WHAT IS AVERAGING WORTH IN SELF-TRAINING? OVERLAPPING CORRELATION BOUNDS ACROSS STEPS, NETWORKS, AND VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most self-training methods use ensemble averaging across training steps (e.g., exponential moving average (EMA) teachers), networks, or input data transformation views to aid semi-supervised learning. However, little is known about how ensemble size and member weighting affect these averages and thus self-training efficacy. Here, we use a prediction error decomposition to show that variance reduction by training step, network, and view averaging overlaps and is bounded by measurable error correlation. We further identify the private and shared variance of each averaging type by inverting measured correlation, and show gains in effective sample size are sublinearly additive across averaging types. To connect variance reduction to increased task performance, we isolate the contributions of effective sample size gains from pseudo-label quality improvements. By decoupling these components, we show that averaging over larger ensembles approaches an upper bound in shared variance reduction and thus provides quickly diminishing task performance returns, suggesting that targeted reductions in private variance could more efficiently improve self-training. Together with extensive experimentation on Pascal VOC and Cityscapes, we apply our findings to optimize the ensemble averaging configurations of state-of-the-art self-training methods, which improves task performance on the ADE20K, COCO, Pancreas-CT, and Left Atrium datasets. This work establishes error correlation as a common budget that unifies step, network, and view averaging in self-training, and will guide the design of averaging configurations in future methods.

1 INTRODUCTION

Most self-training methods for semi-supervised learning build their teacher and their final prediction by averaging over an ensemble comprising either different training steps, networks, or input data views (i.e., data transformations or augmentations; Figure 1). For example, Mean Teacher keeps an exponential moving average (EMA) of student network weights (Tarvainen & Valpola, 2017), cross pseudo supervision averages predictions from multiple networks (Chen et al., 2021), and test-time augmentation approaches average class probabilities over transformed inputs (Wang et al., 2019; Shanmugam et al., 2021). Each of these ensemble average types is controlled by a key hyperparameter determining the ensemble sample size: the number of training steps (rate of EMA decay), networks, or views to average over, respectively. But while classical ensemble theory ties the effectiveness of averages to the correlation between ensemble member errors (Hastie et al., 2009; Wood et al., 2023), these hyperparameters do not account for correlations and instead simply fix the weighted count of averaged members (Kish, 1965). While some papers have acknowledged correlation effects empirically, for example that different checkpoints from a single training run make similar predictions (Fort et al., 2019) and that an EMA teacher stays close to its student (Ke et al., 2019), little is known about the correlation-dependent bounds of these three ensemble average types, whether each average type addresses the same error components, or the degree to which variance reduction contributes to the semi-supervised performance gains achieved by each type.

In this paper, we show that the three average types reduce variance in the same prediction error component in a manner bounded by a measurable error correlation ρ, such that the performance

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.

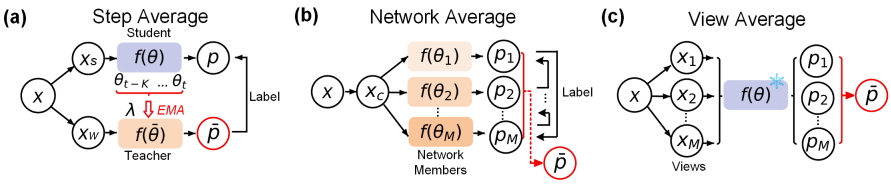


Figure 1: The ensemble average types studied in this paper. (a) Step averaging combines snapshots from a single training run, as in an EMA teacher (Tarvainen & Valpola, 2017). (b) Network averaging combines co-trained networks (Chen et al., 2021). (c) View averaging passes transformed copies of one input through a frozen network and combines their predictions (Shanmugam et al., 2021).

gains produced by each type overlap. By measuring ρ between prediction-error fields during training, we derive the ensemble effective size N_{eff}, a measure distinct from the weighted count M_{kish}, that equates to the (typically much smaller) number of ensemble members that would achieve the same variance reduction were they independent. However many members are averaged, the error variance of the average cannot fall below a fraction ρ of the error variance of a single member, and it approaches this fraction as members are added. We therefore refer to ρ as the lower bound on the variance ratio and to $1/\rho$ as the upper bound on N_{eff}. To compare the three average types, we then model the error field as a crossed random-effects decomposition over training steps, networks, and views (Theorem 4.5). In this decomposition, the lower bound of any combination of average types is the sum of the variance components that none of them removes, and the variance that is private to each type or shared between types is identified by inverting measured correlations. We find that the error components addressed by the three types overlap substantially: adding up the maximum N_{eff} gain that each type can achieve on its own overstates the maximum gain that all three can achieve together by 9–61% across datasets. Because step averaging removes the error fluctuations that are private to each network, two EMA teachers are more strongly correlated with each other than their two raw students are (Corollary 4.6).

To connect variance reduction to task performance, we separate two channels through which a teacher affects a self-training method. First, the teacher's average is the prediction that is read out, and its error variance decreases as N_{eff} increases; we call this the readout channel. Second, the teacher produces the pseudo-labels on which the student trains (Tarvainen & Valpola, 2017; Yang et al., 2025; Yi et al., 2026), and their accuracy affects the quality of the student itself; we call this the target channel. For the readout channel, we calibrate how much mIoU one unit of N_{eff} is worth on the step axis and find that this single readout rate predicts most of the gain of network and view averages without refitting. For the target channel, we replace a fraction of the teacher's accepted pseudo-labels with ground-truth labels while holding the confidence mask and the averaging fixed. This intervention raises mIoU without changing N_{eff}, which shows that gains through the target channel, unlike gains through the readout channel, are not limited by the upper bound that the correlation imposes.

Our main experiments use Pascal VOC 2012 (Everingham et al., 2010; Hariharan et al., 2011) and Cityscapes (Cordts et al., 2016), and experiments on CIFAR-100 (Krizhevsky et al., 2009) confirm that the network-axis bound also holds for semi-supervised classifiers. We then apply the decomposition to two state-of-the-art methods, UniMatch V2 (Yang et al., 2025) on ADE20K (Zhou et al., 2017) and MS COCO (Lin et al., 2014) and DualFete (Yi et al., 2026) on Pancreas-CT (Roth et al., 2015; 2016) and Left Atrium (Xiong et al., 2021), to identify the average types that still hold removable private variance. Expanding the averages along these types improves UniMatch V2 by 0.97–1.60 mIoU and DualFete by 0.68–0.79 Dice. Because these gains are capped by the correlation between members, we then add members of a second network architecture, which are less correlated with the existing members. This raises the upper bound on N_{eff} and increases the gains to 2.41 and 2.59 mIoU on ADE20K and MS COCO and to 0.99 and 0.85 Dice on Pancreas-CT and Left Atrium. In summary, our contributions are as follows:

- **A unified decomposition.** We show that step, network, and view averaging remove variance from one crossed decomposition of the error field, identify the private and shared variance of each average type from measured correlations, and validate the decomposition with a prediction of the correlation between EMA teachers (Sections 4 and 5.1).

- **Measured bounds.** We measure the effective sample size of each average type during training and show that it approaches the upper bound set by its error correlation, and that the EMA decay, the number of members, and the coupling weight change the gain only through this correlation (Section 5.1).
- **Two channels to performance.** We separate the readout and target channels of a teacher, show that a readout rate is shared by step, network and view averages, and show with an oracle intervention that better pseudo-labels raise performance beyond what averaging could still add below its upper bound (Sections 4.3 and 5.2).
- **Application to state-of-the-art methods.** We use the decomposition to locate the average types of UniMatch V2 and DualFete that still hold private variance, and we show that adding less correlated members amplifies the averaging gain by raising the upper bound on the effective sample size (Section 5.3).

2 Related Work

Averages in Self-Training. Temporal Ensembling (Laine & Aila, 2016) averages predictions across epochs and Mean Teacher (Tarvainen & Valpola, 2017) averages weights at every step. Consistency methods perturb inputs or features (Sajjadi et al., 2016; Miyato et al., 2018; Liu et al., 2022), cross pseudo supervision adds a second network (Chen et al., 2021), FixMatch and its successors refine target selection (Sohn et al., 2020; Wang et al., 2022b; Chen et al., 2023), and recent segmentation systems add stronger backbones and label propagation (Sun et al., 2024; Yang et al., 2025; Yi et al., 2026). Their weight average descends from iterate averaging (Polyak & Juditsky, 1992) through SWA (Izmailov et al., 2018) and snapshot ensembles (Huang et al., 2017), and recent work scales it with batch size (Busbridge et al., 2023), reconstructs it post hoc (Karras et al., 2024), asks when averaging helps (Ajroldi et al., 2025) and folds it into the optimizer (Defazio et al., 2024; Song et al., 2026), while model soups average across runs (Wortsman et al., 2022). None of this work measures the error component that an average cannot remove or how that component is shared between averages of different types.

Correlation, Diversity, and Variance Components. That ensemble error depends on member correlation is classical (Krogh & Vedelsby, 1994; Breiman, 2001; Hastie et al., 2009). Wood et al. (2023) unify bias, variance and diversity, Lobacheva et al. (2020) fit power laws in ensemble size, Abe et al. (2023) show that diversity need not justify its cost in member quality, and Jeffares et al. (2023) find that jointly trained members collude, as our coupled members do. For step averages, the corresponding calculation is the EWMA variance used in control charts (Schmid, 1997; Vermaat et al., 2008). Our multi-axis model builds on crossed random-effects decompositions (Searle et al., 2009), functional ANOVA (Hoeffding, 1992; Sobol', 1990), and the D-study composition rule of generalizability theory (Brennan, 1992). In machine learning, Bouthillier et al. (2021) isolate the variance contributed by individual sources of randomness, where we instead fit the full crossed decomposition, including interactions, to the error fields observed during self-training.

Teacher Quality and the Conversion to a Metric. Wei et al. (2020) and Frei et al. (2022) give conditions under which self-training improves on a pseudo-labeler of bounded error, the weak-to-strong literature quantifies a student's gain from an imperfect teacher (Burns et al., 2023; Charikar et al., 2024), and distillation studies find that students neither match their teacher nor need to (Stanton et al., 2021). Our target channel measures this gain per unit teacher accuracy inside a running system. What a variance reduction is worth on the metric depends on the loss and the aggregation rule (Ortega et al., 2022) and on the choice and weighting of test-time views (Lyzhov et al., 2020; Shanmugam et al., 2021). We follow Lyzhov et al. (2020) and Shanmugam et al. (2021) and use the IoU derivative that its surrogates descend (Rahman & Wang, 2016; Berman et al., 2018) to explain why target accuracy converts to mIoU at a dataset-specific rate.

3 Preliminaries

Let Π be a *probe set* of pixels; the pixelwise *error field* on Π is $e = 1 - p(\text{true class})$. It carries one index per aggregation axis, $e_{i,t,v} \in \mathbb{R}^{|\Pi|}$ for network member i, training step t and input view

v. A snapshot is the model state at a training step, and temporal lags are measured in training steps. Statistical moments are taken over the $|\Pi|$ probe units, with Var and Corr the empirical variance and Pearson correlation on Π (Appendix B.1). An *aggregation rule* on an axis is a weight vector w on that index, $w_k \geq 0$, $\sum_k w_k = 1$, mapping the field to $\bar{e}_w := \sum_k w_k e_{(k)}$ with the other indices held fixed. The reference variance $\sigma^2 := \mathrm{mean}_k \, \mathrm{Var}[e_{(k)}]$ is that of a single member. *Probability averaging* combines predicted class-probability vectors and thus averages their error fields exactly. *Weight averaging* combines model parameters; weight-space EMA is linked to error averaging through the linearized model in Assumption 4.1(ii).

Definition 3.1 (Weighted Count and Effective Sample Size)**.** For members $X_1, \ldots, X_M$ of a common target with weights w_k and average $\bar{X} = \sum_k w_k X_k$,

$$M_{\mathrm{kish}} := 1 \Big/ \sum_k w_k^2 \quad (1) \qquad N_{\mathrm{eff}} := \sigma^2 \big/ \, \mathrm{Var}[\bar{X}], \quad (2) \qquad V := 1/N_{\mathrm{eff}}, \quad (3)$$

where V is the variance ratio.

The two sizes coincide for uncorrelated members of equal variance, whereas with equal variances and non-negative mean correlation, $N_{\mathrm{eff}} \leq M_{\mathrm{kish}}$ (Lemma A.2). For members of one axis, ρ denotes their common pairwise error correlation (ρ_{network}, ρ_{view}). For one network, the training trajectory is the sequence of weights $\theta_1, \ldots, \theta_T$ that a single run visits from one random initialization, and a snapshot is its prediction at one step t. For snapshots of one training trajectory, $\rho(\Delta) := \mathrm{Corr}(e_t, e_{t+\Delta})$ denotes the temporal correlation curve, with limit $\rho_\infty := \lim_{\Delta \to \infty} \rho(\Delta)$. The single-axis model uses a per-training-step decay r. Across two different network members, $\rho_{\mathrm{cross}}(\Delta) := \mathrm{Corr}(e_{i,t}, e_{j,t+\Delta})$. We reserve w_k for aggregation weights and use w_{cpl} for the coupling weight and w_{scale} for the scale half-width.

4 Lower Bounds, Their Overlap, and the Conversion to Performance

Assumption 4.1 (Single-Axis Covariance)**.** (i) Network or view members have equal variance σ^2, a common pairwise correlation ρ, and equal weights. (ii) Motivated by a common first-order linearization in the parameters, we assume that the EMA teacher's error equals the w_k-weighted average of the snapshot errors for snapshots of one training trajectory. Snapshot errors are variance-stationary over the analysis window with $\rho(\Delta) = \rho_\infty + (1 - \rho_\infty) r^{|\Delta|}$, $\rho_\infty, r \in [0, 1)$.

Assumption 4.2 (Crossed Components)**.** Let $\mathcal{A}$ be the set of axes. The centered error field is a sum $e = \sum_{S \subseteq \mathcal{A}} c_S$ of mutually uncorrelated components, where c_S is constant along every axis outside S and has variance share $s_S \geq 0$, $\sum_S s_S = 1$. The correlation of c_S between units u and u' factorizes over the axes in S, $\mathrm{Corr}(c_S(u), c_S(u')) = \prod_{a \in S} \kappa_a(u_a, u'_a)$, with $\kappa_a = \mathbf{1}\{u_a = u'_a\}$ for a network or view index and $\kappa_{\mathrm{s}} = r_S^{|\Delta|}$, $r_S \in [0, 1)$, for two steps Δ apart.

