Understanding the Limits of Vision Test-Time Scaling: Path Redundancy, Instance Difficulty, and Adaptive Compute

Anonymous Author(s)

Abstract

Test-time scaling has shown strong gains in language reasoning, yet its behavior in vision remains poorly understood. We present one of the first systematic studies of vision test-time scaling through CLIP-based multi-path inference, where computation is increased via prompt ensembles and test-time augmentations. Our results show that additional inference paths improve accuracy in early regimes but rapidly exhibit diminishing returns. Through correlation analysis, we demonstrate that strong path redundancy limits the effective value of additional computation. We further show that compute gains concentrate on high-uncertainty samples, motivating adaptive inference strategies. Although entropy-based adaptive stopping approaches favorable compute-accuracy trade-offs, our analysis reveals substantial remaining efficiency headroom. Overall, our findings suggest that the primary bottleneck of vision test-time scaling is not computation itself, but the lack of informational diversity across inference paths.

Keywords: Vision Test-Time Scaling, Test-Time Compute, Multi-Path Inference, CLIP, Zero-Shot Classification, Adaptive Inference, Path Diversity, Inference Redundancy, Compute-Accuracy Trade-offs, Information Scaling

1. Introduction

Recent progress in large language models suggests that increasing test-time computation can significantly improve reasoning performance. Such improvements are often attributed to exploring multiple reasoning trajectories and aggregating their outputs. Whether similar scaling behavior exists for vision models, however, remains unclear.

In vision, test-time scaling is typically realized through prompt ensembling, test-time augmentation, or multi-path inference. While these techniques are widely used in practice, we still lack a principled understanding of how performance scales with additional compute, why gains eventually saturate, and which factors determine the value of extra inference paths. This gap limits our ability to design efficient and scalable vision inference systems.

In this work, we present a systematic study of vision test-time scaling using CLIP multi-path inference. Unlike language reasoning, where additional computation may explore diverse reasoning chains, vision inference paths often remain highly correlated. This raises a central question: is performance determined primarily by the amount of compute or by the diversity of information introduced by additional paths?

We investigate three fundamental questions: (i) what scaling behavior emerges as test-time compute increases; (ii) what mechanisms drive diminishing returns; and (iii) whether compute can be allocated adaptively according to instance difficulty. Our empirical analysis reveals that scaling gains are real but strongly limited by path redundancy, and that additional computation mainly benefits difficult, uncertain samples.

These findings motivate a shift in perspective from *compute scaling* toward *information scaling*, where the diversity of inference paths becomes the key factor governing efficiency and performance.

Contributions

- We provide one of the first systematic empirical studies of vision test-time scaling using CLIP multi-path inference.
- We show that performance improvements diminish due to strong path redundancy and high inter-path correlation.
- We demonstrate that additional compute primarily benefits difficult, high-uncertainty instances.
- We analyze adaptive inference and show that uncertainty-based compute allocation approaches favorable compute–accuracy trade-offs.
- We highlight informational diversity as a central factor for future progress in vision test-time scaling.

2. Method

2.1. Problem Formulation

We formulate vision test-time scaling as prediction under a test-time compute budget. Given an input image x, inference is performed through a set of computation paths $\mathcal{P} = \{p_1, \ldots, p_N\}$, where each path corresponds to a different combination of prompt template and image augmentation. Each path produces a predictive distribution $p^{(i)}(y|x)$

over labels.

The goal of test-time scaling is to improve prediction quality by increasing the number of inference paths while respecting a compute budget. Formally, we study the trade-off between accuracy and expected compute:

$$\max \mathrm{Acc} \quad \text{s.t.} \quad \mathbb{E}[N] \leq C.$$

This formulation allows us to analyze when additional computation provides new information versus when it becomes redundant.

