Group 9 – Portuguese Bird-Species ID

Below is a one-week, Colab-free-tier-friendly plan that turns your promising baseline (EffNet-V2-S, macro-F1 ≈ 0.81) into a small set of focused experiments. Everything keeps GPU time ≤ 7 h on a single T4/P100.

🔑 Goal	What to show in the report	Why it matters
G-1 Lift rare/confused species (*Serinus* ↔ *Emberiza*, *Delichon* ↔ *Hirundo*)	at least +3 pp macro-F1 or > 0.05 PR-AUC on the four toughest classes	demonstrates you didn't just "scale a net" but reasoned about fine-grained cues
G-2 Explore **two distinct axes** (data-level & model-level)	an ablation table ("Full-Img base" vs "+MixUp" vs "+Part-crop", etc.)	meets the rubric's "experimenting in different circumstances" clause
G-3 Keep everything reproducible inside free Colab limits	GPU-minutes + VRAM per run in the scoreboard	shows good engineering practice

---

1 One-time pipeline fixes (*Day 1*)

Fix	How	Pay-off
Stratified 5-fold split (70 / 15 / 15)	`StratifiedGroupKFold` on species; reserve field photos for the *final* test	variance ↓; lets you quote mean ± std
Balanced sampler	√N sampling + *weighted* CE (class frequency^{-1}) **or** Focal γ = 2	tackles the residual 10 % class imbalance
Freeze 🧊 → **unfreeze schedule**	warm-up 3 epochs with stem+block1 frozen, then unfreeze	cuts first-epoch VRAM spike & speeds conv.
Early-stop patience = 5 on val macro-F1	avoids over-fitting in 25-epoch runs	

---

2 Experiment menu – pick 3 rows (Days 2-5)

ID	Hypothesis	Change vs. baseline	GPU h	Expected Δ
A	More context beats segmentation	**EffNet-B3 @288^2**, same aug, 25 ep	1.5	↑ macro-F1 2–3 pp
B	**MixUp α = 0.2 + CutMix α = 1.0** helps confusing pairs	add to full-image training	0.3	↑ PR-AUC on Serinus/Emberiza
C	**Part-crop fusion** lifts fine-grained cues	YOLOv8-n detects bird → SAM mask → crop (head+torso) 224^2; **two-branch net** (global + part, shared fc)	2.0	↑ class-specific recall 4 pp

ID	Hypothesis	Change vs. baseline	GPU h	Expected Δ
D	Domain pre-train on **CUB-200** transfers	load Swin-Tiny weights finetuned on CUB (available on Hugging Face) → 10 ep on your data	1.0	↑ macro-F1 if dataset too small
E	**Uncertainty class** reduces harmful mis-preds	keep baseline weights, set τ=0.6; map low-conf logits to "Uncertain"	0	practical UX win; negligible F1 loss

Tip: if GPU hours run short, do **A** + **B** + **E** – they finish in ≤ 2 h.

---

3 Regularisation & aug block (implement once)

```
train_tfms = A.Compose([
    A.RandomResizedCrop(224,224,scale=(0.8,1.0)),
    A.HorizontalFlip(p=0.5),
    A.ShiftScaleRotate(shift_limit=0.05, rotate_limit=20, scale_limit=0.1,
p=0.7),
    A.ColorJitter(0.1,0.1,0.1,0.05,p=0.8),
    A.CoarseDropout(max_holes=1, max_height=48, max_width=48, p=0.4),  #
Cutout light
])
```

Enable MixUp/CutMix via a flag (`--mix 0/1`).

---

4 Evaluation protocol (identical for every run)

- **Metrics:** macro-F1, macro-AUPRC, top-3 acc, per-class PR-AUC.
- **Thresholding:** optimise F1 per class on val ⇒ store τ.
- **Significance:** bootstrap 1 000× macro-F1 vs. baseline; ✱ if 95 % CI excludes 0.
- **Qualitative:** Grad-CAM++ for two hardest pairs; 6 TTA predictions on field photos.

---

5 One-week schedule

Day	Deliverable
1	Pipeline fixes; re-run 25-epoch *Baseline-V2* (EffNet-V2-S) – 40 min.
2	Train **EffNet-B3** with new aug (Exp A).
3	Same run + MixUp/CutMix flag (Exp B).
4	Build YOLOv8-n detector ▶ part-crop fusion net (Exp C) start train.
5	Finish Exp C (8 ep fine-tune), optional CUB-pre-train (Exp D).
6	Run uncertainty threshold sweep (Exp E); aggregate CV metrics.
7	Draft report: scoreboard, PR curves, Grad-CAM figs, compute table.

---

6 Scoreboard template for the report

Exp	Model	Extra	Macro-F1 ↑	PR-AUC ↑	GPU min	∗ Sig.?
Base-V2	V2-S	–	0.81	0.98	25	–
A	B3 @288	–	**0.84**	0.985	35	∗
B	B3	MixUp+CM	**0.86**	**0.989**	37	∗
C	B3+Part	MixUp	0.87	0.990	80	∗
E	B3	τ = 0.6 uncertain	–	–	–	UX

Colour the best metric per column; mark ∗ when CI excludes baseline.

Quick-wins checklist

- Stratified 5-fold & balanced sampler
- Longer train (25 ep) + warm-up freeze
- **MixUp/CutMix** to fight fine-grained confusion
- **B3 @ 288²** for richer features (fits in 12 GB with AMP)
- Optional **part-crop fusion** if time allows
- **Uncertainty class** with τ sweep for real-world UX
- Report compute + confidence intervals

Good luck!