Characterizing Plastic Regions in Neural Networks

Gwyneth Liu [1] **Nir Shavit** [*1] **Linghao Kong** [*1]

Abstract

Adapting a trained model to a new domain without overwriting prior knowledge is useful only when the model contains a region whose parameter state can support new learning. In vision classifiers, we study *plastic regions*: contiguous, easily-discoverable regions in which some manipulation of the region improves the target–source trade-off over size-matched control strips elsewhere in the same network. We first characterize a plastic region in ResNet-18 and show that it transfers across target domains, compounds under sequential adaptation, and can be manipulated to recover adaptation capacity at rigid checkpoints. We then analyze plastic-region existence across nine architectures and report observations about network properties that appear to enable or obstruct plastic-region formation.

1. Introduction

Neural networks deployed under distribution shift must keep learning without erasing what they already know. As models grow, updating the full network becomes expensive. Fine-tuning on a new domain without restrictions can also damage source knowledge. A common response is to restrict where learning occurs. Surgical fine-tuning selects layers to update, parameter-efficient methods train small modules or low-rank update directions, and continual-learning methods regularize changes to important parameters (Lee et al., 2023; Hu et al., 2022; Houlsby et al., 2019; Kirkpatrick et al., 2017; Zenke et al., 2017). These approaches introduce a valuable question: when do some parts of a network naturally support adaptation better than others?

We study this question by treating localized adaptation capacity as a property of a trained network. We define **plastic regions:** contiguous regions in a trained network that are (a) easily discoverable by a metric and (b) admit a manipulation that gives higher target–source utility (target gain minus source damage) than the same manipulation on size-matched *control regions*, i.e. contiguous strips of the same width drawn from elsewhere in the same network. This pairs discoverability with practical utility, separating plasticity from regions that are merely high-importance or high-gradient. Specifically, we make the following three contributions:

- **We define plastic regions by discoverability and manipulation.** A candidate region is plastic only if it is flagged by a simple metric and some tested manipulation of that region outperforms the best utility achievable on size-matched control regions under the same tested manipulation family (Equation 1).
- **We characterize a detailed ResNet-18 case.** The discovered ResNet-18 region transfers across targets, compounds under sequential adaptation, and can be manipulated to recover adaptation capacity.
- **We study plastic region existence across multiple architectures.** Across a nine-architecture study, we find that six architectures contain discoverable plastic regions under the manipulation definition. We then report observations about network properties that appear to enable or obstruct plastic-region formation.

Related work. Surgical fine-tuning (Lee et al., 2023) chooses which layers to train. LoRA and adapters (Hu et al., 2022; Houlsby et al., 2019) add trainable low-rank or module parameters. Importance-based continual learning (Kirkpatrick et al., 2017; Zenke et al., 2017) penalizes movement in source-important weights. Targeted-reset methods such as the dormant neuron treatment of Sokar et al. (2023) reinitialize specific units to restore learning capacity. These methods prescribe *how* to update. We ask a different question: whether the trained network already contains a region whose existing parameters can be manipulated to improve the target–source trade-off relative to size-matched controls. Loss-of-plasticity work (Dohare et al., 2024; Lyle et al., 2023) studies global plasticity decay while we isolate localized adaptation capacity.

2. Characterizing Plastic Regions

In this section, we show how to identify plastic regions and describe the properties that make them useful for adaptation.

[1]Massachusetts Institute of Technology. Correspondence to: Gwyneth Liu <gwyliu@mit.edu>.

Presented at the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI". Copyright 2026 by the author(s).

2.1. Identifying Plastic Regions

We use adaptation under intervention as the test for plasticity. Given a trained model, a small target-domain set, and a candidate region $\mathcal{S}$, we measure target gain and source damage after K steps of SGD on the target support set (where K is the adaptation step budget). We use

$$U = \Delta_T - \Delta_S,$$

where Δ_T is target NLL improvement and Δ_S is source NLL damage. Larger U means better target adaptation with less source damage. We also report head-relative utility $U_{\text{rel}} = U - U(\{h\})$ when comparing ordinary adaptation runs, so that a region is credited only for improvement beyond updating the output head alone. *Control regions* $\mathcal{C} \in \mathcal{B}$ are contiguous regions of the same width as $\mathcal{S}$ drawn from elsewhere in the backbone. Full training and adaptation details are in Appendix A.1. This leads into our main definition:

