Variant Generation and Diversity Strategies for Enhanced Variant Parallelism in Edge AI

Navidreza Asadi
Chair of Communication Networks
Technical University of Munich
Munich, Germany
navidreza.asadi@tum.de

Sagnik Dutta
Chair of Communication Networks
Technical University of Munich
Munich, Germany
sagnik.dutta@tum.de

Wolfgang Kellerer
Chair of Communication Networks
Technical University of Munich
Munich, Germany
wolfgang.kellerer@tum.de

***Abstract*—Serving deep learning models on IoT edge devices faces considerable challenges because of their restricted computational capabilities. Variant Parallelism (VP) mitigates this by distributing lightweight model variants, but has two main limitations: (i) variant generation typically lacks a systematic approach to ensure diversity, and (ii) ensembles composed of identical-sized variants underperform compared to heterogeneous ensembles. To address these issues, we propose Variant-Improved Parallelism (VIP), which advances VP through two main innovations. First, VIP introduces a structured variant generation pipeline that leverages knowledge distillation and targeted model slicing to efficiently produce diverse variants. Second, it employs diversity-inducing methods, such as data augmentation and class-imbalanced training, to enable effective homogeneous ensembles, expanding VP's utility in environments with similar device capabilities. Experiments on ResNet50 with CIFAR10, CIFAR100, and ImageNet-1K demonstrate that VIP achieves consistent accuracy gains of up to 10%, while preserving VP's advantages in fault tolerance and communication efficiency, supporting scalable and adaptive edge AI deployments.**

***Index Terms*—Distributed computing, Distributed vision networks, Machine learning algorithms, Diversity schemes, ensemble learning, parallel processing, Knowledge distillation, Distributed machine learning, variant parallelism, Edge AI, Collective intelligence, Edge computing**

I. Introduction

Deploying deep learning models on resource-constrained Internet of Things (IoT) devices in edge computing environments presents a significant challenge due to limitations in memory, processing power, and network bandwidth [1]–[6]. While techniques like model and class parallelism attempt to distribute computational load, they introduce substantial communication overhead and single points of failure, limiting their effectiveness in dynamic edge networks.

Variant Parallelism (VP) [2] offers a more robust alternative by deploying multiple, complete lightweight model variants that operate independently. This approach enhances fault tolerance and reduces communication by sharing only final predictions. Variant Parallelism bases its effectiveness on the ensemble learning principle, where the diversity among model variants leads to improved accuracy and robustness.

Challenge. While particularly effective in heterogeneous IoT environments such as smart home scenarios, VP faces two primary limitations: (1) variant generation is often empirical, and therefore, it lacks a systematic approach to ensure diversity, and (2) ensembles of same-sized (homogeneous) variants perform poorly compared to those with different-sized (heterogeneous) variants, restricting scalability in edge environments where devices have similar characteristics or available resources.

Contribution. To address these limitations, we introduce Variant-Improved Parallelism (VIP), a methodology that enhances VP with two improvements. First, we systematize variant creation by combining knowledge distillation with strategic model slicing for efficient generation of diverse variants. Second, we utilize diversity-inducing strategies, namely data augmentation and class-imbalanced training. These help generating effective homogeneous ensembles. Our evaluation across multiple datasets shows that VIP consistently improves accuracy over baseline models while retaining the core benefits of VP, enabling a more scalable and adaptive approach to edge AI.

II. Background and Related Work

A. Distributed Inference Approaches

Deploying deep neural networks on edge devices with limited resources has led to the development of various distributed inference methods. Each approach offers different balances between accuracy, communication requirements, and system robustness.

Model Parallelism. Traditional model parallelism partitions neural network layers across multiple devices, requiring sequential computation and intermediate activation exchange [7]–[11]. While this approach can handle models larger than any single device's capacity, it suffers from several limitations: (1) *Communication bottlenecks:* large activation tensors must be transmitted between devices, often exceeding the computational savings; (2) *Sequential dependencies:* the pipeline nature creates idle time as devices wait for inputs from previous stages; (3) *Fault vulnerability:* failure of any device in the pipeline halts the entire inference process; and (4) *Load balancing complexity:* uneven computational loads across layers lead to device under-utilization.