2.2. Multi-Path Test-Time Scaling

Each inference path combines (i) a prompt template and (ii) an image augmentation, producing diverse views of the same input. Predictions are aggregated by averaging:

$$\bar{p}(y|x) = \frac{1}{N} \sum_{i=1}^{N} p^{(i)}(y|x),$$

where N denotes the test-time compute budget. Increasing N corresponds to scaling inference-time computation.

We hypothesize that scaling effectiveness depends not only on the number of paths but also on their informational diversity. When paths are highly correlated, additional computation yields limited gains.

2.3. Adaptive Compute

Uniformly allocating the same number of inference paths to all samples may be inefficient, since easy examples often require less computation than difficult ones. To address this, we adopt an adaptive inference strategy that dynamically determines the number of paths per sample.

After each aggregation step n, we compute predictive entropy:

$$H(\bar{p}_n) = -\sum_{y} \bar{p}_n(y|x) \log \bar{p}_n(y|x).$$

Inference stops early when uncertainty falls below a threshold:

$$H(\bar{p}_n) < \delta \Rightarrow \text{stop}.$$

This uncertainty-based policy aims to allocate more compute to hard samples while reducing redundant computation on easy ones. Although simple, this formulation allows us to study how uncertainty interacts with scaling behavior and compute efficiency.

2.4. Effective Diversity (Conceptual Definition)

While increasing the number of inference paths increases compute, the usefulness of additional paths depends on how much new information they contribute. We therefore introduce the notion of *effective diversity*, defined conceptually as the amount of non-redundant information between inference paths.

Intuitively, when paths are highly correlated, effective diversity remains low even as compute grows, leading to diminishing returns. This perspective suggests that scaling behavior is governed not only by path count but also by the diversity of predictions produced by those paths.

3. Related Work

Test-time scaling in language models. Recent advances in large language models show that increasing test-time computation through sampling, self-consistency, or explicit reasoning trajectories can substantially improve performance [12, 17–20]. These approaches rely on the idea that additional computation explores diverse reasoning paths, and aggregation reduces reasoning errors. Our work is inspired by this perspective but investigates whether similar scaling behavior holds in vision models, where inference paths may exhibit stronger correlations.

Vision-language models and prompting. Large-scale vision–language models such as CLIP [13] enable zero-shot recognition through text prompts. Subsequent work studies prompt engineering and prompt learning for improving adaptation and robustness [11, 21, 22]. These methods demonstrate that multiple prompts can provide performance gains, but they do not analyze scaling behavior as test-time compute increases or examine redundancy between inference paths.

Test-time augmentation and ensembling in vision. Test-time augmentation (TTA) and ensembling are widely used to improve model robustness and accuracy [5, 14, 15]. Recent work explores augmentation policies and multi-view aggregation for improved generalization [2, 6]. However, prior studies primarily report empirical gains rather than investigating the limits of scaling with additional inference paths.

Adaptive inference and dynamic compute allocation. Dynamic inference aims to reduce computation by allocating resources based on prediction difficulty. Early-exit architectures and confidence-based policies [8, 10, 16] allow easier samples to exit earlier, improving efficiency. More recent approaches consider adaptive computation in transformers and efficient vision architectures [1, 4]. Our work differs in focusing on adaptive inference within multi-path test-time scaling and analyzing how path diversity affects the compute–accuracy trade-off.

Understanding scaling and diversity. Scaling laws have been studied extensively in language and vision models [7, 9]. In ensemble learning, diversity is known to be critical for performance improvements [3]. Our work connects these ideas by showing that test-time scaling in vision is fundamentally constrained by path correlation, suggesting a shift from compute scaling toward information scal-

ing.

Position of our work. Unlike prior studies that primarily optimize performance through prompting, augmentation, or adaptive inference, we provide a systematic analysis of vision test-time scaling. We highlight path redundancy as a key mechanism behind diminishing returns and introduce informational diversity as a conceptual lens for understanding scaling efficiency.

4. Experiments

4.1. Setup

Model. We use CLIP (*ViT-B/32*, OpenAI pretrained) as the base vision–language model.

