TANGENT KERNELS MEASURE THE GAUGE: NON-IDENTIFIABILITY AND FULL-SPACE AUDITS OF TRAINING-FREE NAS PROXIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Architectures are increasingly compared *without* training them, by reading spectral summaries—effective rank, condition number, kernel–target alignment—off an empirical tangent kernel as architectural properties. For softmax classifiers they are not. We prove that a function- and training-preserving *softmax gauge* $z \mapsto z + \gamma h_\phi(x)\mathbf{1}_C$ injects arbitrary PSD contamination into the trace NTK while leaving the softmax, cross-entropy loss, gradient and training trajectory untouched; its same-architecture form $W \mapsto W + \gamma \mathbf{1}_C a^\top$ leaves *every* gradient unchanged. The class-quotient NTK—the class-projected kernel used to analyse cross-entropy dynamics—is exactly invariant under both forms; sample centering removes the injected block only when its tangent features are constant across the probe set.

Auditing whole search spaces—all $15{,}625$ NAS-Bench-201 cells and the $4{,}096$-cell TransNAS-Bench-101 micro space—shows the consequence is practical. Editing one weight matrix leaves the spoofed top-10 shortlist holding 2 of its 10 architectures, and an adversary choosing the gauge strength per candidate forces the selection of a network 34.6 accuracy points worse than the proxy's choice, while predictions, loss and gradients are preserved to 10^{-3} in float32. On the quotient kernel the same sweep leaves every shortlist and the selected architecture fixed, with drift below 10^{-4}. The effect persists on a 3.0×10^8-parameter vision transformer and across graph architectures, where sample centering amplifies rather than repairs drift; a pre-specified control on softmax-free heads bounds the claim.

Consequently, a tangent-kernel number reported for a softmax model is a claim about a gauge choice unless it is quotiented, so proxy comparisons built on one are not falsifiable as stated. The repair is cheap—one extra evaluation of the diagnostic under a random function-preserving gauge, released as an audit—but precision-critical: default reduced-precision GPU matmuls alone make an exact invariance look violated.

1 INTRODUCTION

The empirical Neural Tangent Kernel (Jacot et al., 2018; Lee et al., 2019a; Arora et al., 2019) is routinely used to compare and select architectures through the spectrum of a *trace* NTK—a sum of per-logit Jacobian inner products—via summaries such as effective rank, condition number and kernel–target alignment (Cristianini et al., 2001). The kernel in question is the finite-width, finite-checkpoint one—the regime in which the linearization is a lazy-training approximation (Chizat et al., 2019; Yang, 2020) rather than an identity—and standard tooling has made it cheap to obtain (Novak et al., 2020). TE-NAS ranks architectures directly on NTK spectral indicators (Chen et al., 2021); the wider training-free literature adds activation-overlap (Mellor et al., 2021) and pruning-derived zero-cost scores (Abdelfattah et al., 2021), which we treat separately in §7. For *softmax* classifiers the practice is unsound at its foundation: the softmax is invariant to a class-common shift of the logits, so an entire subspace of tangent directions—the all-ones "logit gauge"—is invisible to the predictive function and to cross-entropy training, yet contributes to the trace NTK. One can add any PSD matrix to the trace NTK on a probe set, moving its spectral summaries anywhere within their

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

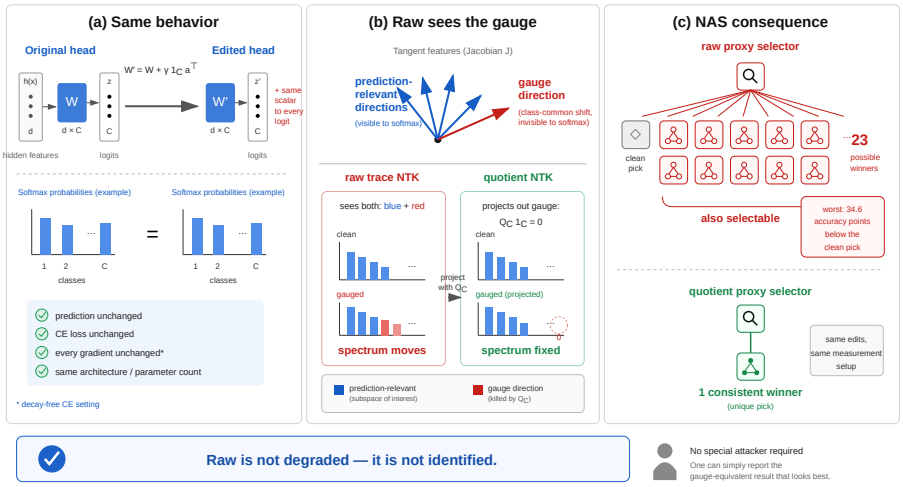


Figure 1: **Same model behavior, different raw-NTK answers.** The edit adds the same scalar to every logit, so predictions, loss and gradients are unchanged; the trace NTK nevertheless sums over the added direction, which the class quotient annihilates. Over all 15,625 NAS-Bench-201 cells on CIFAR-10 the raw d_{PR} admits 23 winners under an attacker choosing the gauge strength per candidate, the quotient one (§4, Appendix L).

attainable ranges, without changing the model's predictions, loss, accuracy or training trajectory (Fig. 1). This is the NTK analogue of the reparameterization non-invariance Dinh et al. (2017) exposed for sharpness: a quantity used as an intrinsic measure of a network is in fact a coordinate—here, gauge—artifact.

Three claims.

1. The trace NTK is **non-identifiable** on the softmax predictive equivalence class: a behaviour-preserving gauge injects arbitrary PSD contamination.
2. The **Quotient NTK**—the class-projected kernel $P\Theta P$ of Plenk et al. (2026)—is exactly gauge-invariant; its invariance, not its dynamics, is what our audit certifies.
3. The problem does **real damage** on full NAS search spaces (§4–§5) and at pretrained-model scale (§6.1), and we delimit where it stops.

