Fine-Grained Bird Species Classification

Domain-Specific Pretraining, Multi-View Detection & Specialist Models

Olha Baliasina & Nima Kamali Lassem

BDMA 07 | CentraleSupélec | February 2026

Presentation Outline

1. Problem & Dataset Analysis
2. Our Approach — Pipeline Overview
3. Method Deep Dive
4. Experiments & Results
5. What Didn't Work
6. Key Takeaways

The Task

Classify 20 bird species from the CUB-200-2011 dataset in a Kaggle competition setting.

1,212	~60	400	20
Training Images	Per Class	Test Images	Species

Key Challenges

- Low-data regime — only ~60 labeled images per class
- Subtle inter-class differences within taxonomic families
- High intra-class variation (pose, lighting, background)
- Only provided labeled data can be used for classification

Why Is This Hard?

The FGVC Paradox: different species look alike, same species looks different

Error Concentration After Best Backbone

American Crow ↔ Fish Crow — 55.6%

Rusty ↔ Brewer Blackbird — 22.2%

Cuckoo pair — 5.6%

Taxonomic Families in Our Dataset

Corvids	American Crow, Fish Crow
Icterids	Rusty/Brewer/Red-winged/Yellow-headed Blackbird, Bobolink, Cowbird
Buntings	Indigo, Lazuli, Painted Bunting
Cuculids	Yellow-billed/Black-billed Cuckoo, Groove-billed Ani

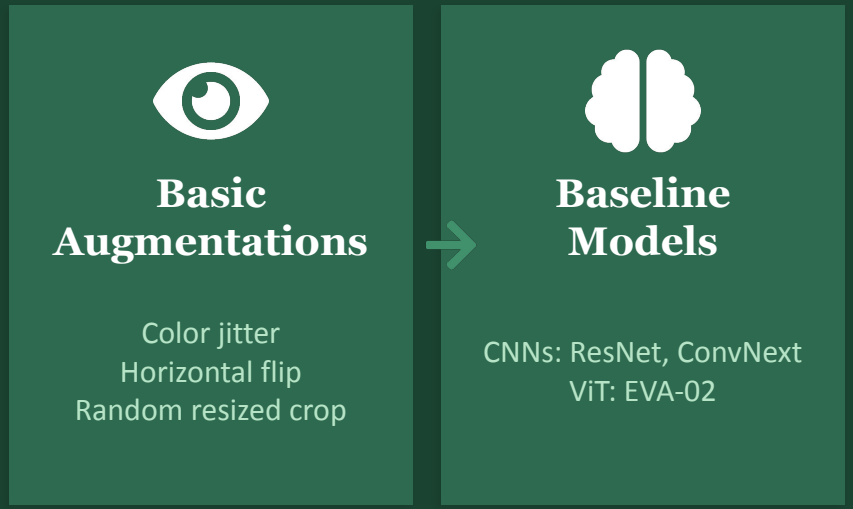
Point the odd one out among these pictures

American Crow?

Fish Crow?

Point the odd one out among these pictures

Our Simple Pipeline (Baseline)



Using the provided train/validation split (1109/103 images)

Our Full Pipeline

Detection

YOLOv8m +
Grounding DINO
(99% coverage)

Multi-View

Full image +
Bird crop
(curriculum)

CNN / VIT Backbone

Pretrained on ImageNet / iNaturalist
+ 5-fold ensemble
+ TTA
+ stronger augmentations
(CutMix, CLAHE etc)

Blending

Confidence-gated
full/crop fusion
(3 modes)

Specialist Override: 3 dedicated binary classifiers (Crow, Blackbird, Cuckoo) with dynamic α blending activated when main model predicts confused pair or combined probability > 0.3.

Cross Validation with Full Dataset: All labeled data were concatenated to be used for 5 fold cross validation training.

Stage 1: Two-Stage Object Detection

Why? Raw images contain background clutter. Localizing the bird before classification reduces noise.

YOLOv8m (Primary)

COCO class 14 = "bird"
Confidence threshold: 0.25
15% bounding box padding
Select largest detection by area

Handles ~97% of all images

Grounding DINO (Fallback)

Zero-shot open-vocabulary detector
Text prompt: "a bird", “a bird perched in a tree”
Box threshold: 0.25
Recovers YOLO misses

Total coverage: >99%

All bounding boxes precomputed and cached — zero runtime overhead during training

Stage 1: Two-Stage Object Detection

Why? Raw images contain background clutter. Localizing the bird before classification reduces noise.

