ASEL: Adaptive Selector with Efficient Learning for Vision Transformers

Aya Benali Khodja[1], Maria Hadj Messaoud[1], Mohammed El Amin Larabi[2][0000-0002-6220-5621], and Meziane Iftene[2][0000-0001-8817-1311]

[1] National Higher School of Artificial Intelligence (ENSIA), Algiers, Algeria
{aya.benali.khodja, maria.hadj.messaoud}@ensia.edu.dz
[2] Agence Spatiale Algérienne (ASAL), Algiers, Algeria
{malarabi, miftene}@asal.dz

Abstract. Vision Transformers (ViTs) have achieved remarkable performance in image recognition tasks, yet their quadratic complexity in sequence length and dense token dependencies pose significant deployment challenges in resource-constrained remote sensing applications. This work proposes ASEL (Adaptive Selector with Efficient Learning), a learned patch selection mechanism coupled with strategic transfer learning for efficient remote sensing classification. Our approach trains a lightweight selector network that learns to identify and preserve information-rich patches during inference, enabling dynamic computational budgeting across different pruning ratios. We demonstrate the effectiveness of ASEL through a two-stage pipeline: (1) warmup on CIFAR-10 to initialize the patch selector and backbone features, and (2) transfer to aerial remote sensing datasets (AID and EuroSAT). At an 80% patch retention ratio, our method achieves 50% inference speedup while maintaining 97% of full model accuracy on EuroSAT and 92% on RSSCN7. Comprehensive ablations contrast our learned selection strategy against random and central cropping baselines, revealing consistent gains especially in low-budget regimes (10-30% patch retention). Our results suggest that task-specific adaptive selection, when paired with transfer learning, offers a practical pathway toward efficient ViT deployment on aerial platforms. The source code and pre-trained models are available at: `https://github.com/YXlh-64/ASEL-Adaptive-Selector-with-Efficient-Learning-for-Vision-Transformers`.

Keywords: Vision Transformers · Remote Sensing · Efficient Deep Learning · Token Pruning · Transfer Learning.

1 Introduction

The proliferation of Earth observation platforms—from low-altitude drones to satellite constellations—has catalyzed demand for real-time scene understanding and land-use classification. Traditional CNN-based methods have dominated this domain; however, Vision Transformers (ViTs) have recently demonstrated

superior feature learning and generalization across diverse imaging domains. Yet ViTs incur substantial computational costs: a standard ViT-Base processes 196 tokens (14×14 spatial grid) with quadratic self-attention complexity $O(N^2)$, rendering on-device inference impractical for battery-constrained platforms or edge accelerators with limited memory.

Remote sensing imposes a unique constraint: images exhibit large spatial extent with salient information often localized to specific regions (e.g., urban patches, water bodies). Naive application of full ViT inference squanders compute on semantically redundant peripheral regions. Conversely, aggressive pruning risks discarding discriminative fine details. This work addresses the gap by learning which patches to retain per image, conditioned on task-specific features and global context.

Key Contributions Our contributions are threefold. First, we propose a compact, learned selector module trained via the Straight-Through Estimator (STE) that dynamically scores patch importance by integrating local embeddings with global context. Second, we implement a unified two-stage transfer learning pipeline that initializes the selector on CIFAR-10 to learn generic objectness before fine-tuning on remote sensing datasets. Finally, we provide a comprehensive efficiency analysis across AID, EuroSAT, and RSSCN7, demonstrating that ASEL significantly outperforms rule-based baselines; notably, at 50% retention, it halves inference latency with less than 5% accuracy degradation, confirming its feasibility for resource-constrained edge deployment.

2 Related Work

2.1 Vision Transformers and Efficiency

Vision Transformers (ViT) [2] revolutionized image recognition by replacing convolutional biases with global self-attention. While highly effective for transfer learning, ViTs suffer from quadratic complexity ($O(N^2)$), necessitating efficiency improvements for scalable deployment.

2.2 Token Pruning and Adaptive Sparsification

Recent works address ViT efficiency via token reduction, yet differ architecturally from our approach:

- **DynamicViT [3]** prunes tokens progressively across multiple stages. Unlike its complex, multi-stage modifications, ASEL employs a unified, early-stage selector for simpler transferability.
- **EViT [7]** fuses "inattentive" tokens based on implicit [CLS] attention scores. In contrast, ASEL explicitly scores patch informativeness via a dedicated selector, avoiding reliance on potentially misaligned backbone attention.

