

LEARNING MORE BY SEEING LESS: STRUCTURE-FIRST LEARNING FOR EFFICIENT, TRANSFERABLE, AND HUMAN-ALIGNED VISION

Anonymous authors

Paper under double-blind review

ABSTRACT

Despite remarkable progress in computer vision, modern recognition systems remain fundamentally limited by their dependence on rich, redundant visual inputs. In contrast, humans can effortlessly understand sparse, minimal representations like line drawings, suggesting that structure, rather than appearance, underlies efficient visual understanding. In this work, we propose a novel structure-first learning paradigm that uses line drawings as an initial training modality to induce more compact and generalizable visual representations. We demonstrate that models trained with this approach develop a stronger shape bias, more focused attention, and greater data efficiency across classification, detection, and segmentation tasks. Notably, these models also exhibit lower intrinsic dimensionality, requiring significantly fewer principal components to capture representational variance, which mirrors observations of low-dimensional, efficient representations in the human brain. Beyond performance improvements, structure-first learning produces more compressible representations, enabling better distillation into lightweight student models. Students distilled from teachers trained on line drawings consistently outperform those trained from color-supervised teachers, highlighting the benefits of structurally compact knowledge. Together, our results support the view that structure-first visual learning fosters efficiency, generalization, and human-aligned inductive biases, offering a simple yet powerful strategy for building more robust and adaptable vision systems.

1 INTRODUCTION

Despite remarkable progress in computer vision, modern recognition systems remain fundamentally constrained by their reliance on dense, high-fidelity visual inputs. State-of-the-art models, particularly those based on deep convolutional and transformer architectures, have demonstrated impressive accuracy across a variety of benchmarks. However, their success is heavily contingent on exposure to vast datasets filled with richly textured, color photographic imagery. This dependency introduces significant inefficiencies—both in terms of data requirements and generalization capabilities—limiting the robustness of these systems when confronted with abstract or atypical visual inputs (Geirhos et al., 2019; Schlegel et al., 2015).

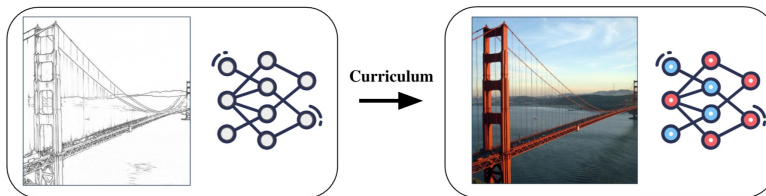


Figure 1: A structure-first curriculum, beginning with line drawings, leads to efficient, transferable, and human-like visual representations.

In contrast, humans can effortlessly interpret line drawings—minimal, structure-preserving representations that strip away redundant surface features. This ability suggests that the human visual

054 system is tuned to extract and reason over efficient structural information such as shape and topol-
055 ogy, rather than relying on unstable texture or color. Neuroscientific fMRI studies support this view,
056 showing that line drawings elicit robust, category-specific activity patterns throughout the visual
057 cortex—comparable to those induced by full-color photographs (Walther et al., 2011). Similarly,
058 Shin et al. (2008) demonstrate that abstract visual symbols like icons also elicits similar neuronal
059 activation in the brain as their corresponding natural images.

060 Psychological evidence further supports the transformative power of minimal representations in vi-
061 sual learning. Studies reveal that drawing practice enhances perceptual skills and object recognition
062 in humans (Kozbelt, 2001; Chamberlain et al., 2019; Perdreau & Cavanagh, 2015). Artists develop
063 specialized visual cognition through prolonged exposure to structural information—improved visual
064 memory, enhanced structural analysis, and refined attention to diagnostic features (Schlegel et al.,
065 2015; Perdreau & Cavanagh, 2015). These enhancements are believed to be associated with learn-
066 ing from minimal representations (Perdreau & Cavanagh, 2014)—by eliminating distracting surface
067 features, if the visual system is encouraged to extract more robust structural relationships.

068 Motivated by these insights, we introduce a structure-first learning paradigm. This approach uses
069 line drawings as an initial training modality to guide models toward better representation learning.
070 Our central hypothesis is that by training vision models to first understand the world through mini-
071 mal structural cues, we aim to induce representations that are more compact, efficient, generalizable,
072 and human-aligned. Thanks to the recent development from Chan et al. (2022), we can now accu-
073 rately convert photographic color images to their minimal line-drawing form automatically at scale.
074 Our results show that models developed through this structure-first approach exhibit stronger shape
075 bias, more focused attention, and lower intrinsic dimensionality—mirroring characteristic findings
076 of human minds. These models also yield substantial gains in data efficiency and distillation, out-
077 performing color-supervised teachers in student performance even when matched on classification
078 accuracy.

079 Together, our findings suggest that a minimal, structure-first learning curriculum is a powerful strat-
080 egy for fostering robust, efficient, and transferable vision systems—grounded in the computational
081 principles that underlie human perception.

082 2 RELATED WORK

083 **Line Drawing Synthesis and Structural Representation** Recent advances in line drawing genera-
084 tion have focused on balancing geometric fidelity and semantic preservation. The work by Chan
085 et al. (2022) pioneered an unsupervised approach using geometry and semantic losses to translate
086 photographs into sparse yet informative sketches, circumventing the need for paired datasets. Earlier
087 methods relied on supervised training with stroke annotations or 3D geometry, limiting their applica-
088 bility to arbitrary photographs (Das et al., 2020; Ganin et al., 2018; Ha & Eck, 2018; Smirnov et al.,
089 2020). Concurrent work in unpaired translation, such as CycleGAN (Zhu et al., 2017) and UPDG
090 (Yi et al., 2019), achieved domain adaptation but struggled with structural coherence in sparse out-
091 puts. Building on the foundations of modern line drawing synthesis from authors like Chan et al.
092 (2022), our work leverages these semantically coherent yet sparse sketches for a novel purpose. We
093 use these minimal representations as the basis for a learning curriculum that demonstrates robust
094 visual understanding can be achieved by prioritizing structural information over dense, high-fidelity
095 detail.

096 **Visual Pre-training** Self-supervised pre-training has emerged as a dominant paradigm for learning
097 robust visual representations. Methods like Masked Autoencoders (MAE) (He et al., 2021) learn
098 by reconstructing masked patches, while contrastive and distillation-based approaches like DINO
099 (Caron et al., 2021b) and SwAV (Caron et al., 2020) learn invariant representations by matching
100 different views. A common thread in these methods is their reliance on dense, texture-rich color
101 photographs as the sole training input. This paper demonstrates that a 'structure-first' approach is
102 not mutually exclusive with these pre-training methods, but is a powerful, complementary strategy.
103 We show that introducing line drawings into their training processes—for example, by using them as
104 an input modality in the initial phase of frameworks like DINO and SwAV—can massively increase
105 performance and computational efficiency, instilling a robust structural bias that enhances existing
106 pre-training frameworks.
107

Methods to Improve Shape Bias The texture bias of convolutional neural networks has motivated efforts to instill shape-centric reasoning. Work by Geirhos et al. (2019) showed that stylized training data improves shape sensitivity, but this did not scale well. Other approaches from Hermann et al. (2020) and Lee et al. (2022) respectively used human-like and shape-guided data augmentation techniques to induce more shape bias. Work by Dapello et al. (2023) focused on architectural improvements and induced shape bias through sparsity constraints in transformer architectures. Our work bridges these insights by using line drawings—which are naturally sparse representations—to explicitly steer models toward geometric regularization. Unlike augmentation strategies, our method leverages the inherent spatial regularity of line art to improve cross-domain generalization.

