BUDGET-DISTILLED ES-SSM: CROSS-BUDGET KNOWLEDGE DISTILLATION FOR ELASTIC SPECTRAL STATE SPACE MODELS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Elastic Spectral State Space Models (ES-SSM) enable runtime budget adaptation through ordered spectral truncation, allowing a single model to operate at any spectral budget K by using only the first K channels. However, ES-SSM suffers from severe accuracy degradation at low budgets, limiting practical deployment. We propose Budget-Distilled ES-SSM (BD-ES-SSM), which applies cross-budget KL distillation to align truncated-budget predictions with full-budget teacher distributions during training. By using the full-budget forward pass as an in-place teacher, BD-ES-SSM encourages shared spectral channels to approximate the full model's decision boundary at all truncation levels. On LRA Text, BD-ES-SSM improves low-budget accuracy by +22.61 percentage points at $K = 2$ (80.67% vs 58.06%) and achieves near-flat accuracy curves with only 0.53 pp variation from $K = 2$ to $K = 32$, compared to 19.39 pp degradation for the baseline. Full-budget accuracy is preserved and improved (+2.69 pp), demonstrating that cross-budget distillation enables budget-elastic inference with minimal accuracy loss.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

State space models (SSMs) have emerged as efficient alternatives to Transformers for long-sequence modeling, offering linear complexity in sequence length while maintaining strong performance on tasks requiring long-range dependencies (Gu et al., 2021; Gu & Dao, 2023; Dao & Gu, 2024). Spectral approaches to SSMs (Agarwal et al., 2023) provide additional benefits through interpretable, ordered representations where spectral channels are naturally ranked by importance. This ordering enables Elastic Spectral SSMs (ES-SSM) (Song & Wang, 2026) to support runtime budget adaptation: a single trained model can operate at any spectral budget K by truncating to the first K channels, trading compute for quality without retraining.

However, ES-SSM suffers from severe accuracy degradation at low budgets. While the model maintains reasonable performance at high budgets, accuracy drops sharply when truncating to small K, limiting the practical utility of budget-elastic inference. This degradation stems from *budget inconsistency*: the shared spectral channels must serve multiple truncation levels, but standard training provides insufficient guidance for low-budget configurations where capacity is most constrained.

Existing elastic network methods (Yu et al., 2018; Yu & Huang, 2019; Cai et al., 2019; Yu et al., 2020) address width and depth elasticity in convolutional and transformer architectures, often using in-place distillation to improve small subnetworks. However, these approaches do not directly apply to spectral truncation, where budgets correspond to ordered prefixes of spectral channels rather than arbitrary width choices. Knowledge distillation has been explored for SSM compression (Shah et al., 2025; Chahine et al., 2025), but not for cross-budget alignment within a single elastic model.

We propose **Budget-Distilled ES-SSM (BD-ES-SSM)**, which applies cross-budget KL distillation to align truncated-budget predictions with full-budget teacher distributions during training. By using

[1] https://gitlab.com/fars-a/budget-distilled-spectral-ssm

the full-budget forward pass as an in-place teacher, BD-ES-SSM encourages the shared low-index spectral channels to approximate the full model's decision boundary when operating at small K. Our contributions are:

- A cross-budget distillation objective for elastic spectral SSMs that uses the full-budget prediction as an in-place teacher for truncated-budget configurations.
- Dramatic improvement in low-budget accuracy: +22.61 percentage points at $K = 2$ and +10.96 pp average across low budgets on LRA Text.
- Near-flat accuracy curves (0.53 pp variation from $K = 2$ to $K = 32$) compared to 19.39 pp degradation for the baseline, reducing the minimum budget sweet spot from $K = 12$ to $K = 2$ ($6\times$ reduction).

2 Related Work

2.1 State Space Models

State space models (SSMs) have emerged as efficient alternatives to Transformers for long-sequence modeling. The Structured State Space (S4) model (Gu et al., 2021) introduced a principled approach to parameterizing continuous-time state spaces with HiPPO initialization, achieving strong performance on the Long Range Arena benchmark while maintaining linear complexity in sequence length. Mamba (Gu & Dao, 2023) extended this line of work by introducing selective state spaces with input-dependent dynamics, enabling content-aware reasoning while preserving computational efficiency. Mamba-2 (Dao & Gu, 2024) further unified SSMs with attention mechanisms through structured state space duality, revealing deep connections between these architectures.

