EasyChair Preprint

№ 16024

What Moves, and What Misleads, the BirdCLEF+ 2026 Leaderboard: An Empirical Study Behind a 13th Place Multi-Taxon Soundscape System

Guanrui Lu and Ethan Tsai

July 31, 2026

What Moves, and What Misleads, the BirdCLEF+ 2026 Leaderboard: An Empirical Study Behind a 13th Place Multi-Taxon Soundscape System*

Guanrui Lu[1,*,†], Ethan Tsai[2,†]

[1]University of Southern California
[2]University of California, Irvine

Abstract

We describe the 13th place solution to BirdCLEF+ 2026, a multi-taxon passive-acoustic classification task covering 234 bird, amphibian, insect, and mammal taxa in Pantanal soundscapes. Our system averages six convolutional models, three CNN classifiers and three SED networks on EfficientNetV2, HGNetV2, and ECA-NFNet backbones. We blend the six 0.5/0.5 with a public Perch/ProtoSSM pipeline, scoring 0.957 public and 0.955 private. We use the system to study a practical problem: in the competitive region, offline validation does not predict the leaderboard. From 67 logged *(offline AUC, public LB)* pairs, Pearson correlation falls from 0.68 overall to 0.09 above LB 0.92 and turns negative above 0.925. We trace this to two causes: a saturated labeled-soundscape surface with an AUC spread of only 0.00089 across twelve strong models, and pseudo-label leakage that silently inflates soundscape validation scores. The dominant lever moving the actual leaderboard is pseudo-label quality: strong public pseudo-labels lift a single model from 0.881 to 0.934, while a weak self-teacher lowers it. Architectural diversity and the complementary Perch pipeline supply the remaining gains.

Keywords

Bioacoustics, Soundscape classification, Semi-supervised learning, Pseudo-labeling, Validation reliability, Sound event detection, Model ensembling, BirdCLEF

1. Introduction

Passive acoustic monitoring (PAM) lets ecologists track species presence at a scale manual surveys do not reach. Continuous audio works as a signal only after automatic annotation. The BirdCLEF series, run under the LifeCLEF umbrella [1, 2], drives much of the recent progress on this problem. Recent editions moved from birds alone in the Western Ghats [3] to multi-taxon sound identification in Colombia [4]. Its 2026 edition, BirdCLEF+ 2026 [5], pushes two old difficulties to an extreme. First, the target vocabulary now spans 234 taxa: birds, amphibians, insects, and mammals, recorded in the Pantanal wetlands of Brazil. Of these, 28 taxa have zero focal training audio. Second, the gap between the clean, single-source focal training recordings and the low-SNR, multi-source PAM soundscapes used for evaluation is large.

A participant who runs dozens of experiments here hits a worse problem than domain shift or the long tail. Offline validation stops predicting the leaderboard. Strong models score the same on local metrics. A change improving a held-out AUC often fails to improve the public score, and sometimes lowers it. Many teams report this. Few quantify the effect from one team's controlled record. Our working notes take this record as the object of study.

We make three contributions. First, we give a complete, reproducible description of a six-model CNN and SED ensemble, blended with a public Perch/ProtoSSM pipeline, placing 13th (Sections 3 to 4). Second, we study validation reliability from 67 of our own *(offline AUC, public LB)* pairs. We measure

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

*Working notes describing team *vogel*'s solution to the BirdCLEF+ 2026 challenge, hosted on Kaggle as part of LifeCLEF 2026.

*Corresponding author.

†These authors contributed equally.

stevelu@usc.edu (G. Lu); tsaie6@uci.edu (E. Tsai)

0009-0000-7948-8579 (G. Lu)

the decorrelation. We trace it to a saturated validation surface and to pseudo-label leakage, which contaminates the obvious fix of validating on soundscapes (Section 5). Third, we isolate what does move the score: pseudo-label quality first (Section 6), then diversity and a complementary pipeline (Section 7). We give a diagnosed account of the failed ideas (Section 8).