Class Parallelism. SensAI ConvNets [12] introduced class parallelism (CP), where the model is divided into subnets,

each specializing in different classes or class groups. Each device processes the full input but only predicts a subset of classes, reducing computational load per device. However, CP has fundamental scalability limitations: (1) *Class-dependent scalability:* the maximum number of devices is bounded by the number of classes in the dataset; (2) *Load imbalance:* uneven class distributions in real datasets lead to computational imbalances across devices; (3) *Accuracy degradation:* each subnet sees only partial class information during training, potentially reducing individual accuracy; and (4) *Deployment inflexibility:* adding new classes requires retraining and redistribution of all subnets.

Feature-based Distributed Inference. DISCO [10] proposed a method where devices compute local features and selectively communicate sparse representations to reduce bandwidth requirements. While promising for reducing communication costs, DISCO faces the challenge that determining optimal feature communication patterns is computationally complex (NP-hard), and the approach still maintains dependencies between devices that can create communication bottlenecks.

B. Variant Parallelism Fundamentals

Variant Parallelism [2], [13] represents a paradigm shift from partitioning-based approaches by generating multiple complete, lightweight model variants rather than dividing a single large model. Each variant is a fully functional model capable of predicting all classes independently, eliminating the communication dependencies and failure points of traditional parallelism approaches.

Architecture and Benefits. In VP, variants are usually generated through architectural modifications, layer pruning, or different initialization strategies from a shared base (teacher) model. Key advantages include:

(1) *Independence:* each variant operates autonomously without requiring intermediate communication;

(2) *Fault tolerance:* device failures reduce ensemble size but do not halt inference;

(3) *Flexible aggregation:* predictions can be combined using various ensemble strategies (averaging, voting, weighted combination);

(4) *Graceful degradation:* system performance scales with the number of available devices; and

(5) *Minimal communication:* only final predictions (typically small vectors) need to be shared.

Performance Characteristics and Challenges. Experimental results from the original VP work demonstrate significant improvements in inference speed, communication overhead, and accuracy compared to single-model deployments and other distributed inference approaches. The ensemble effect enables VP systems to achieve accuracy levels larger than a single device model and close to a much larger model while using substantially fewer computational resources per device. However, practical application in homogeneous settings is constrained by two main challenges. First, variant generation is often *empirical and ad-hoc*, lacking systematic methods for inducing diversity. This can lead to highly correlated variants that provide diminishing returns in ensemble accuracy and makes the generation process computationally expensive and unpredictable. Second, VP struggles with *homogeneous ensemble inefficiency*, where ensembles of same-sized variants show minimal accuracy gains compared to heterogeneous ensembles. This limits VP's applicability in common edge scenarios with resource-similar devices. Our work directly targets these challenges by introducing systematic and principled methods for creating diverse and effective homogeneous ensembles.

C. Knowledge Distillation and Ensemble Diversity

Our approach builds upon established techniques in knowledge distillation and ensemble learning to address VP's limitations.

Knowledge Distillation. Popularized by Hinton et al. [14], knowledge distillation (KD) enables efficient transfer of knowledge from large teacher models to smaller student models by training students to match the teacher's output distributions rather than just the hard labels. The distillation loss combines traditional cross-entropy loss with the Kullback-Leibler divergence [15] between teacher and student predictions, enabling students to learn from the teacher's "soft" knowledge about class relationships and uncertainty.

Ensemble Diversity Principles. The effectiveness of ensemble learning critically depends on diversity among component models [16], [17]. Research by Allen-Zhu et al. [18] demonstrates that model diversity significantly affects ensemble performance, while Wood et al. [19] provide a unified theoretical framework treating diversity as a quantifiable component alongside bias and variance in ensemble error analysis. Their work reframes diversity not as an ad-hoc property to be maximized, but as a fundamental element of the bias-variance-diversity trade-off. This perspective suggests that the goal is not merely to increase diversity, but to manage the interplay between bias, variance, and diversity to minimize the overall ensemble error. Diversity can be induced through various mechanisms including different architectures, training procedures, data subsets, or input transformations. Our work leverages this insight by employing principled strategies to generate diverse variants.

III. VIP Methodology

A. Overview and Architecture

VIP addresses the fundamental limitations of current VP approaches through a systematic methodology that enhances both variant generation efficiency and ensemble diversity. Our approach focuses on variant generation that combines knowledge distillation with strategic diversity-inducing techniques to create meaningful variant pools.

Our two core components are:

(1) Enhanced Variant Generation Pipeline. Integrates knowledge distillation with strategic model slicing to accelerate training while maintaining performance quality.