YOLOv8m (Primary)

Detected

05d54ffe-c3ef-41c8-94eb-44bdd2c77311.jpg

Train: 1212/1212
Test: 387/400

Undetected

583b57ee-62c5-4441-8af7-b4536bdd3395.jpg

Test: 13/400

Grounding DINO (Fallback)

Detected

GDINO recovered
583b57ee-62c5-4441-8af7-b4536bdd3395.jpg

Test: 12/13

All bounding boxes precomputed and cached — zero runtime overhead during training

Stage 2: EVA-02 Large — The Key Ingredient

Architecture

Type	Vision Transformer (ViT-L/14)
Parameters	304M
Resolution	336 × 336 px
Features	1024-dim patch tokens
Pooling	Mean over spatial tokens
Head	Dropout(0.3) → Linear(20)

Why iNaturalist Pretraining?

CLIP (2B pairs) → **iNat21 (10k spp.)**

ImageNet teaches: edges, textures, object shapes

iNaturalist teaches: plumage patterns, bill morphology, body proportions

Result: backbone already near-optimal for bird species — converges in 4-5 epochs

+4% accuracy over ConvNeXt-Base (IN-22k) with identical pipeline — the single largest gain

Pretraining domain matters!

Training Configuration for EVA-02

Parameter	Value	Why?
Backbone LR	**3×10^{-6}**	Near-optimal features, minimal adjustment
Head LR	**5×10^{-4}**	Random init, must learn from scratch
Warmup	**2 epochs**	Prevent catastrophic forgetting
Label smoothing	**$\varepsilon = 0.1$**	Calibration + regularization
Dropout	**$p = 0.3$**	304M params vs ~1200 images
Weight decay	**0.05**	Standard for ViT fine-tuning
Mixup / CutMix	**$p = 0.2$ each**	Smoother decision boundaries
Epochs	**20 (early stop: 8)**	Converges by epoch 4-10
CV strategy	**5-fold stratified**	All data used for training + validation

Data augmentation: RandomResizedCrop, HFlip, ShiftScaleRotate, CLAHE, ColorJitter, GaussianBlur/Noise, CoarseDropout (Albumentations)

Inference and Confidence-Gated Blending

Test-Time Augmentation (4 views)

Original + Horizontal flip + Scale×1.15 + Flip+Scale×1.15

Logit averaging (not probability) for better calibration

5-Fold CV

Accuracy-cubed weighting: $w_k = a_k^3 / \Sigma a_j^3$

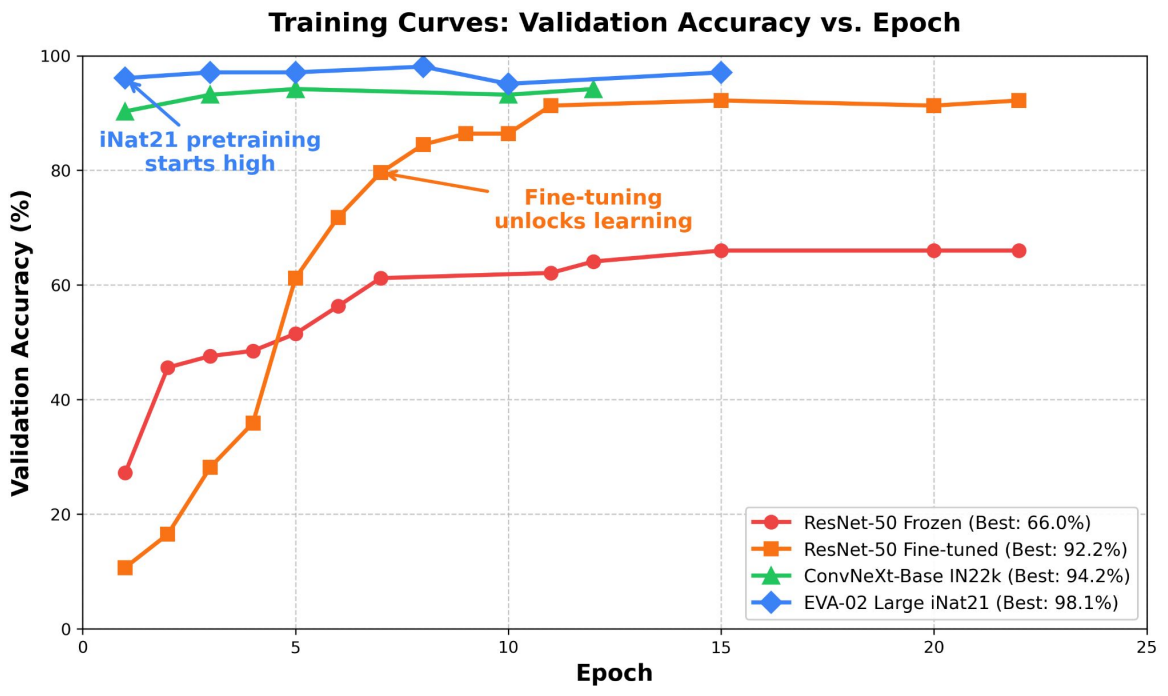
Higher-accuracy folds get disproportionate influence