Spectral approaches to SSMs offer complementary advantages. Spectral State Space Models (Agarwal et al., 2023) parameterize the state transition in the frequency domain, providing interpretable and ordered representations where spectral channels are naturally ranked by importance. Flash STU (Liu et al., 2024) accelerates spectral transform units through hardware-efficient implementations. Elastic Spectral SSMs (ES-SSM) (Song & Wang, 2026) leverage this ordered structure to enable runtime budget adaptation through spectral truncation, though they suffer from accuracy degradation at low budgets. Our work addresses this limitation through cross-budget knowledge distillation.

2.2 Elastic Networks

Elastic networks enable runtime adaptation of model capacity to varying computational budgets. Slimmable Neural Networks (Yu et al., 2018) introduced switchable batch normalization to train a single network executable at different width multipliers, while Universally Slimmable Networks (Yu & Huang, 2019) extended this to arbitrary widths through sandwich training and inplace distillation. Once-for-All (OFA) (Cai et al., 2019) further generalized elastic training to jointly support variable depth, width, kernel size, and resolution. BigNAS (Yu et al., 2020) scaled these ideas to large single-stage models with weight sharing across sub-networks.

These approaches primarily address width and depth elasticity in convolutional and transformer architectures. In contrast, spectral SSMs offer a fundamentally different elasticity mechanism: ordered spectral truncation, where lower-indexed channels capture more important signal components. This ordered structure enables principled budget reduction but requires specialized training strategies to maintain accuracy across truncation levels. Our cross-budget distillation approach adapts the inplace distillation concept from elastic networks to the spectral truncation setting.

2.3 Knowledge Distillation

Knowledge distillation (Hinton et al., 2015) transfers knowledge from a teacher model to a student by matching soft probability distributions, with temperature scaling to reveal dark knowledge in the teacher's predictions. Born-Again Networks (Furlanello et al., 2018) demonstrated that self-distillation, where a model distills into an identical architecture, can improve performance through iterative refinement. Teacher Assistant (Mirzadeh et al., 2019) addressed the capacity gap problem by introducing intermediate-sized models to bridge large teacher-student gaps.

Recent work has explored distillation specifically for SSMs. SpectraLDS (Shah et al., 2025) provides provable distillation guarantees for linear dynamical systems, establishing theoretical foundations for SSM compression. The Curious Case of In-Training Compression (Chahine et al., 2025) investigates compression dynamics during SSM training, revealing unique properties of state space architectures. Retrieval-Aware Distillation (Bick et al., 2026) addresses distillation for Transformer-SSM hybrids by preserving retrieval capabilities. Our work differs by focusing on cross-budget distillation within a single elastic model, using the full-budget forward pass as an in-place teacher to guide truncated-budget predictions.

3 METHOD

3.1 BACKGROUND: ELASTIC SPECTRAL STATE SPACE MODELS

Elastic Spectral State Space Models (ES-SSM) (Song & Wang, 2026) build on spectral filtering approaches to sequence modeling (Agarwal et al., 2023). Each layer contains $\bar{K}$ spectral channels corresponding to eigenmodes of a fixed Hankel matrix, ordered by decreasing eigenvalue magnitude. This ordering provides a natural importance ranking: lower-indexed channels capture more significant signal components, enabling principled budget reduction through prefix truncation.

At inference time, ES-SSM supports runtime budget adaptation by truncating to the first $K \leq \bar{K}$ spectral channels. The model computes per-channel gate logits and applies a masked softmax that normalizes over channels $1, \ldots, K$ while setting channels $k > K$ to zero. This mechanism allows a single trained model to operate at any budget K without retraining.

ES-SSM trains with **budget dropout**: each training step samples a budget K_{train} from a discrete set (e.g., $\{2, 3, 4, 6, 8, 12, 16, 24, 32\}$) and computes the cross-entropy loss only on the truncated model at that budget. While this provides direct supervision for each budget level, it does not explicitly enforce consistency between different budgets. The result is **budget inconsistency**: for the same input x, the predicted distributions $p_K(y \mid x)$ and $p_{\bar{K}}(y \mid x)$ can diverge because different budgets activate different subsets of spectral channels, and small-K updates are capacity-limited and may be noisier.