(2) Diversity-Inducing Training Strategies. Employs data augmentation and imbalanced training to systematically create variants with complementary prediction patterns, grounded in the theoretical findings of the bias-variance-diversity trade-off [19].

Our preliminary evaluation in this paper focuses on validating the effectiveness of our variant generation strategies for the ResNet50 architecture across multiple datasets, namely CIFAR10, CIFAR100, ImageNet-1K.

B. Systematic Model Variant Generation

To overcome the ad-hoc nature of variant generation in traditional VP, we introduce a structured pipeline that leverages knowledge distillation and targeted model slicing to efficiently produce diverse and high-performing variants.

Teacher Model Construction. For each target architecture, we construct teacher models using a backbone base model, e.g., ResNet50. We then add task-specific classification heads appropriate for each target dataset. We leverage transfer learning to provide strong teacher models while maintaining computational feasibility. The teacher model is then frozen after training, and its weights are used to initialize the shared parts of the student variants. This approach allows us to benefit from the teacher's knowledge while ensuring that each student variant can be trained independently as well as in a diversified manner.

Strategic Model Slicing. The purpose of model slicing is to generate variants of different sizes that are smaller than the base teacher model, allowing for efficient deployment on resource-constrained edge devices. In VIP, rather than arbitrary architectural modifications, we employ systematic model slicing at chosen points to generate student backbones of varying sizes. Our analysis reveals that slicing location significantly impacts performance:

After Convolution Layers: Produces raw, unnormalized feature maps, resulting in poor student performance due to unstable gradient flows. *After Batch Normalization:* Provides normalized feature maps with improved stability but suboptimal activation patterns. *After ReLU Activation:* Yields activated feature maps that capture essential non-linearities, providing optimal performance among slicing options.

We, therefore, slice models after ReLU layers to generate seven size segments per architecture, ranging from 0.4MB to 90MB depending on the dataset. Each size segment serves as a backbone variant for generating multiple diverse variants.

Knowledge Distillation Integration. For each generated backbone, we employ knowledge distillation to accelerate training while maintaining performance quality. Student models are trained using a combined loss function:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{student} + (1 - \alpha)\mathcal{L}_{distillation}$$

where $\mathcal{L}_{student}$ is the standard cross-entropy loss with ground truth labels, and $\mathcal{L}_{distillation}$ is the KL divergence between teacher and student output distributions. This approach significantly reduces training time compared to training variants from scratch while providing better initialization through teacher knowledge transfer.

C. Diversity-Inducing Training Strategies

A key innovation in VIP is the systematic introduction of diversity among variants of the same size, enabling effective homogeneous ensembles. We develop two complementary strategies for inducing diversity through data manipulation rather than architectural changes. These strategies are grounded in the principle of managing the bias-variance-diversity trade-off, as formalized by Wood et al. [19]. By altering the data distribution each variant is exposed to, we aim to create a set of models whose errors are uncorrelated, thereby reducing the variance term of the ensemble's error without disproportionately increasing the bias of individual models.

1) Data Augmentation-Based Diversity: Our first same-sized variant generation strategy employs data augmentation to create diverse training sets for each model variant. This approach ensures that each model learns from distinct visual perspectives of the same data distribution, generating variants that develop complementary feature representations and decision boundaries. From a theoretical standpoint, each augmentation strategy can be viewed as a way to explore a different part of the function space around the optimal solution, leading to a diverse set of models.

(a) Photometric Transformations. We employ color and intensity-based augmentations including random saturation, random hue adjustment, brightness modification, and contrast variation. Each variant is assigned a specific subset of these transformations, ensuring exposure to different photometric variations while maintaining overall data distribution integrity.

(b) Geometric Transformations. Spatial augmentations include horizontal/vertical flips and rotations. These compute-efficient and simple transformations help variants develop invariance to different geometric perspectives, contributing to ensemble robustness across spatial variations.

(c) Spatial Cropping Strategy. We implement systematic spatial crops that reduce input image dimensions by 6%, 12%, 15%, and 20% respectively. This approach forces variants to focus on different spatial scales and regions, encouraging diverse attention patterns and feature extraction strategies.

Augmentation Assignment Protocol. To enhance diversity, we use fixed global and varying per-variant random seeds for augmentation assignments. Each variant within a size segment receives a unique combination of augmentation strategies, creating a diverse pool of models trained on different data perspectives.