Sparsity in Neural Representations Sparse coding has long been theorized as a mechanism for efficient visual processing. Recent work (Dapello et al., 2023) demonstrates that Top-K activation sparsity enhances shape bias by emphasizing part-based representations, while masked autoencoders (He et al., 2021) use spatial sparsity to improve feature learning. Line drawings naturally instantiate this principle, distilling scenes to their essential contours. Our structure-first curriculum leverages this inherent input sparsity; by beginning with an initial phase of learning on line drawings, our method encourages models to develop sparse, structure-focused representations, compelling them to discard textural noise in favor of essential invariants, a process analogous to an artist’s mental decomposition of complex forms (Perdreau & Cavanagh, 2014).

3 METHOD

We employ a two-stage supervised learning curriculum using the STL-10 dataset with stratified sampling across data subsets ranging from 5% to 100% of the full training corpus. Our methodology compares this structure first learning approach against a baseline control where models train directly on natural color images, establishing the performance ceiling for direct supervised learning.

We convert the original STL-10 (Coates et al., 2011) training images to line drawings using the method from Chan et al. (2022), which extracts the most crucial features in line form. For the artistic style variants, we apply style transfer techniques from Huang & Belongie (2017) to generate stylized versions of the original images, creating distinct visual abstractions that test semantic transfer across diverse visual domains. We also use Canny edge detection to generate images as a naive control.



Figure 2: Different augmentations of original image

In our two stage curriculum, models first undergo a foundational training phase on source domains including line drawings and various artistic styles, all using identical class labels as the target domain. Line drawings serve as our primary focus, providing structural information without color, while the artistic styles function as arbitrary augmentations to compare against. We then train the models on natural color images, allowing them to leverage learned semantic representations while adapting to target visual characteristics.

The training configuration employs adaptive learning rates (0.001-0.01) and batch sizes (4-32) that scale with dataset size, along with early stopping (patience=10). Both training stages utilize identical data augmentation strategies including random resized cropping, horizontal flipping, and rotation. Models are evaluated on the STL10 test sets using center-cropped images, with comprehensive testing across all data percentage levels under fixed random seeds for reproducibility.

To demonstrate the broad applicability of our structure-first curriculum, we conducted extensive validation beyond the initial STL-10 experiments. We evaluated our approach across diverse architectures and datasets, training both a YOLOv8 backbone on ImageNet-1K and a Vision Transformer (ViT) on STL-10 (detailed setup provided in the appendix). Additionally, we assessed the quality and transferability of learned representations through evaluation on multiple downstream tasks, including segmentation, detection, and knowledge distillation. These comprehensive experiments

confirm that our method produces robust, generalizable features that perform well across varied computer vision applications.

4 RESULTS

4.1 BENEFITS OF A STRUCTURE-FIRST CURRICULUM

The results presented in Table 1 reveal important insights into the effects of different training curricula on model performance. Training with the two-stage curriculum of line drawings (STL10-Line) followed by color images (STL10) leads to an improvement in generalization to line drawings, with accuracy rising from 16.67% to 29.53%. At the same time, color image classification accuracy also increases from 68.79% to 73.58%, indicating that structural features learned from line drawings provide a beneficial inductive bias for learning from color representations.

| Training Strategy | Color Acc | Line Acc |
|-------------------|-----------|----------|
| Color Only | 68.79 | 16.67 |
| Line Only | 10.67 | 62.62 |
| Line→Color | 73.58 | 29.53 |
| Color→Line | 13.43 | 66.92 |

Table 1: Classification accuracy (%) for different training strategies.

The experiments also demonstrate a pronounced asymmetry in performance depending on the order of training modalities. When models are trained first on color images and then on line drawings, color image classification accuracy drops sharply to 13.43%, despite strong performance on line drawings (66.92%). Conversely, training first on line drawings and then on color images results in high color image accuracy (73.58%) and moderate line drawing accuracy (29.53%). This suggests that features derived from line drawings are more generalizable and facilitate adaptation to color images, whereas color features are more specific and less transferable to line drawings.

These findings have practical implications for the design of image learning systems. A structure-first curriculum offers a robust initialization that is particularly effective when computational resources are limited. The results also highlight the risk of catastrophic forgetting when adapting models trained solely on color images to line drawings. Overall, the results support the notion that structural features learned from line drawings are more transferable and beneficial for generalization across different visual modalities.

| Training Strategy | STL10 Acc. |
|-------------------|------------|
| Color Only | 55.04 |
| Line → Color | 60.97 |

Table 2: STL10 classification accuracy (%) for ViT-Tiny (5.7M parameters) architecture under different training strategies.

| Training Strategy | ImageNet-1k Acc. |
|-------------------|------------------|
| Color Only | 67.87 |
| Line → Color | 68.82 |

Table 3: ImageNet-1k classification accuracy (%) for YOLO-v8 backbone (11M parameters) under different training strategies.

To validate that our findings are not limited to a single architectural class, we conducted additional experiments with both transformer-based and detection-oriented models. As shown in Table 2, applying our structure-first curriculum to a ViT-Tiny architecture on the STL10 dataset yields a substantial performance increase, boosting accuracy from 55.04% to 60.97%. This demonstrates that the benefits of structural learning are highly effective for modern transformer models.

Furthermore, to test our approach on different model types and at a larger scale, we used a YOLO-v8 backbone on the ImageNet-1k dataset. The results in Table 3 show a consistent, albeit more modest, improvement from 67.87% to 68.82%. This confirms that the positive effects of our curriculum extend to modern detection-based backbones and scale to more complex, large-scale datasets.

Notably, these results corroborate our central hypothesis that an initial learning phase on line drawings induces beneficial inductive biases that transfer across visual domains. The convergent evidence from convolutional (ResNet-18, YOLO-v8) and transformer-based (ViT-Tiny) architectures

establishes this structure-first approach as a broadly applicable technique for improving visual representation learning.

4.2 DATA EFFICIENCY

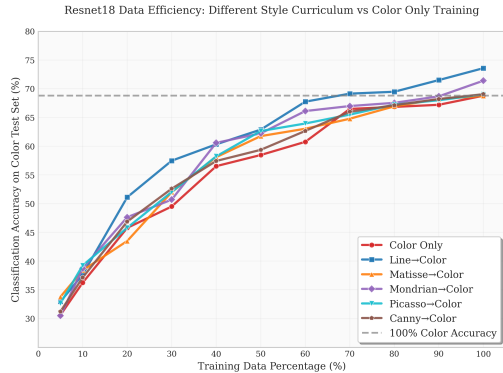


Figure 3: ResNet-18 data efficiency results of training with various augmentations of STL10 and finetuning on the base STL10 data.