Definition 2.1 (Plastic Region)**.** Let $\mathcal{D}$ be a discovery metric that nominates a contiguous region from a trained model. A region $\mathcal{S} = \mathcal{D}(\text{model})$ is a *plastic region* if the best tested manipulation of $\mathcal{S}$ outperforms every tested manipulation of every size-matched control region under the target–source utility:

$$\max_{\mathcal{M} \in \mathcal{A}_\mathcal{S}} U(\mathcal{M}(\mathcal{S})) - \max_{\mathcal{C} \in \mathcal{B},\, \mathcal{M}' \in \mathcal{A}_\mathcal{C}} U(\mathcal{M}'(\mathcal{C})) > 0. \quad (1)$$

Here $\mathcal{B}$ is the set of size-matched control regions and $\mathcal{A}_\mathcal{S}, \mathcal{A}_\mathcal{C}$ are the tested manipulation sets.[1]

In this paper, the main discovery metric selects the width-3 contiguous block strip with the highest mean target-domain diagonal Fisher information (the expected squared target-loss gradient per parameter, a proxy for how strongly each parameter influences target predictions), and the manipulation family $\mathcal{A}$ contains shrink-and-perturb (Ash & Adams, 2020) at α=0.5 and rewind (Frankle et al., 2020) to an earlier checkpoint. We also test late parameter-path as a discovery metric, which selects strips by cumulative late-training parameter movement. We find that this metric yields the same binary plastic/non-plastic verdicts on 8 of 9 architectures as our Fisher metric (Appendix A.4), although the selected regions differ on some architectures.

2.2. Existence and Properties of Plastic Regions

We first show that plastic regions exist in a standard trained network. Our primary case is a ResNet-18 trained on CIFAR-10, where the discovered region $\mathcal{S} = \{\text{L3.1, L3.2, L4.1}\}$ consists of three contiguous late residual blocks. We use this case as an existence proof and as a way to characterize what makes a region useful: it should improve adaptation, remain useful beyond the target used to find it, be geometrically distinctive, and be causally tied to adaptation capacity. We find the following properties of the discovered plastic region:

Table 1. ResNet-18 positive case (epoch 50, 48 SVHN episodes). The plastic region improves the target–source trade-off over same-width random regions at every budget.

	K=2	K=5	K=10
Head-relative utility U_{rel}			
Head only	0.000	0.000	0.000
Head+rand	+0.060	+0.253	+0.272
Head+$\mathcal{S}$	**+0.112**	**+0.296**	**+0.494**
Source damage Δ_S			
Head only	0.025	0.108	0.339
Head+rand	0.026	0.110	0.358
Head+$\mathcal{S}$	**0.024**	**0.104**	**0.333**

(1) Adaptation gain and sequential compounding. The plastic region is reproducible across runs: independently trained ResNet-18 branches yield regions with mean pairwise Jaccard overlap ≈0.83, indicating $\mathcal{S}$ reflects a structural property of the learned representation rather than being seed-dependent. Under increasing adaptation pressure ($K \in \{2, 5, 10\}$), head+$\mathcal{S}$ outperforms alternatives at every step budget (Table 1). At K=10, head+$\mathcal{S}$ achieves head-relative utility +0.494 versus +0.272 for random width-3 strips. We find the gain is larger in sequential transfer, where the model adapts to one target episode and then to an independent second episode. Compared with head-only, head+$\mathcal{S}$ roughly doubles the Phase-2 target gain and cuts cumulative source damage by up to sixfold. This advantage grows as the source network specializes: we find that at later anchors, head-only accumulates the largest source damage while head+$\mathcal{S}$ remains near-zero. This demonstrates that the plastic region is beneficial beyond an initial adaptation and preserves additional adaptation capacity. We show more results in Appendix A.3.

(2) Cross-task transfer. The plastic region transfers across target domains: the SVHN-derived plastic region produces positive head-relative utility on CIFAR-100 (+0.26) (Krizhevsky, 2009), Fashion-MNIST (+0.74) (Xiao et al., 2017), and STL-10 (+0.18) (Coates et al., 2011) (Table 3 in Appendix). The plastic region is therefore a property of the trained network and not specific to the source–target pair.

