Efficient Vision Models for Jetson: Steel Classification via Knowledge Distillation

Yasin Esfandiari[1,2], Karan Rajshekar[1,2,*], Sabine Janzen[2] and Wolfgang Maass[1,2]

[1]Saarland University, Saarbrücken, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany

Abstract

Deploying vision models at the industrial edge requires balancing accuracy against energy and latency constraints that server-oriented models cannot meet, yet prior distilled-ViT work measures efficiency through proxy metrics (FLOPs and server-GPU throughput) rather than actual on-device energy. We close this gap on the DOES non-alloyed steel-scrap benchmark (8,131 test tiles, eight EU steel grades), distilling ViT-Large and Vision-LSTM (ViL-Base, an xLSTM-based backbone) teachers into 5.5M-parameter DeiT-Tiny students and benchmarking every model on an NVIDIA Jetson Xavier NX with onboard power monitoring and latency measurement. Our distilled students achieve 88.81% and 87.07% top-1 with 2.93–3.56% accuracy gaps to their teachers, reducing inference energy by 15.9–34× and latency by 18.1–44×. To the best of our knowledge, this is also the first cross-inductive-bias comparison in knowledge distillation for vision involving a recurrent xLSTM-based teacher: same-family distillation (ViT→DeiT, 88.81%) achieves 1.74 pp higher accuracy than cross-family (ViL→DeiT, 87.07%), yet both deliver equivalent on-device efficiency.

Keywords

Knowledge Distillation, Vision Transformer, xLSTM, Edge Inference, Energy Efficiency, Industrial Image Classification, Steel Scrap

1. Introduction

Edge-deployed industrial perception is energy-bounded: scrapyard inspection stations, sorting-hall drones, and on-truck cameras operate under tight power budgets, so the relevant figure-of-merit is inference energy, not desktop top-1 accuracy [1]. Yet most efficient-ViT research reports FLOPs, parameter counts, or server-GPU throughput, leaving the deployed energy-accuracy trade-off empirically unmeasured for industrial benchmarks.

Three gaps motivate this work. First, ViT-distillation studies measure efficiency through proxy metrics on server hardware rather than measuring actual energy on edge accelerators. Second, steel-scrap-classification methods target scrapyard gantries and embedded inspection systems as their explicit deployment context, yet every published method reports offline accuracy only. Third, cross-architecture KD has not explored distilling a recurrent xLSTM-based vision backbone [2] into a Transformer student.

We address all three gaps on DOES [3]: we distil ViT-Large and Vision-LSTM (ViL-Base, xLSTM-based [4]) teachers into 5.5 M-parameter DeiT-Tiny students [5] and benchmark every model on an NVIDIA Jetson Xavier NX with onboard power monitoring. Our contributions are:

- **First on-device study on DOES.** We report on-device accuracy, energy per inference, and latency on Jetson Xavier NX, providing a replicable profiling protocol for industrial edge vision.
- **Pareto-dominant students.** Distilled students Pareto-dominate their teachers, achieving 15.9–34× energy reduction and 18.1–44× latency speedup with 2.93–3.56% accuracy loss.
- **Cross-inductive-bias KD.** To the best of our knowledge, this is the first distillation from a recurrent xLSTM-based vision teacher into an attention-based student. Same-family distillation achieves 1.74 pp higher accuracy than cross-family distillation (88.81% vs. 87.07%), yet both deliver equivalent on-device efficiency (44–47 mJ and 3.1–3.5 ms).