2) Class-Imbalanced Training Diversity: Our second strategy introduces diversity through class-imbalanced training, where each variant is trained on a subset of the complete dataset with varying class representations. This approach allows us to control the bias of each variant while ensuring that the ensemble can leverage complementary expertise across classes. By manipulating the class distribution each variant is exposed to, we create models that specialize in different

subsets of classes, leading to a more diverse ensemble. This aligns with the bias-variance-diversity framework, where the intentional introduction of bias in individual models can lead to a reduction in the overall ensemble variance, resulting in a lower total error.

Class Grouping Strategy. For each dataset, we create 10 distinct class groups with structured imbalance. Each group contains 100% of samples from its 10% target class subset (1 class for CIFAR10, 10 classes for CIFAR100, 100 classes for ImageNet-1K) and 70% of samples from the remaining classes in the dataset.

Imbalance Ratio Selection. The 70% retention ratio for non-target classes represents an optimal balance discovered through empirical analysis. Lower ratios (e.g., 50%) severely degrade individual variant accuracy, while higher ratios (e.g., 90%) reduce inter-variant diversity. The 70% threshold provides sufficient training data for comprehensive learning while maintaining meaningful diversity across variants.

Diversity Mechanism. This class-imbalanced training method produces variants that are specialized in particular class groups but retain generality across the full class set. When combined in an ensemble, these variants offer complementary strengths; each variant delivers higher confidence for its focus classes while maintaining reasonable predictions for the remaining classes.

D. Ensemble Aggregation Strategy

VIP employs average ensembling as the primary aggregation mechanism, combining predictions by averaging output logits from all participating variants. This approach is aligned with the device homogeneity presumption and provides several advantages: (1) *Simplicity:* no additional training or parameter tuning required; (2) *Robustness:* naturally handles varying numbers of participating devices; (3) *Interpretability:* final predictions remain probabilistic and well-calibrated; and (4) *Efficiency:* minimal computational overhead for aggregation.

The ensemble prediction for input x is computed as:

$$P_{ensemble}(y|x) = \frac{1}{N}\sum_{i=1}^{N} P_i(y|x)$$

where N is the number of available variants and $P_i(y|x)$ represents the prediction from variant i. This averaging approach effectively leverages the diversity induced through our training strategies while maintaining computational efficiency suitable for edge deployment scenarios.

IV. Experimental Evaluation

A. Experimental Setup

Our evaluation is conducted on a high-end GPU server with two A100 GPU accelerators in a containerized Linux environment. We evaluate VIP on the ResNet50 architecture across three datasets, namely CIFAR10, CIFAR100, and ImageNet-1K, to demonstrate the generalizability of our proposal.

Training Protocol: We employ a two-phase training strategy: (1) *Frozen training* (10 epochs for CIFAR, 5 for ImageNet, learning rate=0.01) where backbone weights are frozen and only the classification head is trained, followed by (2) *Fine-tuning* (10 epochs, lr=1e-3) where both backbone and head are trained end-to-end. We use Adam optimizer with categorical cross-entropy loss. Teachers are trained until achieving competitive accuracy, while students use knowledge distillation combining student loss and distillation loss.

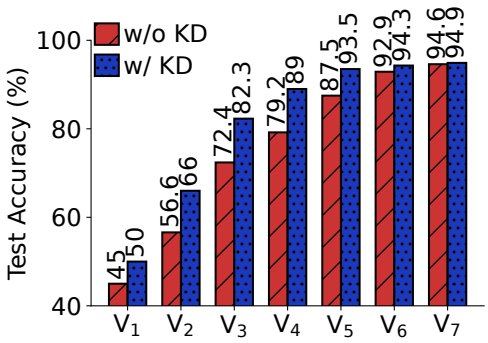


Fig. 1: Knowledge Distillation on different-sized variants of ResNet50 on CIFAR10.