Lemma A.1 establishes the mathematical validity of the temporal correlation function in Assumption 4.1(ii); its applicability to training trajectories is assessed empirically. Assumption 4.2 uses an ANOVA-style decomposition over the product of the axes (Hoeffding, 1992; Sobol', 1990; Owen, 2013), with mutually uncorrelated components modeled as crossed random effects (Searle et al., 2009). It identifies the component each averaging mechanism removes. For network and step, the components are $c_\varnothing$ (shared, constant), $c_{\{\mathrm{n}\}}$ (per member, constant), $c_{\{\mathrm{s}\}}$ (shared, step-varying) and $c_{\{\mathrm{n,s}\}}$ (per member and step), with shares (s_c, s_a, s_g, s_b); Lemma 4.3's $b + \eta_i$ is the one-axis case. We write r_g, r_b for the temporal rates and $B_g = B(\lambda, r_g)$, $B_b = B(\lambda, r_b)$. Since $N_{\mathrm{eff}} = \sigma^2 / \, \mathrm{Var}[\bar{e}]$ is computed from recorded fields (Algorithm S1), the assumptions do not affect the measurement and composition rules. We discuss deviation from the assumptions in Appendix A.1.

4.1 The Lower Bound on Each Axis

Lemma 4.3 (Equicorrelated Decomposition and Lower Bound, Hastie et al., 2009, §15.2)**.** *Under Assumption 4.1(i) with $\rho \geq 0$, the members are distributed, up to second moments, as a part shared by all members plus a part private to each, $X_i - \mathbb{E}[X_i] = b + \eta_i$ with $\mathrm{Var}[b] = \rho\sigma^2$, $\mathrm{Var}[\eta_i] = (1 - \rho)\sigma^2$, and the η_i mutually uncorrelated and uncorrelated with b* (4)*. Averaging acts on η and*

leaves b untouched:

$$\mathrm{Var}[\bar{X}] = \sigma^2\Big(\rho + \frac{1-\rho}{M}\Big), \qquad N_{\mathrm{eff}} = \frac{\sigma^2}{\mathrm{Var}[\bar{X}]} = \frac{M}{1+(M-1)\rho}. \tag{5}$$

For fixed $\rho > 0$, $N_{\mathrm{eff}} \to 1/\rho$ as $M \to \infty$, whereas $M_{\mathrm{kish}} = M$ grows without bound. At $\rho = 0$, $N_{\mathrm{eff}} = M$.

The $(M{+}1)$-th member reduces the variance ratio V by $(1-\rho)/(M(M+1))$, while lowering a positive ρ raises the upper bound itself (Equation (12)). The member in this lemma covers separately initialized networks and transformed copies of one input directly, with $\rho = \rho_{\mathrm{network}}$ or ρ_{view}. Snapshots under the EMA kernel $w_k = (1-\lambda)\lambda^k$ need a separate statement, since the weights are unequal, the correlation between snapshots depends on their lag, and λ sets both $M_{\mathrm{kish}} = (1+\lambda)/(1-\lambda)$ and the mean lag $\lambda/(1-\lambda)$ (Equation (14)).

Lemma 4.4 (Temporal Lower Bound of an EMA Teacher)**.** *Under Assumption 4.1(ii) with unit snapshot variance, the variance ratio V (Definition 3.1) of an EMA teacher of decay λ is*

$$V(\lambda) = \rho_\infty + (1-\rho_\infty)\, B(\lambda, r), \qquad B(\lambda, r) = \frac{1}{M_{\mathrm{kish}}} \cdot \frac{1+\lambda r}{1-\lambda r} = \frac{(1-\lambda)(1+\lambda r)}{(1+\lambda)(1-\lambda r)}, \tag{6}$$

where $B \in (0, 1]$ is the fraction of the decaying part's variance that the kernel leaves, so that $1/B$, the weighted count divided by a penalty for correlation between nearby snapshots, is the number of independent snapshots the kernel is worth. V is strictly decreasing on $[0, 1)$ and $N_{\mathrm{eff}} \to 1/\rho_\infty$ as $\lambda \to 1$, interpreted as infinity when $\rho_\infty = 0$. At $r = 0$, Equation (6) becomes Equation (5) with $(M_{\mathrm{kish}}, \rho_\infty)$ in place of (M, ρ) (Lemma A.2).

4.2 Overlapping Lower Bounds Across Axes

Theorem 4.5 (Multi-Axis Lower Bounds, Identification, and Composition)**.** *Under Assumption 4.2, let $\rho(D)$ be the correlation between two units that differ along the axes $D \subseteq \mathcal{A}$, at infinite step separation when step belongs to D.*

(i) (Lower bounds) $\rho(D) = \sum_{S \cap D = \varnothing} s_S$. *Every finite aggregate of the form in (iii), averaging only axes in D, satisfies $V \geq \rho(D)$ and, when $\rho(D) > 0$, $N_{\mathrm{eff}} \leq 1/\rho(D)$. The bound is approached as all averaged network/view sizes tend to infinity and, if the step axis is averaged, $\lambda \to 1$. A zero lower bound gives no finite upper bound.*

(ii) (Identification) *The $2^{|\mathcal{A}|} - 1$ nontrivial correlations, with $\rho(\varnothing) = 1$, determine the shares by Möbius inversion on the subset lattice, $s_S = \sum_{U \subseteq S} (-1)^{|S \setminus U|} \rho(\mathcal{A} \setminus U)$.*

(iii) (Composition) *For product weights, uniform network and view averages and an EMA of decay λ on the step axis, a finite aggregate has the following variance ratio, with $\phi_a(S) = 1$ on axes not averaged:*

$$V = \sum_{S \subseteq \mathcal{A}} s_S \prod_{a \in S} \phi_a(S), \qquad \phi_a(S) = \begin{cases} 1/M_a, & a \in \{\mathrm{n}, \mathrm{v}\}, \\ B(\lambda, r_S), & a = \mathrm{s}. \end{cases} \tag{7}$$

The theorem identifies the lower bound on the variance ratio for averaging axes D as $\rho(D)$ and the lower bound for averaging all axes as $\rho(\mathcal{A}) = s_\varnothing$. The matching upper bound on N_{eff} is $1/\rho(D)$, and we call $1/\rho(D) - 1$ the *maximum N_{eff} gain* of averaging D. It also makes the component shares identifiable from correlations. A component c_S remains in the lower bound of averaging over D exactly when none of the axes in S is averaged, that is, $S \cap D = \varnothing$. The variance reductions available from two axes overlap through components varying along both axes, and the $2^{|\mathcal{A}|} - 1$ correlations needed to identify the shares are all estimated by Algorithm S1. We write out the two cases the experiments use, network×step in Corollary A.3 and network×step×view in Corollary A.4.

Corollary 4.6 (Teacher-Level Prediction)**.** *Under Assumption 4.2 on network and step axes, model each teacher's error as the EMA of its snapshots' errors (Assumption 4.1(ii)). Two teachers of decay λ are then correlated at $\rho_{\mathrm{cross}}^{\mathrm{EMA}} = (s_c + s_g B_g)/(s_c + s_a + s_g B_g + s_b B_b)$, a function of quantities fitted to the raw students alone.*

Corollary 4.6 implies a counterintuitive result where averaging each network over steps makes the two teachers more correlated than their raw students. The kernel removes from each teacher's variance the share s_b, which is private to each network and varies across steps, but leaves the common

share s_c in their covariance. The correlation therefore moves from $s_c + s_g$ at $\lambda = 0$ to $s_c/(s_c + s_a)$ as $\lambda \to 1$, an increase exactly when $s_c s_b > s_a s_g$ (Appendix A.5). A network ensemble of EMA teachers therefore gains less than one of raw students; we test this corollary in Section 5.1.

4.3 From Variance Reduction to Task Performance

For an error field with probe mean μ, reference variance $\sigma^2 > 0$, and finite positive effective sample size N_{eff}, the bias-variance identity (Geman et al., 1992) gives $S := \mathbb{E}[e^2] = \mu^2 + \frac{\sigma^2}{N_{\text{eff}}}$.

Averaging a Fixed Pool. Two aggregates of the same members, or of members of matched quality, share μ and σ^2 and differ only in effective sample size. For $N_0 \neq N_1$, subtracting the identity for the two aggregates gives $\Delta S = -\sigma^2(N_1 - N_0)/(N_0 N_1)$ and hence

$$b_{01} := \frac{\Delta\text{mIoU}}{N_1 - N_0} = \underbrace{-\frac{\sigma^2}{N_0 N_1}}_{\Delta S/\Delta N_{\text{eff}}\text{: exact}} \underbrace{R_{01}}_{\text{empirical}}, \qquad R_{01} := \frac{\Delta\text{mIoU}}{\Delta S}, \tag{8}$$

where b_{01} is the change in mIoU per unit of N_{eff} and R_{01} is the change in mIoU per unit of S; Appendix A.7 relates R_{01} to the classes whose predictions change. With $N_0 = 1$, this comparison measures the *aggregation gain*, the gain of an average over the mean performance of its own members. The aggregation gain is the only part of performance that the lower bound of Lemma 4.3 limits, since $N_1 \leq 1/\rho$.

Expanding a Pool. New members need not match the existing ones in quality, so expanding a pool moves its mean error and reference variance from $(\bar{\mu}_0, \sigma_0^2)$ to $(\bar{\mu}_1, \sigma_1^2)$, and the same identity gives

$$S_1 - S_0 = \sigma_1^2(V_1 - V_0) + \big[(\sigma_1^2 - \sigma_0^2)V_0 + (\bar{\mu}_1^2 - \bar{\mu}_0^2)\big]. \tag{9}$$

The first term is the variance reduction evaluated at the moments of the new pool, which the lower bound limits. The bracketed term is the change in the members themselves, which no correlation determines (Appendix A.6). When pool 0 is a single reference model, $V_0 = 1$ and the bracketed term compares the mean new member with that model. On the task metric, the gain of an aggregate $\bar{e}$ over the reference model therefore splits into a *member term* and the aggregation gain,

$$P(\bar{e}) - P_{\text{ref}} = \underbrace{\big[\bar{P}_{\text{mem}} - P_{\text{ref}}\big]}_{\text{member term}} + \underbrace{G}_{\text{aggregation gain}}, \qquad G := P(\bar{e}) - \bar{P}_{\text{mem}} \approx \beta_v(N_{\text{eff}} - 1) \leq \beta_v\Big(\frac{1}{\rho} - 1\Big), \tag{10}$$

where $\bar{P}_{\text{mem}}$ is the mean performance of the members and β_v is the readout rate calibrated below. Expanding a pool moves both terms, but only the aggregation gain G is bounded by ρ. A new kind of member that lowers ρ therefore pays off only if the member term it costs is smaller than the aggregation gain it frees, and adding more members of the same kind does not change this trade-off (Proposition A.5). Section 5.3 uses this property to improve saturated averages by constructing less correlated members.

Empirical Calibration Model. In self-training, the two terms of Equation (10) have a common source. The teacher's average determines the aggregation gain, and the pseudo-labels that the teacher supplies determine, through training, the quality of the student and hence the member term. The following model parametrizes these two channels by N_{eff} and by the accuracy of the teacher's pseudo-labels.

Model 4.7 (Local Two-Channel Performance Model)**.** Fix a dataset, a training protocol, a reference configuration, and a specified family and range of interventions. For each intervention c, let ΔP_c, $\Delta N_{\text{eff}\,c}$, and $\Delta a_{T,c}$ denote changes relative to the reference, where P is the task performance and a_T is the teacher's accuracy on accepted pseudo-labels, measured before any oracle replacement. We posit

$$\Delta P_c = \beta_v\, \Delta N_{\text{eff}\,c} + \beta_q\, \Delta a_{T,c} + \varepsilon_c, \qquad \mathbb{E}[\varepsilon_c] = 0 \quad \text{for each } c. \tag{11}$$

The expectation is over training randomness. The *readout rate* β_v and the *target rate* β_q are shared across the specified interventions: they are calibrated from separate readout and target sweeps and held fixed when predicting joint interventions. No additional intercept or interaction term between

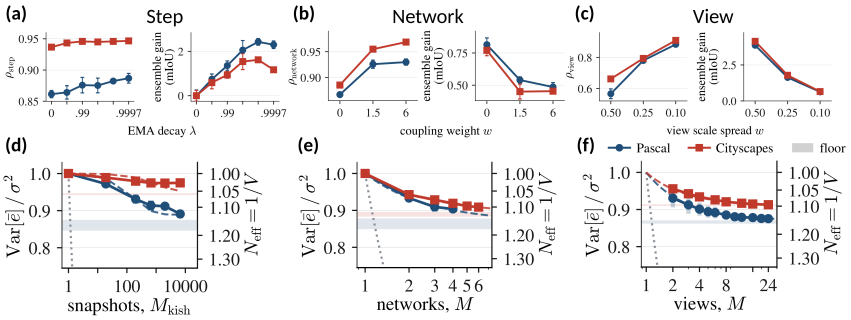


Figure 2: Each average type is capped by its error correlation. (a–c) Error correlation and ensemble gain as a function of each average type's hyperparameter. (d–f) Variance ratio and N_{eff} as a function of the weighted count M_{kish}. Shaded regions mark the measured lower bound on the variance ratio, the gray curve shows $1/M$, and the dashed curve shows the closed form of Lemmas 4.3 and 4.4.

ΔN_{eff} and Δa_T is fitted. The substantive restriction of the model is that the discrepancy ε_c has zero mean under these shared rates. The rates are specific to the calibration setting, and their transfer across average types is evaluated separately.

The interventions provide approximate empirical separation of the channels for calibration. We assess whether constant rates describe the separate sweeps and whether their frozen sum predicts joint interventions without an additional interaction term (Section 5.2; Appendix E). Appendix E explains why a linear readout term suffices over the range of N_{eff} that its upper bound allows. The observed range does not distinguish the linear form from plausible nonlinear alternatives.

5 Experiments

5.1 Lower Bounds on Individual and Combined Axes

Each Average Type Is Capped by Its Error Correlation. The hyperparameter of each average type sets its weighted count, but by Lemma 4.3 its effective sample size cannot exceed $1/\rho$, so a hyperparameter raises the gain only by lowering ρ. Figure 2 tests this on each axis, by varying the hyperparameters in (a–c) and showing the variance reduction against weighted count in (d–f). On the step axis, raising the EMA decay from 0 to 0.9997 enlarges the weighted count by more than three orders of magnitude but barely moves ρ_∞, and even the widest kernel reaches only $N_{\text{eff}} = 1.122$ on Pascal VOC, against $1/\rho_\infty$ of 1.127 at the same run. Its mIoU falls while N_{eff} still rises, because stale snapshots lower member quality, the bracketed term of Equation (9). On the network axis, stronger coupling trains the networks toward the same errors, which raises ρ_{network} and lowers the network-averaging gain, while the temporal correlation of each network falls. Coupling thus moves error between axes rather than removing it, as the decomposition makes explicit. The width of the range of view scales is the only hyperparameter that lowers ρ, and N_{eff} reaches over 99% of the pool's upper bound $1/\bar{\rho}$ at $M = 24$. To test whether these lower bounds extend beyond our segmentation setup, we measure the network-axis lower bound of FixMatch, FreeMatch, and SoftMatch on CIFAR-100 (Sohn et al., 2020; Wang et al., 2022b; Chen et al., 2023) (Figure 3a).