One framing note. We assemble our system largely from public parts: public training notebooks, public pretrained checkpoints, public pseudo-labels, and a public embedding pipeline. The contribution is not a new architecture. The contribution is the empirical and diagnostic work of combining the parts well and explaining why the choices worked. Our code, proxy logs, and the diagnostics behind every figure and table are public [6].

2. Related Work

The standard recipe for bioacoustic classification is the *spectrogram* $\rightarrow$ *CNN* $\rightarrow$ *classification head* pipeline [2]. BirdNET [2] is its most widely deployed form. A common refinement is the sound-event-detection (SED) head of Kong et al. [7]. The SED head pairs a clip-level, weakly labeled prediction with a frame-level attention branch. The frame-level output matches the 5-second-window evaluation of BirdCLEF. ImageNet-pretrained backbones such as EfficientNetV2 [8] and the normalizer-free NFNet and ECA-NFNet family [9] transfer well to mel spectrograms. GPU-friendly backbones such as PP-HGNetV2 [10] suit the CPU-only inference constraint. Beyond ImageNet, the Perch birdsong embeddings [11] give strong in-domain representations usable with simple probes.

Two themes dominate recent winning solutions. The first is semi-supervised learning on the unlabeled soundscapes through pseudo-labeling [12]. The BirdCLEF 2025 1st-place [13] and 2nd-place [14] systems both rely on it. The second is augmentation tuned for domain shift: mixup [15], SpecAugment [16], background mixing, and random equalization, as codified in the BirdCLEF 2023 2nd-place recipe [17]. The closest prior work to ours is the BirdCLEF 2025 2nd-place report [14]. The report also confronts the validation and leaderboard correlation problem. The authors find the correlation between mean ROC-AUC and the public score drops sharply once the public score passes 0.9. Our study reaches the same conclusion from an independent record and pushes further on the causes: surface saturation and pseudo-label leakage. Our system reuses public training notebooks for the CNN [18] and SED [19] families, the 2023 augmentation recipe [17], in-domain pretrained checkpoints [20], public pseudo-labels [21], model souping [22], and a public Perch/ProtoSSM pipeline [23].

3. Task and Data

3.1. Task

BirdCLEF+ 2026 [5] asks for a prediction, for every non-overlapping 5-second window of a 1-minute soundscape, of the presence probability of each of the 234 target taxa. The official metric ranks submissions by a macro-averaged score over the 5-second windows, on a public and private split of a held-out soundscape set. Each submission must run within a 90-minute, CPU-only Kaggle Notebook. This budget bounds ensemble width.

3.2. Data and Domain Shift

The labeled training set holds 35,549 focal recordings over 234 taxa. It comes with 10,592 unlabeled soundscapes and 66 labeled soundscapes. All audio is mono at 32 kHz. Soundscapes run exactly 60 s. The class distribution is long-tailed. A total of 28 taxa, mostly insects and amphibians, have no focal training audio. The model learns these only from soundscape pseudo-labels (Table 1).

The main axis of difficulty is the mismatch between focal recordings and soundscapes. Focal recordings are clean and high-SNR, usually with one foreground individual. Soundscapes are low-SNR, dense, and multi-source. We quantify the multi-source character straight from the labeled-soundscape annotations (Figure 1). Across 1,478 labeled 5-second windows (the 66 labeled soundscapes scored as

Table 1
BirdCLEF+ 2026 data sources. All audio is mono at 32 kHz; soundscapes are 60 s long. The target set contains 234 taxa: 28 have no focal training audio, and 122 have in-domain XC pretraining data. In labeled soundscapes, 5 s windows hold 4.22 species labels on average, and 89.4% hold at least two co-occurring species. Pseudo-label duration is window-equivalent.