3.2 CROSS-BUDGET KL DISTILLATION

We propose **Budget-Distilled ES-SSM (BD-ES-SSM)**, which addresses budget inconsistency through in-place cross-budget knowledge distillation. The key insight is that the full-budget forward pass provides a smoother target distribution than one-hot labels, encouraging the shared low-index spectral channels to approximate the full model's decision boundary when operating at small K.

For each training step, we compute both a full-budget forward pass at $\bar{K} = 32$ and a sampled-budget forward pass at K_{train}. Let $z_{\bar{K}}(x)$ denote the classifier logits from the full-budget pass and $z_K(x)$ the logits from the sampled budget. We define temperature-scaled probability distributions:

$$q = \text{softmax}(z_{\bar{K}}/T), \quad p = \text{softmax}(z_K/T), \tag{1}$$

where T is the temperature parameter that softens the distributions for better knowledge transfer (Hinton et al., 2015).

The **anchored dual-CE baseline** uses both forward passes but only applies cross-entropy supervision:

$$\mathcal{L}_{\text{base}} = \text{CE}(y, z_{\bar{K}}) + \text{CE}(y, z_K). \tag{2}$$

Our **BD-ES-SSM objective** augments this with a KL divergence term that aligns the sampled-budget prediction to the full-budget teacher:

$$\mathcal{L}_{\text{ours}} = \mathcal{L}_{\text{base}} + \lambda T^2 \cdot \text{KL}(\text{sg}(q) \| p), \tag{3}$$

where $\text{sg}(\cdot)$ denotes stop-gradient to prevent backpropagation through the teacher distribution, and λ controls the distillation strength. The T^2 scaling compensates for the reduced gradient magnitude from softened distributions. We use $T = 4.0$ and $\lambda = 2.0$, which we found effective for binary classification where higher temperatures are needed to create informative soft targets.

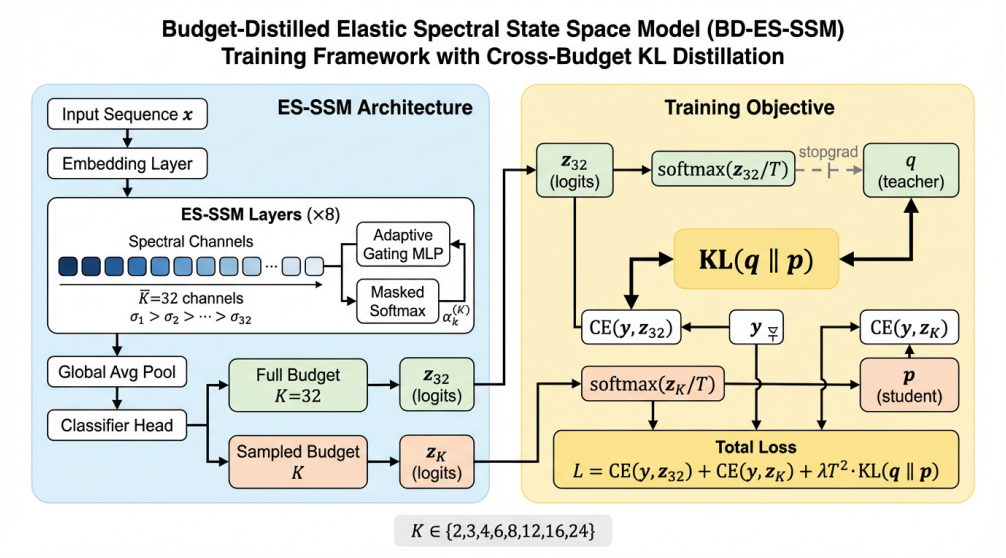

Figure 1: BD-ES-SSM training framework with cross-budget KL distillation. The ES-SSM architecture produces logits at both full budget $K = 32$ and sampled budget K. The training objective combines anchored dual cross-entropy losses with a KL divergence term that aligns the sampled-budget prediction p to the full-budget teacher distribution q (with gradient stopping). The total loss is $\mathcal{L} = \mathrm{CE}(y, z_{32}) + \mathrm{CE}(y, z_K) + \lambda T^2 \cdot \mathrm{KL}(q \| p)$.