The Three Average Types Draw on One Decomposition, So Their Gains Overlap. By Theorem 4.5, every lower bound is a sum over one set of variance components, so two average types remove overlapping error whenever a component varies along both of their axes. Correlations measured along several axes on a shared probe set identify the private and shared components of each axis. We first test this on the network and step axes of Cityscapes. Across the coupling grid, the sum of the two single-axis maximum N_{eff} gains ($1/\rho(D) - 1$ in Theorem 4.5) overstates the joint maximum N_{eff} gain by 22–31%, and Corollary A.3 predicts how much of the network-averaging gain survives step averaging (Tables S5 and S6). The decomposition also passes the parameter-free test of Corollary 4.6: the shares fitted to the raw students alone predict that step averaging makes two EMA teachers more correlated than their students, and they reproduce the measured rise from

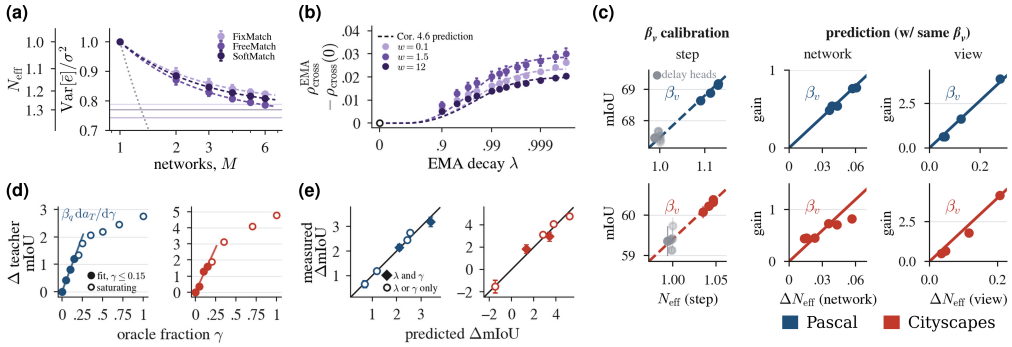


Figure 3: (a) Network-axis variance ratios and lower bounds for three semi-supervised classifiers on CIFAR-100; horizontal lines mark $\bar{\rho}$ and the dotted curve shows $1/M$. (b) Parameter-free prediction of the correlation between EMA teachers on Cityscapes (Corollary 4.6). (c) The readout rate β_v calibrated on the step axis predicts the gains of network and view averages. (d) Calibration of the target rate β_q under oracle replacement of pseudo-labels. (e) Test of the additive prediction of Equation (11) when the EMA decay λ and the ground-truth replacement fraction γ are varied jointly.

0.885 to 0.919 at $w_{\text{cpl}} = 0$ (Figure 3b). We then decompose all three axes of UniMatch V2 and DualFete (Corollary A.4; Figure 4a, b). Which axis holds the private variance differs between the two systems: on Pancreas-CT, most of the variance that view averaging removes is private to views, whereas most of the variance that step averaging removes is also removed by network averaging, and on UniMatch V2, step averaging adds almost nothing beyond the other two axes. In all four systems the axes overlap, and the summed single-axis maximum N_{eff} gains overstate the joint maximum gain by 9–61% (Table S7).

5.2 From Variance to Performance

Readout Rate β_v and Its Transfer across Axes. We calibrate the readout rate on the step axis by comparing each EMA with a single delayed snapshot of the same mean lag. The snapshot has $N_{\text{eff}} = 1$ and nearly constant mIoU across lags (gray markers in Figure 3c), so this contrast isolates ΔN_{eff} and gives $\beta_v = 13.3$ mIoU per unit of N_{eff} on Pascal VOC and 19.0 on Cityscapes. On Pascal VOC, the readout term accounts for 93% of the gain from raising the EMA decay from 0.99 to 0.999 (Table S3). The rate is stable across labeled fractions but changes with the training schedule (Tables S4 and S3). When the rate calibrated on the step axis is applied without refitting to network and view ensembles trained on the same schedule, it predicts the gains of view averages and of coupled-network averages closely (Figure 3c).

Target Rate β_q. To isolate the target channel, we replace a random fraction γ of the accepted pseudo-labels with ground-truth labels after the confidence mask has been computed, so that the mask and the averaging remain unchanged. As γ increases, the mIoU of the teacher rises while N_{eff} shows no trend (Figure 3d), and plotted against the accuracy of the accepted targets measured before replacement, the response is close to linear. This gives $\beta_q = 90.9$ mIoU per unit of target accuracy on Pascal VOC and 1594 on Cityscapes within the entire oracle sweep. The rate is much larger on Cityscapes since the classes that account for 80% of the mIoU gain occupy only 3.3% of the pixels, so a small change in target accuracy is accompanied by a large change in mIoU (Appendix A.7).

Better Targets Reach beyond What Averaging Can Still Add. The upper bound set by the correlation limits the readout channel but not the target channel, and the two calibrated rates show how much each channel still has to give. On Pascal VOC at $\lambda = 0.99$, taking the EMA all the way to its upper bound would add at most 0.92 mIoU at the calibrated readout rate, whereas replacing all accepted pseudo-labels with ground truth adds 2.75 mIoU on Pascal VOC and 4.78 on Cityscapes. The oracle is an upper reference rather than a practical method, but it shows that the error that averaging leaves behind is still reachable through the targets. The two channels also combine as Model 4.7 assumes: when the EMA decay and the oracle fraction are varied jointly, the separately calibrated rates predict the joint gains without an interaction term (Figure 3e; Appendix E).

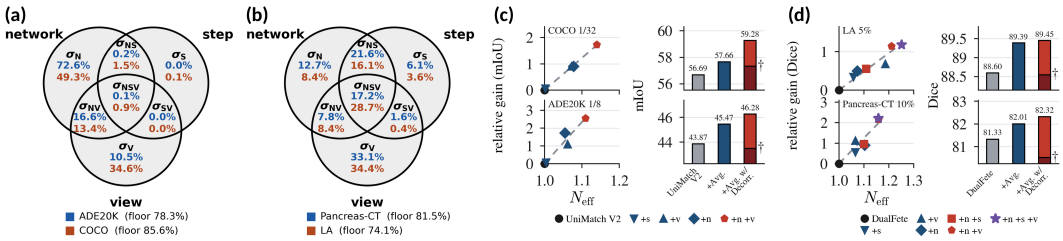


Figure 4: (a, b) Shares of the removable error variance of Theorem 4.5 for UniMatch V2 and DualFete, normalized by $1 - s_{\varnothing}$, with each dataset's lower bound $s_{\varnothing}$ shown below. (c, d) Left: aggregation gain as a function of N_{eff} for step (+s), view (+v), and network (+n) averages and their combinations, measured relative to the mean of each aggregate's own members. Right: performance with expanded members (+Avg.) and with decorrelated members (+Avg. w/ Decorr.); the last bar splits at † into the mean member score and the aggregation gain (Equation (10)).

5.3 Evaluation on State-of-the-Art Methods

Expanding the Averages of Two State-of-the-Art Methods. We use the decomposition to expand the averages of UniMatch V2 and DualFete along the axes that still hold private variance (Appendix F). For UniMatch V2 this is the network axis, whose variance partly overlaps with that of the view axis, and for DualFete it is mostly the view axis (Figure 4a, b). In both methods, a unit of effective sample size adds the same aggregation gain whichever axis supplies it: for each dataset, the gains of all pools over their own members fall on one line in N_{eff} (Figure 4c, d, left). Network averaging improves UniMatch V2 by 0.97 mIoU on COCO and by 1.60 mIoU on ADE20K, and the best multi-axis average improves DualFete by 0.79 Dice on Left Atrium and by 0.68 Dice on Pancreas-CT (Figure 4c, d, right). These gains are modest for two reasons: the remaining axes hold little removable variance, and an added axis improves on the existing pool only if its members are accurate enough that the member term does not cancel the aggregation gain (Equation (10)).

Raising the Upper Bound with Decorrelated Members. Because the upper bound of each pool is set by the correlation between its members (Lemma 4.3; Theorem 4.5(i)), the remaining route to larger gains is to add members with lower correlation, which raises the upper bound on N_{eff} and thereby amplifies the averaging gain. A practical way to obtain such members is to train a second network architecture: for UniMatch V2, we train a UPerNet head with the DPT head, and for DualFete, we train a 3D U-Net with the V-Net. On ADE20K, Pancreas-CT, and Left Atrium, the mixed pool reaches a higher N_{eff} than any single-architecture pool, and the gains rise to 2.41 mIoU on ADE20K and to 0.99 and 0.85 Dice on Pancreas-CT and Left Atrium. On COCO, the correlation decrease is not as significant, and the 2.59 mIoU gain comes mainly from the higher accuracy of the incorporated UPerNet, which raises the member term by 0.65 mIoU and, because the two families differ in mean error, also lifts the aggregation gain (Appendix F). The added members on the other datasets are no more accurate than the reference model, so the member term is zero or negative, but the aggregation gain that the less correlated members provide outweighs it and gives a net gain.

6 Limitations and Conclusion

Our work has three limitations. First, the extrapolated ρ_{∞} of the step axis depends on the fitting window and on the assumed form of the temporal decay. Second, the decomposition assumes members of equal variance and mutually uncorrelated components, and it treats weight averaging through a first-order approximation. Third, the oracle intervention is an upper reference for better targets rather than a method that provides them.

Within these limits, this work provides a unified perspective on step, network, and view averaging. We use a prediction error decomposition to show that their variance reduction overlaps and is bounded by error correlation. We further isolate the contributions of effective sample size gains and show that targeted reductions in private variance could more efficiently improve self-training. We apply our findings to optimize the ensemble averaging configurations of state-of-the-art self-training methods, which improves task performance on four datasets.

AI USE STATEMENT

In this work, AI tools were used strictly as assistants. Specifically, we used generative AI tools to polish writing and exposition of proofs, provide feedback on research methodology and experiments, find related work, assist in writing code, and refine hypotheses. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, designing research methodology or experiments, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results. We reviewed all AI-assisted work, including checking AI-generated content line by line and independently reading and verifying the literature identified through AI-assisted searches. We treated AI-generated feedback solely as suggestions and made all final decisions independently. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research does not involve human subjects or the collection of personally identifiable information.

REPRODUCIBILITY STATEMENT

To support the reproducibility of our results, we provide the following resources. All implementation details, such as model versions, training configurations, and evaluation procedures, are described in Appendix B.2. The design and implementation of our measurement procedure are presented in Algorithm S1, and Appendix B.1. For theoretical results, complete and detailed proofs are provided in Appendix A. We will release the complete source code, configuration files, and scripts upon acceptance.

REFERENCES

Taiga Abe, E Kelly Buchanan, Geoff Pleiss, and John P Cunningham. Pathologies of predictive diversity in deep ensembles. *arXiv preprint arXiv:2302.00704*, 2023.

Niccolò Ajroldi, Antonio Orvieto, and Jonas Geiping. When, where and why to average weights? *arXiv preprint arXiv:2502.06761*, 2025.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4413–4421. IEEE, 2018.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of machine learning and systems*, 3:747–769, 2021.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Robert L Brennan. Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4): 27–34, 1992.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Dan Busbridge, Jason Ramapuram, Pierre Ablin, Tatiana Likhomanenko, Eeshan Gunesh Dhekane, Xavier Suau Cuadros, and Russell Webb. How to scale your ema. *Advances in Neural Information Processing Systems*, 36:73122–73174, 2023.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. *Advances in neural information processing systems*, 37:126474–126499, 2024.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622. IEEE, 2021.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. *Advances in Neural Information Processing Systems*, 37: 9974–10007, 2024.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Spencer Frei, Difan Zou, Zixiang Chen, and Quanquan Gu. Self-training converts weak learners to strong learners in mixture models. In *International Conference on Artificial Intelligence and Statistics*, pp. 8003–8021. PMLR, 2022.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pp. 991–998. IEEE, 2011.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Francis Begnaud Hildebrand. *Introduction to numerical analysis*. Courier Corporation, 1987.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334. Springer, 1992.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Alan Jeffares, Tennison Liu, Jonathan Crabbé, and Mihaela van der Schaar. Joint training of deep ensembles fails due to learner collusion. *Advances in Neural Information Processing Systems*, 36: 13559–13589, 2023.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184. IEEE, 2024.

Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6728–6736, 2019.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4248–4257. IEEE, 2022.

Ekaterina Lobacheva, Nadezhda Chirkova, Maxim Kodryan, and Dmitry P Vetrov. On power laws in deep ensembles. *Advances In Neural Information Processing Systems*, 33:2375–2385, 2020.

Alexander Lyzhov, Yuliya Molchanova, Arsenii Ashukha, Dmitry Molchanov, and Dmitry Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on uncertainty in artificial intelligence*, pp. 1308–1317. PMLR, 2020.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Luis A Ortega, Rafael Cabañas, and Andres Masegosa. Diversity and generalization in neural network ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 11720–11743. PMLR, 2022.

Art B Owen. Variance components and generalized sobol'indices. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):19–41, 2013.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pp. 234–244. Springer, 2016.

Holger Roth, Amal Farag, Evrim B Turkbey, Le Lu, Jiamin Liu, and Ronald M Summers. Data from pancreas-ct. *(No Title)*, 2016.

Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 556–564. Springer, 2015.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.

Wolfgang Schmid. On ewma charts for time series. In *Frontiers in statistical quality control*, pp. 115–137. Springer, 1997.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*. John Wiley & Sons, 2009.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 1194–1203. IEEE, 2021.

Il'ya Meerovich Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608, 2020.

Minhak Song, Beomhan Baek, Kwangjun Ahn, and Chulhee Yun. Through the river: Understanding the benefit of schedule-free methods for language model training. *Advances in Neural Information Processing Systems*, 38:127524–127555, 2026.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in neural information processing systems*, 34:6906–6919, 2021.

Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3097–3107. IEEE, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

MB Vermaat, Ronald JMM Does, and Søren Bisgaard. Ewma control chart limits for first-and second-order autoregressive processes. *Quality and Reliability Engineering International*, 24(5): 573–584, 2008.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, Lan-Zhe Guo, et al. Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35:3938–3961, 2022a.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022b.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359): 1–49, 2023.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. Pmlr, 2022.

Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.

Lihe Yang, Zhen Zhao, and Hengshuang Zhao. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 3031–3048, 2025.

Le Yi, Wei Huang, Lei Zhang, Kefu Zhao, Yan Wang, and Zizhou Wang. Dualfete: revisiting teacher-student interactions from a feedback perspective for semi-supervised medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27852–27860, 2026.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

Appendix Overview. Appendix A proves the results of Section 4. Appendix B describes the implementation details and the experimental configurations. Appendix C tests the measurements with different training schedules and label fractions. Appendix D reports the detail of multi-axis decompositions. Appendix E tests the two-channel model with joint interventions. Finally, Appendix F supplements the evaluation on state-of-the-art methods and qualitative results.

A PROOFS

A.1 PRELIMINARIES

The proofs rely on two covariance models: equicorrelated members for the network and view axes (Assumption 4.1(i)), and a correlation that decays with the lag between snapshots of one training trajectory (Assumption 4.1(ii)). This section checks that both are valid covariance structures and states what changes when the simplification to equal variances and a common correlation fails.

Equicorrelated Members. Under Assumption 4.1(i), M networks or views have covariance $\Sigma = \sigma^2[(1-\rho)I + \rho\,\mathbf{1}\mathbf{1}^\top]$. Its eigenvalues are $\sigma^2[1+(M-1)\rho]$ on $\mathbf{1}$ and $\sigma^2(1-\rho)$ on $\mathbf{1}^\perp$, so for $M \geq 2$ the matrix is positive semidefinite if and only if $\rho \in [-1/(M-1), 1]$. Negative correlations are therefore admissible for finite M but not for large M. We restrict attention to $\rho \in [0,1]$, and every upper bound $1/\rho$ requires $\rho > 0$.

Snapshots of One Training Trajectory. Assumption 4.1(ii) replaces the common correlation with one that decays with the lag Δ between snapshots. Lemma A.1 shows that this correlation function is valid and that its two parameters can be identified.

Lemma A.1 (Admissibility of the Temporal Correlation Model)**.** *For $\rho_\infty, r \in [0,1)$, $\rho(\Delta) = \rho_\infty + (1-\rho_\infty)r^{|\Delta|}$ is a valid correlation function. It is realized by $e_k = \sqrt{\rho_\infty}\,Z + \sqrt{1-\rho_\infty}\,a_k$, where Z is a unit-variance component shared by all snapshots and a_k is an independent unit-variance AR(1) sequence with lag-one correlation r. The parameters are identified by $\rho_\infty = \lim_{\Delta\to\infty}\rho(\Delta)$ and $r = (\rho(1)-\rho_\infty)/(1-\rho_\infty)$.*

Proof. The function is a convex combination of the constant kernel, whose Gram matrix $\mathbf{1}\mathbf{1}^\top$ is positive semidefinite, and the AR(1) kernel, whose spectral density

$$S(\omega) = \sum_{\Delta\in\mathbb{Z}} r^{|\Delta|}e^{-i\omega\Delta} = \frac{1-r^2}{1-2r\cos\omega + r^2}$$

is non-negative. A convex combination of positive semidefinite kernels is positive semidefinite. Since Z and a_k are independent, $\mathrm{Corr}(e_k, e_{k+\Delta}) = \rho_\infty + (1-\rho_\infty)r^{|\Delta|}$. The two identities follow from letting $\Delta \to \infty$ and from setting $\Delta = 1$. □

For $r > 0$, $\log[\rho(\Delta) - \rho_\infty]$ is linear in $|\Delta|$. We nevertheless fit $\rho(\Delta)$ directly by least squares, because the log transform requires ρ_∞ in advance and amplifies noise at lags where $\rho(\Delta)$ is close to ρ_∞.

The model also assumes that the snapshot variance is constant over the analysis window. Restricting the window to late training reduces drift but does not guarantee stationarity, and a drifting variance has the same effect as unequal member variances, which we discuss next. We therefore form $\hat{V}$ on the late half of training and normalize it by the mean snapshot variance (Algorithm S1). For the same reason, we do not compare r, a per-step rate, across configurations whose training steps consume different numbers of images.

Unequal Variances and Correlations. On the view axis the pairwise correlations ρ_{ij} are measured rather than modeled. They form the Gram matrix of centered fields on one probe set and are positive semidefinite by construction. Assumption 4.1(i) is used only to summarize them by a single ρ. Let the members instead have variances σ_i^2, with $\sigma^2 := \frac{1}{M}\sum_i \sigma_i^2$ and equal weights. Then

$$\mathrm{Var}[\bar{e}] = \frac{1}{M^2}\Big[\sum_i \sigma_i^2 + \sum_{i\neq j}\rho_{ij}\sigma_i\sigma_j\Big] = \frac{\sigma^2}{M}\big[1 + (M-1)\,\tilde{\rho}\,h\big],$$

with

$$\tilde{\rho} := \frac{\sum_{i\neq j} \rho_{ij}\sigma_i\sigma_j}{\sum_{i\neq j} \sigma_i\sigma_j}, \qquad h := \frac{\sum_{i\neq j} \sigma_i\sigma_j}{(M-1)\sum_i \sigma_i^2} = \frac{M/(1+\mathrm{CV}_\sigma^2)-1}{M-1} \in [0,1].$$

Here $\tilde{\rho}$ is the mean correlation weighted by $\sigma_i\sigma_j$, and h measures how unequal the variances are, with CV_σ the coefficient of variation of the σ_i computed with population normalization. By Cauchy–Schwarz, $h = 1$ if and only if the variances are equal. Hence $N_{\text{eff}} = M/[1 + (M-1)\tilde{\rho}h]$ exactly, and if CV_σ and a positive $\tilde{\rho}$ converge as the pool grows, $N_{\text{eff}} \to (1+\mathrm{CV}_\sigma^2)/\tilde{\rho}$. Unequal variances therefore raise the upper bound above $1/\tilde{\rho}$. A wide view pool that admits weaker views can gain N_{eff} in this way without lowering its correlation. A finite pool does not by itself establish these limits.

With equal variances ($h = 1$) the expression reduces to Lemma A.2, and with a common correlation it reduces further to Equation (5). For a finite pool, Equation (5) evaluated at the unweighted mean $\bar{\rho}$ misstates the variance ratio $\mathrm{Var}[\bar{e}]/\sigma^2$ by $\frac{M-1}{M}(\bar{\rho} - \tilde{\rho}h)$.

A.2 Proof of Lemma 4.3

Lemma 4.3 (Equicorrelated Decomposition and Lower Bound, Hastie et al., 2009, §15.2)**.** *Under Assumption 4.1(i) with $\rho \geq 0$, the members are distributed, up to second moments, as a part shared by all members plus a part private to each, $X_i - \mathbb{E}[X_i] = b + \eta_i$ with $\mathrm{Var}[b] = \rho\sigma^2$, $\mathrm{Var}[\eta_i] = (1-\rho)\sigma^2$, and the η_i mutually uncorrelated and uncorrelated with b (4). Averaging acts on η and leaves b untouched:*

$$\mathrm{Var}[\bar{X}] = \sigma^2\Big(\rho + \frac{1-\rho}{M}\Big), \qquad N_{\text{eff}} = \frac{\sigma^2}{\mathrm{Var}[\bar{X}]} = \frac{M}{1+(M-1)\rho}. \tag{5}$$

For fixed $\rho > 0$, $N_{\text{eff}} \to 1/\rho$ as $M \to \infty$, whereas $M_{\text{kish}} = M$ grows without bound. At $\rho = 0$, $N_{\text{eff}} = M$.

Proof. Under Assumption 4.1(i) the weights are equal, the M diagonal terms of the covariance equal σ^2 and the $M(M-1)$ off-diagonal terms equal $\rho\sigma^2$, so

$$\mathrm{Var}[\bar{X}] = \frac{1}{M^2}\sum_{i,j}\mathrm{Cov}(X_i, X_j) = \frac{M\sigma^2 + M(M-1)\rho\sigma^2}{M^2} = \sigma^2\Big[\rho + \frac{1-\rho}{M}\Big],$$

$$N_{\text{eff}} = \frac{\sigma^2}{\mathrm{Var}[\bar{X}]} = \frac{M}{1+(M-1)\rho}, \qquad \rho\, N_{\text{eff}} = \frac{M\rho}{M\rho + 1 - \rho} \leq 1.$$

The second line inverts Equation (2). N_{eff} increases in M for $\rho < 1$ with limit $1/\rho$ for $0 < \rho < 1$, where the inequality is strict. At $\rho = 1$, $N_{\text{eff}} = 1$ for every M. For (4), let $Z, Z_1, \ldots, Z_M$ be centered, uncorrelated variables with unit variance and set

$$b = \sqrt{\rho}\,\sigma Z, \qquad \eta_i = \sqrt{1-\rho}\,\sigma Z_i, \qquad \mathrm{Cov}(b+\eta_i,\, b+\eta_j) = \rho\sigma^2 + (1-\rho)\sigma^2\,\mathbf{1}\{i=j\} = \Sigma_{ij}.$$

Adding the original means matches the first moments as well. This is a second-moment representation, not an equality in distribution. □

Lemma A.2 (Heterogeneous Correlations)**.** *Let the members have equal variance σ^2 and pairwise correlations ρ_{ij}. With equal weights,*

$$\mathrm{Var}[\bar{X}] = \frac{\sigma^2}{M}\big[1 + (M-1)\bar{\rho}\big], \qquad \bar{\rho} := \frac{1}{M(M-1)}\sum_{i\neq j}\rho_{ij},$$

and with weights w_k, $\mathrm{Var}[\bar{X}] = \sigma^2[1/M_{\text{kish}} + (1 - 1/M_{\text{kish}})\bar{\rho}_w]$ with $\bar{\rho}_w$ the w_iw_j-weighted off-diagonal mean. For $M_{\text{kish}} > 1$, $N_{\text{eff}} \leq M_{\text{kish}}$ if and only if $\bar{\rho}_w \geq 0$.

Proof. Expanding $\mathrm{Var}[\sum_k w_k X_k] = \sum_{i,j} w_i w_j \mathrm{Cov}(X_i, X_j)$, the diagonal contributes $\sigma^2\sum_k w_k^2 = \sigma^2/M_{\text{kish}}$ and the off-diagonal $\sigma^2\sum_{i\neq j} w_i w_j \rho_{ij} = \sigma^2(1 - 1/M_{\text{kish}})\bar{\rho}_w$, since $\sum_{i\neq j} w_i w_j = 1 - \sum_k w_k^2$. Equal weights give the first display. Dividing σ^2 by the weighted form, $N_{\text{eff}} \leq M_{\text{kish}}$ reads $(1 - 1/M_{\text{kish}})\bar{\rho}_w \geq 0$. □

Equality in Lemma A.2 requires a zero weighted mean covariance, not that every pair of members be uncorrelated. Since Algorithm S1 estimates $\bar{\rho}$ directly, the finite-M statements rest on equal variances and equal weights alone; the upper bound $1/\bar{\rho}$ requires in addition that $\bar{\rho}$ stop drifting as M grows.

The members in Assumption 4.1 are not limited to networks or snapshots. The proof uses only their covariance, so it also bounds an average over transformed copies of one input by $1/\rho$, and the estimator applies unchanged (Section 5.1).

Marginal Returns and Sensitivity to ρ. Subtracting successive instances of Equation (5) and differentiating $N_{\text{eff}} = M/(1 + (M-1)\rho)$ give

$$V_M - V_{M+1} = \frac{1-\rho}{M(M+1)}, \qquad \frac{\partial N_{\text{eff}}}{\partial \rho} = -\frac{M(M-1)}{(1+(M-1)\rho)^2}. \tag{12}$$

Here $V_M = \mathrm{Var}[\bar{X}_M]/\sigma^2$. Thus the marginal reduction in V is $O(M^{-2})$ at fixed ρ, whereas lowering a positive ρ raises the upper bound $1/\rho$.

Adding a View. For M equally weighted views with equal variances and mean pairwise correlation $\bar{\rho}_M$, let an added view of the same variance have mean correlation $\bar{r} = M^{-1}\sum_{i \le M} \rho_{i,M+1}$ with the pool. Counting the $2M$ new ordered pairs and substituting into $V_M = \bar{\rho}_M + (1-\bar{\rho}_M)/M$ gives

$$\begin{aligned} \bar{\rho}_{M+1} &= \bar{\rho}_M + \frac{2(\bar{r} - \bar{\rho}_M)}{M+1}, \\ V_M - V_{M+1} &= \frac{1-\bar{\rho}_M}{M(M+1)} - \frac{2M}{(M+1)^2}(\bar{r} - \bar{\rho}_M). \end{aligned} \tag{13}$$

The added view lowers variance if and only if $\bar{r} - \bar{\rho}_M < (1-\bar{\rho}_M)(M+1)/(2M^2)$, and the threshold vanishes as $M \to \infty$.

A.3 Proof of Lemma 4.4

Lemma 4.4 (Temporal Lower Bound of an EMA Teacher)**.** *Under Assumption 4.1(ii) with unit snapshot variance, the variance ratio V (Definition 3.1) of an EMA teacher of decay λ is*

$$V(\lambda) = \rho_\infty + (1-\rho_\infty)\, B(\lambda, r), \qquad B(\lambda, r) = \frac{1}{M_{\text{kish}}} \cdot \frac{1+\lambda r}{1-\lambda r} = \frac{(1-\lambda)(1+\lambda r)}{(1+\lambda)(1-\lambda r)}, \tag{6}$$

where $B \in (0, 1]$ is the fraction of the decaying part's variance that the kernel leaves, so that $1/B$, the weighted count divided by a penalty for correlation between nearby snapshots, is the number of independent snapshots the kernel is worth. V is strictly decreasing on $[0, 1)$ and $N_{\text{eff}} \to 1/\rho_\infty$ as $\lambda \to 1$, interpreted as infinity when $\rho_\infty = 0$. At $r = 0$, Equation (6) becomes Equation (5) with $(M_{\text{kish}}, \rho_\infty)$ in place of (M, ρ) (Lemma A.2).

Lemma 4.4 is exact within the error-averaging and covariance model of Assumption 4.1(ii), with (ρ_∞, r) fixed within one training trajectory.

