the same normalization and stimuli for both training conditions.

A.13 Depth Estimation from Frozen Features

For the depth-transfer experiment in Section 4.4, we use the official NYU Depth V2 labeled split, holding out validation scenes from the training partition. The resulting sets contain 705 training, 90 validation, and 654 test images, with no scenes shared between sets. RGB inputs are resized to 240×320 and normalized with ImageNet channel means and standard deviations.

We keep one pretrained ResNet-18 checkpoint frozen per condition and train a feature-pyramid depth head with width 128. Training uses an absolute-error loss on valid depth pixels, Adam with weight decay 0.02, a batch size of 8, and ten epochs. A one-cycle learning-rate schedule starts at 10^{-4}, reaches 10^{-3} after the first 30% of steps, and then decays. For each backbone, we train five head instances with seeds 0–4; corresponding seeds use the same head initialization across conditions. The checkpoint with the lowest validation RMSE is selected for each run before evaluating the test set.

At evaluation, predicted depth is bilinearly resized to the native target resolution. We evaluate finite ground-truth depths from 0.1 to 10 metres within the zero-based crop `[45:471, 41:601]`, clipping predictions to the same depth range. For valid pixels V in an image, REL is $|V|^{-1} \sum_{i \in V} |\hat{d}_i - d_i| / d_i$ and RMSE is $\sqrt{|V|^{-1} \sum_{i \in V} (\hat{d}_i - d_i)^2}$. We average each metric across test images, then report the mean and sample standard deviation across the five trained heads in Table 3. These summaries average evaluation scores; they do not ensemble predictions or average model weights. The implementation is provided in `Training/DepthEstimation`.

A.14 Adaptation to Semantic Segmentation

We evaluate transfer to semantic segmentation on ADE20K using an encoder–decoder architecture. Table 4 compares two ResNet-18 initializations: Photo-only and Line→Photo. The broader experimental setup also included simultaneous training on line drawings and photographs; the reported comparison focuses on the two conditions used throughout the main paper.

We modify the semantic segmentation framework from `https://github.com/CSAILVision/semantic-segmentation-pytorch` (Zhou et al., 2019) to support the two encoders. The architecture combines a Pyramid Pooling Module (PPM) decoder with a ResNet-18 backbone that produces 512 output channels, using a segmentation downsampling rate of 32.

Both configurations use 20 epochs, 5,000 iterations per epoch, a batch size of 2 per GPU, and SGD with momentum 0.9 and weight decay 10^{-4}. The initial decoder learning rate is 0.02 in both conditions. The encoder learning rate is 0.02 for Photo-only and zero for Line→Photo, retaining the latter's learned encoder parameters. Nonzero learning rates follow polynomial decay, $lr_{current} = lr_{initial}(1 - iter/max_iter)^{0.9}$. Both configurations set `fix_bn=False`, allowing batch-normalization running statistics to update. This is therefore a comparison under different encoder-update settings.

We train the models on ADE20K with 150 semantic classes, resizing images to multiple scales $(300, 375, 450, 525, 600)$ with a maximum size of 1000 pixels. Data augmentation includes random horizontal flipping. Training uses negative log-likelihood loss with ignore index -1 to exclude invalid labels.

Each model is evaluated using its own checkpoint at epoch 20. We report pixel accuracy and mean intersection-over-union (mIoU) on the ADE20K validation set. The encoder wrapper and configurations are provided in `Training/Segmentation` for use with the upstream training and evaluation framework.

A.15 Knowledge Distillation Experiments

We evaluate knowledge distillation on STL10 with ResNet-18 teachers and MobileNetV1, ResNet-8, and VGG-8 students (Table 5). The primary comparison uses a Photo-only teacher trained on 100% of the data (68.89% accuracy) and a Line→Photo teacher using the same 70% subset in both curriculum stages (69.01%). Their similar recognition accuracies allow us to examine differences in student performance with approximately matched teacher accuracy. We also report a full-data Line→Photo teacher (73.58%) as a higher-accuracy reference. Students distilled from Line→Photo achieve higher accuracy in all three architectures, supporting the utility of the resulting representations for knowledge transfer. The configurations and implementation are provided in `Training/KnowledgeDistillation`.

A.16 Self-Supervised Training and Evaluation

The DINO and SwAV experiments use the implementation and configurations in `Training/SelfSupervised`, based on `solo-learn` (Turrisi da Costa et al., 2022). At the transition from line drawings to photographs, model weights are transferred while the optimizer and scheduler are reset; SwAV's queue is also reinitialized. We evaluate each pretrained backbone with a frozen-feature linear classifier trained for 100 epochs using SGD, batch size 256, learning rate 0.3, and no weight decay. The learning rate decays at epochs 60 and 80. The released configurations specify pretraining augmentations and optimization settings for the schedules reported in Table 2.

A.17 Code

The anonymized code repository is available at `https://anonymous.4open.science/r/Early-Structure-Code-7B32/`.