SuRE'26: Workshop on Sustainability and Resource-Efficiency of Artificial Intelligence, August 17, 2026, Bremen, Germany
*Corresponding author.
y.esfandiari@iss.uni-saarland.de (Y. Esfandiari); karan.rajshekar@uni-saarland.de (K. Rajshekar); sabine.janzen@dfki.de (S. Janzen); wolfgang.maass@dfki.de (W. Maass)

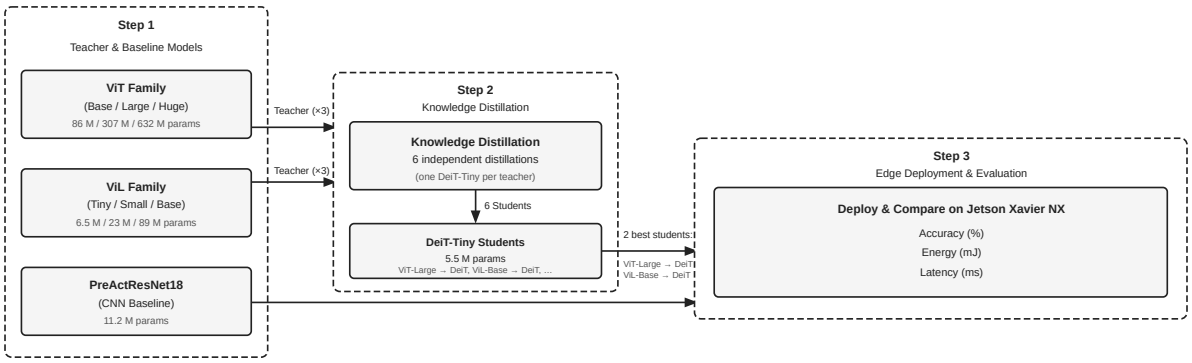

Figure 1: Experimental pipeline. Six DeiT-Tiny students are independently distilled, one per teacher. The two best-performing students (one per teacher family) are selected for on-device profiling alongside their teachers and the PreActResNet18 baseline.

2. Related Work

ViT distillation and edge efficiency. Prior ViT distillation work, including DeiT [5], TinyViT [6], MiniViT [7], Manifold-KD [8], TaT [9] and Co-Advise [10], reports efficiency exclusively through proxy metrics (FLOPs, parameters, desktop-GPU throughput) on server hardware. Latency-aware designs such as EfficientFormer [11] (79.2% top-1 at 1.6 ms on iPhone 12 CoreML) and EdgeViTs [12] measure latency and energy on mobile consumer devices, but not on an industrial embedded accelerator such as the Jetson Xavier NX.

Steel-scrap classification. Steel-scrap classification research has converged on offline accuracy benchmarks. The DOES dataset [3] is paired with ResNet baselines. CLRiuS [13] extends it with contrastive pre-training. LASER [14] applies a layered vision-foundation-model pipeline, and Temur and Karacan [15] compare ResNet and Transformer families. None of these report on-device latency or energy, despite the explicitly stated deployment target of scrapyard gantries and embedded inspection systems. The ESCADE project [16] addresses steel-industry visual AI at the data-centre level. Our work is its embedded-edge complement, providing the first on-device efficiency study on DOES.

Cross-architecture knowledge distillation. Cross-architecture distillation has advanced from CNN-to-Transformer transfer (DeiT [5], Co-Advise [10]) through heterogeneous CNN, Transformer, and MLP pairings (OFA-KD [17]) to Transformer-to-SSM distillation in language (MOHAWK [18]) and in vision (ViT-Linearizer [19], which transfers a ViT teacher into a Mamba-2-based student). Every published work distils *toward* the more efficient architecture. Distilling a recurrent xLSTM-based *teacher* into a Transformer student has not been studied in vision. The only published xLSTM distillation work, Distil-xLSTM [20], uses xLSTM as the *student* for language modelling, the opposite direction.

3. Experimental Setup

Dataset. We use DOES [3], a publicly available steel-scrap benchmark derived from 6,196 smartphone and drone images collected at Saarstahl/Dillinger and local scrapyards. Images are tiled at 720 px with a fixed overlap ($\approx$355 px) and resized to 256×256, yielding 105,523 training and 8,131 test tiles ($\sim$113k total) across eight non-alloyed EU steel grades (E1, E2, E3, E5H, E6, E8, E40, EHRB) plus a background class. Test imagery is collected at a different time from training to avoid scene overlap. The background class is empty in the test split and is excluded from evaluation. The remaining eight classes are severely

imbalanced (E8 accounts for 37.4% of training samples while rare grades such as E40 and E5H contribute less than 1%).