Dataset. Experiments are conducted on CIFAR-10 zero-shot classification, which provides a controlled benchmark for isolating test-time scaling behavior without additional training.

Inference Paths. Each path is defined by a unique combination of prompt template and image augmentation. Unless otherwise stated, paths are sampled uniformly from a predefined path pool.

Path sampling variability. Because inference paths are sampled from a prompt–augmentation pool, different paths may vary in quality. As a result, adding an additional path can occasionally introduce noisy or redundant evidence, causing small non-monotonic fluctuations at low compute budgets. These fluctuations reflect sampling variability rather than contradictions to the overall scaling trend.

Compute Budgets. We evaluate fixed test-time budgets $N \in \{1, 2, 4, 6\}$ corresponding to increasing numbers of inference paths.

Adaptive Policy. Adaptive inference uses entropy-based early stopping with threshold δ.

Evaluation Metrics. We report classification accuracy, average inference steps, path agreement, and pairwise logit correlation to quantify both performance and diversity.

Figure 1 summarizes the overall pipeline and the key hypotheses of this work.

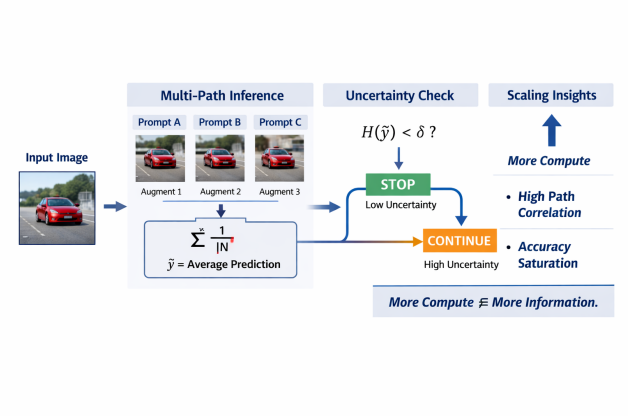

Figure 1. Overview of vision test-time scaling. Multiple inference paths (prompt + augmentation) are aggregated at test time. Increasing compute improves accuracy initially, but high path correlation leads to diminishing returns. Adaptive inference allocates compute based on uncertainty.

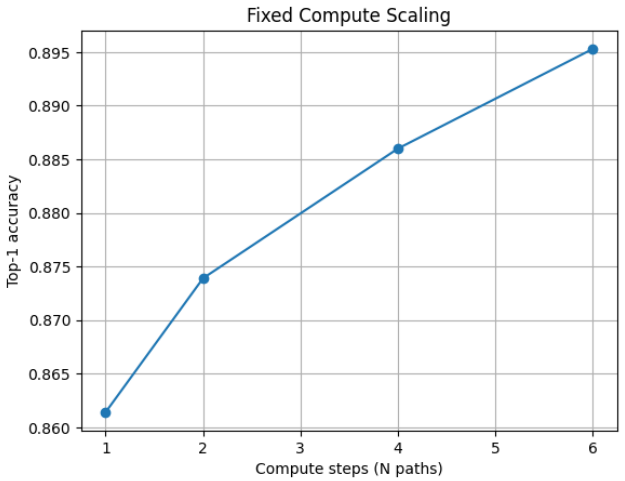

Figure 2. Fixed test-time compute scaling. While performance generally improves with additional paths, small local fluctuations may appear due to path-quality variability. Overall improvements exhibit diminishing returns as compute increases.

5. Results

5.1. Fixed Compute Scaling Improves Accuracy

We first examine how performance changes under fixed test-time compute budgets. Figure 2 shows that increasing the number of inference paths improves accuracy from 0.8614 at $N = 1$ to 0.8953 at $N = 6$. Performance gains are strongest in early scaling regimes and gradually diminish as compute increases.

This behavior suggests that multi-path inference initially provides complementary evidence, but additional computation becomes increasingly redundant at larger budgets.