Confidence-Gated Multi-View Blending

Crop Only (4.5%)	Multi-View (35.5%)	Full Only (30.8%)
Crop conf > 0.85 and gap > 0.08	**Full conf < 0.90**	**Full conf ≥ 0.90**
Crop isolates bird perfectly; full image confused by background	0.75 × full + 0.25 × crop Crop provides second opinion	Model already confident; crop would only add noise

Baseline Experiments

Simple pipeline: no detection, no CV, no TTA, provided train/test split. Isolating the impact of backbone choice.



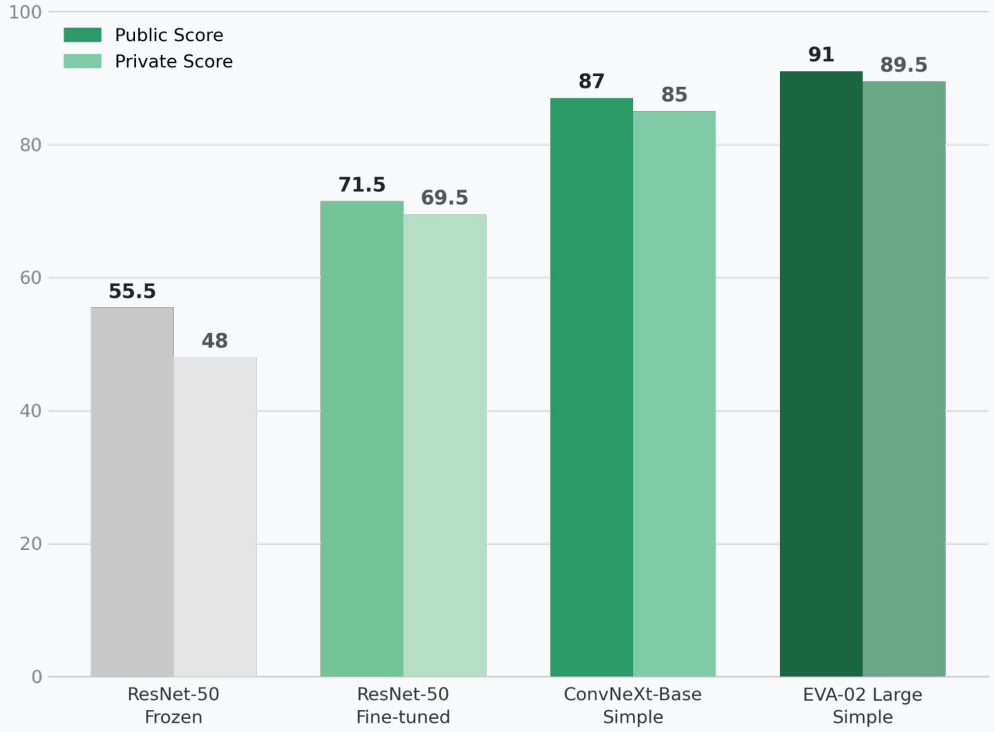
Fine-tuning unlocks learning

Richer pretraining (IN-22k, 384px) + ConvNeXt

Domain-specific pretraining (iNat21) + VIT

Baseline Experiments

Simple pipeline: no detection, no CV, no TTA, provided train/test split. Isolating the impact of backbone choice.