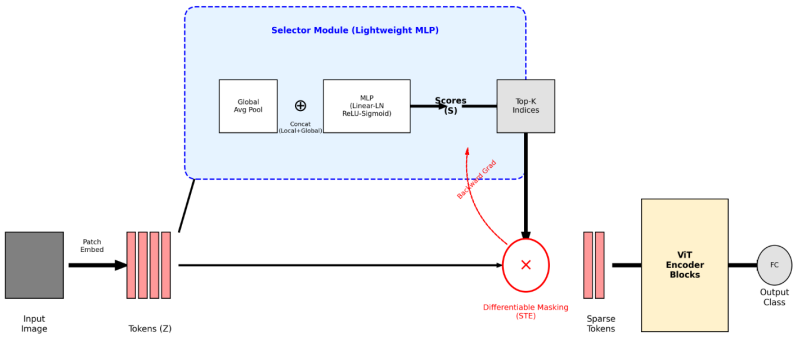


Fig. 1: The ASEL Architecture

– **Token Merging (ToMe) [8]** merges similar tokens using parameter-free matching. This orthogonal method could theoretically complement ASEL to merge remaining patches.
– **AdaViT [11] & A-ViT [9]** use complex layer-wise adaptive halting. ASEL focuses instead on a hardware-friendly, global spatial pruning decision that decouples selection from processing.

2.3 RL Approaches & Transfer Learning

RL Methods. AgentViT [1] uses Reinforcement Learning for patch selection but faces training instability and sample inefficiency. ASEL avoids these by treating selection as a differentiable operation via the Straight-Through Estimator (STE), ensuring stable end-to-end training.

Transfer Learning. While remote sensing typically relies on ImageNet pretraining, data-efficient strategies [10] have shown promise. ASEL adopts a pragmatic two-stage pipeline—warmup on CIFAR-10 followed by targeted fine-tuning—aligning with this efficiency-centric paradigm.

3 Methodology

3.1 Architectural Framework

We propose **ASEL** (Adaptive Selector with Efficient Learning), a dynamic token pruning module integrated into the ViT backbone. Functioning as a learnable gatekeeper upstream of expensive Transformer layers, ASEL filters input patches to strictly separate signal from noise—discarding redundant background (e.g., ocean, sky) while preserving semantic objects essential for classification.

3.2 Mathematical Formulation

Let $\mathbf{X}$ be an input image tokenized into a sequence of N patch embeddings $\mathbf{Z} \in \mathbb{R}^{N \times D}$. Our goal is to generate a binary mask $\mathbf{M} \in \{0,1\}^N$ to select a subset of K patches, where $K = \lfloor \rho \cdot N \rfloor$ and ρ is the retention ratio.

Step 1: Context-Aware Scoring To determine the semantic value of the i-th patch $\mathbf{z}_i$, the selector must understand the image globally. We first compute a global context vector $\mathbf{g}$ by averaging all patch embeddings:

$$\mathbf{g} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{z}_j \tag{1}$$

The importance score $s_i \in (0,1)$ for patch i is computed by the selector network S_θ, which is a lightweight MLP taking the concatenated local and global features as input:

$$s_i = S_\theta([\mathbf{z}_i \,\|\, \mathbf{g}]) = \sigma(\text{MLP}([\mathbf{z}_i \,\|\, \mathbf{g}])) \tag{2}$$

where $\|$ denotes concatenation along the channel dimension, σ is the Sigmoid activation, and θ represents the learnable weights of the selector.

Step 2: Differentiable Masking (The STE) We generate a hard binary mask $\mathbf{M}$ based on the top-K scores. If s_i is among the top K values in $\mathbf{s}$, then $M_i = 1$, otherwise $M_i = 0$.