3.3 Training Procedure

Figure 1 illustrates the BD-ES-SSM training framework. Each training step performs two forward passes through the shared ES-SSM model: one at full budget $K = 32$ and one at a sampled budget K_{train}. The full-budget logits serve as the teacher signal for the KL distillation term, while both passes receive cross-entropy supervision from the ground-truth labels.

We employ **inverse-K weighted budget sampling**, where the probability of sampling budget K is proportional to $1/K$. This prioritizes low-budget training, which is where the accuracy gap is most severe. The full budget $K = 32$ is excluded from the student budget sampling since it already receives direct supervision through the anchored CE term.

For **checkpoint selection**, we use a composite metric that balances full-budget and low-budget performance: $0.5 \cdot \text{val\_acc}_{K=32} + 0.5 \cdot \text{avg}(\text{val\_acc}_{K=2}, \text{val\_acc}_{K=4})$. This ensures the selected checkpoint performs well across the budget spectrum rather than optimizing solely for full-budget accuracy. We apply early stopping with patience of 15 evaluation rounds.

4 Experiments

4.1 Experimental Setup

We evaluate BD-ES-SSM on the Long Range Arena (LRA) Text benchmark (Tay et al., 2020), which consists of IMDb movie reviews processed at the byte level for binary sentiment classification. This benchmark is designed to test long-range dependency modeling, with sequences of length $L = 4096$ tokens and a vocabulary size of 258 (byte values plus special tokens).

Our ES-SSM architecture follows the configuration from Song & Wang (2026): $d_{\text{model}} = 256$, $n_{\text{layers}} = 8$, maximum spectral budget $\bar{K} = 32$, and gate dimension $d_g = 64$. We train with AdamW optimizer using learning rate 10^{-3} (with $0.1\times$ scaling for SSM-specific parameters), weight decay 0.05, and cosine learning rate schedule with 10% warmup. Training uses bfloat16 mixed precision with batch size 16 per GPU across 8 A100 GPUs (effective batch size 128) and gradient clipping at 0.5. Additional implementation details are provided in Appendix A.

Table 1: Test accuracy (%) on LRA Text across spectral budgets K. BD-ES-SSM achieves near-flat accuracy across all budgets, with +22.61 pp improvement at $K = 2$ and +2.69 pp at $K = 32$ over the baseline. Best results in **bold**. Results averaged over 3 seeds ($\pm$std).

Method	K=2	K=3	K=4	K=6	K=8	K=12	K=16	K=24	K=32	$\text{AvgAcc}_{\text{lowK}}$
ES-SSM Baseline	58.06$\pm$6.26	66.36$\pm$7.56	75.26$\pm$1.72	75.42$\pm$1.64	74.72$\pm$2.71	75.95$\pm$0.64	76.98$\pm$1.13	77.44$\pm$0.59	77.45$\pm$0.54	69.96$\pm$3.80
BD-ES-SSM (Ours)	**80.67**$\pm$3.68	**81.04**$\pm$3.70	**81.32**$\pm$3.44	**80.80**$\pm$4.64	**80.76**$\pm$4.42	**80.73**$\pm$4.35	**80.87**$\pm$3.80	**80.20**$\pm$2.77	**80.14**$\pm$2.77	**80.92**$\pm$3.96
Δ (Improvement)	+22.61	+14.68	+6.06	+5.38	+6.04	+4.78	+3.89	+2.76	+2.69	+10.96

We compare two conditions, both compute-matched with two forward passes per training step: (1) **Anchored Dual-CE Baseline**: $\mathcal{L} = \text{CE}(y, z_{32}) + \text{CE}(y, z_K)$, and (2) **BD-ES-SSM (Ours)**: the baseline loss plus cross-budget KL distillation (Equation 3) with $\lambda = 2.0$ and $T = 4.0$. Both methods use identical budget sampling from $K \in \{2, 3, 4, 6, 8, 12, 16, 24, 32\}$ with inverse-K weighting. We report results averaged over 3 random seeds (42, 123, 456).