Weight-Only Quantities of the EMA Kernel. For $w_k = (1-\lambda)\lambda^k$, $k \ge 0$, and $\lambda \in [0, 1)$, the geometric sums $\sum_k w_k^2 = (1-\lambda)^2/(1-\lambda^2)$ and $\sum_k k w_k = \lambda/(1-\lambda)$ give

$$M_{\text{kish}}(\lambda) = \frac{1+\lambda}{1-\lambda}, \qquad \text{mean lag} = \frac{\lambda}{1-\lambda}. \tag{14}$$

Hence $M_{\text{kish}} = 2\,(\text{mean lag}) + 1$: an EMA cannot average more snapshots without including older ones.

Proof of Lemma 4.4. We first evaluate the double sum in closed form, then read off the range of B, and finally differentiate to obtain monotonicity. By Assumption 4.1(ii) and $\sum_j w_j = 1$, with all sums absolutely convergent,

$$V(\lambda) = \sum_{j,k \ge 0} w_j w_k\, \rho(|j-k|) = \rho_\infty + (1-\rho_\infty)\, S, \qquad S := (1-\lambda)^2 \sum_{j,k \ge 0} \lambda^{j+k} r^{|j-k|}.$$

The diagonal of the double sum is $\sum_k \lambda^{2k} = 1/(1-\lambda^2)$, and each off-diagonal half is $\sum_k \lambda^{2k} \sum_{d\geq 1}(\lambda r)^d$. Therefore

$$S = \frac{(1-\lambda)^2}{1-\lambda^2}\Big(1 + \frac{2\lambda r}{1-\lambda r}\Big) = \frac{(1-\lambda)(1+\lambda r)}{(1+\lambda)(1-\lambda r)} =: B(\lambda),$$

which is the stated form. For the range, write $B = (A-C)/(A+C)$ with $A := 1-\lambda^2 r$ and $C := \lambda(1-r)$. On $[0,1)^2$ we have $A > 0$, $C \geq 0$ and $A - C = (1-\lambda)(1+\lambda r) \geq 0$, so $0 \leq B \leq 1$ with $B = 1$ if and only if $\lambda = 0$, and hence $\rho_\infty \leq V \leq 1$. As $\lambda \to 1$, $A - C \to 0$ while $A + C \to 2(1-r) > 0$, so $V \to \rho_\infty$ and $N_{\text{eff}} \to 1/\rho_\infty$. Finally, with $A' = -2\lambda r$ and $C' = 1 - r$,

$$B' = \frac{2(A'C - C'A)}{(A+C)^2} = \frac{-2(1-r)(1+\lambda^2 r)}{(1+\lambda)^2(1-\lambda r)^2} < 0,$$

so $V' = (1-\rho_\infty)B' < 0$ for $\rho_\infty < 1$ and $N_{\text{eff}} = 1/V$ is strictly increasing with supremum $1/\rho_\infty$, not attained on $[0,1)$. At $r = 1$, $B \equiv 1$ and no averaging occurs, and $r < 1$ is assumed for this reason. At $r = 0$, $B = (1-\lambda)/(1+\lambda) = 1/M_{\text{kish}}$ by (14), so $V = \rho_\infty + (1-\rho_\infty)/M_{\text{kish}}$, which is the last claim. □

A.4 Proof of Theorem 4.5

Theorem 4.5 (Multi-Axis Lower Bounds, Identification, and Composition)**.** *Under Assumption 4.2, let $\rho(D)$ be the correlation between two units that differ along the axes $D \subseteq \mathcal{A}$, at infinite step separation when step belongs to D.*

(i) (Lower bounds) $\rho(D) = \sum_{S \cap D = \varnothing} s_S$. *Every finite aggregate of the form in (iii), averaging only axes in D, satisfies $V \geq \rho(D)$ and, when $\rho(D) > 0$, $N_{\text{eff}} \leq 1/\rho(D)$. The bound is approached as all averaged network/view sizes tend to infinity and, if the step axis is averaged, $\lambda \to 1$. A zero lower bound gives no finite upper bound.*

(ii) (Identification) *The $2^{|\mathcal{A}|} - 1$ nontrivial correlations, with $\rho(\varnothing) = 1$, determine the shares by Möbius inversion on the subset lattice, $s_S = \sum_{U \subseteq S}(-1)^{|S \setminus U|}\rho(\mathcal{A} \setminus U)$.*

(iii) (Composition) *For product weights, uniform network and view averages and an EMA of decay λ on the step axis, a finite aggregate has the following variance ratio, with $\phi_a(S) = 1$ on axes not averaged:*

$$V = \sum_{S \subseteq \mathcal{A}} s_S \prod_{a \in S} \phi_a(S), \qquad \phi_a(S) = \begin{cases} 1/M_a, & a \in \{\text{n}, \text{v}\}, \\ B(\lambda, r_S), & a = \text{s}. \end{cases} \tag{7}$$

Proof. We prove the three claims in turn: the lower-bound formula, its inversion, and the composition law.

Lower bounds. Let two units differ in exactly the axes D, at infinite lag if D contains step. Along a differing axis in S the component c_S is uncorrelated across the units, so c_S contributes s_S to their covariance if $S \cap D = \varnothing$ and 0 otherwise. Therefore

$$\rho(D) = \sum_{S \cap D = \varnothing} s_S.$$

This establishes the correlation formula in part (i). The finite-aggregate lower bound and its limiting attainability follow from the composition calculation below.

Identification. With $g(U) := \rho(\mathcal{A} \setminus U) = \sum_{S \subseteq U} s_S$, the map from shares to correlations is triangular with unit diagonal. By Möbius inversion,

$$s_S = \sum_{U \subseteq S} (-1)^{|S \setminus U|}\, \rho(\mathcal{A} \setminus U),$$

so $2^{|\mathcal{A}|} - 1$ measured correlations and the unit-sum constraint determine the shares. Non-negative recovered shares are necessary for this component model, and the algebraic inversion alone does not establish its validity.

Composition. For probability averaging the aggregate is exactly $\sum_S \bar{c}_S$, with $\bar{c}_S$ the average of c_S over the axes in S. The $\bar{c}_S$ are uncorrelated, and since the correlation of c_S factors over the axes in S, averaging multiplies its variance by $\prod_{a\in S}\phi_a(S)$, with $\phi_a(S) = 1/M_a$ for a network or view axis and $\phi_{\mathrm{s}}(S) = B(\lambda, r_S)$ for step (Appendix A.3). Hence

$$V = \sum_{S\subseteq\mathcal{A}} s_S \prod_{a\in S} \phi_a(S),$$

with $\phi_a(S) = 1$ on any axis not averaged. If only axes in D are averaged, $s_S \geq 0$ and $0 < \phi_a(S) \leq 1$ give

$$V = \rho(D) + \sum_{S\cap D\neq\varnothing} s_S \prod_{a\in S\cap D} \phi_a(S) \geq \rho(D).$$

Thus $N_{\mathrm{eff}} = 1/V \leq 1/\rho(D)$ whenever $\rho(D) > 0$, for every finite aggregate under consideration. To approach the bound, send $M_a \to \infty$ for every averaged network/view axis and $\lambda \to 1$ if the step axis is averaged. Since $r_S < 1$, every term with $S \cap D \neq \varnothing$ then tends to zero, giving $V \to \rho(D)$. The inequality holds before taking this limit. For weight-space EMA, the calculation uses the error-averaging approximation in Assumption 4.1(ii), with component covariances from Assumption 4.2. □

Two- and Three-Axis Cases. We write out Theorem 4.5 for the two sets of axes the experiments use. For $\mathcal{A} = \{\mathrm{n}, \mathrm{s}\}$, the case measured on Cityscapes and Pascal VOC (Appendix D.1), we write the error of network member i at step t as

$$e_{i,t} = c + a_i + g_t + b_{i,t}, \tag{15}$$

with $c = c_\varnothing$, $a_i = c_{\{\mathrm{n}\}}$, $g_t = c_{\{\mathrm{s}\}}$ and $b_{i,t} = c_{\{\mathrm{n},\mathrm{s}\}}$ in the notation of Section 4, shares (s_c, s_a, s_g, s_b), temporal rates r_g, r_b, and $\mathrm{Var}[e_{i,t}] = 1$.

Corollary A.3 (Two-Axis Instance of Theorem 4.5)**.** *Under Assumption 4.2, with nonzero denominators in the ratios below, and with* $\rho(\Delta) = \mathrm{Corr}(e_{i,t}, e_{i,t+\Delta})$ *and* $\rho_{\mathrm{cross}}(\Delta) = \mathrm{Corr}(e_{i,t}, e_{j,t+\Delta})$ *for* $i \neq j$*:*

(i) (identification) $\rho_{\mathrm{cross}}(\Delta) = s_c + s_g r_g^{|\Delta|}$ *and* $\rho(\Delta) = s_c + s_a + s_g r_g^{|\Delta|} + s_b r_b^{|\Delta|}$*, so* $s_c = \rho_{\mathrm{cross}}(\infty)$*,* $s_g = \rho_{\mathrm{cross}}(0) - \rho_{\mathrm{cross}}(\infty)$*,* $s_a = \rho_\infty - \rho_{\mathrm{cross}}(\infty)$ *and* $s_b = 1 - s_c - s_a - s_g$*. Further,* $(\rho - \rho_{\mathrm{cross}})(\Delta) = s_a + s_b r_b^{|\Delta|}$ *identifies* r_b *without reference to* r_g*.*

(ii) (composition) *Let* $\bar{E} = \frac{1}{M}\sum_{i=1}^{M}\sum_{k\geq 0} w_k e_{i,t-k}$ *with* $w_k = (1-\lambda)\lambda^k$*,* $\lambda \in [0, 1)$*. Under the component covariances in Assumption 4.2,*

$$V(M, \lambda) := \mathrm{Var}[\bar{E}] = s_c + s_g B(\lambda, r_g) + \frac{1}{M}\big[s_a + s_b B(\lambda, r_b)\big], \tag{16}$$

with $B(\lambda, r) = (1-\lambda)(1+\lambda r)\big/\big((1+\lambda)(1-\lambda r)\big)$ *as in Lemma 4.4. For weight-space EMA, identifying* $\bar{E}$ *with the teacher error additionally uses the error-averaging approximation in Assumption 4.1(ii); no single-rate model for the full temporal correlation is required.*

(iii) (three upper bounds) $V(1, \lambda) \downarrow s_c + s_a$ *as* $\lambda \to 1$*;* $V(M, 0) \downarrow s_c + s_g$ *as* $M \to \infty$*;* $V(M, \lambda) \downarrow s_c$ *as both limits are taken. Equivalently* N_{eff} *is bounded by* $1/\rho_\infty$*, by* $1/\rho_{\mathrm{network}}$ *and by* $1/\rho_{\mathrm{cross}}(\infty)$ *respectively, and* $\rho_{\mathrm{cross}}(\infty) \leq \min\{\rho_\infty, \rho_{\mathrm{network}}\}$*.*

(iv) (overlap) *The variance the network axis removes when the step axis is already averaged at decay* λ *is* $(1 - \frac{1}{M})[s_a + s_b B(\lambda, r_b)]$*, a fraction* $[s_a + s_b B(\lambda, r_b)]/(s_a + s_b)$ *of what it removes alone, decreasing in* λ *to* $s_a/(s_a + s_b)$*. Symmetrically the step axis retains* $\big[s_g(1 - B_g) + \frac{1}{M} s_b(1 - B_b)\big]\big/\big[s_g(1 - B_g) + s_b(1 - B_b)\big]$*, decreasing in* M *to* $s_g(1 - B_g)/[s_g(1 - B_g) + s_b(1 - B_b)]$*.*

Proof. Parts (ii) and (iii) are Theorem 4.5 at $\mathcal{A} = \{\mathrm{n}, \mathrm{s}\}$. With $\phi_{\mathrm{n}} = 1/M$ and $\phi_{\mathrm{s}}(S) = B(\lambda, r_S)$, the composition law gives Equation (16), and letting $B \to 0$, $1/M \to 0$, or both gives the three limits, each approached from above since the dropped terms are non-negative. The inequality in (iii) follows from $s_a, s_g \geq 0$.

(i) The families in Equation (15) are uncorrelated and $\mathrm{Var}[e_{i,t}] = 1$, so covariances are correlations. For $i \neq j$ the a and b terms drop out, and for $i = j$ the a term contributes s_a at every lag. This gives the two curves. Letting $\Delta \to \infty$ removes the geometric terms, $\Delta = 0$ in the cross curve gives $s_c + s_g$, and $\rho(0) = 1$ closes the sum. Subtracting the two curves cancels s_c and $s_g r_g^{|\Delta|}$.

(iv) Both statements are $V(1, \lambda) - V(M, \lambda)$ and $V(M, 0) - V(M, \lambda)$ from (ii), divided by the same quantity at $\lambda = 0$ and at $M = 1$. The monotonicity and the limits are those of B in Lemma 4.4. □

Corollary A.4 (Three Axes)**.** *At $\mathcal{A} = \{\mathrm{n}, \mathrm{s}, \mathrm{v}\}$, the case measured in Section 5.1, write $B_S := B(\lambda, r_S)$. For M members, K views and decay λ,*

$$V(M, \lambda, K) = s_\varnothing + \frac{s_\mathrm{n}}{M} + s_\mathrm{s} B_\mathrm{s} + \frac{s_\mathrm{v}}{K} + \frac{s_\mathrm{ns} B_\mathrm{ns}}{M} + \frac{s_\mathrm{nv}}{MK} + \frac{s_\mathrm{sv} B_\mathrm{sv}}{K} + \frac{s_\mathrm{nsv} B_\mathrm{nsv}}{MK},$$

with lower bounds $\rho(\{\mathrm{n}\}) = s_\varnothing + s_\mathrm{s} + s_\mathrm{v} + s_\mathrm{sv}$ and its permutations, $\rho(\{\mathrm{n}, \mathrm{s}\}) = s_\varnothing + s_\mathrm{v}$ and its permutations, and $\rho(\mathcal{A}) = s_\varnothing$. The seven non-trivial lower bounds and $\sum_S s_S = 1$ determine the eight shares.