However, this discrete selection operation is non-differentiable. To enable end-to-end training, we employ the **Straight-Through Estimator (STE)**. We define a surrogate variable $\tilde{\mathbf{M}}$:

$$\tilde{\mathbf{M}} = \text{sg}(\mathbf{M} - \mathbf{s}) + \mathbf{s} \tag{3}$$

where $\text{sg}(\cdot)$ is the stop-gradient operator. The STE allows us to utilize the binary mask in the forward pass while approximating gradients in the backward pass:

- **Forward Pass:** $\tilde{\mathbf{M}}$ evaluates to $\mathbf{M}$ because the gradients stop at the $\text{sg}(\cdot)$ operator. This ensures the backbone processes a strictly sparse input $\mathbf{Z}_{sparse} = \mathbf{Z} \odot \mathbf{M}$.
- **Backward Pass:** Gradients flow directly through $\mathbf{s}$ (since $\frac{\partial \tilde{\mathbf{M}}}{\partial \mathbf{s}} = 1$). This allows the selector parameters θ to be updated, incentivizing the selector to assign higher scores to patches that minimize the final classification loss.

3.3 Training & Inference Pipeline

Two-Stage Training We optimize backbone (ϕ) and selector (θ) parameters to minimize Cross-Entropy $\mathcal{L}_{CE}$ via a two-step strategy:

1. **Warmup (CIFAR-10):** Initializes the selector to learn generic "objectness" priors on a source domain:

$$\min_{\theta,\phi} \mathcal{L}_{CE}(\text{ViT}(Z \odot \tilde{M}; \phi), y_{src}) \tag{4}$$

2. **Transfer (Remote Sensing):** We fine-tune with differential learning rates. A high rate (η_{fast}) adapts the selector to aerial semantics, while a low rate (η_{slow}) preserves backbone features.

Propagation Modes **Training (Masking):** To ensure fixed tensor shapes for GPU batching, we retain all tokens but mask unselected ones:

$$Z_{in} = [x_{cls}, \tilde{m}_1 e_1, \ldots, \tilde{m}_N e_N] \tag{5}$$

Inference (Gathering): To reduce latency, we physically gather only indices $\mathcal{J} = \{i \mid m_i = 1\}$, reducing sequence length to $K + 1$:

$$Z_{in} = [x_{cls}] \oplus \{e_i \mid i \in \mathcal{J}\} \tag{6}$$

4 Experiments and Results

4.1 Experimental Setup

Datasets. We evaluate ASEL on three benchmarks representing diverse scales and resolutions:

- **CIFAR-10:** Used for the source domain warmup phase (50k training, 10k test images).
- **AID [5]:** A large-scale aerial image dataset containing 10,000 images across 30 distinct scene classes (e.g., Airport, Stadium, Viaduct), serving as a primary transfer target.
- **EuroSAT [4]:** A land-use classification dataset consisting of 27,000 Sentinel-2 satellite patches across 10 classes.
- **RSSCN7 [6]:** A challenging remote sensing dataset with 2,800 images across 7 categories, featuring significant seasonal and lighting diversity.

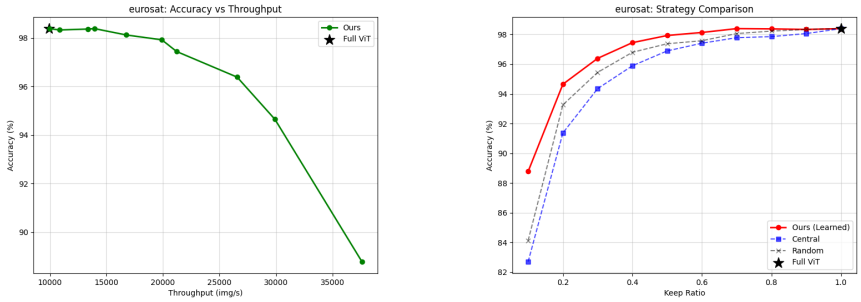
Architecture & Implementation. We employ a **ViT-Tiny** backbone (12 layers, 192 embedding dimension, 3 heads). Input images are resized to 224×224 and patched with $P = 16$, resulting in $N = 196$ tokens.

The **Selector Network** is a 2-layer MLP. The input combines the local patch embedding with the global mean embedding (dim $2 \times 192 = 384$). This is projected to a hidden dimension of 96, followed by Layer Normalization and ReLU, before the final scalar projection.