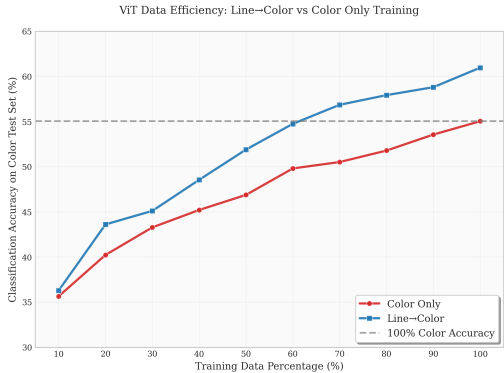


Figure 4: ViT-Tiny data efficiency results of training on STL10 with our Line→Color curriculum vs Color only

Our experimental results provide compelling evidence for the data efficiency advantages of line drawing curriculum learning over conventional color-only training approaches. As illustrated in Figure 3, the Line-Color method demonstrates superior learning efficiency across all training data regimes, achieving the baseline color accuracy (69%) with only 70% of the training data—representing a substantial 30% reduction in labeled data requirements. This efficiency gain manifests early in the training process, with the Line-Color approach consistently outperforming the color-only baseline by 3-5 percentage points at equivalent data volumes. Notably, the learning trajectory exhibits a steeper ascent during the initial training phases, suggesting that the line drawing curriculum provides a more structured optimization landscape that facilitates faster convergence to target performance levels.

The comparative analysis with style controls (Matisse-Color, Mondrian-Color, and Picasso-Color, Canny-Color) further validates the specificity of line drawing advantages. We specifically include Canny edge images as a naive control, as they represent a classic, low-level approach to structure extraction that lacks the semantic coherence of learned line drawings. While some of these alternative visual abstractions yield improvements over the color-only baseline, they are significantly less effective than line drawings in achieving data efficiency. The other methods require substantially more training data to reach baseline performance compared to the line drawing approach. This differential performance strongly supports our theoretical framework that line drawings tap into fundamental shape and topology processing mechanisms inherent to visual recognition systems. The reduced effectiveness of alternative abstractions demonstrates that the efficiency gains are not merely artifacts of stylistic variation, but rather reflect the unique capacity of line drawings to provide structured, shape-preserving visual representations that align with the computational principles underlying human visual processing.

Furthermore, these data efficiency benefits extend beyond convolutional architectures. As demonstrated with a Vision Transformer (ViT-Tiny) in Figure 4, the advantages of the structure-first curriculum are even more pronounced. The ViT model trained with the Line-Color curriculum achieves the final performance of the color-only baseline while using approximately 40% less training data, a greater reduction than observed with ResNet-18. This consistent outperformance across both CNN and transformer-based models underscores the general applicability of our approach, confirming that the induced structural priors provide a robust, cross-architecture benefit for efficient visual learning.

4.3 COMPLEMENTING EXISTING PRETRAINING METHODS

To demonstrate that our structure-first approach is not merely an alternative to pre-training but can actively enhance it, this work tests its impact on powerful, modern pre-training paradigms. The

investigation explores whether using line drawings as an input modality during the pre-training phase can improve performance and efficiency.

As case studies, our 'structure-first' phase is integrated with DINO (Caron et al., 2021a) and SwAV (Caron et al., 2020), two state-of-the-art pre-training methods. All experiments are implemented using the `solo-learn` library (da Costa et al., 2022). A strong baseline was established for each method by training a ResNet-18 model for 400 epochs on color images from the ImageNet-100 dataset. These baselines are compared against mixed-modality curricula that use a matched (400) or reduced (250) total epoch budget. The results are presented in Table 4.

| Method | Line Epochs | Color Epochs | Total Epochs | Top-1 (%) | Top-5 (%) |
|-----------------|-------------|--------------|--------------|---------------|---------------|
| DINO (Baseline) | 0 | 400 | 400 | 74.92 | 92.78 |
| DINO (Ours) | 150 | 250 | 400 | 77.22 (+2.30) | 94.18 (+1.40) |
| DINO (Ours) | 150 | 100 | 250 | 75.54 (+0.62) | 93.18 (+0.40) |
| SwAV (Baseline) | 0 | 400 | 400 | 74.28 | 92.84 |
| SwAV (Ours) | 150 | 250 | 400 | 77.62 (+3.34) | 94.34 (+1.50) |
| SwAV (Ours) | 150 | 100 | 250 | 74.40 (+0.12) | 92.96 (+0.12) |

Table 4: Performance of DINO and SwAV baselines versus our mixed-modality curriculum on ImageNet-100 (ResNet-18). Our structure-first phase significantly boosts performance with matched compute and achieves superior or comparable results with a 37.5% compute reduction.

The results from these experiments yield several critical insights. Our method provides significant gains across different pre-training frameworks. With an identical 400-epoch compute budget, our curriculum boosts DINO's Top-1 accuracy by +2.30% and SwAV's by an even larger +3.34%. This demonstrates that the performance gain is not an artifact of a larger compute budget; rather, introducing structural data as an initial input modality enhances the final representation.

The benefits are also clear in the reduced-compute scenarios. With a drastically reduced 250-epoch budget (37.5% less compute), our curriculum still outperforms both 400-epoch baselines. The DINO (150+100) model achieves 75.54% Top-1, surpassing its baseline (74.92%), and the SwAV (150+100) model achieves 74.40%, slightly edging out its own 400-epoch baseline (74.28%). These case studies confirm that our structure-first approach and pre-training are not mutually exclusive. Using line drawings as an input modality is a complementary strategy that can effectively supplement existing pre-training methods, improving both their final performance and their computational efficiency.

4.4 MORE FOCUSED ATTENTION

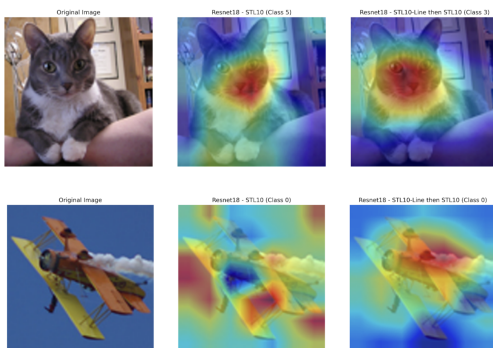


Figure 5: GradCAM visualizations comparing attention heat maps for models trained on color images vs trained with our curriculum.