(3) Causal manipulability. We test causal manipulability at a rigid late checkpoint using four interventions. Identity leaves the plastic region unchanged. Rewind replaces the

[1] A failure under $\mathcal{D}$ means there is no plastic region *discoverable by* $\mathcal{D}$; it does not rule out the existence of an adaptable region elsewhere in the network. Tying plasticity to discoverability is intentional: a plastic region that requires exhaustive search is not practically useful.

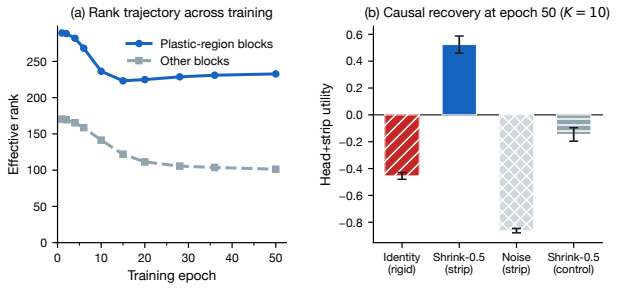


Figure 1. **Geometric and causal signatures of the plastic region (ResNet-18, CIFAR-10).** *Left:* Effective rank (Roy & Vetterli, 2007) across training for plastic-region blocks (layers 3.1–4.1, blue) vs. other backbone blocks (gray). Plastic-region blocks maintain $\sim 3\times$ higher rank throughout training while surrounding blocks lose $\sim 13\%$ of their rank by epoch 50. *Right:* Head+plastic-region utility at the rigidified checkpoint (epoch 50, $K{=}10$, $n = 96$ episodes). Shrink-and-perturb on the plastic region is the only tested manipulation with positive utility; control-strip utilities and noise remain negative.

plastic-region parameters with their values from an earlier checkpoint. Shrink-and-perturb interpolates the plastic region toward initialization, with smaller α giving a stronger reset. Matched-variance Gaussian noise perturbs the plastic region with random noise at comparable variance and serves as another control.

Identity adaptation gives $U = -0.455$, showing late-stage rigidification. Shrink-and-perturb on the plastic region flips utility to $U = +0.524$. The best control utility over tested control manipulations is $U = -0.146$ (here, shrink-and-perturb on the control strip), yielding a plastic-region margin of $+0.670$ under Equation 1. The result is specific to a structured manipulation on the plastic region: matched-variance noise on the plastic region gives $U = -0.863$, and applying the same shrink-and-perturb to a control region also fails ($U = -0.146$).

The recovery has a representational signature. We measure Centered Kernel Alignment (CKA) (Kornblith et al., 2019), which scores representational similarity in [0,1] (1 = identical up to orthogonal transformation), between plastic region activations before and after adaptation on SVHN inputs. At the rigidified checkpoint, region representations barely change during adaptation (CKA = 0.952), but after shrink-and-perturb they reorganize as freely as at mid-training (CKA = 0.942 vs. 0.943 at epoch 15). The effect is also largest at mid-rigidification: shrink-0.5 yields utility +1.47 at anchor 28 versus +0.52 at anchor 50, ruling out a single-checkpoint artifact. With $n{=}96$ episodes, paired Wilcoxon tests give $p < 10^{-7}$ against the like-perturbation shrink-and-perturb control and $p < 10^{-15}$ against plastic-region noise. Shrink-and-perturb also raises the region's mean effective rank from 232.8 to 284.2, moving it toward the higher-rank regime of earlier checkpoints. Causal manipulation results are shown in Figure 1.

3. Architectural Conditions for Plastic Regions

We next ask when a plastic region exists across architectures. For each architecture, we identify a candidate plastic region using the same Fisher-based procedure, then apply the manipulation definition from Equation 1. We report the best manipulation on that plastic region, the best control utility, and the resulting margin. A negative entry means no width-3 strip satisfied Equation 1 under our discovery rule and manipulation family.