Proof. Substitute $\mathcal{A} = \{\mathrm{n}, \mathrm{s}, \mathrm{v}\}$ into Theorem 4.5: the composition law with $\phi_\mathrm{n} = 1/M$, $\phi_\mathrm{v} = 1/K$ and $\phi_\mathrm{s}(S) = B_S$ gives $V(M, \lambda, K)$, Equation (7) gives the lower bounds, and Möbius inversion gives the shares. □

A.5 Proof of Corollary 4.6

Corollary 4.6 (Teacher-Level Prediction)**.** *Under Assumption 4.2 on network and step axes, model each teacher's error as the EMA of its snapshots' errors (Assumption 4.1(ii)). Two teachers of decay λ are then correlated at $\rho_{\mathrm{cross}}^{\mathrm{EMA}} = (s_c + s_g B_g)/(s_c + s_a + s_g B_g + s_b B_b)$, a function of quantities fitted to the raw students alone.*

Proof. Each teacher is $T_i = c + a_i + \bar{g} + \bar{b}_i$ with $\bar{g} = \sum_k w_k g_{t-k}$ and $\bar{b}_i = \sum_k w_k b_{i,t-k}$, so its variance is $s_c + s_a + s_g B_g + s_b B_b$ by part (ii) of Corollary A.3 at $M = 1$. For $i \neq j$ the a and b families are uncorrelated across members, leaving $\mathrm{Cov}(T_i, T_j) = s_c + s_g B_g$. Dividing by the common variance gives the stated ratio, whose inputs $s_c, s_a, s_g, s_b, r_g, r_b$ are all fitted to the raw students' two curves by part (i). At $\lambda = 0$ it returns $\rho_{\mathrm{cross}}(0) = s_c + s_g$. For $s_c + s_a > 0$, its limit as $\lambda \to 1$ is $s_c/(s_c + s_a)$, and

$$\frac{s_c}{s_c + s_a} - (s_c + s_g) = \frac{s_c s_b - s_a s_g}{s_c + s_a}.$$

Hence the limiting correlation exceeds the raw correlation if and only if $s_c s_b > s_a s_g$. For finite λ, with $D_\lambda := s_c + s_a + s_g B_g + s_b B_b > 0$,

$$\rho_{\mathrm{cross}}^{\mathrm{EMA}} - (s_c + s_g) = \frac{s_c s_b (1 - B_b) - s_a s_g (1 - B_g) + s_g s_b (B_g - B_b)}{D_\lambda}.$$

The numerator determines the sign. If $r_g = r_b$ and $\lambda > 0$, this reduces to the same condition $s_c s_b > s_a s_g$. For unequal decay rates, neither this condition nor $s_b > 0$ alone guarantees an increase at every finite decay or monotonicity in λ. □

A.6 Squared Error, Unequal Member Means, and Expanding a Pool

Equation (8) applies the identity $\mathbb{E}[e^2] = (\mathbb{E}[e])^2 + \mathrm{Var}(e)$ to two aggregates evaluated on the same probe pixels. It is exact when the two aggregates share the mean error μ and the reference variance σ^2. We first examine probability averaging against these two conditions and then drop them to obtain Equation (9).

Probability Averaging. Averaging predictions is linear, $\mathbb{E}[\sum_i w_i e_i] = \sum_i w_i \mathbb{E}[e_i]$, so members with a common mean give an aggregate with the same mean. The member experiments, however, compare the aggregate with the mean score of its members, whose means $\mu_i = \mathbb{E}[e_i]$ may differ.

With equal weights, $\bar{\mu} = M^{-1}\sum_i \mu_i$, $\sigma^2 = M^{-1}\sum_i \mathrm{Var}(e_i)$ and $V = \mathrm{Var}[\bar{e}]/\sigma^2 = 1/N_{\mathrm{eff}}$, the exact difference is

$$S(\bar{e}) - \frac{1}{M}\sum_i S(e_i) = \sigma^2(V-1) - \frac{1}{M}\sum_i (\mu_i - \bar{\mu})^2.$$

The first term is the variance reduction of Equation (8). The second term is non-negative, so unequal member means make the aggregate improve on its members by more than the variance term predicts.

Expanding a Pool. Equation (8) compares two aggregates with common moments, whereas adding snapshots, networks, or views changes the moments of the pool. The following proposition records what survives. For $k \in \{0,1\}$, let pool k have mean of member means $\bar{\mu}_k$, reference variance σ_k^2, and variance ratio V_k, let $\bar{e}_k$ be its equal-weight probability average, and write $S_k := S(\bar{e}_k)$.

Proposition A.5 (Expanding a Pool)**.** (i) $S_k = \bar{\mu}_k^2 + \sigma_k^2 V_k$ *for* $k \in \{0,1\}$*, and Equation* (9) *follows.* (ii) *If pool* 1 *satisfies Assumption 4.1(i) with correlation* ρ_1*, then* $S_1 \geq \bar{\mu}_1^2 + \sigma_1^2\rho_1$*, so* $S_1 < S_0$ *only if*

$$\bar{\mu}_1^2 + \sigma_1^2\rho_1 < S_0. \tag{17}$$

Proof. (i) Probability averaging is linear, so $\mathbb{E}[\bar{e}_k] = \bar{\mu}_k$, and $\mathrm{Var}[\bar{e}_k] = \sigma_k^2 V_k$ by Definition 3.1. Hence $S_k = \mathbb{E}[\bar{e}_k]^2 + \mathrm{Var}[\bar{e}_k] = \bar{\mu}_k^2 + \sigma_k^2 V_k$. Subtracting the two cases and adding $\sigma_1^2 V_0 - \sigma_1^2 V_0$ gives Equation (9). *(ii)* By Lemma 4.3, $V_1 = \rho_1 + (1-\rho_1)/M_1 \geq \rho_1$, and $\sigma_1^2 \geq 0$, so (i) gives $S_1 \geq \bar{\mu}_1^2 + \sigma_1^2\rho_1$. If $S_1 < S_0$, this bound is below S_0. □

Equation (17) states that the lower bound of the enlarged pool, its squared error as $M_1 \to \infty$, must lie below the squared error that the smaller pool already attains. The lower bound of Lemma 4.3 limits the first term of Equation (9), whereas the bracketed term is the change in the pool's mean error and reference variance, which no correlation determines, since an added member can lower ρ and raise $\bar{\mu}$ at the same time. For members of equal mean, M_1 does not enter Equation (17), so enlarging a pool that fails the condition does not repair it. With common moments, Equation (9) reduces to $\Delta S = \sigma^2(V_1 - V_0)$, the comparison underlying Equation (8). Comparing an aggregate with the mean score of its own members instead introduces the dispersion term of the preceding paragraph.

A.7 Metric Conversion and Class Composition

This section explains why the target rate β_q need not transfer across datasets. Let a_Π be the student's pixel accuracy on the probe set Π and a_T the teacher's accuracy on accepted pseudo-labels (Model 4.7). Pixel accuracy weights every pixel equally, whereas mIoU weights every class equally, so correcting a pixel of a rare class moves mIoU more than correcting a pixel of a common class.

Proposition A.6 (Metric Conversion and Class Composition)**.** *Let* n, n' *be confusion matrices on the same probe with a fixed set of evaluated classes and positive class unions* D_c *along the segment joining them.* (i) *The class-c entries of* $J = \partial\mathrm{mIoU}/\partial n$ *scale as* $1/D_c$ *(Equation* (18)*), and with* $\bar{n} = (n+n')/2$*,* $\Delta\mathrm{mIoU} = \langle J(\bar{n}), n'-n\rangle + O(\|n'-n\|^3)$*.* (ii) *Hence for* $\Delta a_\Pi \neq 0$ *the local ratio* $\kappa := \Delta\mathrm{mIoU}/\Delta a_\Pi$ *depends on which classes change and is not a universal conversion. If also* $\Delta a_T \neq 0$*, the target rate factors as* $\Delta\mathrm{mIoU}/\Delta a_T = \kappa\,\Delta a_\Pi/\Delta a_T$*, whose second factor is a training response that* J *does not determine.*

The Jacobian in (i) is the gradient that differentiable IoU surrogates follow (Rahman & Wang, 2016; Berman et al., 2018), and the midpoint form is the standard second-order rule (Hildebrand, 1987).

Proof of (i). We first differentiate mIoU with respect to one confusion count, then compare the endpoint and midpoint forms. Let n_{ij} be the confusion count of true class i predicted as j, $R_i = \sum_j n_{ij}$, $P_j = \sum_i n_{ij}$ and $D_c = R_c + P_c - n_{cc}$, so that $\mathrm{IoU}_c = n_{cc}/D_c$ and mIoU is the mean over $\mathcal{P} = \{c : R_c > 0\}$, on which $D_c \geq R_c > 0$. Since $\partial D_c/\partial n_{ij} = \mathbf{1}\{i{=}c\} + \mathbf{1}\{j{=}c\} - \mathbf{1}\{i{=}j{=}c\}$,

$$\frac{\partial\,\mathrm{mIoU}}{\partial n_{ij}} \;=\; \frac{1}{|\mathcal{P}|}\sum_{c\in\mathcal{P}} \frac{D_c\,\mathbf{1}\{i{=}j{=}c\} \;-\; n_{cc}\big(\mathbf{1}\{i{=}c\} + \mathbf{1}\{j{=}c\} - \mathbf{1}\{i{=}j{=}c\}\big)}{D_c^2}. \tag{18}$$

Hence the class-c term of the Jacobian is

$$\frac{D_c - n_{cc}}{|\mathcal{P}|\, D_c^2} \text{ for a correct pixel of class } c, \qquad -\frac{n_{cc}}{|\mathcal{P}|\, D_c^2} \text{ for an error with } i = c \text{ or } j = c,$$

so both scale as $1/D_c$, and the map from pixel counts to mIoU is not proportional to pixel accuracy. This rules out universal proportionality but does not preclude a useful local approximation along a particular path. Each IoU_c is a ratio of affine functions of n, so along the segment from n to n', with $\Delta n = n' - n$, $\bar{n} = \frac{1}{2}(n + n')$ and H the Hessian at n,

$$\mathrm{mIoU}(n') - \mathrm{mIoU}(n) = \langle J(n), \Delta n \rangle + \tfrac{1}{2}\, \Delta n^\top H\, \Delta n + O(\|\Delta n\|^3) = \langle J(\bar{n}), \Delta n \rangle + O(\|\Delta n\|^3),$$

since $J(\bar{n}) = J(n) + \frac{1}{2} H \Delta n + O(\|\Delta n\|^2)$. The endpoint form therefore carries a second-order error and the midpoint form a third-order one. □

Proof of (ii). Write $P = |\Pi|$, $K = |\mathcal{P}|$, and $d = n' - n$. Along the segment between the two confusion matrices, the fundamental theorem of calculus gives

$$\Delta\mathrm{mIoU} = \int_0^1 \langle J(n + td), d \rangle\, dt, \qquad \Delta a_\Pi = P^{-1} \sum_c d_{cc}.$$

If $\Delta a_\Pi \neq 0$, dividing the first identity by the second and applying part (i) yields

$$\kappa = \frac{\langle J(\bar{n}), d \rangle}{\Delta a_\Pi} + O\!\left(\frac{\|d\|^3}{|\Delta a_\Pi|} \right).$$

The denominator can amplify the midpoint error when the accuracy change is small.

For an elementary correction of δ pixels of true class c, from prediction $k \neq c$ to prediction c, take $d_{cc} = \delta$ and $d_{ck} = -\delta$. Then

$$\frac{\langle J(\bar{n}), d \rangle}{\delta / P} = \frac{P}{K} \left(\frac{1}{\bar{D}_c} + \mathbf{1}\{k \in \mathcal{P}\} \frac{\bar{n}_{kk}}{\bar{D}_k^2} \right).$$

The indicator drops class k when it is not among the evaluated classes. Because rare classes have small unions, correcting their pixels generally carries more weight. A general change combines signed corrections with transfers between two wrong predictions. A transfer can change mIoU while leaving accuracy unchanged, so the general ratio need not be a convex combination of correction rates. The conversion therefore depends on the confusion matrices and on the direction of the change, not on accuracy alone.

Finally, a_T and a_Π are measured on different pixels and for different predictors. If $\Delta a_T \neq 0$ as well, multiplying and dividing by Δa_Π gives $\Delta\mathrm{mIoU}/\Delta a_T = \kappa\, \Delta a_\Pi / \Delta a_T$. The Jacobian determines the first factor, the metric conversion, but not the second, $\Delta a_\Pi / \Delta a_T$, which is a training response. The fitted β_q is therefore an empirical description of the full response.

□

B Experimental Details

B.1 Measurement and Estimation

Algorithm S1 records the probe error field and the model state every τ steps, and all aggregates are evaluated on the same probe pixels. Moments are taken over the $|\Pi|$ probe pixels, $\mathrm{Var}[x] := \frac{1}{|\Pi|} \sum_{u \in \Pi} (x(u) - \langle x \rangle)^2$, with $\mathrm{Corr}(x, y)$ the corresponding Pearson correlation, and pair statistics are averaged over the analysis window, which for the decay sweeps is a matched lag window of 3,330 training steps (333 recording intervals). Centering each field on Π removes its mean error level, so a trend in accuracy does not by itself inflate a correlation. The raw-field branches of the algorithm measure the variance reduction of the evaluated predictor directly, and the single-axis and crossed-model fits only summarize the measured covariances; they do not replace the direct measurements. The measured N_{eff} has a standard deviation across seeds of 0.013 on Pascal VOC and 0.004 on Cityscapes.

Algorithm S1 Probe measurement of ρ and N_{eff}

Input: probe set Π; units $k \in \mathcal{K}$ (snapshots in the analysis window, networks, views) with parameters θ_k and error fields $e_k = 1 - p_k(\text{true class})$ on Π; rules $R \in \mathcal{R}$ with pools $\mathcal{K}_R \subseteq \mathcal{K}$ and weights $w_{R,k}$

Output: correlations $\hat{\rho}_{kl}$; $N_{\text{eff}}(R)$ for each R; $(\hat{\rho}_\infty, \hat{r})$; shares $\{\hat{s}_S\}$

for each required pair of distinct units (k, l) **do**
  $\hat{\rho}_{kl} \leftarrow \mathrm{Corr}_\Pi(e_k, e_l)$
end for
for each $R \in \mathcal{R}$ **do**
  $\sigma_R^2 \leftarrow \mathrm{mean}_{k \in \mathcal{K}_R} \mathrm{Var}_\Pi[e_k]$
  if R averages probabilities **then**
    $\bar{e}_R \leftarrow \sum_{k \in \mathcal{K}_R} w_{R,k}\, e_k$
  else
    $\bar{\theta}_R \leftarrow \sum_{k \in \mathcal{K}_R} w_{R,k}\, \theta_k$; $\bar{e}_R \leftarrow 1 - p_{\bar{\theta}_R}(\text{true class})$ on Π
  end if
  $N_{\text{eff}}(R) \leftarrow \sigma_R^2 / \mathrm{Var}_\Pi[\bar{e}_R]$
end for
$\hat{\rho}(\Delta) \leftarrow$ mean $\hat{\rho}_{kl}$ over within-trajectory pairs at lag Δ; fit $(\hat{\rho}_\infty, \hat{r})$ of Assumption 4.1
$\hat{\rho}(D) \leftarrow$ mean $\hat{\rho}_{kl}$ over pairs differing exactly on D, using fitted asymptotes if step $\in D$
$\{\hat{s}_S\} \leftarrow$ Möbius inversion of $\{\hat{\rho}(D)\}$ by Theorem 4.5

Table S1: Experimental settings at the published schedules unless noted. † Joint interventions and oracle sweeps use five seeds. ‡ Classification network ensembles use six runs on one labeled split.