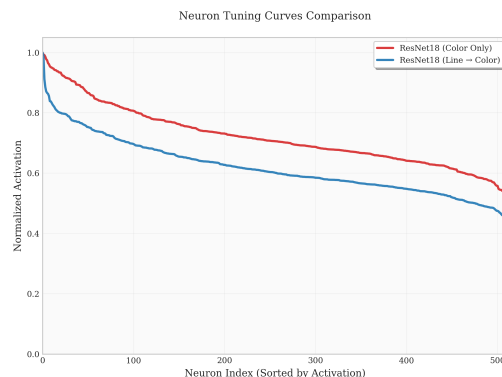


Figure 6: Neuron tuning curves of base ResNet-18 vs ResNet-18 with our curriculum. (Normalized activations over 500 test images)

To understand the mechanisms driving these performance improvements, we analyzed the attention patterns of models trained under different learning curricula. Figure 5 presents GradCAM (Gildenblat & contributors, 2021) visualizations comparing models trained exclusively on color images versus those trained with our structure-first curriculum (STL10-Line followed by STL10 color).

Models initialized with structure-first learning on line drawings demonstrate markedly more focused attention patterns, with heat maps concentrated on the primary subject. For the cat image, the structure-first model concentrates on facial features and body contours, while the color-only model’s activation is more diffuse. Similarly, for the airplane, the structure-first model focuses on the aircraft’s structural elements, unlike the scattered activation of the color-only model. A more quantitative analysis of these GradCAM activations is included in the supplementary material.

4.5 SPARSE NEURONAL TUNING CURVES

The neuron tuning curve presented in Figure 6 provides quantitative support for these qualitative observations. We evaluate the activation tuning curve for the last layer of the trained ResNet-18. Each neuron in Figure 6 represents one channel. We take the average firing activations from each channel and normalize them according to their maximum firing activation within each setting. The resulting activation tuning curve shows that ResNet-18 first trained on line drawings and then adapted to natural images (blue curve) exhibits a much sharper tuning curve in activation compared to the baseline counterpart (red curve). This pattern indicates that strong responses are concentrated among a small subset of neurons, while most units remain weakly or sparsely engaged. Such a rapid fall-off shows an increased sparsity and efficiency in neural representations.

Instead of distributing activation broadly, the network from the structure-first curriculum allocates capacity to the most informative neurons, silencing irrelevant or noisy features. The base model, by contrast, shows a more gradual decrease, activating a broader collection of neurons, including many that likely encode non-discriminative or redundant information.

This selective focus echoes principles of efficient coding: the model learns to emphasize features that matter for object recognition while suppressing spurious or noisy activations. This transition to more selective, quieter representations improves performance without any trade-off. The structure-first model achieves stronger accuracy on both minimal and natural images.

These results encapsulate the representational transformation induced by an initial phase of minimal, structural learning, which encourages the model to specialize its detectors for geometric cues.

4.6 INCREASED SHAPE BIAS

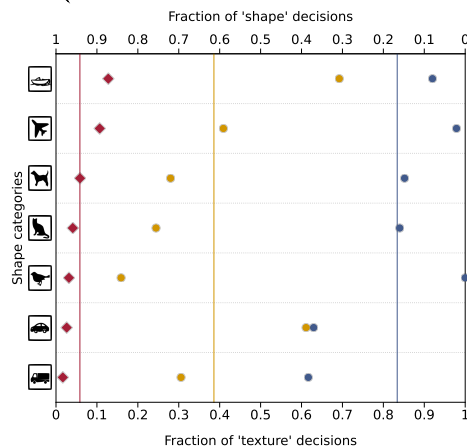


Figure 7: Shape versus texture decision analysis across object categories. Red diamonds indicating human comparison data from lab experiments, yellow circles indicate ResNet-18 Line→Color, and blue circles represent base ResNet-18 results.

To further assess the inductive biases, we evaluated shape versus texture preferences using the model-vs-human benchmark (Geirhos et al., 2021). This benchmark presents cue-conflict images in which shape and texture cues are intentionally mismatched, allowing for a direct comparison of classification strategies between models and human observers.

Figure 7 displays the distribution of shape and texture decisions across object categories (these are the 7 categories in the cue-conflict dataset that coincide with STL10 classes). Models first trained on line drawings and subsequently adapted to color images demonstrate a pronounced shift toward shape-based decisions, as indicated by the clustering of yellow circles toward higher fractions of shape choices. In contrast, models trained only on natural images show a greater reliance on texture cues, with more blue circles distributed toward the texture-dominant end of the spectrum. This pattern suggests that an initial phase of learning on line drawings encourages the model to prioritize global structural features over local textural information when resolving ambiguous visual input.

These results indicate that our structure-first approach not only enhances classification performance and attention focus, but also aligns model decision-making more closely with human perceptual strategies. This increased shape bias likely contributes to the improved robustness and generalization observed in downstream tasks, as models become less reliant on superficial texture cues and more attuned to the underlying structure of visual objects.

4.7 BENEFITS IN SEMANTIC SEGMENTATION

We further show that an initial phase of learning on line drawings enhances generalization to dense prediction tasks such as semantic segmentation. To evaluate this, we adapt ResNet-18 models—initialized either from color-only training or from our structure-first curriculum—on the ADE20K benchmark.

As shown in Figure 8, the model initialized with line drawings consistently outperforms the color-only baseline across multiple semantic categories. The improvement is especially pronounced for object-centric classes (e.g., “chair”, “car”), where shape plays a dominant role in visual discrimination. This demonstrates that structural priors learned from minimal representations translate effectively to tasks requiring precise spatial reasoning.

These results extend our findings beyond classification and detection, highlighting that our structure-first curriculum induces representations that generalize across both sparse and dense vision tasks.

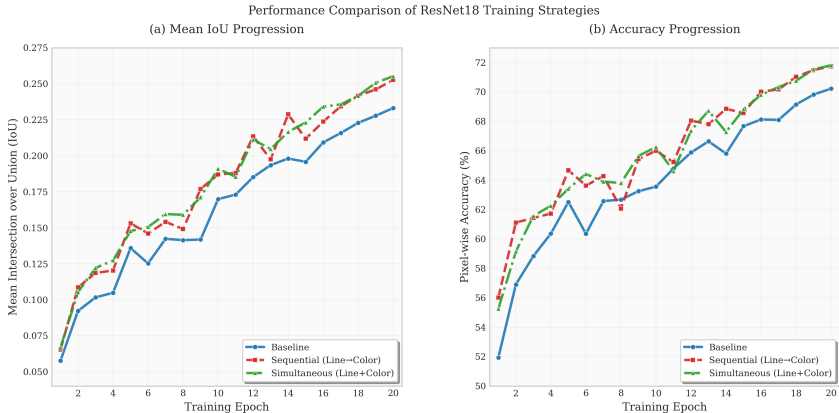
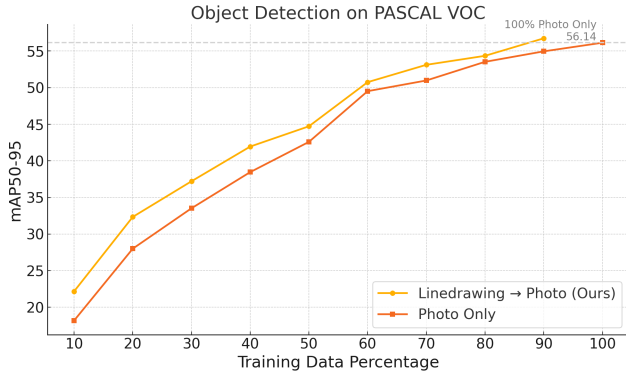


Figure 8: Different learning curricula on ResNet-18 generalizing to segmentation on ADE20K.