Source	Items	Duration	Labels	Use
Focal recordings	35,549	–	recording	training
Unlabeled soundscapes	10,592	176.5 h	none	pseudo-labeling
Labeled soundscapes	66	1.1 h	5 s windows	validation
Public pseudo-labels	∼92,700 win.	128.8 h	5 s windows	training

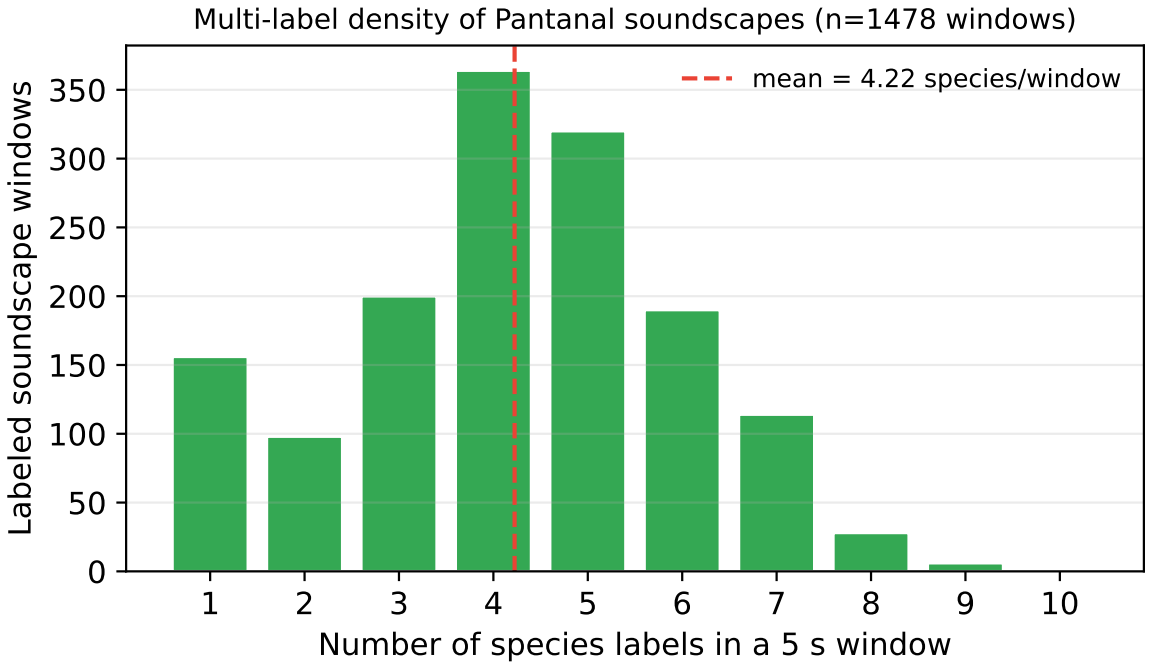

Figure 1: Multi-label density of Pantanal soundscapes: the number of species labeled at once per 5-second labeled-soundscape window (n=1,478, the 66 soundscapes scored on a 2.5 s hop). The mean of 4.22 species per window is the core signature of the focal-to-soundscape domain shift. Source: `analysis/labeled_soundscape_labels_per_window_hist.csv`.

overlapping 5-second windows on a 2.5-second hop, ≈24 per 60-second file), the mean number of species present at once is 4.22. 89.4% of windows hold at least two species. 82.8% hold at least three. A model trained on near single-label focal clips must generalize to a regime where four-way co-occurrence is normal. The training labels are weak, at recording level. The test labels are strong, per window. Semi-supervised learning on the soundscapes aims to close this gap.

4. System Overview

4.1. The Ensemble

Our final system averages six convolutional models with equal weight. Three are CNN classifiers. Three are SED networks. We then blend the result $0.5/0.5$ with a public Perch/ProtoSSM embedding pipeline [23]. We export all six of our models to OpenVINO [24] for fast CPU inference. The models differ by design along backbone, spectrogram resolution, detection head, and context length to raise diversity (Table 2). Section 7 measures the payoff of this diversity.

Table 2
The six convolutional models in the ensemble. "CNN" denotes an image-style classifier head. "SED" denotes a sound-event-detection head with a frame-level attention branch [7]. Backbones marked "in-domain" start from BirdCLEF 2025 pretrained checkpoints [20]. Each weight set is a model soup [22] over training runs. Single-model scores are public leaderboard, with the strong public pseudo-labels included.