Dataset	Labeled	Crop	Schedule	Batch	Seeds
Mean Teacher (Tarvainen & Valpola, 2017), CPS (Chen et al., 2021) and view TTA (Shanmugam et al., 2021); ResNet-101, DeepLab-v2 (Chen et al., 2017)					
Pascal VOC 2012	1/8 (1,323)	321^2	40 epochs	10	3 / 5†
	1/8 (1,323)	321^2	20 epochs	4	3
	1/16 (662), 1/4 (2,646)	321^2	40 epochs	10	3
Cityscapes	1/8 (372)	256×512	40 epochs	4	3 / 5†
	1/16 (186), 1/4 (744)	256×512	40 epochs	4	3
FixMatch, FreeMatch and SoftMatch (Sohn et al., 2020; Wang et al., 2022b; Chen et al., 2023); ViT-S/2, pretrained					
CIFAR-100	1/250 (200)	32^2	204,800 iters	8	3 / 6‡
DualFete (Yi et al., 2026); V-Net, and 3D U-Net as second architecture					
Pancreas-CT	1/10 (6)	96^3	15,000 iters	4	3
Left Atrium	1/20 (4)	$112 \times 112 \times 80$	15,000 iters	4	3
UniMatch V2 (Yang et al., 2025); DINOv2-S/14, fine-tuned, DPT head, and UPerNet head as second architecture					
ADE20K	1/8 (2,526)	518^2	60 epochs	8	3
COCO	1/32 (3,697)	518^2	20 epochs	16	3

B.2 Experimental Configurations and Datasets

Implementation Details. Table S1 lists the experimental configurations and Table S2 the optimization settings. We use the published hyperparameters of each method and vary only the EMA decay λ, the coupling weight w_{cpl}, the view scale range w_{scale}, the oracle fraction γ, and the random seed. Experiments are run on four NVIDIA RTX A5000 GPUs (24 GB) and two NVIDIA RTX A6000 GPUs (48 GB). The error field is recorded every 10 steps on 8,192 pixels drawn once from the validation batches in the segmentation experiments, every 200 steps on 4,095 images in the classification experiments, and every 50 steps on 16,384 voxels on Pancreas-CT and Left Atrium. The view axis is evaluated offline on stored checkpoints.

Table S2: Optimization settings of each method, taken from the configurations of the reported runs. [a] Backbone at $0.1\times$. [b] Layer-wise decay 0.5. [c] Head at $40\times$. [d] Swept around the reference $\lambda = 0.99$.

Method	Optimizer	Learning rate	Schedule	Weight decay	Threshold	EMA
Mean Teacher	Adam	3×10^{-5} [a]	constant	–	0.97	swept[d]
CPS	Adam	3×10^{-5} [a]	constant	–	–	–
FixMatch	AdamW	5×10^{-4} [b]	warm-up	5×10^{-4}	0.95	–
FreeMatch, SoftMatch	AdamW	5×10^{-4} [b]	warm-up	5×10^{-4}	adaptive	–
DualFete	SGD	10^{-2}	poly	10^{-4}	–	–
UniMatch V2	AdamW	5×10^{-6} [c]	poly	0.01	0.95	0.996

Correlations between snapshots are estimated over the second half of training, with lags restricted to at most 3,330 steps in the decay sweeps (Appendix B.1 and Algorithm S1). Since the EMA decay also changes the teacher that supervises the student, we isolate the readout effect of N_{eff} by comparing each EMA, computed post hoc on stored snapshots, with a single delayed snapshot of the same mean lag (Equation (14)), which has $N_{\text{eff}} = 1$ at any lag. The main experiments use three random seeds on the $1/8$ labeled splits of Pascal VOC 2012 and Cityscapes and follow the original implementations; all datasets are described below.

Seeds are independent replicates. Tables report mean $\pm$ standard deviation (SD) across seeds and figures report mean ±1 standard error (SE), unless a caption states otherwise; SEM denotes the SE of a seed mean, computed as $\mathrm{SD}/\sqrt{n}$ with the sample SD and n seeds. Gains are relative to a single member at the same setting unless otherwise stated. mIoU, Dice, and top-1 accuracy are reported as percentages, and target accuracy a_T is a fraction, so β_q is expressed in mIoU points per unit fraction. Derived quantities are computed from unrounded inputs.

Pascal VOC 2012. Pascal VOC 2012 (Everingham et al., 2010) is a semantic segmentation benchmark with 20 object classes and a background class. Its standard segmentation split provides 1,464 training and 1,449 validation images. Following common practice, we add the annotations of Hariharan et al. (2011), which gives an augmented training set of 10,582 images. Our main experiments use the $1/8$ split with 1,323 labeled images and treat the remaining training images as unlabeled.

Cityscapes. Cityscapes (Cordts et al., 2016) is a semantic segmentation benchmark of urban street scenes with high-resolution images and fine pixel annotations. The standard evaluation uses 19 semantic classes, and the finely annotated subset contains 2,975 training, 500 validation, and 1,525 test images of 1024×2048 pixels. Our main experiments use 372 labeled training images, and we additionally evaluate the $1/16$ and $1/4$ splits.

CIFAR-100. CIFAR-100 (Krizhevsky et al., 2009) is an image classification benchmark of 60,000 RGB images of 32×32 pixels spanning 100 classes, with 500 training and 100 test images per class. Our semi-supervised classification experiments use 200 labeled images and follow the USB configurations (Wang et al., 2022a) for FixMatch, FreeMatch, and SoftMatch.

ADE20K. ADE20K (Zhou et al., 2017) is a scene parsing benchmark with 150 classes whose training split contains 20,210 annotated images, a much larger label space than Pascal VOC 2012 or Cityscapes. Following the semi-supervised protocol of UniMatch V2 (Yang et al., 2025), we use the $1/8$ split with 2,526 labeled images and treat the remaining training images as unlabeled.

MS COCO. MS COCO (Lin et al., 2014) is a large-scale benchmark of objects in everyday scenes. We adopt the COCO 2017 semantic segmentation setting of UniMatch V2 (Yang et al., 2025), which contains 118,287 training images and 81 classes including background. Our experiments use the $1/32$ split with 3,697 labeled images, and the remaining training images provide the unlabeled data.

Pancreas-CT. Pancreas-CT (Roth et al., 2015; 2016) contains 82 abdominal CT scans with pancreas annotations, and the task is to segment the pancreas in each three-dimensional scan. Following DualFete (Yi et al., 2026), we use 62 scans for training and 20 for testing, and we adopt the 10%

labeled setting, in which six training scans retain their annotations and the remaining scans serve as unlabeled data.

Left Atrium. Left Atrium (LA) (Xiong et al., 2021) is a cardiac MRI benchmark for three-dimensional segmentation of the left atrium. Following the semi-supervised protocol of DualFete (Yi et al., 2026), we use 100 annotated MRI volumes split into 80 training and 20 test volumes, and we evaluate the 5% labeled setting with four annotated training volumes, with the remaining training volumes treated as unlabeled.

C Sensitivity to Schedule and Label Fraction

This appendix splits the total gain from raising the EMA decay into a variance part, due to the change in N_{eff} of the average, and a training part, due to the change in the trajectory that the teacher induces, and uses this split to ask which of our measurements are properties of the method and which are properties of the particular run on which we calibrated.

Training Schedule. Halving the training schedule leaves the mechanism intact (Table S3): at both schedules, N_{eff} saturates near the fitted upper bound, mIoU peaks before N_{eff} does, the readout term accounts for the gain from raising the decay, and the training gain is close to zero. The schedule does, however, change the constants. The readout rate rises from 13.3 to 17.3 mIoU per unit of N_{eff}, so the same variance reduction buys more performance in a model that is further from convergence. The lower bound is therefore a stable target for measurement, whereas the rate that converts it into mIoU must be recalibrated whenever the schedule changes, which also limits how far the rate transfers in Section 5.2.

Table S3: Effect of training length at different EMA decays on Pascal VOC. Gains come from separate seed-paired runs, not from differences of the rows; the variance gain uses $\Delta N_{\text{eff}} = 0.041$ (half) and 0.039 (full), measured on one trajectory, whereas the N_{eff} column compares separate runs whose trajectories also differ. $\pm$ gives nominal 95% half-widths over seed-paired differences.

	Half schedule			Full schedule		
λ	ρ_∞	N_{eff}	mIoU (%)	ρ_∞	N_{eff}	mIoU (%)
0	.863	1.000	62.71	.862	1.000	66.89
0.9	.860	1.029	64.25	.863	1.029	67.63
0.99	.865	1.086	65.29	.875	1.074	68.25
0.997	.867	1.105	65.64	.875	1.095	68.95
0.999	.868	1.138	**66.21**	.882	1.097	**69.32**
0.9997	.865	**1.166**	65.93	.887	**1.122**	69.19
Readout rate β_v		17.3			13.3	
Total gain (mIoU)		$+0.70 \pm 0.32$			$+0.56 \pm 0.32$	
Variance gain		$+0.71$			$+0.52$	
Staleness gain		$+0.13$			-0.02	
Training gain		-0.14 ± 0.41			$+0.06 \pm 0.27$	

Labeled Fraction. Doubling or halving the labeled fraction changes neither the lower bound nor the rate, and the gain still arrives through the same channel (Table S4). Across a fourfold range of labels, ρ_∞ moves by less than 0.01, N_{eff} stays below its upper bound and approaches it at $\lambda = 0.999$, and the readout rate stays close to its value at the $1/8$ split. The split between the two channels is instead a property of the dataset.

Table S4: Effect of label fraction on correlations, effective sample size, and gains. Gains come from the seed-paired runs of Table S3, not from differences of the columns; the variance gain uses $\Delta N_{\text{eff}} = .035, .039, .029$ (Pascal) and .012 (Cityscapes), measured on one trajectory rather than across the separate runs in the N_{eff} columns. The remainder of the total is staleness (Table S3). ρ_∞ is measured at λ=.999 over the full lag window. $\pm$ gives nominal 95% half-widths over seed-paired differences.

			N_{eff}		mIoU (%)			Gain from λ=.99 to .999		
Split	ρ_∞	$1/\rho_\infty$	.99	.999	λ=0	.99	.999	Total	Variance	Training
Pascal VOC 2012 ($\beta_v = 14.7, 13.3, 13.9$)										
1/16	.883	1.13	1.063	1.090	65.2	67.3	67.9	+.63±.19	+.52	+.11±.12
1/8	.875	1.14	1.074	1.097	66.9	68.3	69.3	+.56±.32	+.52	+.06±.27
1/4	.878	1.14	1.060	1.107	68.3	69.6	70.3	+.75±.47	+.40	+.27±.37
Cityscapes ($\beta_v = 20.9, 19.0, 20.1$)										
1/16	.952	1.05	1.017	1.020	53.5	54.9	54.9	+.06±1.54	+.25	−.09±1.62
1/8	.948	1.05	1.021	1.027	59.1	60.1	60.8	+.68±.40	+.22	+.43±.32
1/4	.947	1.06	1.028	1.034	62.3	63.4	64.0	+.65±.21	+.24	+.33±.21

D Variance in Multi-Axis Decompositions

D.1 Network and Step

Network Gain after Step Averaging. Let $V_n = V(M, 0)$, $V_s = V(1, \lambda)$, and $V_{ns} = V(M, \lambda)$. We measure the fraction of the network-averaging gain that is retained after step averaging as

$$Q_N = \frac{V_s/V_{ns} - 1}{1/V_n - 1} = \frac{s_a + s_b B(\lambda, r_b)}{s_a + s_b} \frac{V_n}{V_{ns}}. \tag{19}$$

In the second form, s_a and s_b are the network-private shares of Equation (15), constant and varying across steps respectively, and $B(\lambda, r_b)$ is the factor by which step averaging with decay λ scales s_b. Step averaging leaves s_a intact and removes part of s_b, so Q_N depends mainly on $s_a/(s_a + s_b)$, the part of the network-private variance that step averaging cannot remove.

Tables S5 and S6 report the four variance shares estimated from the raw student predictions, the model-implied upper bounds, and Q_N at three decays. On Cityscapes, the predictions of Equation (16) track the measured Q_N to within 0.019 at $\lambda \leq 0.999$ and within 0.039 at the widest kernel; the fit is looser on Pascal VOC. Both tables lead to the same conclusions. Coupling moves variance out of s_a and into the shared share s_g, which lowers $s_a/(s_a+s_b)$, so Q_N is lower with coupling than without on both datasets. The same shift lowers the upper bound of network averaging and raises the upper bound of step averaging. Because s_b can be removed along either axis, the two single-axis gains sum to more than the joint gain in every row.

Table S5: Effect of coupling on variance shares, upper bounds and Q_N at $M = 2$ on Cityscapes. Components follow Equation (15). Upper bounds are the model-implied effective sample sizes of averaging steps alone, networks alone, and both.

	Variance share				Upper bound			Q_N, measured / predicted		
w_{cpl}	s_c	s_a	s_g	s_b	Step	Netw.	Joint	λ=.99	.999	.9997
0	.873	.081	.012	.034	1.048	1.130	1.145	.814/.821	.745/.743	.695/.734
1.5	.921	.017	.034	.029	1.066	1.047	1.086	.572/.590	.426/.416	.360/.396
6	.916	.008	.053	.023	1.082	1.032	1.092	.493/.508	.326/.308	.260/.286
12	.914	.006	.059	.021	1.087	1.028	1.094	.480/.499	.308/.290	.243/.266

D.2 Three Axes on UniMatch V2 and DualFete

Table S7 reports the absolute variance shares behind Figure 4a, b, which shows normalized shares, together with the implied upper bounds. Which axis holds the private variance differs between the

Table S6: Effect of coupling on variance shares, upper bounds and Q_N at $M = 2$ on Pascal VOC. Components follow Equation (15). Upper bounds are the model-implied effective sample sizes of averaging steps alone, networks alone, and both.