Local non-monotonicity at small budgets. In practice, we occasionally observe slight performance drops when moving between very small budgets (e.g., from $N = 1$ to $N = 2$). This occurs when the newly added path contributes noisy or highly correlated predictions that perturb the aggregated output. Importantly, performance recovers as compute increases, and the overall diminishing-return trend remains unchanged. This observation further highlights that scaling effectiveness depends on the informational quality of added paths rather than path count alone.

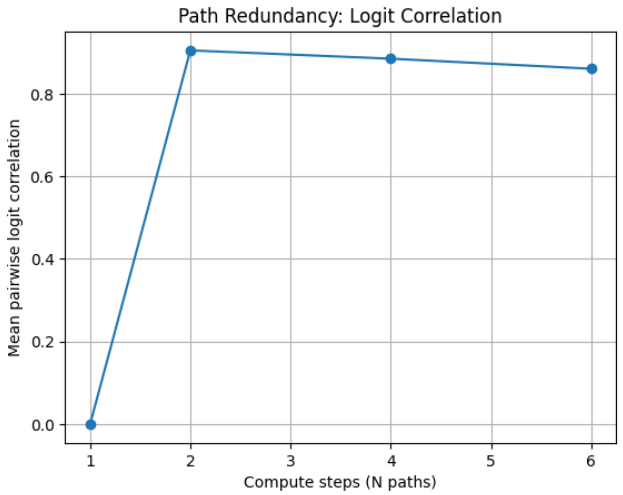

Figure 3. Path redundancy analysis. High pairwise logit correlation indicates limited diversity between inference paths.

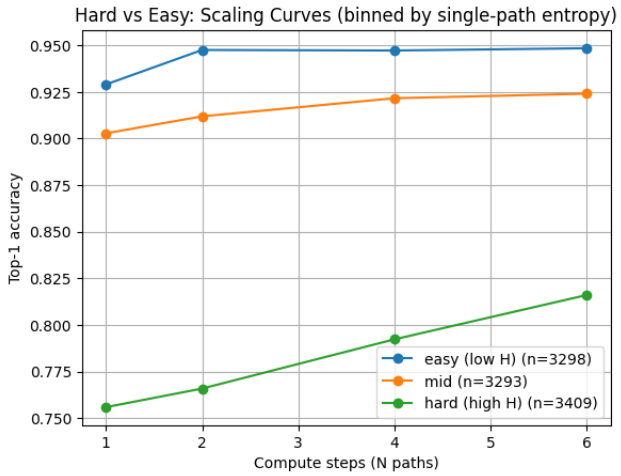

Figure 4. Scaling curves grouped by instance difficulty. Compute improvements are concentrated on hard samples.

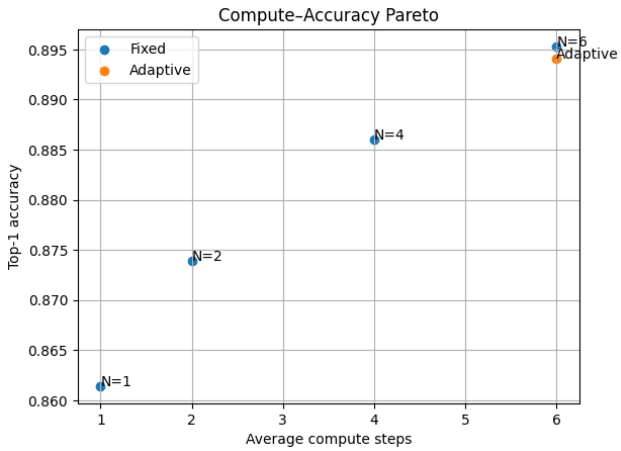

Figure 5. Compute–accuracy Pareto comparison between fixed and adaptive compute.

5.2. Path Redundancy Explains Diminishing Returns

To understand saturation, we analyze pairwise logit correlation across inference paths. Figure 3 shows consistently high correlations ($\sim$ 0.86–0.90), indicating that different paths often produce highly similar predictions.