Scope, stated once and kept. Our results concern the *additive softmax gauge*: a class-common, input-dependent shift of logits consumed by a softmax under cross-entropy, where the invariance statements are exact. Outside it—heads with no softmax, and multiplicative reparameterizations, which do not preserve the Euclidean optimizer geometry—the construction is not expected to apply, and §5 measures that boundary rather than asserting it. Multiplicative symmetries, gauge-fixing and singular learning theory are supporting material (Appendices D, P), not claims.

2 Gauge non-identifiability of tangent diagnostics

Setup. Logits $z_\theta(x) \in \mathbb{R}^C$ (for sequence and dense models, after a fixed output reduction), $p = \mathrm{softmax}(z)$, one-hot targets y, cross-entropy $\mathcal{L}$, and logit gradient $g = p - y$, so $\mathbf{1}_C^\top g = 0$. Stacking m probe inputs, $J \in \mathbb{R}^{mC \times P}$ is the logit Jacobian and $\Theta = JJ^\top$ the matrix-valued NTK; the *trace NTK* is $K^{\mathrm{tr}}_{ij} = \sum_c \langle \nabla_\theta z_{ic}, \nabla_\theta z_{jc} \rangle$. Write $Q_C = I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top$ (class quotient, on the logit axis) and $H_m = I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$ (sample centering, on the example axis); they act on different tensor factors and are not interchangeable.

Which "effective rank". Throughout, for PSD K we use the *spectral participation ratio* $d_{\mathrm{PR}}(K) = (\mathrm{Tr}\, K)^2 / \mathrm{Tr}(K^2)$, the Rényi-2 analogue of the entropy-based effective rank of Roy

& Vetterli (2007) and distinct from the regularized effective dimension $\mathrm{Tr}[K(K+\lambda I)^{-1}]$; every number in this paper is d_{PR}, and "effective rank" in a table header abbreviates it.

Theorem 1 (Softmax gauge attack)**.** *Let $h_\phi(x) \in \mathbb{R}$ be any scalar network with parameters ϕ, and set $z_{\theta,\phi}(x) = z_\theta(x) + \gamma\, h_\phi(x)\mathbf{1}_C$. Then for all γ: (a)* $\mathrm{softmax}(z_{\theta,\phi}) = \mathrm{softmax}(z_\theta)$*, so predictions, CE loss and accuracy are identical; (b) the CE gradient on ϕ is $\propto \mathbf{1}_C^\top(p-y) = 0$ and on θ is unchanged, so the training trajectory is identical; yet (c)*

$$K^{\mathrm{tr}}_{\theta,\phi}(x,x') = K^{\mathrm{tr}}_\theta(x,x') + \gamma^2 C\, k_h(x,x'), \quad k_h(x,x') = \langle \nabla_\phi h_\phi(x), \nabla_\phi h_\phi(x')\rangle.$$

With h_ϕ constant, k_h is rank-one and $d_{\mathrm{PR}}(K^{\mathrm{tr}}) \to 1$ as $\gamma \to \infty$. Arbitrary contamination, at its exact scope: *on the finite probe set, a linear feature gauge $h_\phi(x) = \psi(x)^\top\phi$ gives $k_h = \Psi\Psi^\top$, and every PSD $G \in \mathbb{R}^{m\times m}$ factors that way once the feature dimension is at least* $\mathrm{rank}\, G \le m$*. Hence on any finite probe set the trace NTK can be moved to $K^{\mathrm{tr}} + \gamma^2 C\, G$ for* any *PSD G: trace-NTK spectral summaries are not functions of the predictive model.*

Corollary 1 (Same-architecture, gradient-preserving form)**.** *The gauge needs no extra branch. Let the head be $z = Wh_\theta(x) + b$ and set $W' = W + \gamma\mathbf{1}_C a^\top$. Then (i) $z' = z + \gamma(a^\top h_\theta(x))\mathbf{1}_C$, an* input-dependent *gauge, so predictions, CE loss and accuracy are identical; (ii) every gradient in the network is unchanged, since $\partial\mathcal{L}/\partial W = gh^\top$ and $\partial\mathcal{L}/\partial b = g$ do not involve W, while $\partial\mathcal{L}/\partial h = W'^\top g = W^\top g + \gamma a(\mathbf{1}_C^\top g) = W^\top g$ because $\mathbf{1}_C^\top g = 0$. Any optimizer whose update is a function of the gradient history alone—gradient descent, momentum, Adam without weight decay—therefore produces trajectories differing by the constant offset $\gamma\mathbf{1}_C a^\top$ at every step, with identical predictions and loss throughout; (iii) the model has the* same architecture and the same parameter count. Caveat: *an update that also reads parameter values sees $W' \neq W$. Isotropic L_2 decay with a linear update rule (SGD, momentum, decoupled decay) only shrinks the offset, which stays class-common, so predictions and cross-entropy remain identical while the trajectory differs; Adam with a coupled penalty, or a non-isotropic penalty, breaks the predictive equivalence (Appendix C).*

What the in-place edit does to the kernel. An independent branch adds a block orthogonal to everything else, so K^{tr} moves by exactly $\gamma^2 C k_h$; the in-place edit shares the body parameters θ, so a cross term appears. With $\Phi_{i,:} = \nabla_\theta\big(a^\top h_\theta(x_i)\big)^\top$ the tangent features of the injected shift and $S_{i,:} = \sum_c \nabla_\theta z_{ic}^\top$ the class-summed body Jacobian of the clean network (head blocks are unchanged),

$$K^{\mathrm{tr}}_\gamma = K^{\mathrm{tr}}_0 + \gamma\,(S\Phi^\top + \Phi S^\top) + C\gamma^2\,\Phi\Phi^\top. \tag{1}$$

A quadratic in γ with PSD leading term vanishes at no more than one non-zero γ when $\Phi \neq 0$, but it *can* vanish there (a two-class example with $K_\gamma = K_0$ despite $H_m\Phi \neq 0$ is in Appendix C), so the kernel-level claim H2 of §3 is stated for the sweep, not for one γ. Q_C annihilates every term on the logit axis, so the quotient kernel below is invariant at every γ under both forms. The two are implementations of one symmetry: (a) bounds *how much* contamination is reachable, (b) closes the escape route "you changed the model". Every full-space audit below uses (b); Appendix E also reports (a).