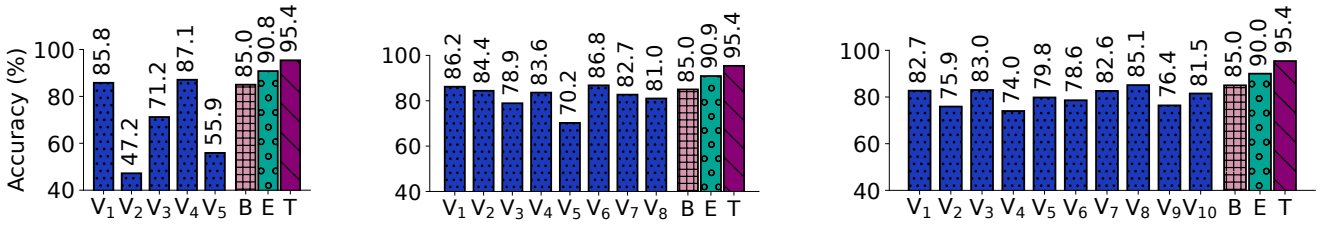
Model Generation: For each architecture, we create seven different size variants by slicing the pre-trained model at the strategic points and after ReLU layers for best performance. Model sizes range from 0.4MB to 90MB depending on the dataset, as detailed in Table I. For each size segment, we generate multiple variants using our three presented diversity-inducing strategies: data augmentation (filters and crops) and imbalanced datasets while applying knowledge distillation for generating and training the variants. Each strategy produces a set of variants that are trained independently but share the same base model architecture.

B. Impact of Knowledge Distillation

We first validate the effectiveness of knowledge distillation (KD) for variant generation by comparing different-sized variants generated in the first stage, and trained with and without KD on ResNet50 with CIFAR10. Figure 1 demonstrates that KD consistently improves individual variant accuracy. The impact is more pronounced for smaller variants, where KD helps them achieve performance closer to larger variants. This confirms that KD is an important component of our variant

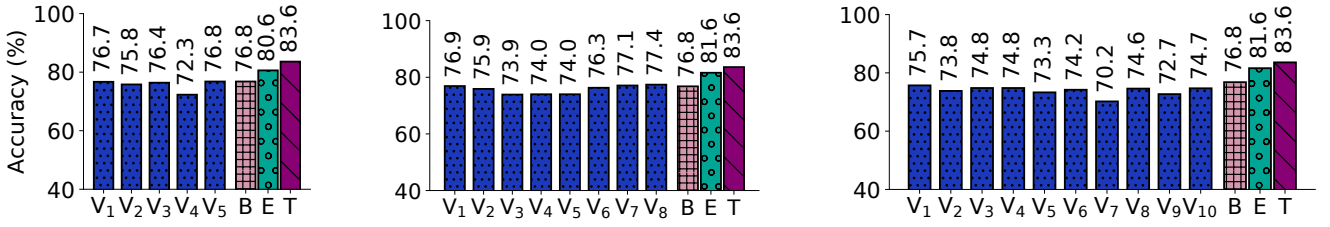
TABLE I: Model size distribution across datasets

Model	CIFAR10/100 (MB)	ImageNet (MB)	Parameters
Teacher	90	90	23.5M
Model 1	0.4-0.5	1	120K
Model 2	3	4	750K
Model 3	12	15	3.2M
Model 4	21	24	5.8M
Model 5	35	37	9.1M
Model 6	58	63	15.2M
Model 7	76	80	19.8M



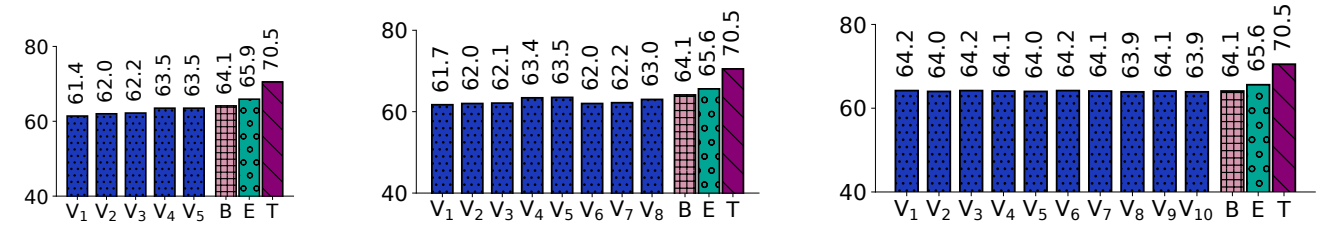
(a) Filter-based (b) Crop-based (c) Class-imbalanced

Fig. 2: Ensemble accuracy improvements using different diversity-inducing strategies - ResNet50 on CIFAR10.



(a) Filter-based (b) Crop-based (c) Class-imbalanced

Fig. 3: Ensemble accuracy improvements using different diversity-inducing strategies - ResNet50 on CIFAR100.



(a) Filter-based (b) Crop-based (c) Class-imbalanced

Fig. 4: Ensemble accuracy improvements using different diversity-inducing strategies - ResNet50 on ImageNet-1K.

generation pipeline, enabling smaller models to learn from the teacher's knowledge and improving their performance.