+16.5% Fine-tuning unlocks learning

+15.5% Richer pretraining (IN-22k, 384px)

+4.0% Domain-specific pretraining (iNat21)

Ablation Study — Every Component Matters

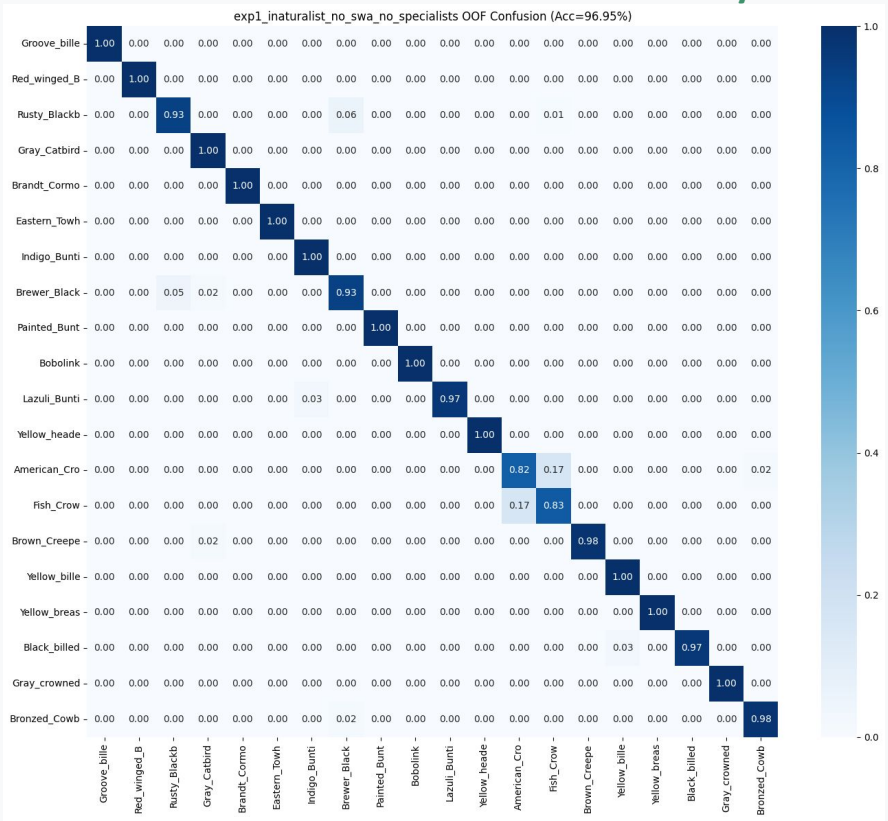
Configuration	CV Acc	Kaggle public	Δ public	Kaggle private	Δ private
ConvNeXt-Base (simple pipeline)	94.17%	0.870	—	0.850	
+ Detection + 5-fold CV + TTA	94.55%	0.865	-0.5%	0.875	+2.5%
EVA-02 iNat21 (simple pipeline)	98.06%	0.910	+4.0%	0.895	+4.5%
+ Detection + 5-fold CV + TTA	96.95%	0.925	+5.5%	0.900	+5.0%

Key Observations

- ConvNeXt + full pipeline actually drops 0.5%! Pipeline overhead hurts without strong backbone
- EVA-02 simple pipeline (0.91) already beats ConvNeXt full pipeline (0.865)
- All ConvNeXt experiments (Swin, EfficientNetV2, EVA ensembles) plateaued at 0.88
- **Pretraining domain is a significant factor**

Error Analysis — Where Mistakes Happen

12 of 20 classes achieve 100% OOF accuracy



5-Fold CV Results

96.95% ± 0.43%

Low variance → stable training
Best fold: 97.53% (epoch 4)

36 Total Errors

Crow pair: 20 errors (55.6%)

Blackbird pair: 8 errors (22.2%)

Cuckoo pair: 2 errors (5.6%)