Family	Backbone	Context	Init	Single-model public LB
CNN	HGNetV2-B0	5 s	ImageNet	0.935
CNN	EfficientNet-B3-NS	5 s	ImageNet (NS)	0.935
CNN	EfficientNetV2-S	5 s	In-domain 2025	0.937
SED	EfficientNetV2-S	5 s	In-domain 2025	0.938
SED	EfficientNetV2-S	10 s	In-domain 2025	0.934
SED	ECA-NFNet-L0	5 s	In-domain 2025	0.935

Table 3
Log-mel spectrogram parameters per model family.

Parameter	CNN family	SED family
`n_fft`	2048	2048
`hop_length`	313	512
`win_length`	626	2048
`n_mels`	256	128
`fmin` (Hz)	20	20
`fmax` (Hz)	16000	16000
`power`	2.0	2.0

4.2. Front-Ends, Sampling, Loss, Augmentation

Both families read single-channel log-mel spectrograms with family-specific parameters (Table 3). The CNNs use a higher mel and time resolution. The SEDs use a coarser representation, which pairs better with the attention branch. Training examples are random 5-second crops, or 10 s for the long-context SED. Random cropping beat energy-aware (RMS) cropping and start-of-recording cropping. Longer windows hurt, in line with the 5-second test granularity, so we keep the 10-second model only for diversity. All six models train with focal binary cross-entropy [25] over 30 epochs, peak learning rate 5×10^{-4}, a one-cycle schedule [26], on the `timm` [27] backbones. CNNs use mel-spectrogram mixup [15], with the 2023 recipe [17] added for the EfficientNetV2-S. SEDs combine mixup with the full 2023 recipe: one of {white, pink, Gaussian} noise at $p{=}0.3$ and random gain at $p{=}0.5$. The in-domain 2025 checkpoints, in place of ImageNet, added about $+0.007$ public LB.

4.3. Inference and Ensembling

At inference we slide non-overlapping 5-second windows over each 60-second soundscape. For the 10-second SED model we follow the strategy of Babych [13]. We slide a 10-second window with a 5-second step and pad each side by 2.5 s, so every window centers on the 5-second row under prediction. For the 5-second models we add a 2.5-second shift as test-time augmentation. We combine the six probability outputs with a simple equal-weight mean. We add no post-processing of our own. The public Perch/ProtoSSM pipeline we blend with already post-processes heavily, and extra smoothing on our average lowered the score. The submitted prediction is a $0.5/0.5$ average of our six-model ensemble and the public pipeline.