Models. We fine-tune six teacher backbones on DOES from ImageNet-pretrained checkpoints: three Vision Transformers ViT-{Base, Large, Huge} (86, 307, 632 M params [21]) with global self-attention, and three Vision-LSTM models ViL-{Tiny, Small, Base} (6.5, 23, 89 M params [4]) with recurrent gating (xLSTM-based, near-linear complexity in sequence length). The student is *DeiT-Tiny-Distilled* [5] (5.5 M params), which extends the standard ViT encoder with a dedicated distillation token and head alongside the classification head. One student is independently distilled per teacher. Figure 1 illustrates the full three-step pipeline.

Training. Teacher and student receive the same augmented tensor (*single-view*): images are resized to 224×224, augmented with RandAugment (magnitude 9, std 0.5) and random erasing ($p = 0.25$), and normalized with ImageNet statistics. We apply *hard distillation* [5]: the teacher pseudo-label is $\hat{y}_t = \arg\max z_t$. Per-class weights $w_c \propto 1/\sqrt{n_c}$ are applied to both branches via weighted cross-entropy $\mathrm{CE}_w(z, y) = -w_y \log \mathrm{softmax}(z)_y$:

$$\mathcal{L} \;=\; \tfrac{1}{2}\,\mathrm{CE}_w\big(z_s^{(\mathrm{cls})},\, y\big) \;+\; \tfrac{1}{2}\,\mathrm{CE}_w\big(z_s^{(\mathrm{dist})},\, \hat{y}_t\big). \tag{1}$$

Students are trained using AdamW (learning rate 2×10^{-5}, cosine decay, weight decay 0.01, 500-step warm-up) with a batch size of 64 for up to 10 epochs, using early stopping with a patience of 3. During training, the student performs forward and backward passes in FP16, while the teacher remains in FP32.

On-device profiling. All models are exported to TorchScript and evaluated on an NVIDIA Jetson Xavier NX. Inference preprocessing uses a 224×224 resize with ImageNet normalization, matching the training pipeline. Energy is measured via `jtop` (a `tegrastats` wrapper) sampled at 10 Hz. Baseline power is measured during system idle state. Differential energy isolates task-specific consumption: $E_{\text{task}} = (P_{\text{avg}}^{\text{inf}} - P_{\text{avg}}^{\text{idle}}) \times T_{\text{inf}}$, following [22]. Each model is run 3 times at batch size 32 on the 8,131-image test set. Batch size 32 was chosen to measure sustained throughput under continuous-stream conditions. Per-image latency is computed as batch wall time divided by batch size. Single-frame (batch-size-1) profiling is identified as future work. We report the mean and standard deviation for inference energy (mJ) and latency (ms).

4. Results and Analysis

4.1. KD Results: Teacher Comparison and Head Selection

Table 1 shows top-1 accuracy for all six teacher–student pairs. We focus on the highest-performing teachers in each family, ViT-Large (90.60%) and ViL-Base (90.01%), which distil into highly capable students reaching 88.65% and 86.48% combined-head accuracy. The DIST head outperforms CLS across all teacher families. The margin is largest for ViL-Base (87.68% vs. 85.45%, $\Delta = 2.23$ pp) and ≈1.1–1.2 pp for the three ViT variants. We use the combined (averaged) head for all Jetson measurements, ensuring consistency across teacher families and avoiding the selection of the best individual head per model.

4.2. Edge Deployment Efficiency

Table 2 reports on-device measurements for the baseline, both primary teachers, and their distilled students on an NVIDIA Jetson Xavier NX. Where accuracy values differ from Table 1, this reflects the evaluation pipeline: Table 2 uses TorchScript export with the unified ImageNet normalization applied at training, while Table 1 uses the HuggingFace evaluation pipeline with each model's native preprocessor config.