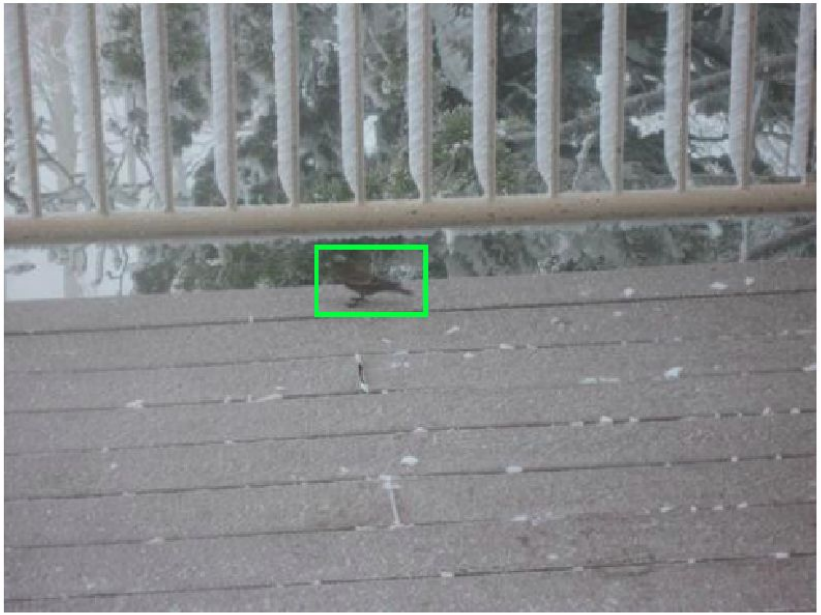
Other scattered: 6 (16.7%)

Error Analysis — Where Mistakes Happen

Examples of low confidence predictions:

4bffce44-78aa-43d9-b28b-2de0a1caffeb.jpg
Pred: Fish_Crow
Conf: 0.278 | crow_spec | YOLO/GDINO

4e599817-3e0f-456a-8a72-f13267dadc2a.jpg
Pred: Rusty_Blackbird
Conf: 0.295 | bb_spec | YOLO/GDINO



Specialist Binary Classifiers

Dedicated models for the 3 most confused species pairs, triggered by main model uncertainty

Crow Specialist

Fish vs. American

86.67%

±6.12%

Blackbird Specialist

Brewer vs. Rusty

93.15%

±2.18%

Cuckoo Specialist

Yellow vs. Black-billed

98.33%

±2.04%

Override Logic

Trigger: prediction ∈ pair classes OR combined probability > 0.3

Blending: $\alpha = \min(0.5, 0.2 + 0.3 \times \text{specialist\_confidence})$

Effect: cautious (α=0.2) when uncertain → assertive (α=0.5) when confident. Only 3 predictions changed (all crows).

Ablation Study — Every Component Matters

Configuration	CV Acc	Kaggle public	Δ public	Kaggle private	Δ private
ConvNeXt-Base (simple pipeline)	94.17%	0.870	—	0.850	
+ Detection + 5-fold CV + TTA	94.55%	0.865	-0.5%	0.875	+2.5%
EVA-02 iNat21 (simple pipeline)	98.06%	0.910	+4.0%	0.895	+4.5%
+ Detection + 5-fold CV + TTA	96.95%	0.925	+5.5%	0.900	+5.0%
+ Specialist classifiers	—	**0.930**	+6.0%	**0.920**	+7.0%

Key Observations

- ConvNeXt + full pipeline actually drops 0.5%! Pipeline overhead hurts without strong backbone
- EVA-02 simple pipeline (0.91) already beats ConvNeXt full pipeline (0.865)
- All ConvNeXt experiments (Swin, EfficientNetV2, EVA ensembles) plateaued at 0.88
- **Pretraining domain is a significant factor**

What Didn't Work

Stochastic Weight Averaging

Result: 0.925 (no improvement)

- Insufficient averaging window — early stopping at epoch 10-12, only 5-7 checkpoints
- LR already near 10^{-7} — consecutive checkpoints nearly identical
- Pre-converged backbone explores narrow weight-space region