Theorem 2 (Quotient NTK is gauge-invariant and drives the projected dynamics)**.** *Define $\Theta_Q = (I_m \otimes Q_C)\,\Theta\,(I_m \otimes Q_C)$. Since $Q_C\mathbf{1}_C = 0$, the gauge of Thm. 1 and Cor. 1 leaves Θ_Q exactly invariant. As $Q_C g = g$, gradient flow $\dot z = -\Theta g$ gives*

$$\tfrac{d}{dt}(Q_C z) = -\Theta_Q\, g, \qquad \dot{\mathcal{L}} = -g^\top\Theta_Q g \le 0.$$

The dynamical half is not ours: with $P = Q_C$ it is the observation of Plenk et al. (2026, §3, App. E.3) that for the standard softmax the gradient flow of the projected network Pf is driven by $P\Theta P$; it is the *invariance* half that the audits certify. They audit the class trace $K^Q_{ij} = \sum_c (\Theta_Q)_{ic,jc}$, an $m\times m$ kernel that inherits the exact invariance of Θ_Q but not its full dynamics—no scalar summary of K^Q determines a trajectory. Both orders are confirmed on a finite-step sweep in Appendix G.

Proposition 1 (Exact criterion for sample centering)**.** *Stack the gauge* tangent *features $(\Phi_h)_{i,:} = \nabla_\phi h_\phi(x_i)^\top$, so that $k_h = \Phi_h\Phi_h^\top$. Sample centering acts on the example axis, $K \mapsto H_m K H_m$, and*

$$H_m k_h H_m = (H_m\Phi_h)(H_m\Phi_h)^\top, \qquad \textit{hence} \qquad H_m k_h H_m = 0 \iff H_m\Phi_h = 0.$$

Sample centering therefore removes exactly those gauges whose tangent features *are constant across the probe set; input dependence of h_ϕ is neither necessary nor sufficient ($h_\phi(x) = \phi + r(x)$ varies*

with x yet has $\nabla_\phi h_\phi \equiv 1$ and is removed). For the in-place gauge of Cor. 1 the statement is quantified over the sweep: by (1)*, $H_m K_\gamma^{\mathrm{tr}} H_m$ is unchanged for* every *γ iff $H_m\Phi = 0$ (the γ^2 coefficient $C(H_m\Phi)(H_m\Phi)^\top$ must vanish, and then so does the centered cross term), while for a single γ it can be unchanged although $H_m\Phi \neq 0$. By contrast $Q_C\mathbf{1}_C = 0$ removes every class-common gauge whatever h_ϕ is, and Q_C is the maximal projector on the logit axis that does so.*

Two distinct invariants, and a distinct artifact. The Fisher/GGN sample kernel (Martens, 2020) also annihilates the softmax null direction but is probability-weighted: proportional to Θ_Q at uniform predictions, otherwise a different gauge-invariant diagnostic (Appendix F). Mean pooling with an output bias injects a rank-one *sample-axis* mode that Q_C does not remove and H_m does (Appendix J). Proofs are in Appendix C.

3 Audit protocol and falsifiable predictions

We fix intervention, measurements and predictions before any audit, so that §4–§6 read as tests, not descriptions.

Intervention. The same-architecture head gauge of Cor. 1: for a linear head $W \in \mathbb{R}^{C\times d}$, $W \mapsto W + \gamma\,\mathbf{1}_C a^\top$ with $a \in \mathbb{R}^d$ a random unit direction; for a dense (1×1 convolutional) head the same edit with $a \in \mathbb{R}^{d\times k\times k}$ broadcast over the output channels. No branch is added; no parameter count changes. The strength is reported not as γ—meaningless across models—but as the scale-free *gauge ratio* $r_\gamma = \mathrm{Std}_x[\gamma\, a^\top h_\theta(x)] \,/\, \mathrm{Std}_{x,c}[z_c(x)] \in \{0, 0.1, 0.3, 1, 3, 10\}$, the injected shift in units of the model's own logit scale (each section states the subset it ran).

Diagnostics. One suite of five, defined once (Appendix A, with the conventions for a vanishing centered kernel and for ties): the spectral participation ratio d_{PR}, the ridge-regularized log condition number and log determinant (both at $\lambda = 10^{-6}\lambda_{\max}$, eigenvalues clipped at 0), the top-eigenvalue energy fraction, and the kernel–target alignment (both arguments sample-centered before the inner product). Trace is recorded but never ranked on. Each is computed on up to three kernels—*raw*, *sample-centered* H_mKH_m, *quotient*—and drift is $|D(K_\gamma) - D(K_0)|/(|D(K_0)| + \varepsilon)$, $\varepsilon = 10^{-12}$. One degeneracy matters below: $H_mKH_m\mathbf{1}_m = 0$, so $\lambda_{\min} = 0$ exactly and $\log\kappa_\lambda(H_mKH_m) = \log(1{+}10^6)$ for every non-zero K—the centered log condition number is a constant, and its $\approx 10^{-12}$ "drift" in §5–§6.2 is rounding around that constant, not a repair. Table 1 (Appendix A) lists which diagnostics were computed on which kernels in each audit.