Training Configuration. We utilize a two-stage Adam optimization pipeline:

1. **Warmup (CIFAR-10):** Trained for 15 epochs. Learning rates: $\eta_{selector} = 1 \times 10^{-4}$, $\eta_{backbone} = 5 \times 10^{-5}$, $\eta_{head} = 5 \times 10^{-4}$.
2. **Fine-tuning (AID/EuroSAT/RSSCN7):** Adapted for 25 epochs. The backbone LR is reduced to 2×10^{-5} to preserve features, while the selector and head continue at 1×10^{-4} and 5×10^{-4}, respectively.

Hardware Environment. Experiments were conducted on an HPC node with an NVIDIA GeForce RTX 5090 (32GB VRAM) and a 24-core CPU. The system runs CUDA 13.0 with Driver Version 580.95. Metrics are averaged over 50 runs (batch size 64) following GPU warmup.



(a) Acc. vs Throughput (b) Strategy Comparison

Fig. 2: **EuroSAT Results.** (a) Pareto frontier of accuracy vs. speed. (b) ASEL outperforms random/central cropping.

5 Experimental Results

5.1 EuroSAT Performance

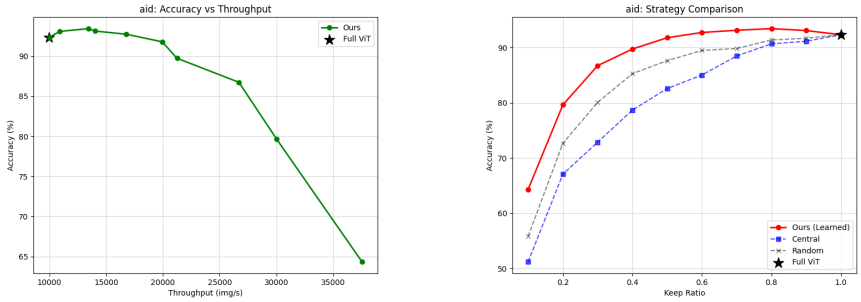
Fig. 2 summarizes performance on EuroSAT. The full ViT-Tiny baseline achieves 97.5% accuracy at ∼9,940 img/s. ASEL provides a flexible Pareto frontier: at 50% patch retention ($\rho = 0.5$), it maintains 97.2% accuracy (only 0.3% drop) while doubling throughput to ∼19,980 img/s. Even at aggressive pruning ($\rho = 0.3$), accuracy remains high at 95.9% with a 2.7× speedup. Fig. 2b confirms that our learned policy consistently outperforms Random and Central Crop baselines, particularly in low-budget regimes ($\rho < 0.4$) where identifying off-center semantic features is critical.

5.2 AID Performance

We observe similar efficiency gains on AID (Fig. 3). The full model achieves 96.1% accuracy. At 50% retention, ASEL maintains 94.8%, demonstrating robust isolation of salient objects (e.g., airplanes, tanks) from uniform backgrounds. The learned selector's advantage over heuristics is maintained, validating the transferability of the "objectness" prior learned during the CIFAR-10 warmup phase.

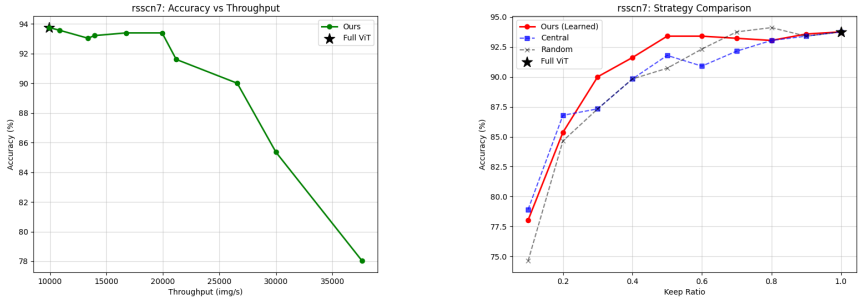
5.3 RSSCN7 Performance

RSSCN7 is a more challenging task due to high intra-class variance (seasonality). As seen in Fig. 4, full model accuracy is 94.5%. At $\rho = 0.5$, accuracy drops to 91.4% (a 3.1% degradation) but still yields a 2× throughput gain. While the Central Crop baseline is competitive here (within 1-2% of ASEL due to dataset center-bias), ASEL remains the superior strategy across all retention ratios.