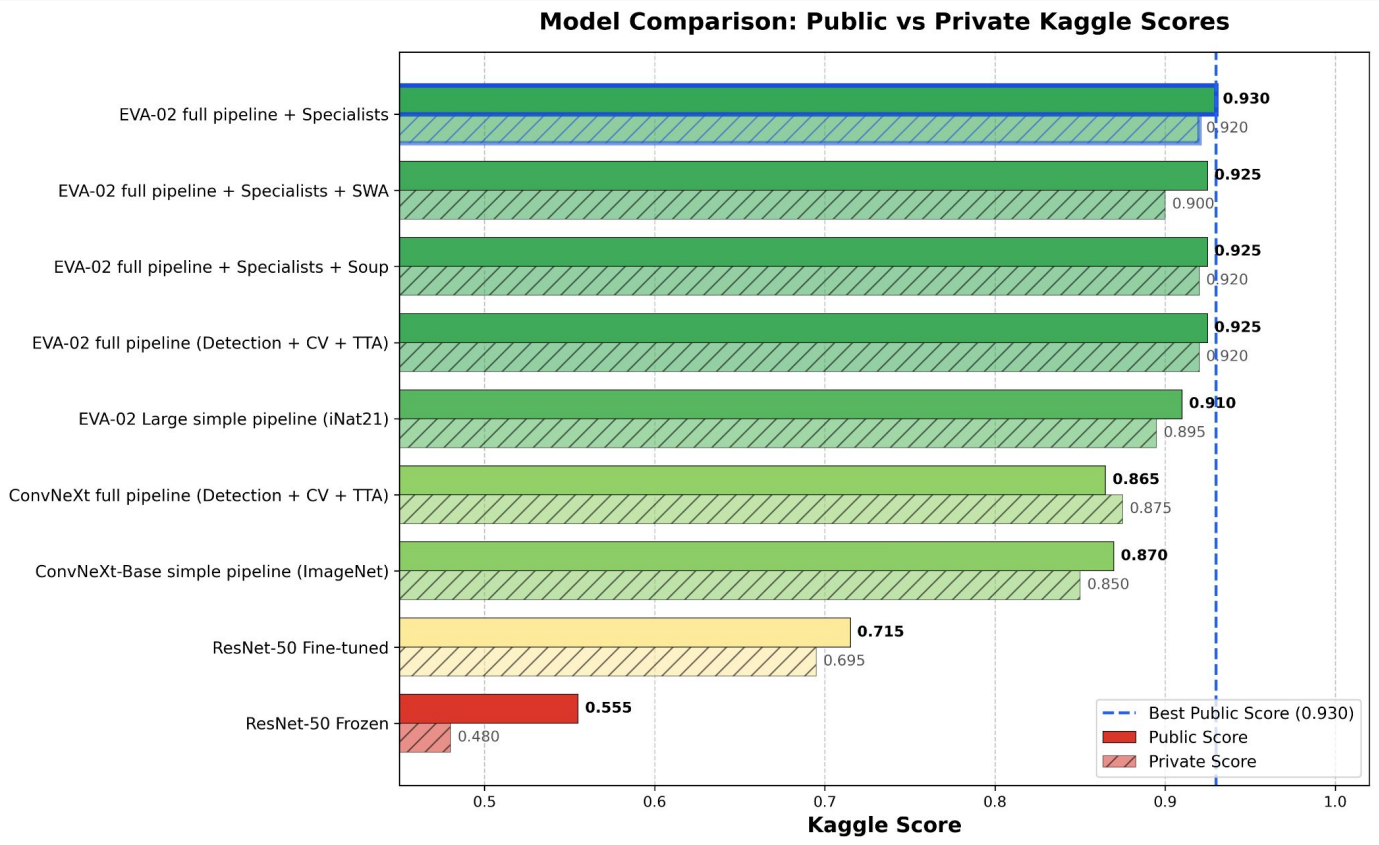
Model Soup

Result: 0.925 (no improvement)

- 25 runs (5 folds × 5 restarts) with greedy selection
- HP changes increased variance (±0.43% → ±1.24%) — training too sensitive
- ~1,200 images too few for diverse weight-space exploration
- Soup averaged away lucky variance of the best individual run

Lesson: Weight averaging assumes diverse optima. On small datasets, models converge similarly → no benefit from averaging.

Complete Score Progression



Key Takeaways

1

Pretraining Domain Matters

iNaturalist pretraining gave +5.5% — more than all other components combined. For domain-specific FGVC, backbone selection is the #1 priority.

2

Error Analysis Drives Improvement

Confusion matrix revealed 80% of errors in just 2 species pairs → directly motivated specialist classifiers. Without this, we'd waste effort on solved classes.

3

Targeted > General-Purpose

Specialist classifiers (+0.5%) with minimal effort outperformed SWA and model soup (0%) despite significant engineering investment.

4

Diminishing Returns Are Real

First decisions (backbone, pretraining) gave biggest gains. Later refinements yield ever smaller improvements at exponentially higher cost.

Thank You

55% → 93%

From frozen ResNet-50 to EVA-02 + Specialists in systematic steps

Questions?