Estimator. We never materialize the $mC \times P$ Jacobian: with Rademacher v_s, forward-mode JVPs (Baydin et al., 2022) give $\widehat{K}_{ij} = \frac{1}{S}\sum_s \langle J_i v_s, J_j v_s\rangle$, the quotient version applying Q_C to each $J_i v_s$ first. Base and gauged models are driven with the *same* directions, so $Q_C J_\gamma v_s = Q_C J_0 v_s$ holds *per draw* and quotient invariance is asserted exactly rather than statistically (Appendix H).

Pre-specified predictions. Let D be any diagnostic, K_γ the raw kernel at gauge ratio r_γ, and $\mathrm{Drift}_D(\gamma) = |D(K_\gamma) - D(K_0)|/(|D(K_0)| + \varepsilon)$.

- **H1** *Behaviour preservation.* $p_\gamma = p_0$ and $\mathcal{L}_\gamma = \mathcal{L}_0$ up to arithmetic noise, and every body gradient is unchanged. What is *measured* differs (Appendix B): the full-space audits record $\max|\Delta p|$ only; the ViT sweep also records the loss and every body gradient; Appendix E records predictions and the hidden gradient. Where a quantity is not measured, its preservation is the theorem's.
- **H2** *Raw non-identifiability.* The raw kernel moves by the block of (1), non-zero for all but at most one non-zero γ whenever $\Phi \neq 0$. Whether a scalar summary D moves, and whether monotonically in r_γ, is not implied by that and is an *empirical* endpoint per diagnostic.
- **H3** *Quotient invariance.* $\mathrm{Drift}_D(\gamma) = 0$ for the quotient kernel, to arithmetic precision.
- **H4** *Centering criterion.* Sample centering removes the injected block at every gauge strength iff its tangent features are sample-constant, $H_m\Phi = 0$ (Prop. 1); for a general gauge direction it does not. Its effect on a scalar diagnostic is not predetermined—unchanged, amplified, or (for one diagnostic) replaced by a constant—and we measure it per proxy.

H5 *Boundary / negative control.* The construction holds *only* where the head's output is consumed by a softmax; on a regression head the identical edit must *violate* H1 and move the outputs (§5). Behaviour preservation there would expose our invariance measurements as a numerical artifact rather than the softmax symmetry. It would not falsify Thm. 1, whose hypotheses exclude such heads—which is why the control is needed: it tests that the theorem's assumptions are doing work, and blocks the reading that class quotienting is a universal remedy for tangent-kernel diagnostics.

Arithmetic precision is part of the claim. The gauge writes entries of order $\gamma|a_j|$ into a head whose own weights are $O(1/\sqrt{d})$, so every mantissa bit the matmul gives up is taken out of the signal under test. Predicted error in $Q_C z$ at ViT scale: 10^{-15} (fp64), 6×10^{-7} (fp32), 5×10^{-3} (TF32), 4×10^{-2} (bf16). On a stock NVIDIA NGC container our first run measured 3.9×10^{-2} on an exact identity, because the image enables `allow_tf32`, reduced-precision bf16/fp16 reductions and `fp32_precision=tf32` by default. We force IEEE fp32 through every reachable flag, record the flags as found and as used, and the per-direction residual then tracks the mantissa (§6.1). "We computed it in fp32" is not a complete description of an experiment.

Reproducibility record. Per-run records and per-experiment settings (Appendix B) are described in the reproducibility statement; the anonymised code is at `https://anonymous.4open.science/r/gauge-audit-005B/`.

4 Full-space audit: all 15,625 NAS-Bench-201 cells

We evaluate *all* 15,625 *NAS-Bench-201 (Dong & Yang, 2020) cells, instantiated in the official benchmark macro-skeleton* ($C = 16$ initial channels, $N = 5$ cells per stage), with real-image probes and the benchmark's own CIFAR-10 and CIFAR-100 test accuracies. Enumerating the space removes architecture-sampling error entirely; what remains is seed variation, addressed below. We enumerate, and audit a second space in full (§5), because benchmark-suite studies report that conclusions about zero-cost proxies transfer poorly between tabulated spaces (Mehta et al., 2022)—across NAS-Bench-101 (Ying et al., 2019), NAS-Bench-201 and its NATS-Bench extension (Dong et al., 2022).

What was run. 31,250 records (15,625 architectures $\times$ two datasets) over 64 shards, *zero* failed, median 9.0 s per record, 0.058 GB peak: 78.2 GPU-hours; $m = 32$ real CIFAR probe images and $S = 32$ sketch directions per record. Kernels accumulate in float64 from a float32 forward pass with the reduced-precision flags forced off. The grid is the raw $\gamma \in \{0, 0.01, 0.1, 1, 10, 100\}$ rather than r_γ, since one fixed macro-skeleton makes γ comparable across the space. The sample-centered kernel was not computed here; that evidence is in §5–§6.2.

Primary endpoints. The claim here is about diagnostics, not proxy quality, so the primary endpoints do not depend on the proxy predicting anything: the distribution of $\mathrm{Drift}_D(\gamma)$ over the full space, the movement of the induced ordering, and the exactness of quotient invariance. Accuracy is *secondary*, deliberately: on our probes these proxies are weakly and sometimes negatively correlated with benchmark accuracy (τ from -0.196 to $+0.203$), a fact about training-free proxies rather than about the intervention.

H1: the intervention is a no-op. The median prediction change over the whole space is $\approx 10^{-9}$ and no architecture exceeds 10^{-3} at $\gamma \leq 1$. At $\gamma = 10$ and 100 a few near-degenerate cells do exceed it—the float32 cancellation §3 predicts, not a failure of Thm. 1 (Appendix K). We therefore report drift and shortlist survival at $\gamma = 1$, inside the certified band, where the damage is already large (top-10 overlap 0.2 for three of four proxies, 0 for two at $\gamma \geq 10$, Fig. 2(b)); only the selection-attack columns of Table 3 draw on the whole grid, and say so.