	Variance share				Upper bound			Q_N, measured / predicted		
w_{cpl}	s_c	s_a	s_g	s_b	Step	Netw.	Joint	λ=.99	.999	.9997
0	.852	.053	.019	.076	1.105	1.149	1.174	.672/.746	.505/.498	.415/.458
1.5	.848	.019	.076	.057	1.154	1.082	1.180	.524/.553	.346/.323	.274/.295
6	.845	.015	.088	.052	1.163	1.072	1.183	.507/.545	.315/.297	.236/.266
12	.865	.017	.077	.041	1.133	1.062	1.156	.549/.577	.380/.369	.293/.344

two methods: on DualFete, the view axis holds a large private share, whereas on UniMatch V2, the step axis is empty once the released EMA has converged. Across the two methods and four datasets, the summed single-axis maximum N_{eff} gains exceed the joint maximum gain by 9–61%.

Table S7: Variance shares and joint upper bounds across three averaging axes on Pancreas-CT, Left Atrium, ADE20K and COCO. s_S is removable along any axis in S.

Component or upper bound	Pancreas-CT	Left Atrium	ADE20K	COCO
$s_{\varnothing}$ (common)	.815	.741	.783	.856
s_{n} (network only)	.024	.022	.158	.071
s_{s} (step only)	.011	.009	.000	.000
s_{v} (view only)	.061	.089	.023	.050
s_{ns}	.040	.042	.000	.002
s_{nv}	.014	.022	.036	.019
s_{sv}	.003	.001	.000	.000
s_{nsv}	.032	.074	.000	.001
Network upper bound	1.123	1.190	1.241	1.104
Step upper bound	1.094	1.145	1.001	1.004
View upper bound	1.124	1.229	1.063	1.076
Joint upper bound	1.228	1.350	1.278	1.168
Excess of summed maximum N_{eff} gains	50%	61%	10%	9%

E JOINT INTERVENTIONS ON BOTH CHANNELS

Local Linear Approximation. Model 4.7 is linear in N_{eff}, although smooth readout responses such as $1 - 1/N_{\text{eff}}$, $1 - N_{\text{eff}}^{-1/2}$, or $\log N_{\text{eff}}$ are curved. The upper bound on N_{eff} makes the linear form adequate: for fixed $\rho > 0$, the available range $N_{\text{eff}} - 1$ is at most $1/\rho - 1$, which is 0.16 at the fitted step upper bound on Pascal VOC and 0.08 at the coupled-network upper bound, and over such a range the three responses are indistinguishable from a straight line given our contrasts. We anchor the gain at zero by comparing an aggregate with itself, since perfectly correlated predictors can still differ in quality.

Factorial Test. We test Model 4.7 with a 2×2 factorial design in the EMA decay λ and the oracle fraction γ, with reference cell $(\lambda_0, \gamma_0) = (0.99, 0)$. For a quantity X measured in the four cells, the interaction is

$$\mathcal{D}X = X_{\lambda_1,\gamma_1} - X_{\lambda_1,\gamma_0} - X_{\lambda_0,\gamma_1} + X_{\lambda_0,\gamma_0}, \qquad \widehat{I} = \widehat{\beta}_v \, \mathcal{D}N_{\text{eff}} + \widehat{\beta}_q \, \mathcal{D}a_T.$$

The model allows $\mathcal{D}\,\text{mIoU} \neq 0$, since λ and γ can interact in their effects on N_{eff} and a_T, but it predicts $\mathcal{D}\,\text{mIoU} \approx \widehat{I}$. We also re-estimate the target rate at λ_1 as $(\Delta\text{mIoU} - \widehat{\beta}_v \Delta N_{\text{eff}})/\Delta a_T$. Table S8 confirms both predictions in all four cells: the observed interaction lies within 2SE of $\widehat{I}$, and the target rate at λ_1 lies within 2SE of its value at λ_0. Changing the EMA decay therefore leaves the target rate unchanged, and the interaction in mIoU is carried by N_{eff} and a_T.

Table S8: Effect of EMA decay on factorial interactions and target rates. 2SE is twice the SE of each difference, including rate uncertainty for interactions.

		Interaction (mIoU)			Target rate β_q (mean $\pm$ SEM)		
Dataset	γ	Observed	$\widehat{I}$	2SE	Reference	Alternative	2SE
Pascal VOC	0.25	-0.27	-0.43	0.42	101.4 ± 12.1	113.7 ± 8.7	29.9
	1	-0.10	-0.06	0.54	98.3 ± 5.1	96.2 ± 5.6	15.2
Cityscapes	0.35	$+0.20$	-0.02	1.16	2327 ± 652	1986 ± 174	1349
	0.7	$+0.35$	$+1.14$	0.94	1831 ± 186	1491 ± 151	478

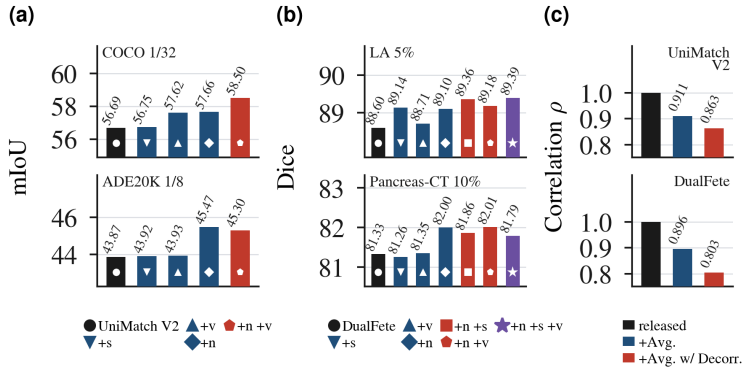


Figure S1: Expanded averages of UniMatch V2 and DualFete. (a) Performance of UniMatch V2 on COCO $1/32$ and ADE20K $1/8$. (b) Performance of DualFete on Left Atrium 5% and Pancreas-CT 10%. Markers identify each pool by the axes it averages over (s: step, n: network, v: view). (c) Mean pairwise correlation ρ of the released model ($\rho = 1$ by definition), of the network average ($+$Avg.), and of the network pool with members of a second architecture ($+$Avg. w/ Decorr.).

F Evaluation on State-of-the-Art Methods

Figure S1 summarizes the performance of the expanded averages on top of UniMatch V2 and DualFete. On UniMatch V2, the net gains closely match the aggregation gains of Figure 4c because the member term of Equation (10) is near zero. On DualFete, the member term dominates instead: step averaging shares 84% and 92% of its removable variance with network averaging on Pancreas-CT and Left Atrium (Table S7), so adding snapshots to the network pool adds only 0.06 Dice of aggregation gain, and the net gain moves by -0.14 and $+0.26$ Dice depending on the quality of the snapshots. A second architecture lowers the mean correlation of the network pool from 0.91 to 0.86 on UniMatch V2 and from 0.90 to 0.80 on DualFete (Figure S1c).

F.1 Qualitative Results

Figures S2 and S3 show where the gains of Section 5.3 come from. From left to right, the member-specific errors (orange) shrink while the errors shared by every member (dark blue) stay nearly unchanged, which is the mechanism of Theorem 4.5: averaging removes the variance that its members do not share and leaves the shared part as a lower bound. The effect is clearest on Pancreas-CT and Left Atrium, where the dark blue regions are almost identical across the three columns. Expanding the network average with the decorrelated members further shrinks the orange errors.

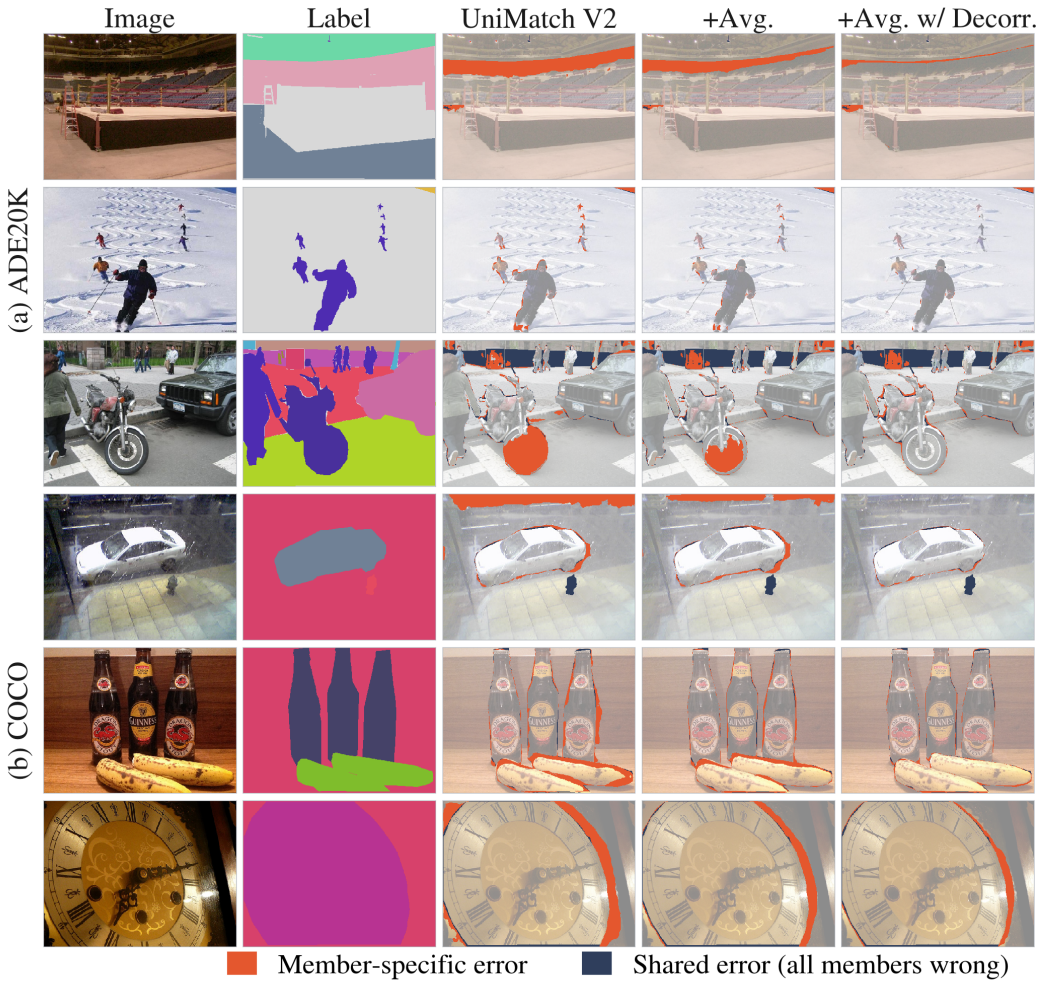


Figure S2: Error maps of UniMatch V2 on (a) ADE20K and (b) COCO. "+Avg." denotes UniMatch V2 with expanded members, and "+Avg. w/ Decorr." denotes the pool whose upper bound on N_{eff} is raised with decorrelated members. Orange marks errors that not all members share, and dark blue marks errors that every member makes.

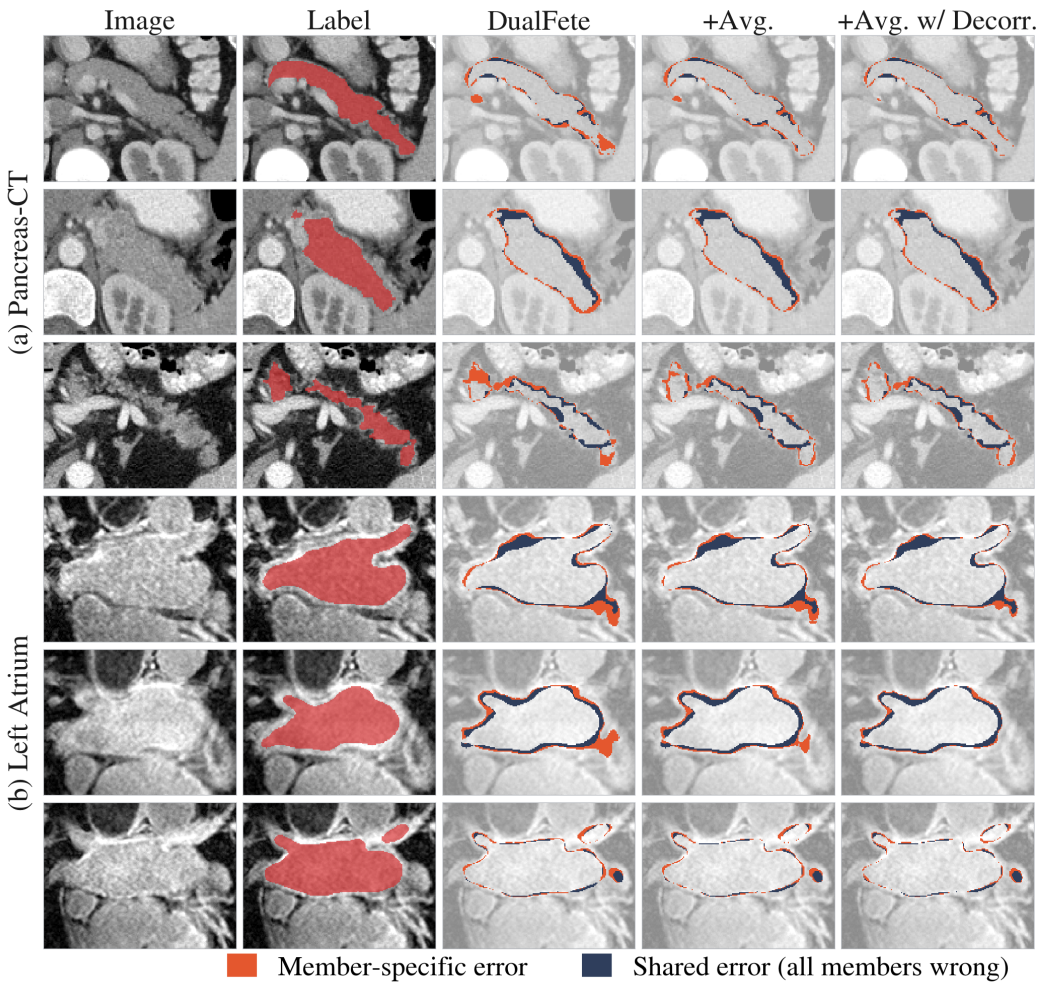


Figure S3: Error maps of DualFete on (a) Pancreas-CT and (b) Left Atrium. "+Avg." denotes DualFete with expanded members, and "+Avg. w/ Decorr." denotes the pool whose upper bound on N_{eff} is raised with decorrelated members. Orange marks errors that not all members share, and dark blue marks errors that every member makes.