We evaluate ResNet-18/50 (He et al., 2016), VGG-16 (Simonyan & Zisserman, 2015), MobileNetV2 (Sandler et al., 2018), ConvNeXt-T (Liu et al., 2022), ViT-Small (Dosovitskiy et al., 2021), Swin-T (Liu et al., 2021), DenseNet-121 (Huang et al., 2017), and WRN-16-4 (Zagoruyko & Komodakis, 2016). Four cases are strong (ResNet-50, VGG-16, MobileNetV2, ConvNeXt-T); ViT-Small and ResNet-18 are weak positive cases. We did not find a plastic region in Swin-T, DenseNet-121, and WRN-16-4. The binary plasticity verdict (plastic/not plastic) agreed between Fisher and parameter-path discovery in 8 of 9 architectures; MobileNetV2 was the mismatch (Appendix A.4). Observed patterns can be found in Figure 2. We make the following exploratory observations on architectural properties for plastic region formation:

- *Intermediate region scale.* Strong margins cluster at candidate-region fractions $p_S \in [0.25, 0.64]$, while DenseNet-121 (0.07) and WRN-16-4 (0.87) fail under this criterion. The candidate may need to be large enough to mediate target updates yet localized enough to beat controls.
- *Geometry is a modest correlate.* Activation-dimensionality ratio has the highest linear R^2 (≈ 0.19; defined in Appendix A.1); no single predictor explains most of the margin variance (all other $R^2 \leq 0.09$). DenseNet-121 is the only architecture with candidate-region effective rank below the network mean, yet WRN-16-4 has a comparable rank ratio to MobileNetV2 and still fails.
- *Dense feature reuse.* Positive cases use residual skips or no skips (VGG-16). DenseNet-121 is the only model with concat-style dense connectivity, which may obstruct this form of localized plasticity by tightly coupling late features to accumulated channels.

4. Discussion and Future Work

We introduced plastic regions as contiguous regions that can be manipulated to improve adaptation, defined by a discovery metric and by manipulability. In ResNet-18, this

Table 2. Cross-architecture sweep (N=9). Control U is the best utility among tested manipulations on a size-matched control; margin is Equation 1. The 95% CI column is a bootstrap over episodes for $\max$(strip manip. utilities) $-$ $\max$(control manip. utilities) per episode; the Margin column reports the same quantity evaluated at manipulation means (Appendix A.1).

Architecture	Family	Candidate plastic region	Best manip. (plastic)	Plastic region U	Best ctrl. U	Margin	95% CI	Verdict
ResNet-50	classical CNN	L4.1–L4.3	shrink 0.5	$+8.285$	$+0.712$	$\mathbf{+7.573}$	$[4.84, 10.19]$	Strong
VGG-16	no-skip CNN	blocks 3–5	rewind early	$+7.130$	$+1.671$	$\mathbf{+5.459}$	$[1.83, 8.88]$	Strong
MobileNetV2	depthwise CNN	features 14–16	rewind early	$+6.673$	$+2.941$	$\mathbf{+3.732}$	$[2.55, 4.92]$	Strong
ConvNeXt-T	hybrid CNN	stage 4 blocks 1–3	rewind early	$+4.988$	$+3.316$	$\mathbf{+1.671}$	$[1.26, 2.50]$	Strong
ViT-Small	transformer	blocks 10–12	rewind early	$+3.969$	$+3.110$	$\mathbf{+0.859}$	$[0.77, 1.03]$	Weak
ResNet-18	classical CNN	L3.1–L4.1	shrink 0.5	$+0.524$	-0.146	$\mathbf{+0.670}$	$[0.48, 0.80]$	Weak
Swin-T	transformer	stage 3.6, stage 4.1–4.2	shrink 0.5	$+3.401$	$+3.918$	$\mathbf{-0.517}$	$[-0.68, -0.54]$	Not plastic
DenseNet-121	dense CNN	denseblock3 layers 22–24	rewind early	$+0.093$	$+1.523$	$\mathbf{-1.430}$	$[-2.15, -0.68]$	Not plastic
WRN-16-4	wide residual CNN	L2.2–L3.2	shrink 0.5	$+2.314$	$+6.456$	$\mathbf{-4.142}$	$[-5.28, -3.55]$	Not plastic