H2 and H3, over all 15,625 cells. Fig. 2(a) is the whole result in one picture: six to seven orders of magnitude between the two families of curves, and no overlap outside one exact exception. Raw diagnostics move by a median of 2–28% with a heavy tail; quotient diagnostics by 10^{-10}–10^{-8}, never exceeding 6.2×10^{-5} over the whole space, the five proxies recorded and every gauge strength inside the certified band. The exception is 341 cells (2.2%) whose raw drift is exactly zero:

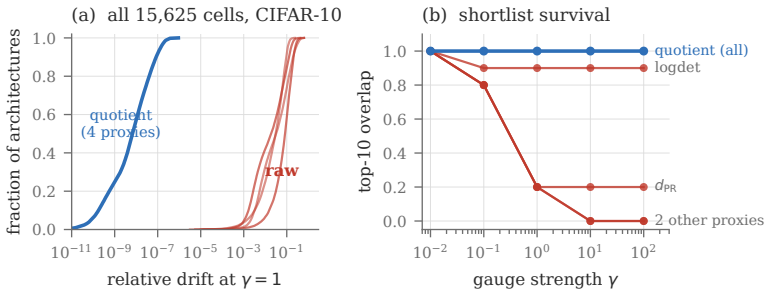


Figure 2: **All** 15,625 **NAS-Bench-201 cells, CIFAR-10.** (a) Cumulative distribution of per-architecture relative drift at $\gamma = 1$: four red curves, one per raw proxy (d_{PR}, log condition number, alignment, top-eigenvalue fraction), and one blue curve pooling the quotient drifts of the same four proxies (architecture×proxy mixture), six to seven orders of magnitude below. The 341 disconnected cells, whose raw drift is exactly zero, cannot appear on a log axis. (b) Survival of the clean top-10 shortlist under the raw kernel against gauge strength; log condition number and alignment coincide and reach 0 at $\gamma \geq 10$, the quotient shortlist is unchanged at every γ, and logdet is the one raw proxy whose shortlist largely survives (Table 3 shows logdet in place of the top-eigenvalue fraction).

the disconnected ones—every path severed by `none`—whose kernel is already rank one, at 10.0% chance accuracy. Appendix L gives the distribution numerically, per proxy.

The ordering, not just the number. At $\gamma = 1$ the top-10 shortlist keeps 2 of 10 members for d_{PR}, log condition number and alignment on CIFAR-10; the quotient keeps all 10 (top-50 and top-100 overlaps 1.00 too), with Kendall τ against accuracy unchanged to four decimals in all 8 (dataset, proxy) cells—residual τ changes of order 10^{-8} are sketch rounding.

Selection impact, and where the attack does *not* work. The attacker of Table 3 is *targeted*: per candidate it picks the γ from the whole grid $\{0.01, \ldots, 100\}$ that flatters that candidate most, gauging either the candidate alone (*promote-only*) or every candidate (*full-control*); definitions, tie rule and regret convention are in Appendix L. The picture is proxy-dependent, and we report the negative half. For d_{PR} an adversary can force any of 23 architectures on CIFAR-10, the worst sitting 52.9 accuracy points below the benchmark's best against 18.3 for the clean winner: a 34.6-point gain from an edit that changes no prediction in exact arithmetic. For logdet the gain is 6.7 points over 148 architectures. For log condition number and alignment the promote-only attack fails outright—one promotable architecture, 0.00 gain—though their shortlists are still destroyed; under full control log condition number becomes promotable for 9,695 of 15,625 cells. Since the attacker may use $\gamma \in \{10, 100\}$, where a few cells leave the float32 certificate (Appendix K), these counts are those of an attacker not bound by H1; we did not recompute them under a $\gamma \leq 1$ restriction (§8). *The gauge always moves the number; whether it moves the winner depends on the proxy, so both should be reported.* The quotient gives one promotable architecture and zero gain throughout, for any choice of grid.

Seeds. Full enumeration removes architecture-sampling error but not variation in initialization seed, probe batch, gauge direction a and sketch seed. The full space was run at one of each; only ViT-L/16 repeats the sweep (Appendix I, three probe subsets × three gauge directions), so seed sensitivity of the NAS ranking results is not quantified here (§8).

5 Full TransNAS-Bench-101 micro audit, and a falsification test

NAS-Bench-201 answers whether a real training-free selection benchmark can be corrupted. TransNAS-Bench-101 (Duan et al., 2021), whose seven tasks are drawn from the Taskonomy suite (Zamir et al., 2018), answers a different question: whether the prediction generalizes to another cell space and to several tasks, and where it stops. We enumerate the *full micro (cell) search space*, all

$4^6 = 4{,}096$ cells, parsed from the benchmark's own encoding. This is a controlled scope test, not a re-evaluation of the authors' trained networks: TransNAS gives each task its own decoder and loss, whereas we reconstruct the cell as a common backbone and attach controlled heads so that only the head differs across tasks—which makes the boundary control interpretable and is why we report no accuracy correlation. We do not use the macro space.

What was run, and what was excluded. Four tasks $\times$ 3,927 cells in float64, 2.4 s per task–architecture record, 10.5 GPU-hours; $m = 16$ Gaussian-noise probes, $S = 32$, dense heads flattened to one kernel row per pixel (Appendix B). The 3,927 is not 4,096: in 169 cells the head input barely varies along a ($\mathrm{Std}_x[a^\top h] \le 10^{-4}\,\mathrm{Std}[h]$), so the γ realising the requested ratio lies outside the numerical range we allow and the calibration skips the cell—identically for all four tasks, since degeneracy is a property of the cell and of a, not of the head. A further $\approx$ 1,520 cells per task (39%) hit the γ cap, so their achieved ratio is below the nominal $r_\gamma = 10$; their drifts are reported *at the achieved strength* and, since a scalar summary need not be monotone in γ (H2), are not lower bounds on the $r_\gamma = 10$ drift (Appendix M). An earlier float32 pass without this guard let γ reach $\sim 10^{13}$ and reported a quotient drift of 0.226 on semantic segmentation—a false counterexample to Thm. 2 produced by cancellation alone, and the reason §3 treats arithmetic as part of the claim.