4.8 BENEFITS IN OBJECT DETECTION

We also run experiments similar to those shown in Figure 3, but on the object detection task using the PASCAL VOC dataset. Figure 9 illustrates that models trained with our structure-first curriculum consistently outperform color-only models across all training data regimes. Notably, our Linedrawing→Photo (Ours) model trained with just 90% of the data achieves an mAP of 56.74, surpassing the Photo Only model trained with the full 100% dataset, which reaches 56.14. This highlights a key advantage of our approach: better performance with less data. The performance gap is particularly

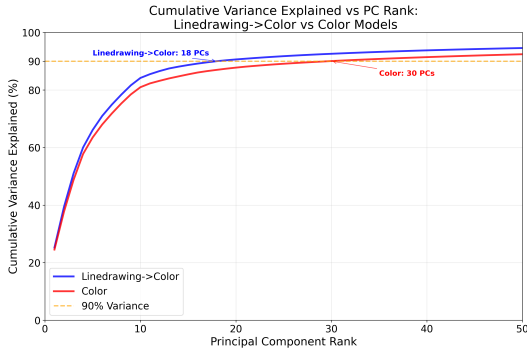
432 prominent in the low-data regime—for example, at 10% training data, our model achieves 22.16
 433 mAP compared to 18.17 for the color-only baseline. These findings suggest that an initial learning
 434 phase on structure-focused line drawings enables models to extract more efficient and transferable
 435 features for object detection, leading to significant gains in data efficiency and generalization.
 436



449 Figure 9: Improved Data Efficiency with Linedrawing→Color Training on PASCAL VOC for Ob-
 450 ject Detection.
 451

452 4.9 COMPACT AND TRANSFERABLE REPRESENTATIONS
 453

454 One of the core advantages of a structure-first curriculum is the emergence of compact and struc-
 455 turally focused internal representations. To investigate this, we first analyze the intrinsic dimension-
 456 ality of neural activations in a ResNet-18 model trained under two regimes: color-only vs. line-
 457 drawing → color. We extract penultimate layer activations across STL10 test images and compute
 458 principal components (PCs) via PCA.
 459



471 Figure 10: The structure-first Linedrawing→Color model learns a lower-dimensional representa-
 472 tional manifold. This is demonstrated by performing Principal Component Analysis (PCA) on the
 473 model’s \mathbb{R}^{512} latent space activations for the STL10 test set. To capture 90% of the variance, the
 474 structure-first model requires only 18 PCs, whereas a standard color-only model needs 30 PCs, in-
 475 dicating a more compact representation.
 476

477 As shown in Figure 10, the model from the structure-first curriculum requires only 18 PCs to explain
 478 90% of the variance, compared to 30 for the color-only model—suggesting that the model organizes
 479 information more economically and filters out redundant features. This compressed manifold struc-
 480 ture aligns with principles of efficient coding and complements earlier findings on activation sparsity
 481 (Figure 6).

482 We further examine whether this structural compactness benefits model distillation. Using a stan-
 483 dard teacher-student framework, we compare the performance of student models (MobileNetV1,
 484 ResNet-8, VGG-8) distilled from ResNet-18 teachers trained with either color-only or line-drawing
 485 → color supervision. As shown in Table 5, students distilled from teachers developed with the
 structure-first curriculum consistently outperform those from color-only teachers. For example, a

486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

| Teacher | Teacher Acc | Student | Distillation Acc |
|-----------------------|-------------|-------------|------------------|
| Linedrawing→Color (1) | 73.58 | MobileNetV1 | 77.06 |
| Linedrawing→Color (2) | 69.01 | MobileNetV1 | 75.48 |
| Color (3) | 68.89 | MobileNetV1 | 73.83 |
| Linedrawing→Color (1) | 73.58 | ResNet-8 | 72.75 |
| Linedrawing→Color (2) | 69.01 | ResNet-8 | 71.08 |
| Color (3) | 68.89 | ResNet-8 | 68.81 |
| Linedrawing→Color (1) | 73.58 | VGG-8 | 77.25 |
| Linedrawing→Color (2) | 69.01 | VGG-8 | 75.41 |
| Color (3) | 68.89 | VGG-8 | 73.15 |

Table 5: This table compares knowledge distillation from a line-drawing → color teacher (2) and a standard color teacher (3). For a fair comparison, their accuracies are matched by training (2) on 70% of the data while (3) uses 100%. The line-drawing teacher (2) yields superior student models. For reference, we also show the performance of an unmatched, higher-performing line-drawing teacher (1), which was trained on 100% of the data.

MobileNetV1 distilled from a teacher from the structure-first curriculum (trained with 70%-data to reduce the confounding factor of a stronger teacher) achieves 75.48% accuracy—exceeding the 73.83% achieved by the color-only teacher while the two teachers share almost the same object recognition performance.

These results suggest that the structure-first curriculum not only reduces the complexity of internal representations, but also produces knowledge that is easier to transfer, yielding more performant student models.

5 CONCLUSION

Our findings show that a structure-first curriculum using line drawings induces compact, structure-aware representations that generalize across tasks and enable more effective distillation into lightweight models. This structure-first learning approach improves performance, enhances data efficiency, and yields representations that are easier to compress and transfer. More broadly, our results offer a computational lens on the longstanding observation that humans can easily understand line drawings—suggesting that prioritizing geometry over appearance may support more general and efficient visual reasoning, both in artificial and biological systems.