Table 1
Teacher vs. distilled student results (desktop GPU, test set). Student columns show head-wise and combined accuracy.

	Teacher		Student (DeiT-Tiny)		
Model	**Params (M)**	**Accuracy (%)**	**CLS (%)**	**DIST (%)**	**Combined (%)**
ViL-Tiny	6.5	86.77	86.00	86.88	86.82
ViL-Small	23	86.38	84.34	85.06	84.81
ViL-Base	**89**	**90.01**	**85.45**	**87.68**	**86.48**
ViT-Base	86	87.16	87.73	88.83	88.34
ViT-Large	**307**	**90.60**	**87.43**	**88.55**	**88.65**
ViT-Huge	632	87.37	87.00	88.24	87.63

Table 2
Key efficiency metrics on Jetson Xavier NX, held-out test set (3-run average). Students achieve 15.9–34× energy reduction and 18.1–44× latency speedup with 2.93–3.56% accuracy loss.

Model	**Accuracy (%)**	**Energy (mJ)**	**Latency (ms)**
PreActResNet18	72.29	62.29	3.68
Teachers			
ViT-Large	92.37	700.59 ± 12.15	62.71 ± 0.98
ViL-Base	90.00	1594.80 ± 25.40	137.34 ± 3.30
ViL-Tiny	86.75	324.67 ± 4.26	30.44 ± 2.66
Distilled Students (Ours)			
DeiT (ViT-Large)	**88.81**	$\mathbf{44.02 \pm 1.24}$	**3.46**
vs. teacher	*−3.56%*	*15.9×*	*18.1×*
DeiT (ViL-Base)	**87.07**	$\mathbf{46.96 \pm 5.12}$	**3.12**
vs. teacher	*−2.93%*	*34.0×*	*44.0×*

Both teachers are too energy-intensive for practical edge deployment: ViT-Large consumes ∼700 mJ and ViL-Base ∼1595 mJ per inference, with latencies of 62.7 ms and 137.3 ms, respectively. Both draw over 10 W during inference (ViT-Large 10.7 W, ViL-Base 11.4 W), roughly 2.3–2.8× the power of the distilled students (4.1–4.7 W). Knowledge distillation transforms these into 5.5 M-parameter students running at 44–47 mJ and 3.1–3.5 ms, making real-time inference ($\geq$30 FPS) on constrained hardware practical: ViT-Large at 62.7 ms per image yields only 16 FPS, while students at $\leq$3.5 ms exceed 285 FPS. Both students are also more energy-efficient than PreActResNet18 (62.29 mJ, 72.29%), while exceeding it by 14–17 pp in accuracy. The larger ratios for ViL-Base (34× energy, 44× latency) vs. ViT-Large (15.9× energy, 18.1× latency) directly reflect ViL-Base's higher absolute inference cost (∼1595 mJ, 137 ms) compared to ViT-Large (∼700 mJ, 63 ms) on the same hardware. ViL-Tiny, the lightest ViL variant (86.75%, 324.67 mJ, 30.44 ms), still consumes 6.9–7.4× more energy and is 8.8–9.8× slower per inference than the distilled students. Figure 2 illustrates the energy–accuracy trade-off: students occupy the high-accuracy, low-energy region while teachers cluster at the energy-expensive end.

Cross-architecture finding. Same-family distillation achieves 1.74 pp higher top-1 accuracy than cross-family distillation (88.81% vs. 87.07%), yet both deliver equivalent efficiency. The 1.74 pp gap suggests that architectural alignment between teacher and student aids knowledge transfer, though both students reach comparable deployment efficiency (44–47 mJ and 3.1–3.5 ms). This paper studies the distillation of a recurrent xLSTM-based vision backbone [4] into an attention-based student. Prior cross-architecture work (e.g. OFA-KD [17]) addresses CNN, Transformer, and MLP pairings, and the only prior xLSTM distillation (Distil-xLSTM [20]) transfers in the opposite direction.