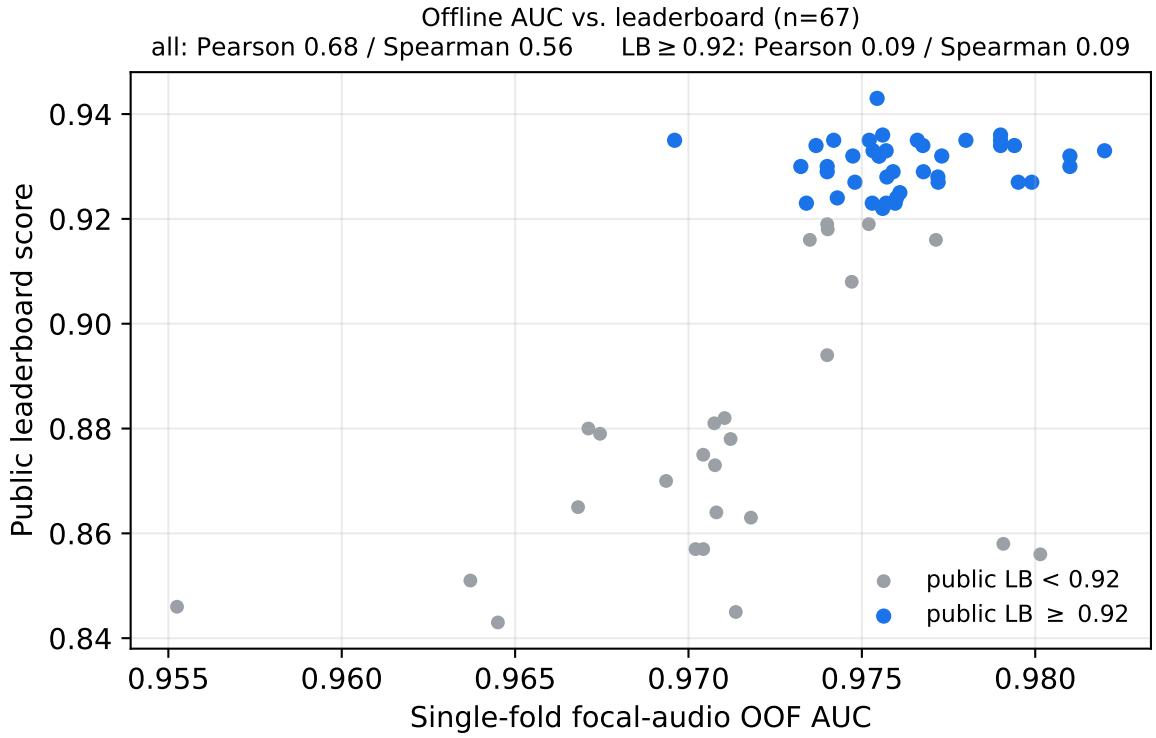

Figure 2: Single-fold focal-audio OOF AUC versus public leaderboard score for 67 of our submissions. The link is weak overall and vanishes in the competitive region (blue, LB ≥ 0.92). Source: `corr_proxy/lb_log.csv`.

Table 4
Correlation between single-fold focal-audio OOF AUC and public leaderboard score, by leaderboard region (`corr_proxy/lb_log.csv`, 67 submissions). The signal disappears where model selection matters.

Subset	n	Pearson	Spearman
All submissions	67	0.68	0.56
Public LB ≥ 0.90	47	0.22	0.23
Public LB ≥ 0.92	41	0.09	0.09
Public LB ≥ 0.925	34	-0.09	-0.07

5. Validation and Leaderboard Proxies

This section is the empirical core of the paper. Over the competition we logged 67 submissions with both a single-fold focal-audio out-of-fold (OOF) AUC and the resulting public leaderboard score (`corr_proxy/lb_log.csv`). We use this record to ask how well offline validation predicts the leaderboard. When it does not, we ask why.

5.1. Offline AUC Decorrelates From the Leaderboard

Figure 2 plots all 67 pairs. Overall the link is positive but weak (Pearson 0.68, Spearman 0.56). In the competitive region the link collapses. Above public LB 0.90 the Pearson correlation is 0.22 (n=47). Above 0.92 it falls to 0.09 (n=41). Above 0.925 it turns to -0.09 (n=34). Once a model is good, its offline AUC carries almost no information about its leaderboard rank, and sometimes points the wrong way. Table 4 reports the breakdown. The effect is concrete. Our highest offline AUC, 0.9801 from a 20-second model, scored only 0.856 on the leaderboard. Our best single submission on the leaderboard, 0.943 from the six-model ensemble, had a middling offline AUC of 0.9754. This reproduces, from an independent record, the high-region decorrelation the BirdCLEF 2025 2nd-place team reported [14].