REFERENCES

- 540
541
542 Mathilde Caron, Ishan Misra, Julien Brown, Saurabh Suri, Anton Osokin, and Pawel Bojanowski.
543 Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neu-*
544 *ral Information Processing Systems*, volume 33, pp. 9912–9924. NeurIPS, 2020.
- 545 Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and
546 Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of*
547 *the International Conference on Computer Vision (ICCV)*, 2021a.
- 548 Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and
549 Armand Joulin. Emerging properties in self-supervised vision transformers, 2021b.
- 550
551 Rebecca Chamberlain, Aaron Kozbelt, Shane Hickey, and Michael Sinatra. Artists as experts in
552 visual cognition: An update. *Psychology of Aesthetics, Creativity, and the Arts*, 13(4):395–406,
553 2019.
- 554 Caroline Chan, Fredo Durand, and Phillip Isola. Learning to generate line drawings that
555 convey geometry and semantics. [https://carolineec.github.io/informative_](https://carolineec.github.io/informative_drawings/)
556 [drawings/](https://carolineec.github.io/informative_drawings/), 2022.
- 557
558 Adam Coates, Honglak Lee, and Andrew Y. Ng. An analysis of single-layer networks in unsuper-
559 vised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial*
560 *Intelligence and Statistics*, volume 15, pp. 215–223. PMLR, 2011.
- 561 Victor Guilherme Turrissi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-
562 learn: A library of self-supervised methods for visual representation learning. *Journal of Machine*
563 *Learning Research*, 23(56):1–6, 2022. URL [http://jmlr.org/papers/v23/21-1155.](http://jmlr.org/papers/v23/21-1155.html)
564 [html](http://jmlr.org/papers/v23/21-1155.html).
- 565
566 Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo.
567 Emergence of shape bias in convolutional neural networks through activation sparsity. In *Ad-*
568 *vances in Neural Information Processing Systems*, volume 36, 2023.
- 569 Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. Béziersketch: A
570 generative model for scalable vector sketches. In *European Conference on Computer Vision*, pp.
571 632–647. Springer, 2020.
- 572
573 Mark Everingham, Luc Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The
574 pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):
575 303–338, 2010.
- 576 Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, S. M. Ali Eslami, and Oriol Vinyals. Synthe-
577 sizing programs for images using reinforced adversarial learning. In *Proceedings of the 35th*
578 *International Conference on Machine Learning*, volume 80, pp. 1666–1675. PMLR, 2018.
- 579
580 Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and
581 Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias im-
582 proves accuracy and robustness. In *International Conference on Learning Representations*, 2019.
- 583 Robert Geirhos, Kantharaju Narayanappa, Bernhard Mitzkus, Tizian Thieringer, Matthias Bethge,
584 Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and
585 machine vision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23885–
586 23899, 2021.
- 587
588 Jacob Gildenblat and contributors. Pytorch library for cam methods. [https://github.com/](https://github.com/jacobgil/pytorch-grad-cam)
589 [jacobgil/pytorch-grad-cam](https://github.com/jacobgil/pytorch-grad-cam), 2021.
- 590 David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference*
591 *on Learning Representations*, 2018.
- 592
593 Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked
autoencoders are scalable vision learners, 2021.

- 594 Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias
595 in convolutional neural networks. In *Advances in Neural Information Processing Systems*, vol-
596 ume 33, pp. 19000–19015, 2020.
- 597 Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normal-
598 ization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- 600 Aaron Kozbelt. Artists as experts in visual cognition. *Visual Cognition*, 8(6):705–723, 2001.
- 602 Sungnyun Lee, Inseop Hwang, Gyu Chul Kang, and Byoung-Tak Zhang. Improving robustness to
603 texture bias via shape-focused augmentation. In *Proceedings of the IEEE/CVF Conference on*
604 *Computer Vision and Pattern Recognition*, pp. 4323–4331, 2022.
- 605 Florian Perdreau and Patrick Cavanagh. Drawing skill is related to the efficiency of encoding object
606 structure. *i-Perception*, 5(2):101–119, 2014. doi: 10.1068/i0635.
- 608 Florian Perdreau and Patrick Cavanagh. Drawing experts have better visual memory while drawing.
609 *Journal of Vision*, 15(5):5, 2015. doi: 10.1167/15.5.5.
- 611 Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection
612 with yolov8. *arXiv preprint arXiv:2305.09972*, 2023.
- 613 Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng
614 Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual
615 recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- 616 Alexander Schlegel, Prescott Alexander, Samuel V Fogelson, Xinyang Li, Zhongming Lu, Peter J
617 Kohler, Erin Riley, Jessica B Dennison, John R Vokey, and Peter U Tse. The artist emerges:
618 visual art learning alters neural structure and function. *NeuroImage*, 105:440–451, 2015. doi:
619 10.1016/j.neuroimage.2014.11.014.
- 621 Yong-Wook Shin, Jun Soo Kwon, Ki Won Kwon, Bon Mi Gu, In Chan Song, Dong Gyu Na, and
622 Sohee Park. Objects and their icons in the brain: The neural correlates of visual concept formation.
623 *Neuroscience Letters*, 436(3):300–304, 2008.
- 624 Dmitriy Smirnov, Matthew Fisher, Vladimir G Kim, Richard Zhang, and Justin Solomon. Deep
625 parametric shape predictions using distance fields. In *Conference on Computer Vision and Pattern*
626 *Recognition (CVPR)*, 2020.
- 628 Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas
629 Beyer. How to train your vit? data, augmentation, and regularization in vision transformers, 2022.
630 URL <https://arxiv.org/abs/2106.10270>.
- 631 Dirk B Walther, Barry Chai, Eamon Caddigan, Diane M Beck, and Li Fei-Fei. Simple line draw-
632 ings suffice for functional mri decoding of natural scene categories. *Proceedings of the National*
633 *Academy of Sciences*, 108(23):9661–9666, 2011.
- 635 Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Apdrawinggan: Generating artistic portrait
636 drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference*
637 *on Computer Vision and Pattern Recognition*, pp. 10743–10752, 2019.
- 638 Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene
639 parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and*
640 *Pattern Recognition*, 2017.
- 642 Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba.
643 Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer*
644 *Vision*, 2018.
- 645 Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation
646 using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference*
647 *on Computer Vision*, pp. 2223–2232, 2017.

A APPENDIX

A.1 SUPERVISED LEARNING STATISTICAL ANALYSIS

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

Table 6: Aggregated results with std dev across all seeds showing performance comparison between Color Only training and different initial learning modalities.

To rigorously evaluate the statistical significance of our findings, we conducted Wilcoxon signed-rank tests comparing each initial learning modality against the Color Only baseline. The Wilcoxon signed-rank test is a non-parametric statistical test that assesses whether there is a significant difference between paired observations without assuming a normal distribution, making it particularly appropriate for our experimental setup where we have matched pairs of performance measurements across different data percentages.

For each comparison, we tested the null hypothesis H_0 : median difference = 0 against the alternative hypothesis H_1 : method performance > Color 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%).

Statistical Results: Line→Color vs Color Only shows a median improvement of +4.33 percentage points (W-statistic: 66, p-value: 0.002, highly significant), with Line→Color outperforming Color Only in all 11 data conditions (11/11 wins). Other stylistic augmentations showed: Mondrian→Color +1.47 percentage points (p = 0.037, significant), Picasso→Color +0.24 percentage points (p = 0.195, not significant), and Matisse→Color -0.02 percentage points (p = 0.500, not significant).

Implications: The statistical analysis provides strong evidence for several key conclusions. The Line→Color approach demonstrates highly significant improvement over the Color Only baseline (p = 0.002), with consistent gains across all tested data percentages. This statistical significance, combined with the perfect win rate (11/11), indicates that the performance benefit is both reliable and substantial. Critically, our results show that Line→Color’s benefits cannot be attributed merely to stylistic augmentation effects. While Mondrian→Color shows modest improvement (+1.47%, p = 0.037), both Matisse→Color and Picasso→Color fail to achieve statistical significance. The substantial difference between Line→Color’s improvement (+4.33%) and that of other stylistic methods suggests that line drawings possess unique structural properties that facilitate more effective learning. The consistent performance gains across all data percentages (5% to 100%) demonstrate that an initial phase of learning on line drawings provides benefits in both low-data and high-data scenarios. The median improvement of 4.33 percentage points represents a practically meaningful enhancement in classification performance that could translate to significant real-world benefits. These statistical findings support our central hypothesis that line drawings provide structurally advanta-

geous representations for subsequent learning phases, offering benefits that extend beyond those achievable through general stylistic augmentation approaches.