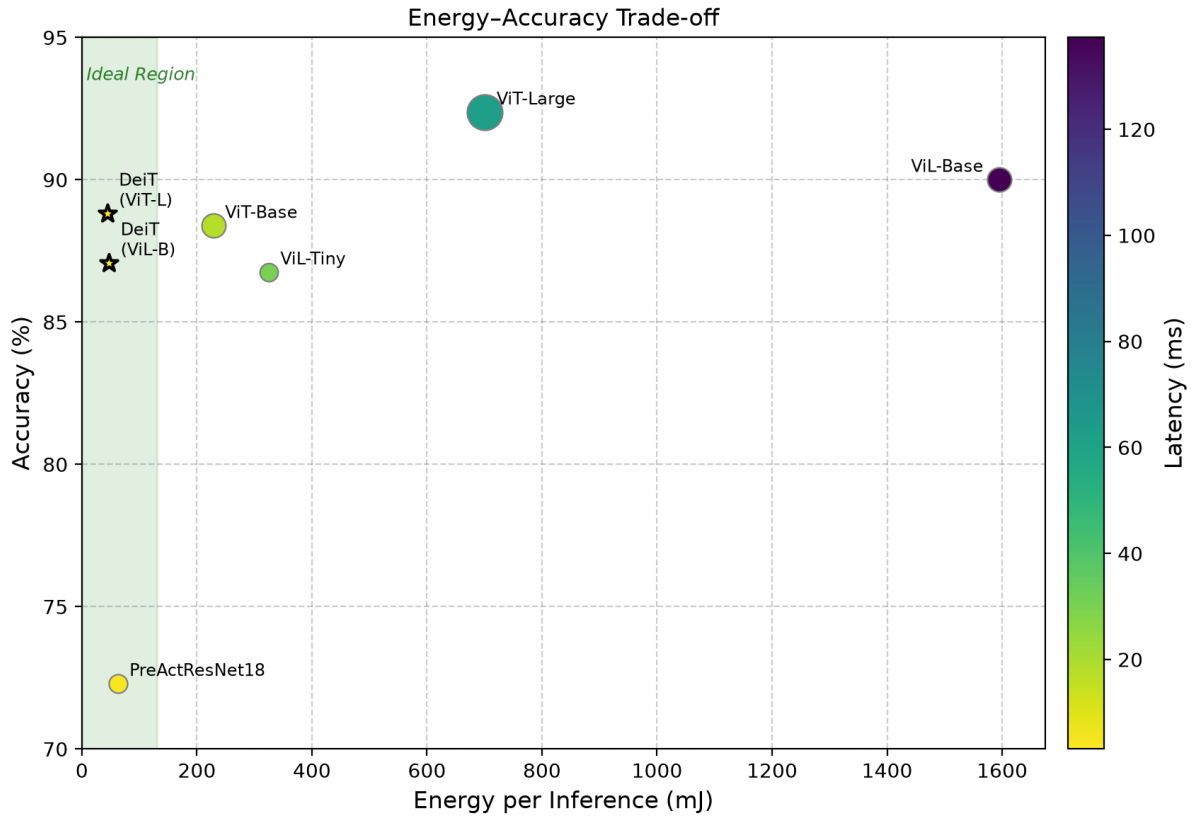

Figure 2: Energy–accuracy trade-off on Jetson Xavier NX. Colour encodes latency and marker size encodes parameter count. Distilled students cluster in the low-energy, high-accuracy region, Pareto-dominating their teachers.