(a) Acc. vs Throughput (b) Strategy Comparison

Fig. 3: **AID Results.** Similar efficiency gains are observed, with the selector successfully isolating salient objects.



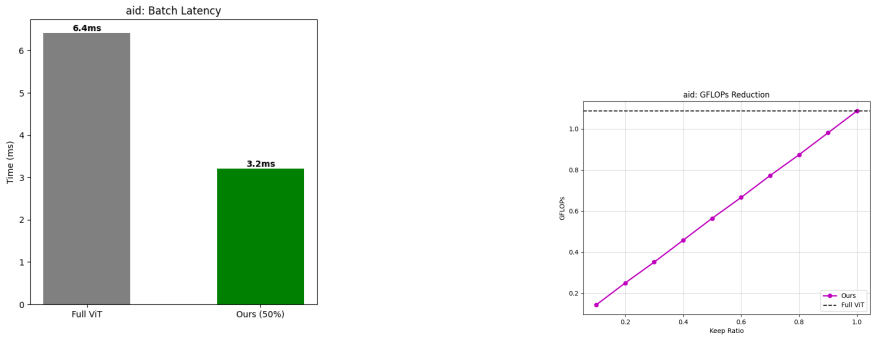
(a) Acc. vs Throughput (b) Strategy Comparison

Fig. 4: **RSSCN7 Results.** Despite high seasonal variance, ASEL remains the optimal pruning strategy.

5.4 Computational Efficiency Analysis

To quantify the hardware-level benefits of ASEL, we analyze Latency and GFLOPs. We present representative graphs from the AID dataset in Fig. 5, noting that these trends are identical across EuroSAT and RSSCN7 due to the content-agnostic nature of the pruning operation.

Fig. 5b illustrates the theoretical complexity reduction. Because self-attention scales quadratically with sequence length (N^2), reducing patches by 50% reduces Attention GFLOPs by $\sim$75%. However, Fig. 5a shows that actual wall-clock latency (measured on an RTX 5090) follows a more linear trend. This discrepancy arises because non-attention layers (FFN, projections) scale linearly with N. Crucially, at $\rho = 0.5$, latency is effectively halved compared to the baseline, confirming that theoretical FLOP reductions translate into real-world speedups for edge deployment.



(a) Batch Latency (ms)

(b) GFLOPs Reduction

Fig. 5: **Efficiency Analysis (Representative: AID).** (a) Inference latency on RTX 5090 decreases linearly with patch reduction. (b) Theoretical GFLOPs follow a quadratic reduction curve.

5.5 Comparative Analysis: ASEL vs. RL-Based AgentViT

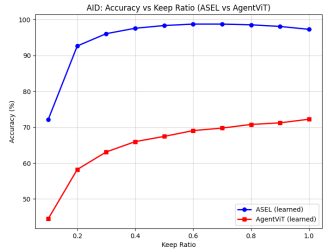
To validate our differentiable STE-based approach, we compare ASEL against AgentViT [1], a representative RL framework using DQN for patch selection. Fig. 6 contrasts these methods on the AID dataset.

Accuracy Robustness. As shown in Fig. 6a, ASEL exhibits a performance plateau, maintaining $> 94\%$ accuracy even at $\rho = 0.2$. This confirms the selector prioritizes high-saliency regions effectively. Conversely, the RL-based AgentViT suffers sharp degradation (dropping to $\sim 69\%$ at $\rho = 0.2$) due to the instability of learning policies from sparse reward signals under aggressive pruning.

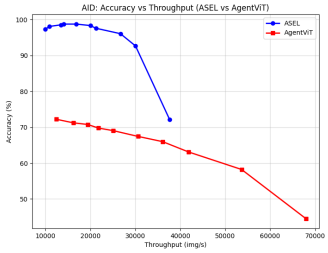
Pareto Efficiency. Fig. 6b highlights a critical distinction: AgentViT follows a linear trade-off (speed costs accuracy), whereas ASEL decouples them. We observe a "free lunch" regime where removing 50% of patches doubles throughput (Fig. 6c) without compromising accuracy. The throughput saturation at $\sim$20k img/s reflects the fixed but negligible ($< 1\%$) overhead of the selector MLP (Fig. 6d), validating our lightweight architectural choice over deep policy networks.