4.2 Main Results

Table 1 presents test accuracy across all spectral budgets. BD-ES-SSM achieves substantial improvements over the baseline at all budget levels, with the most dramatic gains at low budgets where the baseline struggles most severely.

At the most constrained budget ($K = 2$), BD-ES-SSM achieves 80.67% accuracy compared to 58.06% for the baseline, an improvement of +22.61 percentage points. The average low-budget accuracy ($\text{AvgAcc}_{\text{lowK}}$, computed over $K \in \{2, 3, 4, 6, 8\}$) improves from 69.96% to 80.92% (+10.96 pp). Critically, BD-ES-SSM produces **near-flat accuracy curves**: accuracy varies by only 0.53 pp from $K = 2$ (80.67%) to $K = 32$ (80.14%), compared to 19.39 pp degradation for the baseline (58.06% to 77.45%). This demonstrates that cross-budget distillation successfully enables budget-elastic inference with minimal accuracy loss.

Full-budget accuracy ($K = 32$) is not only preserved but improved by +2.69 pp (80.14% vs 77.45%), indicating that the distillation objective provides regularization benefits even at full capacity. This is consistent with findings from self-distillation literature (Furlanello et al., 2018), where matching soft targets can improve generalization.

4.3 Analysis

Figure 2 visualizes the accuracy-budget curves for both methods. The baseline exhibits the characteristic steep degradation at low budgets that motivates this work, while BD-ES-SSM maintains remarkably flat performance across the entire budget range.

We define the **budget sweet spot** K^* as the minimum budget achieving $\geq$98% of full-budget accuracy. For the baseline, $K^* = 12$ (75.95% vs 77.45% at $K = 32$), while BD-ES-SSM achieves $K^* = 2$ (80.67% vs 80.14% at $K = 32$). This represents a **6$\times$ reduction** in the minimum budget required for near-full accuracy, enabling more aggressive compute savings in deployment scenarios.

To verify the distillation mechanism, we measured **budget inconsistency**: $\mathbb{E}_x[\text{KL}(p_{32}(\cdot|x) \| p_K(\cdot|x))]$, the KL divergence between full-budget and truncated-budget predictions. For 2 of 3 seeds, BD-ES-SSM reduces budget inconsistency by 70–85% compared to the baseline, confirming that cross-budget distillation successfully aligns truncated-budget predictions to the full-budget teacher. One seed showed higher inconsistency despite achieving the highest accuracy, suggesting that when low-budget predictions become highly confident and accurate, they may diverge from the full-budget distribution while still being correct.

We observe moderate seed variance in BD-ES-SSM results (std 2.77–4.64% across budgets), higher than the baseline (std 0.54–7.56%). This variance reflects sensitivity to initialization in the distillation dynamics, though all seeds substantially outperform the baseline at low budgets. The worst BD-ES-SSM seed (77.00% at $K = 2$) still exceeds the best baseline seed (65.41% at $K = 2$) by a large margin.