Training configuration. Table 3 compares SET 1 (RandAugment + class-weighted CE) against SET 2 (no augmentation, uniform loss). Augmentation improves the DeiT (ViT-Large) student by 3.11 pp (85.70% to 88.81%) at around 4% energy overhead. For the DeiT (ViL-Base) student, accuracy improves by 1.59 pp while energy means differ by around 12% (46.96 ± 5.12 mJ vs. 41.97 ± 3.21 mJ), with the uncertainty intervals of both configurations overlapping. These results confirm that augmentation and class-weighted loss are important for distillation effectiveness on severely imbalanced industrial data.

Table 3
Ablation: SET 1 (RandAugment + class-weighted CE) vs. SET 2 (no augmentation, uniform loss). Accuracy on held-out test set. Energy averaged over 3 Jetson runs.

Student	Config	Accuracy (%)	Energy (mJ)
DeiT (ViT-Large)	SET 1 (with augmentation)	**88.81**	44.02 ± 1.24
	SET 2 (no augmentation)	85.70	42.33 ± 0.55
DeiT (ViL-Base)	SET 1 (with augmentation)	**87.07**	46.96 ± 5.12
	SET 2 (no augmentation)	85.48	41.97 ± 3.21

4.3. Comparison to State-of-the-Art

Table 4 contextualises accuracy against methods that report only offline metrics. Our ViT-Large teacher (92.37%) nearly matches the best reported result on DOES, LASER (92.80%, ∼86 M params). Through distillation, our best student (88.81%, 5.5 M params) outperforms ResNet34 (84.29%) [15] and CLRiuS (75.89%) [13] while being 56× smaller than our teacher, 16× smaller than LASER's encoder, and the

first model in this literature evaluated on actual edge hardware.

Table 4
Comparison on DOES. *LASER parameter count estimated from its ViT-B/16 encoder. All prior-work numbers are desktop/server accuracy only. Edge energy is not reported.

Method	Params (M)	Accuracy (%)
LASER [14]	∼86*	**92.80**
ViT-Large (our teacher)	307	92.37
ResNet34 [15]	21.8	84.29
CLRiuS [13]	11.7	75.89
PreActResNet18 [3]	11.2	72.29
DeiT (ViT-Large) (ours)	**5.5**	**88.81**
DeiT (ViL-Base) (ours)	**5.5**	**87.07**

5. Conclusion

We benchmarked knowledge distillation from ViT-Large and Vision-LSTM (ViL-Base, xLSTM-based) teachers into 5.5 M-parameter DeiT-Tiny students on an NVIDIA Jetson Xavier NX, providing the first on-device energy and latency measurements for the DOES steel-scrap benchmark. Students Pareto-dominate their teachers, achieving 15.9–34× energy reduction and 18.1–44× latency reduction with 2.93–3.56% top-1 accuracy loss. Same-family distillation (ViT→DeiT, 88.81%) yields 1.74 pp higher accuracy than cross-family distillation (ViL→DeiT, 87.07%), yet both students deliver equivalent on-device efficiency (44–47 mJ and 3.1–3.5 ms), demonstrating that inductive-bias mismatch between teacher and student does not compromise the student's energy-accuracy trade-off. These results establish a replicable on-Jetson measurement protocol for industrial vision and suggest that KD with hardware-aware evaluation is a practical route to sustainable edge deployment in steel-scrap sorting and related domains.

Acknowledgments

This work was part-funded within the research project ESCADE (grant number: 01MN23004A, www.escade-project.de), funded by the German Federal Ministry of Research, Technology and Space (BMFTR), supported by the DLR project management agency.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) and Gemini (Google) in order to: Improve writing style and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green AI, Communications of the ACM 63 (2020) 54–63. doi:10.1145/3381831.

[2] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter, xLSTM: Extended long short-term memory, in: Advances in Neural Information Processing Systems, volume 37, 2024. Spotlight; arXiv:2405.04517.

[3] M. Schäfer, U. Faltings, B. Glaser, DOES: A multimodal dataset for supervised and unsupervised analysis of steel scrap, Scientific Data 10 (2023) 780. doi:10.1038/s41597-023-02662-6.