These results reveal that increasing the number of paths does not necessarily increase informational diversity. Instead, additional compute frequently revisits similar prediction trajectories, leading to diminishing marginal gains.

5.3. Compute Benefits Concentrate on Hard Samples

We further analyze whether scaling benefits different samples equally. Using single-path entropy, we partition samples into easy, medium, and hard subsets. Figure 4 shows that additional compute mainly improves performance on hard, high-uncertainty samples, while easy samples saturate quickly.

This observation suggests that uniform compute allocation is inherently inefficient and motivates adaptive inference strategies.

5.4. Adaptive Compute and Pareto Trade-offs

We next evaluate uncertainty-based adaptive inference. Figure 5 compares adaptive inference with fixed compute budgets and shows that adaptive inference achieves accuracy close to the largest fixed budget (0.8941 vs. 0.8953), approaching the compute–accuracy Pareto frontier.

The adaptive-step histogram (Figure 6) indicates that the current entropy threshold operates in a conservative regime, with many samples still reaching the maximum budget. Rather than indicating a failure of adaptive inference, this suggests that uncertainty alone may be insufficient as a stopping signal when inference paths are highly correlated.

Although adaptive inference does not yet reduce average compute under the current threshold, this result itself is informative: it suggests that uncertainty alone may be insufficient as a stopping signal when paths are highly correlated, motivating adaptive policies that jointly consider uncertainty and path diversity.

5.5. Effective Diversity (Conceptual Interpretation)

While increasing the number of inference paths increases compute, the usefulness of additional paths depends on how much new information they contribute. We therefore introduce the notion of *effective diversity*, defined conceptually

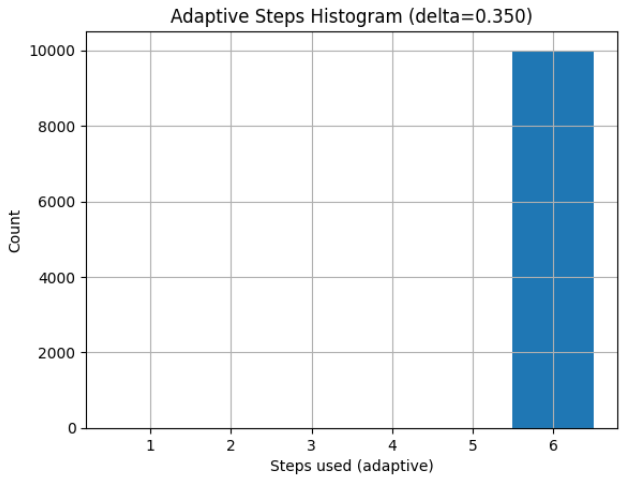

Figure 6. Distribution of adaptive inference steps. Most samples reach the maximum number of paths under the current entropy threshold, indicating a conservative operating point and motivating diversity-aware adaptive policies.

as the amount of non-redundant information between inference paths.

As a practical approximation, effective diversity can be related to prediction correlation, e.g., through a proxy such as $D_{\text{eff}} = 1 - \rho$, where ρ denotes mean pairwise logit correlation. This interpretation links diminishing returns directly to measurable redundancy between inference paths.

Intuitively, when paths are highly correlated, effective diversity remains low even as compute grows, leading to diminishing returns. This perspective suggests that scaling behavior is governed not only by path count but also by the diversity of predictions produced by those paths.

5.6. Quantitative Summary

Table **??** summarizes overall statistics. As compute increases, accuracy improves while path agreement decreases, indicating partial diversity between inference paths. However, persistently high logit correlation confirms that redundancy remains dominant, explaining the diminishing returns observed at larger compute budgets. These results suggest that scaling effectiveness is governed not only by the number of paths but also by the informational novelty introduced by each additional path.