Falsification test (H5), run first. Three tasks—room layout (9-dim regression), surface normals and autoencoding (dense regression)—have no softmax. The identical edit is applied to all seven tasks with backbone, search space and code path held fixed; only the head differs.

Proposition 2 (The gauge exists only where a softmax does)**.** *The edit $W \mapsto W + \gamma \mathbf{1}_C a^\top$ (broadcast over output channels for a dense head) is a* gauge *only because* $\mathrm{softmax}(z + c\mathbf{1}_C) = \mathrm{softmax}(z)$*; on a head whose output is consumed directly it shifts every output coordinate by $\gamma\langle a, h\rangle$ and changes the function. On a benchmark that varies only the head, the same edit must therefore preserve the function on softmax heads and move it on regression heads.*

The prediction is confirmed, and the cleanest statement is that *the logits move in every case*: at $r_\gamma = 10$ the raw head output shifts by 12.8–115.4$\times$ its own scale on the four softmax tasks and by 5.19–111.0$\times$ on the three regression tasks, "shift" being the relative Frobenius norm $\|z_\gamma - z_0\|_F / \|z_0\|_F$ over the probe batch (not the standard-deviation ratio r_γ). Only what happens next separates them—where a softmax follows, predictions move by 1.3×10^{-17} to 2.5×10^{-16} in absolute probability; where none does, the output *is* the shifted quantity. One edit, one magnitude of logit movement, a sixteen-order-of-magnitude split decided by the head alone. Had room layout held still, our attribution of the invariance to the softmax symmetry would have failed.

Within-scope audit, and the sharpest evidence for Prop. 1. Fig. 3(a) puts the question directly: each point is one (task, proxy) pair plotted as centered drift over raw drift, so the dashed line is "no better than doing nothing". For d_{PR} and the top-eigenvalue fraction every task sits above it; logdet straddles it (0.39–2.03); and the log condition number sits ten orders of magnitude below it—not repaired, but constant on centered kernels (§3).

Quotient invariance holds at the float64 floor everywhere: median 2.2×10^{-16} to 9.2×10^{-14}, maximum 1.5×10^{-10} over all 15,708 (task, architecture) pairs, four proxies and three gauge strengths (four, not five: alignment needs labels this kernel-only audit did not load). The raw kernel moves as expected. The result worth stating plainly is the middle column. *Sample centering is not merely insufficient here; for two of the four proxies it is far worse than doing nothing*—up to 17$\times$ the raw drift for the participation ratio and up to 76$\times$ for the top-eigenvalue fraction on the three image-classification tasks (3.1$\times$ and 4.8$\times$ on segmentation, where the raw drift is already an order of magnitude larger). For the log determinant it is a wash, and for the log condition number the $\approx 10^{-12}$ centered drift is the degeneracy of §3, not a repair. This is H4 as stated: Prop. 1 says when the injected block survives centering, not what a scalar summary of it will do, and one sweep shows amplification, no change and collapse to a constant. A practitioner reaching for $H_m K H_m$ cannot tell from the number alone which case they are in. Per task (Appendix M), the raw d_{PR} drift is flat across a 20$\times$ range in C (0.020–0.021 for 47-, 75- and 1000-way heads—a larger label set is no defence) and larger only for the dense segmentation head (0.160), where the shift enters once per pixel.

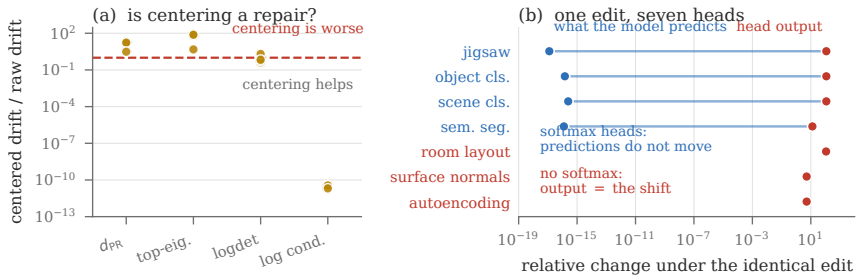


Figure 3: **TransNAS-Bench-101 micro space**, 3,927 cells per task, float64, $r_\gamma = 10$ (achieved strength where the cap binds). (a) Centering is not a repair: above the line it leaves *more* drift than no correction at all; which side a proxy lands on is Prop. 1 seen from outside, and the log condition number's position far below the line is the degeneracy of §3. (b) Boundary control: the identical edit moves the head's own output by 5–115× (relative Frobenius norm) on all seven tasks, but where a softmax follows, what the model predicts does not move ($\leq 2.5 \times 10^{-16}$ absolute); where none does, the output *is* the shift. All other changes are relative.

6 Scaling and Domain Stress Tests

6.1 ViT-L/16: The Phenomenon at Pretrained-Model Scale

The two audits establish scale in *architecture* space; this one, in *parameter* space, with no ranking claim attached. We take an ImageNet-pretrained ViT-L/16 (Dosovitskiy et al., 2021) (3.04×10^8 parameters, `torchvision IMAGENET1K_V1`), probe it with ImageNet validation images, and measure the full-model tangent under the same head gauge with a JVP sketch; no retraining is involved, and ViT-L/16 carries the gauge because its head is a plain *untied* linear layer.