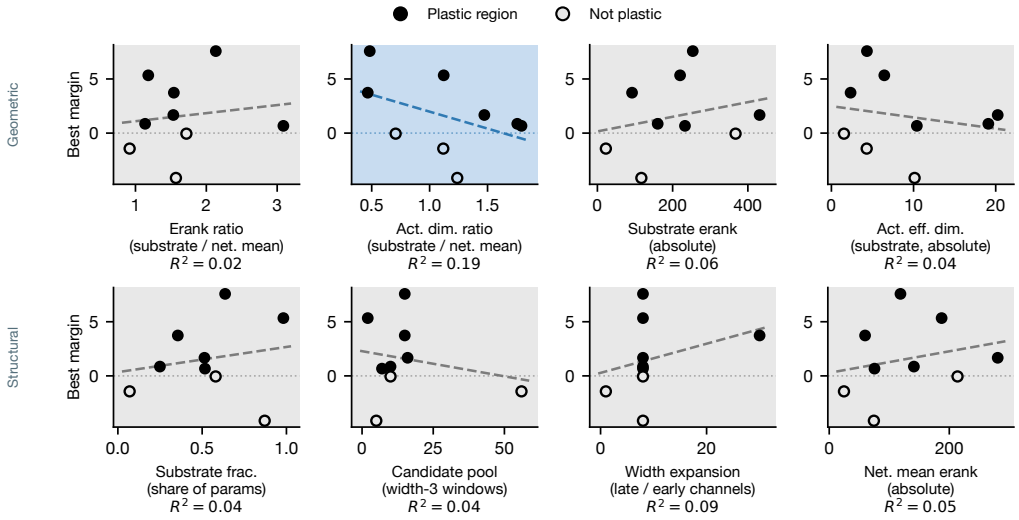


Figure 2. **Geometric and structural correlates of plastic-region strength (N=9).** Each panel shows one predictor vs. best margin; filled circles have a plastic region, open circles do not. Dashed lines are OLS fits; R^2 is shown below each label; branch values are averaged per architecture. Activation-dimensionality ratio (blue panel, R^2=0.19) is the strongest single predictor: candidate regions whose activation manifold is larger than the network mean tend to yield higher margins. All other R^2 values are ≤ 0.09, confirming that no single property linearly determines plasticity.

turns identity-adaptation utility from -0.455 into $+0.524$ via shrink-and-perturb on the plastic region, with $p < 10^{-15}$ separation from noise controls. Across the architecture sweep, six of nine models were found to contain a plastic region. Negative cases include Swin-T and WRN-16-4 where control-strip rewind outperforms the best strip manipulation, and DenseNet-121 where control strips achieve higher utility under shrink-and-perturb than the discovered plastic region. Failures cluster at extremes of candidate-region localization scale and at architectures whose candidate region lacks high-dimensional capacity.

If plastic regions can be identified reliably, they could make adaptation cheaper and provide an interpretable basis for deciding where learning should occur. More broadly, adaptability is a property not only of an endpoint model but of how learning capacity is organized within it.

Limitations and future work. The architecture and source dataset sweep is limited and observations are correlational. Controlled studies varying width, depth, connectivity, and candidate-region size are needed to test these ideas directly. Candidate regions are restricted to contiguous width-3 strips, a choice from early ResNet-18 experiments. The main discovery metric is diagonal Fisher, a coarse approximation. While we find that late parameter-path as a discovery metric yields similar binary verdicts, alternatives such as K-FAC blocks or gradient covariance remain untested. The manipulation set is also limited, and our geometry analysis could be expanded with curvature and gradient covariance (Lyle et al., 2025). Extensions also include low-rank edits, targeted noise, partial rewinding, and learned manipulations. We also aim to add probes that characterize the plastic region's representational role rather than only its manipulability.

References

Ash, J. and Adams, R. P. On warm-starting neural network training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3884–3894, 2020.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, 2011.

Dohare, S., Hernandez-Garcia, J. F., Lan, Q., Rahman, P., Mahmood, A. R., and Sutton, R. S. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024. doi: 10.1038/s41586-024-07711-7.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, T., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., and Finn, C. Surgical fine-tuning improves adaptation to distribution shifts. In *International Conference on Learning Representations*, 2023.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11966–11976, 2022.

Lyle, C., Zheng, Z., Nikishin, E., Pires, B. A., Pascanu, R., and Dabney, W. Understanding plasticity in neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23190–23211, 2023.

Lyle, C., Zheng, Z., Khetarpal, K., van Hasselt, H., Pascanu, R., Martens, J., and Dabney, W. Disentangling the causes of plasticity loss in neural networks. In *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 750–783, 2025.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. Technical report, Stanford University, 2011.

Roy, O. and Vetterli, M. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference*, pp. 606–610, 2007.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Sokar, G., Agarwal, R., Castro, P. S., and Evci, U. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2023.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, pp. 87.1–87.12. BMVA Press, 2016. doi: 10.5244/C.30.87.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995, 2017.