5.7. Why Does Scaling Saturate? A Variance Reduction Perspective

Let $p_i(y|x)$ denote predictions from each path and

$$\bar{p}(y|x) = \frac{1}{N} \sum_{i=1}^{N} p_i(y|x).$$

If paths were independent, averaging would reduce variance proportionally to $1/N$. However, when paths are correlated, the variance becomes

$$\text{Var}(\bar{p}) \propto \frac{1}{N} + \rho \frac{N-1}{N},$$

where ρ denotes average pairwise correlation.

This formulation explains the observed saturation: once correlation remains high, additional paths provide limited new information. Therefore, scaling effectiveness depends not only on compute but also on the diversity of inference paths. Improving scaling efficiency may therefore depend more on reducing inter-path correlation than simply increasing the number of paths.

5.8. Summary of Findings

Our experiments reveal three key insights:

1. Vision test-time scaling improves accuracy primarily in early compute regimes.
2. High path correlation causes diminishing returns by limiting informational diversity.
3. Additional compute mainly benefits hard, high-uncertainty samples, motivating adaptive allocation.

These findings collectively suggest that the effectiveness of test-time scaling is determined more by information diversity than by raw compute.

6. Discussion

Our findings suggest that vision test-time scaling follows a fundamentally different regime from language reasoning. In language models, additional computation often explores diverse reasoning trajectories through stochastic decoding, whereas vision inference paths generated from prompts and augmentations frequently remain highly correlated.

One possible explanation is that language generation naturally expands the space of reasoning trajectories, while vision inference paths often explore nearby regions of the same decision space. As a result, additional compute in vision may increase redundancy more rapidly than informational diversity, leading to earlier saturation.

This difference implies that increasing compute alone is insufficient for sustained gains. Instead, the key bottleneck is informational diversity across inference paths. Future progress may therefore depend on diversity-aware path generation, learned path selection, or mechanisms that explicitly maximize complementarity between paths.

Our results also highlight the role of instance-level heterogeneity: additional compute primarily benefits difficult, high-uncertainty samples. This suggests that adaptive allocation is not only a systems-level optimization problem but also a statistical one, where compute should be treated as a resource conditioned on expected information gain.

An additional perspective is that test-time scaling may expose a mismatch between model capacity and inference diversity. Even when models possess sufficient representational power, inference strategies that repeatedly sample highly correlated paths may fail to fully utilize this capacity. Improving diversity at inference time may therefore be as important as scaling model size itself.

From a practical perspective, these results imply that future systems may benefit more from constructing diverse inference paths than from simply increasing the number of sampled paths under fixed strategies. Under realistic compute constraints, diversity-aware inference could offer a more efficient route to performance gains.

Beyond efficiency considerations, our findings also suggest implications for evaluation practices. If additional test-time compute primarily revisits similar evidence, performance gains may overestimate genuine improvements in robustness or reasoning ability. This highlights the importance of measuring not only accuracy, but also the diversity of evidence contributing to predictions.

More broadly, our study suggests a shift from *compute scaling* toward *information scaling*, where the effectiveness of additional computation is determined by how much new information it contributes rather than by compute alone. This perspective suggests that diversity should be treated as a first-class objective in both inference design and evaluation.

An additional practical insight is that small local fluctuations in scaling curves can arise from path-quality variability. This indicates that future systems may benefit from selecting inference paths based on expected information gain rather than uniformly sampling from a fixed pool.

7. Limitations

This study focuses on CLIP-based zero-shot inference on CIFAR-10, a controlled benchmark chosen to isolate test-time scaling behavior. Although this setting enables clean analysis, scaling dynamics may differ for larger vision–language models, higher-resolution datasets, or tasks requiring richer spatial reasoning and long-range dependencies.

Moreover, CIFAR-10 contains relatively low-resolution images and limited semantic complexity. In more challenging settings involving compositional reasoning or fine-grained recognition, the relationship between compute and effective diversity may differ. Evaluating such scenarios remains an important direction for future work.