Top-1 agreement is exact at every gauge strength and $\max|\Delta p| \leq 1.6 \times 10^{-6}$ at $r_\gamma = 10$. Over that sweep d_{PR} falls $20.486 \to 13.555$ on the raw kernel and $33.522 \to 18.058$ on the sample-centered one—centering again moving further than no correction—while the quotient stays at 20.490 to three decimals (relative Frobenius drift 2.0×10^2 raw, 9.7×10^{-8} quotient). At this scale the arithmetic, not the theorem, sets the floor: the per-direction residual $\|Q_C J_\gamma v - Q_C J_0 v\| / \|Q_C J_0 v\|$ is 3.21×10^{-3} under the container's default reduced-precision matmul, 3.54×10^{-6} under enforced IEEE fp32 and 1.00×10^{-14} in float64. Nine runs over three probe subsets and three gauge directions are in Appendix I.

6.2 Graph Neural Architectures: A Cross-Domain Stress Test

Not a benchmark audit—no accuracy metadata is available here—but a check that the phenomenon belongs to the softmax rather than to vision models. On the NAS-Bench-Graph space (Qin et al., 2022) over Cora, CiteSeer and PubMed (Kipf & Welling, 2017), 200 architectures each, Fig. 3's pattern reproduces: quotient drift at fp32 rounding, centering worse than raw for d_{PR} (1.3–2.5×) and the top-eigenvalue fraction (3.4–5.7×), and the centered log condition number again pinned at its constant ($\approx 10^{-12}$ drift, the degeneracy of §3). Node classification couples the examples, so the forward must see the whole graph and probe nodes be selected *after* it. One case is specific here: centering leaves alignment drift *exactly* unchanged (ratio 1.00 on all three graphs), because that alignment is already sample-centered, so H_m is idempotent on it (Appendix N). One domain thus shows drift unchanged, drift amplified and a diagnostic collapsed to a constant: *centering is not a repair*.

7 Related Work

NTK theory for classification. Plenk et al. (2026) extend NTK theory to cross-entropy, identifying when wide networks remain in the lazy regime. They treat the softmax redundancy

$\mathrm{softmax}(z) = \mathrm{softmax}(z+\alpha\mathbf{1})$ in two ways: with a *reference class* (one logit fixed, their App. C.3) and, for the standard head, by *class centering*—their $P = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ is our Q_C, applied to the logits (their eqs. 17–18), and their App. E.3 proves lazy training for the projected network Pf, whose gradient flow is driven by $P\Theta P$. The projected kernel and dynamics of Thm. 2 are therefore theirs. Our question is different: they ask when the NTK is constant along training; we ask whether the *spectral summaries* of the raw trace NTK, which practitioners report from the standard head, are functions of the predictive model at all. Our delta: arbitrary PSD contamination (Thm. 1); a same-architecture, gradient-preserving edit with its kernel formula (Cor. 1); the exact criterion for when centering fails (Prop. 1); full-space audits with a selection attacker; a boundary control. We claim neither class centering nor the projected dynamics as novel; we claim what they are needed *for*. Dinh et al. (2017) showed sharpness is not reparameterization invariant; Appendix D gives the tangent kernel's version, including that Euclidean orbit projection is *vacuous* for any exact symmetry.

Training-free NAS. NTK-based and zero-cost proxies (Chen et al., 2021; Mellor et al., 2021; Abdelfattah et al., 2021) and studies of their reliability (White et al., 2021; Mehta et al., 2022) ask whether such proxies work; Mok et al. (2022) argue the NTK-based ones track accuracy only under conditions easy to violate. Our contribution is orthogonal: even a proxy that works can be *driven* by an edit that changes nothing observable—a benchmark-integrity claim, not a proxy-quality one—and the same over-parameterized softmax head ends the architectures in language search spaces (Klyuchnikov et al., 2022).

Saliency proxies. SNIP (Lee et al., 2019b), GraSP (Wang et al., 2020) and SynFlow (Tanaka et al., 2020) are sometimes assumed gauge-safe because they are built from the loss gradient. Gradient preservation decides neither way; the aggregation does: the in-place edit changes the aggregated SNIP score (an absolute value) and the SynFlow score (built from $|\theta|$, no softmax), but leaves the signed-sum GraSP score of the reference zero-cost implementation *exactly* unchanged, since a fixed gauge direction lies in the null space of the cross-entropy Hessian (Appendix O).

8 Limitations

The invariance is exact only for the additive softmax gauge under cross-entropy; §5 measures rather than assumes that boundary, and under weight decay Cor. 1(ii) weakens to the case analysis given there. At ViT scale quotient invariance is reported at fp32 rounding, not as an exact zero. The selection attacker of Table 3 draws on the whole γ grid, including strengths at which a few cells leave the float32 H1 certificate, and we did not recompute it under a $\gamma \leq 1$ restriction. NAS-Bench-201 and the graph audit were run at one seed, and 39% of TransNAS cells hit the γ cap, so their drifts are at a lower achieved strength and bound the nominal one in neither direction; a multi-seed subset and a capped/uncapped split are the missing controls. The NAS audits measure prediction preservation only; loss and body-gradient preservation are measured on ViT-L/16 and the small models. The clean proxies are weakly predictive of accuracy on our probes—hence accuracy as a secondary endpoint—and the graph experiments carry no accuracy metadata. The TransNAS backbone is a reconstruction from the published encoding, faithful in the cell but not the authors' trained network, probed with Gaussian noise. We analyze first-order dynamics only (Appendix G), and the quotient's coarse spectral summaries do not predict length generalization (Appendix Q).

9 Conclusion

For softmax models the trace NTK partly measures a quantity with no functional meaning; the class quotient removes it exactly. Across two enumerated NAS spaces, a 3×10^8-parameter transformer and a graph domain, a behaviour-preserving edit moves raw diagnostics while quotient diagnostics do not, and sample centering is not a repair. Auditing a diagnostic's drift under a random gauge costs one more evaluation.