C. Diversity-Induced Variant Performance

We evaluate three diversity-inducing strategies, namely data augmentation (filters and crops), class-imbalanced training, and their combinations. For conciseness, we present representative results from ResNet50 experiments.

Experimental Configuration. We create variants from the variant V_3 using different augmentation techniques: geometric transformations (`flip_left_right`, `rotate_90`), photometric adjustments (`random_contrast`, `random_hue`), and spatial crops (reducing image size by 6%, 12%, 15%, and 20%). For class imbalance, we create 10 variants where each sees 100% of one class group and 70% of others. Table II summarizes our diversity-inducing approaches.

Cross-Dataset Performance Analysis. Figures 2, 3, and 4 illustrate the performance of ResNet50 ensembles using our three diversity-inducing strategies across the CIFAR10, CIFAR100, and ImageNet-1K datasets. In each Figure, B represents the baseline, T represents the large teacher model, and E represents the ensemble of diverse variants. The results consistently demonstrate the effectiveness of our approach. In some cases, the ensembles even reach close to the teacher model's performance, indicating that our diversity strategies effectively enhance the ensemble's predictive power. Across all datasets and model sizes, the ensembles (E) outperform their corresponding baseline (B) models, achieving accuracy improvements of 2-5%. This highlights the benefit of ensembling diverse variants, even when they are of the same size. All the strategies yield comparable accuracy gains, and there is no single strategy that predominantly outperforms the others across all datasets. This highlights that the choice of diversity-inducing strategy can be tailored to specific deploy-

TABLE II: Diversity-inducing strategies for variant generation

Strategy	Technique	Variants
Filters	Geometric/Photometric	5 variants
Crops	Spatial reduction	8 variants
Class Imbalance	Data distribution	10 variants

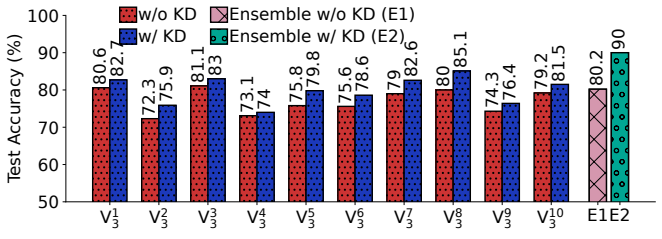


Fig. 5: Impact of knowledge distillation on ensemble performance. The ensemble accuracy is shown for variants trained with and without distillation on CIFAR10.

ment scenarios or resource constraints without significantly sacrificing performance.

While the ensembles narrow the performance gap with the much larger teacher model, in most cases an accuracy difference remains. This suggests that while our diversity strategies are effective, there is still room for improvement. Note that each device is running a variant that is **6**$\times$ smaller than the teacher model. The ensembles achieve competitive performance while maintaining a fraction of the computational overhead. More advanced diversity techniques or combining multiple strategies could potentially close this gap further. The consistent performance gains across diverse datasets validate the generalizability of the VIP methodology.

D. Impact of Knowledge Distillation on Ensemble Performance

We further investigate the impact of knowledge distillation on ensemble performance by comparing ensembles of ten variants generated from V_3, and trained with and without distillation. The results are shown in Fig. 5.

As expected, ensembles of variants trained with knowledge distillation outperform those trained without it. This leads to an accuracy gain of up to 10% over the baseline ensemble that is not trained with distillation.

V. Conclusion and Future Work

In this paper, we introduced Variant-Improved Parallelism (VIP), a systematic methodology that takes a step forward in addressing the main limitations of Variant Parallelism (VP) for distributed inference in edge environments. By integrating knowledge distillation with strategic model slicing and introducing three diversity-inducing strategies, VIP helps generating more effective variants. Our evaluation demonstrates that VIP significantly enhances ensemble accuracy, achieving 10% higher accuracy compared to vanilla ensemble, and up to 5% improvements over single baseline models, while preserving the inherent fault tolerance and communication efficiency of the VP paradigm. We demonstrated that knowledge distillation is a key enabler for efficiently training high-quality variants. Furthermore, our diversity-inducing strategies proved effective in creating complementary variants that lead to more powerful homogeneous ensembles, a key requirement for scalable deployment on resource-similar edge devices.