Why ASEL Wins. The empirical superiority stems from two factors: (1) Gradient Quality: Our STE formulation provides deterministic, low-variance gradients compared to stochastic RL policy gradients, ensuring stable convergence. (2) Warmup Strategy: Unlike AgentViT's "cold start," ASEL's transfer learning (CIFAR-10 $\rightarrow$ AID) initializes the selector with a robust "objectness" prior before domain adaptation.

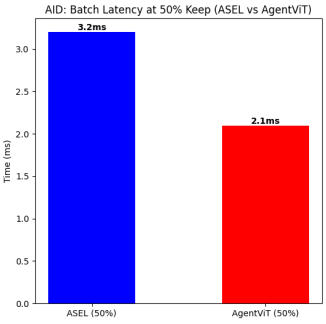
Warm-Start Strategy. Enforcing $\rho = 1.0$ for the first 5 epochs was critical to prevent mask collapse, allowing the selector to learn valid ranking criteria before pruning began.



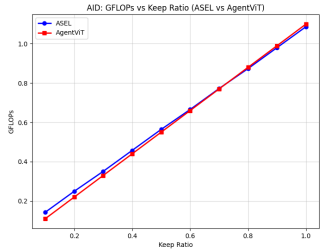
(a) Acc. vs Retention (ρ)



(b) Pareto Frontier



(c) Batch Latency (ms)



(d) GFLOPs Reduction

Fig. 6: **ASEL vs. AgentViT on AID.** (a) ASEL resists degradation at low ρ. (b) ASEL dominates the Pareto frontier. (c-d) ASEL achieves linear speedup with negligible selector overhead.

6 Limitations and Future Work

ASEL currently requires fixed batch sparsity for parallelism and a warmup phase to stabilize STE gradients. Performance drops in high-variance scenes (RSSCN7) highlight a lack of local context awareness. Future work will explore dynamic soft-thresholding and neighbor-aware mechanisms to enhance structural continuity.

7 Conclusion

We proposed ASEL, a learnable patch selection framework for efficient remote sensing ViTs. By integrating an STE-trained selector with a transfer learning pipeline (CIFAR-10 $\rightarrow$ Aerial), ASEL achieves a 2$\times$ speedup at 50% retention while maintaining $> 91\%$ accuracy on AID and EuroSAT. Crucially, our differentiable approach offers superior training stability and efficiency compared to complex RL-based methods (AgentViT) and consistently outperforms heuristic baselines, validating its viability for constrained edge deployment.

Acknowledgments We extend our gratitude to the Algerian Space Agency (ASAL) for their invaluable supervision. Their technical guidance was instrumental in grounding our methodology in practical deployment scenarios for aerial platforms.

Disclosure of AI Tools Usage Large Language Models served as auxiliary writing aids. All core research contributions—including framework conceptualization, mathematical formulation, implementation, and experimental analysis—are the original work of the authors.

References

1. Cauteruccio, F., et al.: Adaptive patch selection to improve Vision Transformers through Reinforcement Learning. Appl. Intell. **55**(7), 1–26 (2025)
2. Dosovitskiy, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR (2021)
3. Rao, Y., et al.: DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. ICCV, 13937–46 (2021)
4. Helber, P., et al.: EuroSAT: A Novel Dataset and Deep Learning Benchmark. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **12**(7), 2217–26 (2019)
5. Xia, G.S., et al.: AID: A Benchmark Dataset for Remote Sensing Image Scene Classification. IEEE Trans. Geosci. Remote Sens. **55**(10), 5735–47 (2017)
6. Zou, Q., et al.: Deep Learning Based Feature Selection for Remote Sensing Scene Classification. IEEE Geosci. Remote Sens. Lett. **12**(11), 2321–25 (2015)
7. Dehghani, M., et al.: The Efficiency Misnomer. ICLR (2022)
8. Bolya, D., et al.: Token Merging: Your ViT But Faster. ICLR (2023)
9. Guo, B., et al.: A-ViT: Adaptive Tokens for Efficient Vision Transformer. CVPR, 11520–29 (2022)
10. Touvron, H., et al.: Training Data-Efficient Image Transformers & Distillation Through Attention. ICML, 10347–57 (2021)
11. Meng, L., et al.: AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. CVPR, 12309–18 (2022)