AI USE STATEMENT

Generative AI tools were used in three ways: to aid and polish writing; for retrieval and discovery when locating related work; and for coding assistance, which was substantial—the released `gaugeaudit` package, the experiment entry points and the cluster job templates were all written with AI assistance. All formal claims, proofs, implementations, citations and reported measurements were checked and approved by the authors. No reported measurement was generated by an AI system—every number comes from that code run on the hardware described in §3, and the released repository is exactly what produced them—and every citation was verified against its official venue, publisher, or arXiv record, as applicable.

ETHICS STATEMENT

This paper describes a construction that can be used to move a training-free architecture-selection proxy anywhere in the range reachable by adding a PSD block to its kernel, while leaving the model's behaviour unchanged to the arithmetic floor. That is a benchmark-integrity concern with an obvious dual use, and we have weighed publishing it. Three considerations decided it. The construction is a two-line consequence of a symmetry that has been in the open literature for decades, so withholding it protects no one for long. The defence is published with the attack, is exact rather than heuristic, and costs one projection to apply. And the audit that detects the problem—measuring a diagnostic's drift under a random function-preserving gauge—costs one further evaluation of the diagnostic, so a benchmark maintainer can adopt it immediately. We use no human subjects and no personal data; all datasets and checkpoints are public research artifacts used under their own licences.

REPRODUCIBILITY STATEMENT

The proofs of Thm. 1, Cor. 1, Thm. 2 and Prop. 1 are in Appendix C, with all assumptions stated in §2 and the scope stated once, and kept, in §1. The intervention, the diagnostics, the estimator and the predictions H1–H5 are fixed in §3 *before* any audit is reported, so the results can be read as tests. §3 also records the arithmetic-precision protocol, which is load-bearing: the invariance under test is exact, so an unnoticed TF32 or bf16 reduction is enough to make a correct implementation look like a violated theorem, and we report the flag set as found and as used for every run. Anonymized source code for every experiment in the paper, the cluster job templates, and a CPU-only one-minute reproduction of the central claim are available at `https://anonymous.4open.science/r/gauge-audit-005B/`. Every run records GPU-hours, peak memory, failed-run count, sketch count, the precision-flag set before and after forcing, the probe source (real images or noise—never silently mixed), whether attention ran fused, and a code revision hash. The per-experiment settings—probe count m, sketch count S, gauge grid, arithmetic, seeds, the γ cap and the degeneracy threshold—are tabulated in Appendix B. All appendices referred to in the text are provided as supplementary material.

REFERENCES

Mohamed S. Abdelfattah, Abhinav Mehrotra, Łukasz Dudziak, and Nicholas D. Lane. Zero-cost proxies for lightweight NAS. In *International Conference on Learning Representations (ICLR)*, 2021.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Atılım Güneş Baydin, Barak A. Pearlmutter, Don Syme, Frank Wood, and Philip Torr. Gradients without backpropagation. *arXiv preprint arXiv:2202.08587*, 2022.

Wuyang Chen, Xinyu Gong, and Zhangyang Wang. Neural architecture search on ImageNet in four GPU hours: A theoretically inspired perspective. In *International Conference on Learning Representations (ICLR)*, 2021.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, 2017.

Xuanyi Dong and Yi Yang. NAS-Bench-201: Extending the scope of reproducible neural architecture search. In *International Conference on Learning Representations (ICLR)*, 2020.

Xuanyi Dong, Lu Liu, Katarzyna Musial, and Bogdan Gabrys. NATS-Bench: Benchmarking NAS algorithms for architecture topology and size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3634–3646, 2022. doi: 10.1109/TPAMI.2021.3054824.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Yawen Duan, Xin Chen, Hang Xu, Zewei Chen, Xiaodan Liang, Tong Zhang, and Zhenguo Li. TransNAS-Bench-101: Improving transferability and generalizability of cross-task neural architecture search. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Nikita Klyuchnikov, Ilya Trofimov, Ekaterina Artemova, Mikhail Salnikov, Maxim Fedorov, Alexander Filippov, and Evgeny Burnaev. NAS-Bench-NLP: Neural architecture search benchmark for natural language processing. *IEEE Access*, 10:45736–45747, 2022. doi: 10.1109/ACCESS.2022.3169897.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. SNIP: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations (ICLR)*, 2019b.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Yash Mehta, Colin White, Arber Zela, Arjun Krishnakumar, Guri Zabergja, Shakiba Moradian, Mahmoud Safari, Kaicheng Yu, and Frank Hutter. NAS-Bench-Suite: NAS evaluation is (now) surprisingly easy. In *International Conference on Learning Representations (ICLR)*, 2022.

Joe Mellor, Jack Turner, Amos Storkey, and Elliot J. Crowley. Neural architecture search without training. In *International Conference on Machine Learning (ICML)*, 2021.

Jisoo Mok, Byunggook Na, Ji-Hoon Kim, Dongyoon Han, and Sungroh Yoon. Demystifying the neural tangent kernel from a practical perspective: Can it be trusted for neural architecture search without training? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations (ICLR)*, 2020.

Jonathan Plenk, Sergio Calvo-Ordoñez, Álvaro Cartea, Yarin Gal, Mark van der Wilk, and Kamil Ciosek. The neural tangent kernel for classification. *arXiv preprint arXiv:2605.17606*, 2026.

Yijian Qin, Ziwei Zhang, Xin Wang, Zeyang Zhang, and Wenwu Zhu. NAS-Bench-Graph: Benchmarking graph neural architecture search. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations (ICLR)*, 2020.

Colin White, Arber Zela, Binxin Ru, Yang Liu, and Frank Hutter. How powerful are performance predictors in neural architecture search? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Greg Yang. Tensor programs II: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, and Frank Hutter. NAS-Bench-101: Towards reproducible neural architecture search. In *International Conference on Machine Learning (ICML)*, 2019.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.