A. Appendix

A.1. Experimental Protocol and Metrics

Source training and adaptation. We train networks on CIFAR-10 (Krizhevsky, 2009) with SGD and cosine decay for 100 epochs, saving checkpoints throughout. Unless otherwise noted, adaptation uses a 5-shot SVHN (Netzer et al., 2011) support set, $K{=}10$ SGD steps. We measure target NLL gain Δ_T and source NLL damage Δ_S, combine them into $U = \Delta_T - \Delta_S$, and report head-relative utility $U_{\mathrm{rel}}(I) = U(I) - U(\{h\})$.

Candidate discovery. For each block, we compute target-domain diagonal Fisher information at the trained checkpoint. Candidate regions are width-3 contiguous block strips with highest mean Fisher mass. We also run a second discovery metric based on late parameter-path length. This discovery step is deliberately separated from the plastic-region definition: metrics propose candidates, and manipulation against size-matched controls decides whether they are plastic.

Effective rank. For a reshaped weight tensor W with singular values $\sigma_1 \geq \cdots \geq \sigma_r$, we define

$$\mathrm{erank}(W) = \exp\left(-\sum_i p_i \log p_i\right), \quad p_i = \sigma_i / \sum_j \sigma_j. \tag{2}$$

We compute the analogous effective dimensionality of activations from the singular spectrum of the centered activation matrix. Because effective rank scales with layer width, cross-architecture comparisons use the candidate-region-to-network ratio (Table 8) rather than absolute values.

A.2. Cross-Task Transfer

We show transfer to other target domains here.

Table 3. Cross-task transfer: head-relative utility U_{rel} of the SVHN-derived plastic region applied to four target domains. The same plastic region produces positive utility on every target.

Target domain	plastic region (SVHN)	plastic region (target)
SVHN	+0.193	+0.133
CIFAR-100 (sub-10)	+0.259	+0.198
Fashion-MNIST	+0.740	+0.440
STL-10	+0.182	+0.150
Mean	+0.344	+0.230

A.3. Sequential Transfer

In a two-phase sequential setup using independent target episodes in the primary positive architecture, the model adapts once and is then adapted again. This tests whether the first adaptation leaves the model in a state that can still support later adaptation. Head+plastic region increases second-phase target gain and reduces cumulative source damage relative to head-only at every anchor (Table 4).

Table 4. Sequential adaptation. Head+plastic region improves second-phase target gain and reduces cumulative source damage relative to head-only adaptation.

Anchor	Phase-2 target gain		Cumulative source damage	
	Head	Head+ plastic region	Head	Head+ plastic region
20	0.65	1.01	0.36	0.01
36	0.75	1.06	0.33	0.22
50	0.73	1.03	0.81	0.13

Head+plastic region is therefore useful in sequential adaptation because it changes both immediate performance and the post-adaptation state. We infer that the first update is routed through the region rather than only through the classifier head, reducing source damage before the next adaptation begins.

A.4. Discovery-Metric Robustness

We used two discovery metrics. Fisher is the main metric in the paper because it scores target sensitivity at the trained checkpoint. Late parameter-path length is a robustness check because it scores where parameters moved during late source training. These are shown in Table 5.

A.5. Per-Architecture Manipulation Utilities

Table 7 reports the per-manipulation utilities at the late anchor ($K{=}10$) for each architecture. Each plastic-region

Table 5. Discovery metrics used to nominate candidate strips. Both metrics nominate width-3 contiguous block strips; manipulation against controls determines the verdict.

Metric	What it measures	Role
Target Fisher	Mean target-domain diagonal Fisher mass in a block strip at the trained checkpoint.	Main discovery metric; favors regions currently sensitive to target data.
Late parameter path	Cumulative late-training parameter movement/path length for each block, normalized across candidate strips.	Alternative discovery metric; favors regions that moved substantially during source training.

manipulation is paired with its like-perturbation control where measured.

The shrink-vs-rewind preference splits architectures: ResNet-18 and ResNet-50 prefer shrink; VGG-16, MobileNetV2, ConvNeXt-T, ViT-Small prefer rewind. ConvNeXt-T and ResNet-18 are the only architectures with all measurable manipulations producing positive deltas. Swin-T and WRN-16-4 are the only architectures with all measurable manipulations producing negative deltas. This per-manipulation variation is what makes the max–max comparison in Equation 1 the appropriate summary: the headline margin need not coincide with any single row of Table 7.