Future Directions. Despite these advances, a performance gap between our 6$\times$ smaller variant ensembles and the teacher models remains, indicating avenues for future research. We propose the following directions:

(1) *Advanced Diversity Generation:* Future work should explore more sophisticated diversity-inducing techniques. Adversarial training could be used to generate variants robust to specific input perturbations, while multi-objective optimization could explicitly balance the trade-off between individual variant accuracy and inter-variant diversity during training.

Further, in Edge-Cloud Continuum scenarios where Split Computing and Early Exit are applicable [20]–[25], these techniques could be adapted to generate diverse variants that optimize not only the end-to-end performance but also both edge and cloud-based inference costs. This would require taking into account the considerable differences in computational resources, latency requirements, and network conditions between edge devices, edge clouds, and central clouds. For instance, on-demand variant generation based on the traffic load and the autoscaling benchmarks [26] of the cloud services is a promising direction to explore, that may not only improve the overall system performance but also mitigate potential cost-oriented attacks [27].

(2) *Enhanced Knowledge Distillation:* Moving beyond output-level distillation, feature-level distillation could enable student models to learn richer, more nuanced representations from intermediate teacher layers, potentially closing the performance gap.

(3) *Broader Applications and Real-World Deployment:* Applying VIP to other domains, such as natural language processing, would validate its generalizability. Further, real-world deployment studies on homogeneous hardware are needed to evaluate VIP's performance under practical constraints, including network variability, energy limitations, and dynamic device availability. Such studies would provide invaluable insights for developing adaptive and resilient edge AI systems.

(4) *Dynamic Variant Management:* Another avenue for future work involves dynamic variant management strategies that adaptively select and deploy the most effective variants based on real-time performance metrics, device availability, and network conditions. This would enhance the scalability and adaptability of VIP in diverse edge environments.

By addressing these areas, we can further enhance the capabilities of VIP, paving the way for more powerful, scalable, and adaptive distributed AI solutions at the network edge.

Acknowledgment

We thank the anonymous reviewers for their valuable feedback. This work was supported by the Federal Ministry of Research, Technology and Space of Germany (BMFTR) in the programme of "Souverän. Digital. Vernetzt." joint project 6G-life, project identification number 16KISK002.

References

[1] T. N. Manjunath, S. Pushpa, R. S. Hegadi, and D. Yogish, "A study on edge computing through machine learning for iot devices," in *2021 International Conference on Forensics, Analytics, Big Data, Security (FABS)*, vol. 1, 2021, pp. 1–6.

[2] N. Asadi and M. Goudarzi, "Variant parallelism: Lightweight deep convolutional models for distributed inference on iot devices," *IEEE Internet of Things Journal*, vol. 11, no. 1, p. 345–352, Jan. 2024. [Online]. Available: http://dx.doi.org/10.1109/JIOT.2023.3285877

[3] L. Wulfert, N. Asadi, W.-Y. Chung, C. Wiede, and A. Grabmaier, "Adaptive decentralized federated gossip learning for resource-constrained iot devices," in *Proceedings of the 4th International Workshop on Distributed Machine Learning*, 2023, pp. 27–33.

[4] N. Asadi, H. I. Bengü, L. Wulfert, H. Wöhrle, and W. Kellerer, "Poster: Road to tiny reality: Digital twins for decentralized ai on microcontrollers," in *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking (MOBICOM '25)*. New York, NY, USA: ACM, Nov. 2025, p. 3. [Online]. Available: https://doi.org/10.1145/3680207.3765668

[5] H. Katebi, N. Asadi, and M. Goudarzi, "Fullpack: Full vector utilization for sub-byte quantized matrix-vector multiplication on general purpose cpus," *IEEE Computer Architecture Letters*, vol. 23, no. 2, pp. 142–145, 2024.

[6] N. Asadi, H. I. Bengü, L. Wulfert, H. Wöhrle, and W. Kellerer, "Gist - Optimizing Segmentation for Decentralized Federated Learning on Tiny Devices," in *Federated Learning and Edge AI for Privacy and Mobility (FLEdge-AI '25)*, ser. FLEdge-AI '25. New York, NY, USA: Association for Computing Machinery, 2025.

[7] E. Samikwa, A. Di Maio, and T. Braun, "Disnet: Distributed micro-split deep learning in heterogeneous dynamic iot," *IEEE internet of things journal*, vol. 11, no. 4, pp. 6199–6216, 2023.

[8] Z. Zhao, K. M. Barijough, and A. Gerstlauer, "Deepthings: Distributed adaptive deep learning inference on resource-constrained iot edge clusters," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2348–2359, 2018.