A.2 COMPUTING INFRASTRUCTURE

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

A.3 DATASETS PREPARATION

To generate the line drawings for our initial structural learning phase, the pre-existing “Anime style” model from <https://github.com/carolineec/informative-drawings> (Chan et al., 2022) was used to create line drawing versions of STL10 (Coates et al., 2011) and ImageNet-1K (Russakovsky et al., 2015).

For the other augmentations, code from <https://github.com/xunhuang1995/AdaIN-style> (Huang & Belongie, 2017) was used to generate Matisse, Mondrian, and Picasso style STL10 images with degree of stylization $\alpha = 0.6$.

The base ADE20K (Zhou et al., 2017) and Pascal VOC (Everingham et al., 2010) are also used downloaded for evaluation on segmentation and detection.

A.4 SUPERVISED LEARNING EXPERIMENTS

We conduct experiments comparing two training approaches: a baseline Color-Only Training (direct training on STL10 color images) and our Line→Color curriculum (an initial phase of learning on line drawings followed by a second phase on color images). As a control, we replicated this same curriculum with 3 different arbitrary augmentations (Matisse, Mondrian, Picasso) to show that performance benefits came from the intrinsic properties of line drawings and not just the effects of using an alternative visual modality for an initial learning stage. The dataset is STL10 with 10-class image classification, using 80% train / 20% validation split with stratified sampling. We test data percentages from 5% to 100% in increments of 5% or 10%.

The model architecture uses ResNet-18 trained from scratch, modified with a final linear layer (512 → 10) for 10-class classification, totaling 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 1e-5 to 1e-3. Final hyperparameters use adaptive learning rates and batch sizes based on dataset size: 0.001/4 for small datasets (≤ 300 samples), 0.003/8 for medium datasets (≤ 1000 samples), 0.005/16 for larger datasets (≤ 2000 samples), and 0.01/32 for full datasets (> 2000 samples).

Training configuration uses SGD optimizer with momentum 0.9, ReduceLROnPlateau scheduler with patience 3, early stopping with patience 10 epochs, and maximum 100 epochs. The primary evaluation metric is test accuracy on held-out test sets.

All experiment runs use system-generated randomness with no fixed seeds, reflecting natural variation and real-world training variability. Experimental runs use 5 runs per configuration with different random seeds. The training code is available in the code appendix in `Training/train_resnet18.py`.

We also conducted experiments using Vision Transformer Tiny (ViT-Tiny) architecture for comparison with the ResNet-18 baseline. The model uses the `vit_tiny_patch16_224` configuration from the `timm` library, trained from scratch on STL10 with 10-class classification, containing approximately 5.7M parameters with patch size 16x16 and input resolution 224x224. The architecture includes 12 transformer layers, 3 attention heads, hidden dimension 192, learnable position embeddings, and a classification head (192 → 10) for 10-class classification.

The training configuration was optimized specifically for small datasets based on recommendations from (Steiner et al., 2022). We used AdamW optimizer with learning rate 0.0005, weight decay 0.01,

and betas (0.9, 0.999), combined with cosine annealing learning rate schedule and 10% warmup epochs. The batch size was set to 32 (reduced from standard 64 for better gradient estimates on small datasets) with light augmentation strategy including RandomResizedCrop (scale 0.85-1.0), RandomHorizontalFlip (p=0.5), and mild ColorJitter. Regularization included conservative mixup ($\alpha = 0.1$, applied 30% of the time), label smoothing (0.05), and gradient clipping (max_norm=0.5). Training used maximum 150 epochs with early stopping patience of 20 epochs.

The configuration specifically addresses challenges of training ViTs on small datasets through reduced regularization, conservative augmentation, smaller batch sizes, and extended patience periods. The curriculum learning approach follows the same Line→Color transfer strategy as the ResNet-18 experiments, with Xavier uniform initialization for linear layers and standard initialization for Layer-Norm components following best practices for training ViTs from scratch. Training code is available in the code appendix in `Training/train_tiny_vit.py`

A.5 GRADCAM DISTRIBUTION ANALYSIS

We analyze the spatial distribution of model attention by counting distinct high-activation regions and average size of said regions in GradCAM heatmaps. This quantifies whether models focus on concentrated areas or distribute attention across multiple regions.

For each GradCAM heatmap, we compute a dynamic threshold as the 85th percentile of heatmap values. We then create a binary mask where pixels above the threshold are marked as attention regions. To remove noise, we filter out regions smaller than 10 pixels and count distinct connected components using 8-connectivity labeling.