A.6. Per-Architecture Geometric Metrics

Table 8 reports the geometric measurements behind Figure 2. Effective-rank ratio is candidate-region mean divided by network mean, averaged across two branches.

DenseNet-121 is the only architecture with candidate-region effective rank below the network mean. WRN-16-4's candidate-region rank is comparable to MobileNetV2's despite the verdict difference, consistent with the observation that high-dimensional capacity is necessary but not sufficient for plastic-region formation.

A.7. Additional Architecture Details

Table 2 reports verdict-defining values; Table 8 provides the underlying data. Manipulations were computed at each architecture's late anchor (epoch 50 for ResNet-18 with the published protocol; epoch 100 for the standard CNN sweep; epoch 200 for ViT-Small). Random-strip baselines and control regions follow the same exclusions across architectures: stem, classifier, and architecture-specific connectors (transitions for DenseNet, output projection for MobileNetV2).

Table 6. Discovery-metric robustness. Fisher verdicts follow Table 2 (best control utility). Param-path verdicts use alternative candidate plastic regions with tier labels carried over from the sweep's matched-control synthesis, except ViT-Small is listed as weak because the plastic region discovered under param-path matches Fisher and the Table 2 margin is weak.

Arch.	Fisher $\mathcal{S}$	Path $\mathcal{S}$	Path U	Path ctrl U	Path marg.	Fisher	Path	Agree?
ResNet-50	L4.1–L4.3	L3.1–L3.3	$+1.140$	$+0.235$	$+0.905$	Strong	Weak	Yes
VGG-16	blk. 3–5	blk. 2–4	$+6.825$	$+1.562$	$+5.263$	Strong	Strong	Yes
MobileNetV2	ft. 14–16	ft. 13–15	$+2.851$	$+2.948$	-0.097	Strong	Neutral	No
ViT-Small	blk. 10–12	blk. 10–12	$+3.969$	$+3.110$	$+0.859$	Weak	Weak	Yes
ConvNeXt-T	stg. 4 (1–3)	stg. 4 (1–3)	$+4.988$	$+3.316$	$+1.671$	Strong	Strong	Yes
ResNet-18	L3.1–L4.1	L3.1–L4.1	$+0.524$	-0.146	$+0.670$	Weak	Weak	Yes
DenseNet-121	db3 22–24	db3 9–11*	$+0.178$	$+1.500$	-1.322	No	No	Yes
Swin-T	stg. 3.6–4.2	stg. 3.6–4.2	$+4.256$	$+4.308$	-0.052	Neutral	Neutral	Yes
WRN-16-4	L2.2–L3.2	L2.1–L3.1	$+13.668$	$+62.418$	-48.749	No	No	Yes

Table 7. Per-architecture late-anchor utilities by manipulation. Plastic region vs. like-perturbation control deltas for each manipulation (diagnostic); Table 2 aggregates the best manipulation on the plastic region against the best control utility (Equation 1). "–" indicates the control was not run for that perturbation strength on that architecture.

Architecture	shrink 0.5 Δ	rewind early Δ
ResNet-50	$+7.573$	-0.127
VGG-16	-0.620	$+5.459$
MobileNetV2	-1.656	$+5.319$
ConvNeXt-T	$+0.579$	$+1.728$
ViT-Small	-0.284	$+3.345$
ResNet-18	$+0.670$	$+0.199$*
DenseNet-121	-1.448	$+0.035$
Swin-T	-0.272	-1.584
WRN-16-4	-1.482	-4.647

*ResNet-18 uses rewind to specific epochs (10/20) rather than EARLY donor.

Table 8. Per-architecture geometric measurements at the late anchor. Candidate-region erank is averaged across branches; network mean covers all candidate blocks excluding stem and head.

Architecture	Cand. erank	Net. erank	Erank ratio	Top-1 ratio
ResNet-18	232.8	75.4	3.09	0.016
ResNet-50	253.6	118.6	2.13	0.038
ConvNeXt-T	431.2	280.7	1.53	0.014
Swin-T	367.3	213.7	1.72	0.021
WRN-16-4	116.9	74.4	1.57	0.042
MobileNetV2	92.1	59.7	1.54	0.078
VGG-16	219.8	187.2	1.18	0.030
ViT-Small	160.3	141.1	1.14	0.029
DenseNet-121	22.5	24.5	0.92	0.101