[9] G. Bartolomeo, N. Asadi, W. Kellerer, J. Ott, and N. Mohan, "{On-Demand} container partitioning for distributed {ML}," in *2025 USENIX Annual Technical Conference (USENIX ATC 25)*, 2025, pp. 1481–1500.

[10] M. Qin, C. Sun, J. Hofmann, and D. Vucinic, "Disco: Distributed inference with sparse communications," 2023. [Online]. Available: https://arxiv.org/abs/2302.11180

[11] M. Zhang, X. Shen, J. Cao, Z. Cui, and S. Jiang, "Edgeshard: Efficient llm inference via collaborative edge computing," *IEEE Internet of Things Journal*, 2024.

[12] G. Wang, Z. Liu, B. Hsieh, S. Zhuang, J. Gonzalez, T. Darrell, and I. Stoica, "sensai: Convnets decomposition via class parallelism for fast inference on live data," in *Proceedings of Machine Learning and Systems*, A. Smola, A. Dimakis, and I. Stoica, Eds., vol. 3, 2021, pp. 664–679. [Online]. Available: https://proceedings.mlsys.org/paper_-files/paper/2021/file/e8a21a93f244b29e4da50ccbe409c28f-Paper.pdf

[13] N. Asadi and M. Goudarzi, "An ensemble mobile-cloud computing method for affordable and accurate glucometer readout," *arXiv preprint arXiv:2301.01758*, 2023.

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[15] J. Shlens, "Notes on kullback-leibler divergence and likelihood," 2014. [Online]. Available: https://arxiv.org/abs/1404.2000

[16] M. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.engappai.2022.105151

[17] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, Eds., vol. 7. MIT Press, 1994. [Online]. Available: https://proceedings.neurips.cc/paper_-files/paper/1994/file/b8c37e33defde51cf91e1e03e51657da-Paper.pdf

[18] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," 2023. [Online]. Available: https://arxiv.org/abs/2012.09816

[19] G. Wood, G. Brown, M. Lujan, and M. Augasta, "A unified theory of diversity in ensemble learning," *Journal of Machine Learning Research*, 2024.

[20] Y. Matsubara, M. Levorato, and F. Restuccia, "Split computing and early exiting for deep learning applications: Survey and research challenges," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–30, 2022.

[21] G. Bartolomeo, N. Asadi, W. Kellerer, J. Ott, and N. Mohan, "{On-Demand} container partitioning for distributed {ML}," in *2025 USENIX Annual Technical Conference (USENIX ATC 25)*, 2025, pp. 1481–1500.

[22] A. Bakhtiarnia, N. Milošević, Q. Zhang, D. Bajović, and A. Iosifidis, "Dynamic split computing for efficient deep edge intelligence," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] G. Bartolomeo, N. Asadi, W. Kellerer, J. Ott, and N. Mohan, "Beyond Layers: Container Registries for Files Distribution and On-Demand Image Partitioning," in *20th European Conference on Computer Systems (EuroSys '25) - Posters*. ACM, 2025. [Online]. Available: https://2025.eurosys.org/posters/final/eurosys25posters-final13.pdf

[24] S. Laskaridis, S. I. Venieris, M. Almeida, I. Leontiadis, and N. D. Lane, "Spinn: Synergistic progressive inference of neural networks over device and cloud," in *Proceedings of the 26th annual international conference on mobile computing and networking*, 2020, pp. 1–15.

[25] H. Guo, N. Asadi, G. Bartolomeo, P. Laufer, J. Ott, and W. Kellerer, "Comparative performance and cost analysis of computer vision in edge-cloud continuum," in *Proceedings of the 2025 ACM CoNEXT Workshop Edge-Cloud Collaboration for AI (ECCAI '25)*. ACM, 2025.

[26] N. Asadi, R.-M. Ursu, L. Wong, and W. Kellerer, "Simurgh: Multi-agent adversarial benchmarking for proactive microservice observability," in *Proceedings of the 1st Workshop on Next-Generation Network Observability*, 2025, pp. 21–28.

[27] N. Asadi, R.-M. Ursu, J. Zerwas, L. Wong, and W. Kellerer, "Wavesurfer-scheduling irregular pulsing attacks on microservice autoscaling," in *Proceedings of the ACM SIGCOMM 2025 Posters and Demos*, 2025, pp. 55–57.