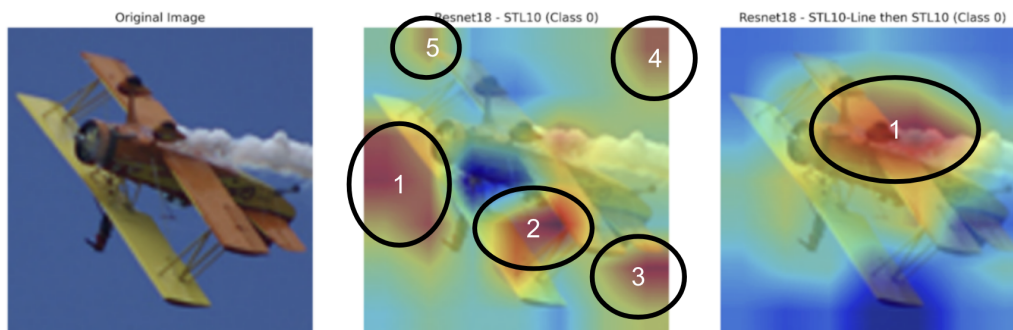


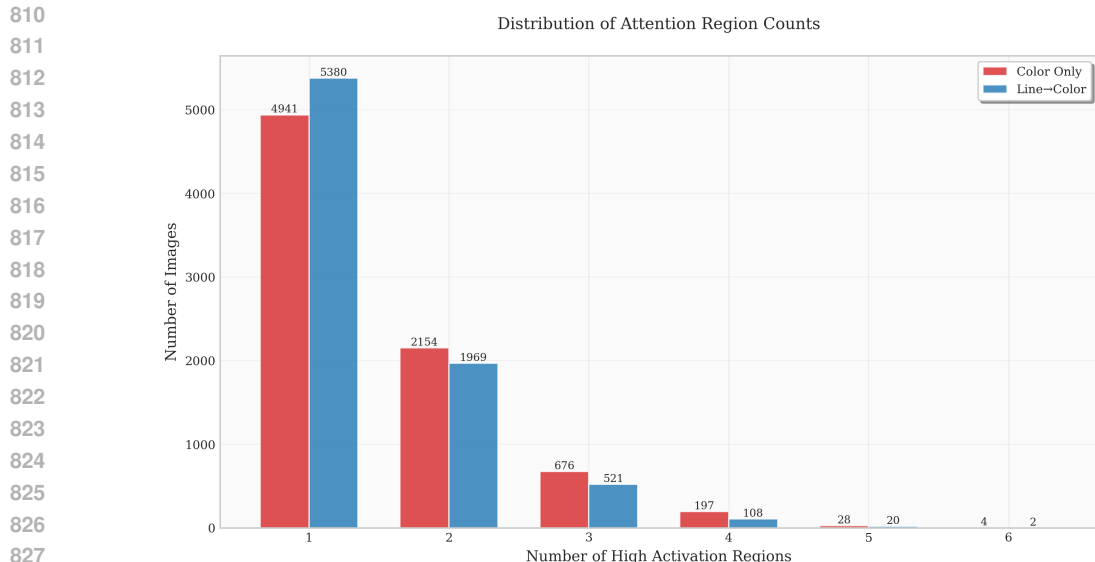
Figure 11: Example of counting number of high activation regions from GradCam. The GradCAM heatmap for base ResNet-18 has 5 distinct activation regions on this image while the ResNet-18 from the structure-first curriculum has one larger focused activation region.

The attention scatter metric is defined as the number of distinct attention regions per image. We visualize the distribution of region counts as a histogram to show the frequency of different attention scatter patterns.

For mean region size, we calculate the area of each connected component in pixels and compute the average region size per image. This metric indicates whether attention regions tend to be large and concentrated or small and scattered. Quantitative analysis confirms that the structure-first curriculum leads to a larger number of images with a single, compact attention region, whereas color-only training results in more fragmented attention (Figure 12). The GradCam analysis is available in the code appendix in `Evaluation/gradcam_activations.py`.

A.6 NEURON TUNING CURVE ANALYSIS

The neuron tuning curve analysis examines how different initial learning modalities affect the activation patterns of individual channels in the final convolutional layer of ResNet-18 models. In this analysis, each "neuron" corresponds to one of the 512 channels in the final convolutional layer (layer4) of ResNet-18. The activation of each neuron is computed by averaging the spatial dimensions (height and width) of the corresponding feature map.



828
829
830
831
832

Figure 12: Number of distinct high activation regions across all STL10 Test images. High activation regions are marked as the number of regions above the 85th percentile in GradCam activation.

833
834
835
836

The analysis processes a dataset of test images (up to 500 images) through both models: ResNet-18 trained on color images only (Color Only), and ResNet-18 first trained on line drawings followed by a second phase of training on color images (Line→Color). For each image, we extract the 512-dimensional neuron activation vector using forward hooks registered on the layer4 module.

837
838
839
840

For each model, we compute the mean activation across all processed images for each of the 512 neurons. The neurons are then sorted by their mean activation values in descending order, creating a tuning curve that shows the distribution of neuron response strengths. The curves are normalized by dividing by the maximum activation value to facilitate comparison between models.

841
842
843
844
845
846
847

The resulting tuning curves reveal the degree of specialization in the learned representations. A steeper, more concentrated curve indicates that fewer neurons are highly active, suggesting more specialized, sparse representations. By comparing the curves between Color Only and Line→Color models, we can assess whether an initial phase of learning on line drawings leads to more efficient, specialized neuronal representations. The code for this is available in the code appendix in `Evaluation/neuron_tuning_curves.py`.

848 849 A.7 GENERALIZING TO SEGMENTATION

850
851
852
853
854
855
856

To evaluate how well ResNet-18 models initialized with different visual modalities generalize to semantic segmentation, we conducted experiments on the ADE20K dataset using a standard encoder-decoder architecture. We compared two ResNet-18 encoders developed through different curricula: one trained on color images and another trained sequentially on line drawings followed by color images (Line→Color). We also include another ResNet-18 that is simultaneously trained on line and color images to investigate whether training order affects generalization ability.

857
858
859
860
861

We modified the semantic segmentation framework from <https://github.com/CSAILVision/semantic-segmentation-pytorch> (Zhou et al., 2018) to support our custom encoders. The training configurations closely match the original framework’s settings to ensure a fair comparison. We employed a Pyramid Pooling Module (PPM) decoder architecture with ResNet-18 as the encoder backbone.

862
863

The training setup was identical for all models to ensure fair comparison. We trained for 20 epochs with a batch size of 2 per GPU using SGD optimization. The learning rate was set to 0.02 for both encoder and decoder, with a polynomial decay schedule ($lr_{current} = lr_{initial} \times (1 - \frac{iter}{max.iter})^{0.9}$).

864 Both the encoder and decoder were trained end-to-end, allowing the learned representations to be
865 adapted for the segmentation task.

866
867 The models were trained on ADE20K with 150 semantic classes, using images resized to multiple
868 scales (300, 375, 450, 525, 600) with a maximum size of 1000 pixels. We employed standard data
869 augmentation including random horizontal flipping and used the negative log-likelihood loss with
870 ignored index -1 for handling invalid labels.

871 Both models were evaluated using the same checkpoint at epoch 20. The evaluation metrics include
872 pixel accuracy and mean intersection-over-union (mIoU) on the ADE20K validation set. This exper-
873 imental setup allows us to directly compare the representational quality of the two learning curricula
874 by measuring their performance on a downstream semantic segmentation task. The code for this is
875 available in the code appendix in `Training/Segmentation`.

876 877 A.8 OBJECT DETECTION IN PASCAL VOC

878 We utilize the PASCAL VOC dataset to evaluate data efficiency on object detection. For our cur-
879 riculum, we first train the model on line drawings for 200 epochs and then train on color images
880 for another 200 epochs. The base model is trained on photographic images alone for 200 epochs.
881 We follow the training setup from ultralytics Reis et al. (2023) and use the YOLO-v8-n architecture.
882 The code to run experiments in this section is located under `Training/ObjectDetection`.

883 884 A.9 KNOWLEDGE DISTILLATION EXPERIMENTS

885 We utilize the STL10 classification task for knowledge distillation. The codebase is included in the
886 supplementary folder `Training/KnowledgeDistillation` as well as the experiment setup.
887 The experiments used three types of teachers: (1) The best performing model developed through our
888 Line→Color curriculum; (2) The model developed with the Line→Color curriculum, where both
889 learning stages are conducted using only 70% of the data. With less data, the model performance
890 drops but still matches the baseline training with raw color images on 100% data. Using such a
891 teacher ensures a fair comparison between the model from our structure-first curriculum and the
892 baseline model. The experimental results in Table 5 in the main manuscript suggest that the line
893 drawings produce a better representation that is superior when used for knowledge distillation. (3)
894 The baseline model trained on 100% of the training data.

895 896 A.10 CODE

897 Source code for this project can be found at the following anonymized repository: [https://](https://anonymous.4open.science/r/ICLR2026-Line-Drawing-FA81)
898 `anonymous.4open.science/r/ICLR2026-Line-Drawing-FA81`.

900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917