Our adaptive inference strategy is based on a simple entropy threshold. While effective as a proof of concept, it remains conservative and does not explicitly account for path redundancy or expected information gain. More advanced policies that jointly model uncertainty and diversity may provide substantially larger efficiency improvements.

In addition, diversity in our experiments is induced through manually designed prompt templates and image augmentations. Such heuristic path construction may not fully explore the space of informative inference trajectories. Learning mechanisms that explicitly optimize path complementarity or effective diversity remains an important direction for future work.

Finally, our analysis focuses primarily on prediction-level statistics such as correlation and uncertainty. A deeper understanding of representation-level diversity and its relation to scaling behavior could provide stronger theoretical grounding. Furthermore, our evaluation emphasizes accuracy, which may not fully capture effects on calibration or robustness introduced by test-time scaling. Future studies should investigate how test-time scaling impacts uncertainty estimation, out-of-distribution generalization, and robustness under distribution shifts.

Our analysis is also conducted in a zero-shot setting. Whether similar scaling behavior persists under fine-tuning or instruction-tuned vision-language models remains an open question. We also note that our findings are based on a single model family, and validating these conclusions across newer multimodal architectures is an important direction for future work.

8. Conclusion

We present a systematic study of vision test-time scaling using CLIP multi-path inference. Our experiments show that additional test-time compute improves performance in early scaling regimes but rapidly exhibits diminishing returns. Through correlation analysis and variance-based reasoning, we identify path redundancy as a central bottleneck limiting scalability.

We further demonstrate that compute benefits concentrate on difficult, high-uncertainty samples, motivating adaptive allocation strategies. Although entropy-based stopping approaches favorable compute–accuracy trade-offs, substantial efficiency headroom remains when paths are highly correlated.

Taken together, our results indicate that the limiting factor of vision test-time scaling is not merely computation budget, but the quality and diversity of information introduced by additional inference paths. This perspective reframes test-time scaling as an information aggregation problem rather than purely a compute expansion problem.

More broadly, our findings suggest that future progress in vision test-time scaling may depend less on increasing raw computation and more on increasing the informational diversity of inference paths. This points toward a shift from *compute scaling* to *information scaling*, where path diversity and adaptive allocation become central principles for efficient vision inference.

We hope this work provides a foundation for future re-

search on diversity-aware inference, learned path planning, and adaptive test-time computation in vision models.

References

[1] Daniel Bolya et al. Token merging: Your vit but faster. In *ICLR*, 2023.

[2] Ekin D. Cubuk et al. Randaugment: Practical automated data augmentation. In *CVPRW*, 2020.

[3] Thomas Dietterich. Ensemble methods in machine learning. *Multiple Classifier Systems*, 2000.

[4] Yizeng Han et al. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021.

[5] Kaiming He et al. Deep residual learning for image recognition. In *CVPR*, 2016.

[6] Dan Hendrycks et al. Augmix: A simple data processing method to improve robustness. In *ICLR*, 2020.

[7] Jordan Hoffmann et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[8] Gao Huang et al. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2018.

[9] Jared Kaplan et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[10] Yigitcan Kaya et al. Shallow-deep networks. In *ICML*, 2019.

[11] Muhammad Uzair Khattak et al. Maple: Multi-modal prompt learning. *CVPR*, 2023.

[12] Takeshi Kojima et al. Large language models are zero-shot reasoners. *NeurIPS*, 2022.

[13] Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[14] Connor Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 2019.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[16] Surat Teerapittayanon et al. Branchynet: Fast inference via early exiting from deep neural networks. In *ICPR*, 2016.

[17] Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[18] Jason Wei et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[19] Shunyu Yao et al. Tree of thoughts: Deliberate problem solving with large language models. *NeurIPS*, 2023.

[20] Eric Zelikman et al. Star: Self-taught reasoner. *NeurIPS*, 2022.

[21] Kaiyang Zhou et al. Conditional prompt learning for vision-language models. *CVPR*, 2022.

[22] Kaiyang Zhou et al. Learning to prompt for vision-language models. *IJCV*, 2022.