[4] B. Alkin, M. Beck, K. Pöppel, S. Hochreiter, J. Brandstetter, Vision-lstm: xlstm as generic vision backbone, in: International Conference on Learning Representations (ICLR), 2025. URL: https://arxiv.org/abs/2406.04303.

[5] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 10347–10357. URL: https://arxiv.org/abs/2012.12877.

[6] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, L. Yuan, Tinyvit: Fast pretraining distillation for small vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 68–85.

[7] J. Zhang, H. Peng, K. Wu, M. Liu, B. Xiao, J. Fu, L. Yuan, Minivit: Compressing vision transformers with weight multiplexing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12145–12154.

[8] Z. Hao, J. Guo, D. Jia, K. Han, Y. Tang, C. Zhang, H. Hu, Y. Wang, Learning efficient vision transformers via fine-grained manifold distillation, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 9164–9175.

[9] S. Lin, H. Xie, B. Wang, K. Yu, X. Chang, X. Liang, G. Wang, Knowledge distillation via the target-aware transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10941–10950.

[10] S. Ren, Z. Gao, T. Hua, Z. Xue, Y. Tian, S. He, H. Zhao, Co-advise: Cross inductive bias distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16773–16782.

[11] Y. Li, G. Yuan, W. Wen, et al., Efficientformer: Vision transformers at mobilenet speed, in: Advances in Neural Information Processing Systems, 2022.

[12] J. Pan, A. Bulat, F. Tan, X. Zhu, Ł. Dudziak, H. Li, G. Tzimiropoulos, B. Martinez, EdgeViTs: Competing light-weight CNNs on mobile devices with vision transformers, in: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI, Springer, 2022, pp. 294–311. doi:10.1007/978-3-031-20083-0_18.

[13] M. Schäfer, U. Faltings, B. Glaser, Clrius: Contrastive learning for intrinsically unordered steel scrap, Machine Learning with Applications 17 (2024) 100573.

[14] J. Lei, S. Huang, Z. Shen, L. Wu, D. Liu, Laser: Layered scene recognition for visual intelligent recycling of steel scrap, IEEE Transactions on Industrial Informatics (2025).

[15] S. Temur, L. Karacan, Performance of resnet and transformer-based models in the classification of steel scrap, in: 2024 9th International Conference on Computer Science and Engineering (UBMK), IEEE, 2024, pp. 1–6.

[16] S. Janzen, H. Stein, K. Trinley, et al., ESCADE: Energy-efficient artificial intelligence for cost-effective and sustainable data centers, in: Proceedings of the Research Projects Exhibition at the 37th International Conference on Advanced Information Systems Engineering (RPE@CAiSE 2025), 2025. CEUR Workshop Proceedings, Vol. 4050.

[17] Z. Hao, J. Guo, K. Han, Y. Tang, H. Hu, Y. Wang, C. Xu, One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation, in: Advances in Neural Information Processing Systems, 2023. ArXiv:2310.19444.

[18] A. Bick, K. Y. Li, E. P. Xing, J. Z. Kolter, A. Gu, Transformers to SSMs: Distilling quadratic knowledge to subquadratic models, in: Advances in Neural Information Processing Systems, 2024. ArXiv:2408.10189.

[19] G. Wei, R. Chellappa, ViT-Linearizer: Distilling quadratic knowledge into linear-time vision models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025. ArXiv:2504.00037.

[20] A. M. O. Thiombiano, B. Hnich, A. Ben Mrad, M. W. Mkaouer, Distil-xlstm: Learning attention mechanisms through recurrent structures, 2025. ArXiv:2503.18565. Uses xLSTM as the student for language modelling — opposite direction from our setting.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[22] I. Mavromatis, K. Katsaros, A. Khan, Computing within limits: An empirical study of energy consumption in ml training and inference, arXiv preprint arXiv:2406.14328 (2024).