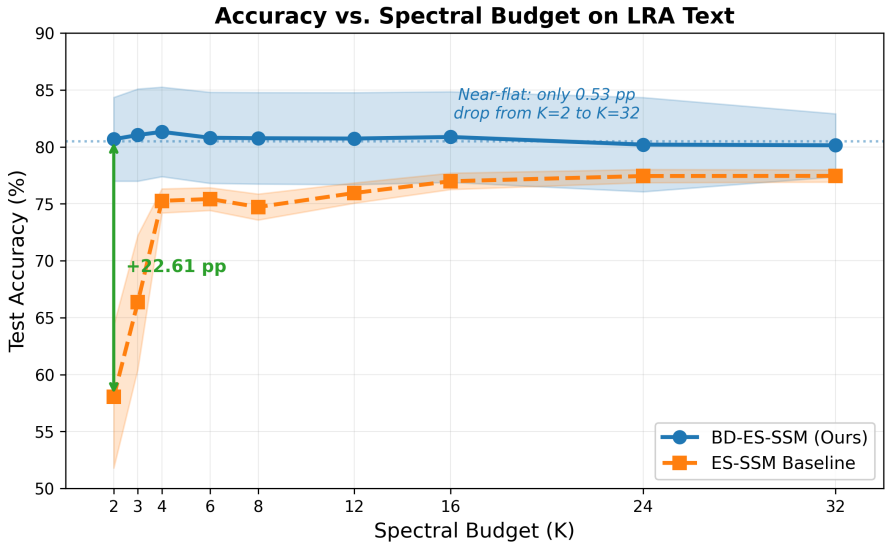

Figure 2: Test accuracy vs. spectral budget K on LRA Text. BD-ES-SSM (blue, solid) maintains near-flat accuracy across all budgets (80.67% at $K = 2$ to 80.14% at $K = 32$, only 0.53 pp drop), while the ES-SSM baseline (orange, dashed) shows steep degradation at low budgets (58.06% at $K = 2$ to 77.45% at $K = 32$, 19.39 pp drop). Shaded regions indicate ± 1 standard deviation across 3 seeds.

5 Conclusion

We presented BD-ES-SSM, a cross-budget knowledge distillation approach for Elastic Spectral State Space Models that enables budget-elastic inference with minimal accuracy degradation. By using the full-budget forward pass as an in-place teacher for truncated-budget predictions, BD-ES-SSM achieves +22.61 pp improvement at $K = 2$ and produces near-flat accuracy curves (0.53 pp variation vs 19.39 pp for the baseline), reducing the minimum budget sweet spot from $K = 12$ to $K = 2$. Our results demonstrate that cross-budget distillation is an effective strategy for improving low-budget performance in elastic spectral models.

Limitations. Our evaluation is limited to a single benchmark (LRA Text), and we observe moderate seed variance in the distillation dynamics. Future work should evaluate on additional tasks and SSM architectures, explore adaptive distillation weights, and investigate the interaction between distillation temperature and task complexity.

References

Naman Agarwal, Daniel Suo, Xinyi Chen, and Elad Hazan. Spectral state space models. *ArXiv*, abs/2312.06837, 2023.

Aviv Bick, Eric P. Xing, and Albert Gu. Retrieval-aware distillation for transformer-ssm hybrids. 2026.

Han Cai, Chuang Gan, and Song Han. Once for all: Train one network and specialize it for efficient deployment. *ArXiv*, abs/1908.09791, 2019.

Makram Chahine, Philipp Nazari, Daniela Rus, and T. Konstantin Rusch. The curious case of in-training compression of state space models. *ArXiv*, abs/2510.02823, 2025.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *ArXiv*, abs/2405.21060, 2024.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, L. Itti, and Anima Anandkumar. Born again neural networks. pp. 1602–1611, 2018.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, abs/2312.00752, 2023.

Albert Gu, Karan Goel, and Christopher R'e. Efficiently modeling long sequences with structured state spaces. *ArXiv*, abs/2111.00396, 2021.

Geoffrey E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

Y. I. Liu, Windsor Nguyen, Yagiz Devre, Evan Dogariu, Anirudha Majumdar, and Elad Hazan. Flash stu: Fast spectral transform units. *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 165–171, 2024.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant. pp. 5191–5198, 2019.

Devan Shah, Shlomo Fortgang, Sofiia Druchyna, and Elad Hazan. Spectralds: Provable distillation for linear dynamical systems. *ArXiv*, abs/2505.17868, 2025.

Dachuan Song and Xuan Wang. Elastic spectral state space models for budgeted inference. 2026.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, J. Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *ArXiv*, abs/2011.04006, 2020.

Jiahui Yu and Thomas S. Huang. Universally slimmable networks and improved training techniques. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1803–1811, 2019.

Jiahui Yu, L. Yang, N. Xu, Jianchao Yang, and Thomas S. Huang. Slimmable neural networks. *ArXiv*, abs/1812.08928, 2018.

Jiahui Yu, Pengchong Jin, Hanxiao Liu, Gabriel Bender, Pieter-Jan Kindermans, Mingxing Tan, Thomas Huang, Xiaodan Song, and Quoc V. Le. Bignas: Scaling up neural architecture search with big single-stage models. pp. 702–717, 2020.

A Implementation Details

Additional implementation details including hyperparameter settings and training configurations are provided here for reproducibility.