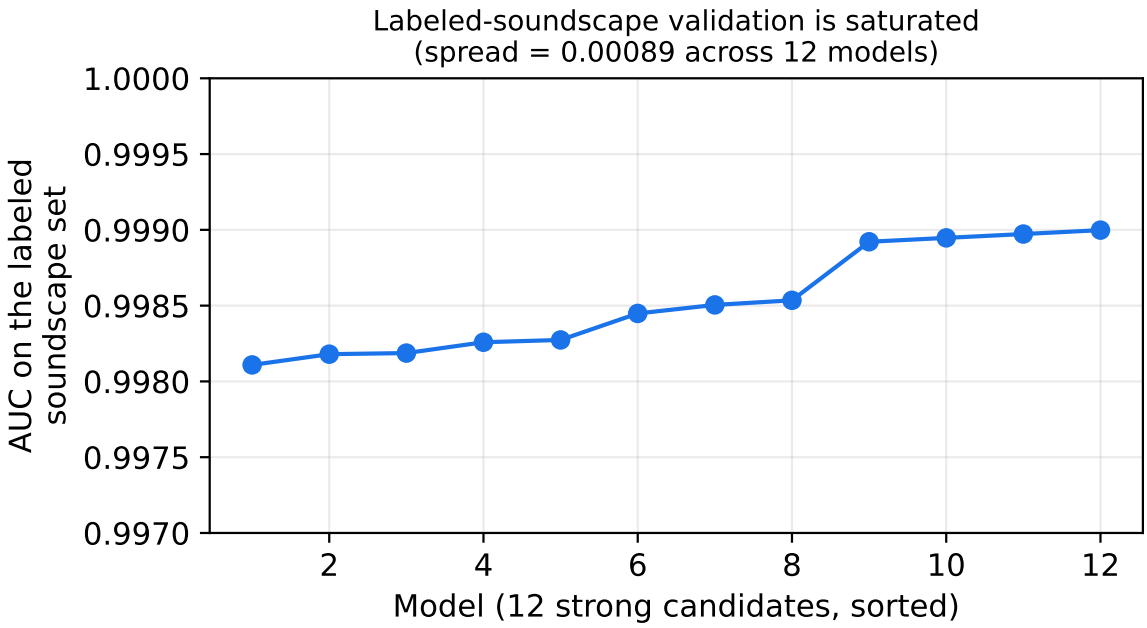

Figure 3: Labeled-soundscape AUC for twelve strong models, sorted. The total spread is 0.00089. The surface does not separate good models, so offline correlations against it are uninformative. Source: `corr_proxy/diagnostic_result.json`.

The drop is a regime change, not gradual decay, and the cause is a shift in where model differences live. Below LB 0.90, gains come from fixing gross errors: mislabeled focal clips, weak augmentation, poor cropping. The focal OOF surface sees these same errors, so the two metrics move together. Above LB 0.92, the gross errors are gone. The differences left between models sit in soundscape-specific behavior: calibration on dense multi-label windows, and the rare and no-focal-audio taxa. The focal OOF surface does not measure this behavior, and it saturates there (Section 5.2). Soundscape validation would measure it, but pseudo-label leakage contaminates the signal (Section 5.3). Both clean surfaces go blind at the point where model selection starts to depend on soundscape behavior. The correlation has nothing left to track.

5.2. Why: a Saturated Validation Surface

A single-fold noise story is incomplete. The deeper problem is a saturated clean validation surface. We scored twelve strong, genuinely different models, CNNs and SEDs across backbones, on the labeled-soundscape set. Their AUCs span only 0.99811 to 0.99900, a total spread of 0.00089 (Figure 3, `corr_proxy/diagnostic_result.json`). When every good model scores about 0.999, no offline metric on this surface ranks them. Any correlation against it measures noise. A leave-one-out consensus "teacher" proxy correlates with this target at $\rho \approx -0.66$. The number describes the degeneracy of the target, not the proxy. The practical lesson follows. Chasing decimal places of offline AUC in this region is worse than unhelpful. The chase misleads.

5.3. Why: Pseudo-Label Leakage Contaminates Soundscape Validation

The natural response is to validate on soundscapes, not focal clips. A subtle leak undermines this. We train on pseudo-labeled soundscape windows. Any held-out soundscape used for validation shares its source file with a pseudo-labeled training window. Our proxy code detects this and flags the rows `blocked_by_pseudo_overlap`. For the affected experiments, 13 source files overlap. They account for all 146 held-out soundscape rows. Zero rows remain for a clean strict-soundscape AUC (`reports/leaderboard_proxy_20260525.csv`). A naïvely computed soundscape AUC for such a

Table 5
Pseudo-label ablations on a single model, public leaderboard. The bottleneck is teacher quality and the mixing scheme, not the pseudo-labeling machinery. Sources: `corr_proxy/lb_log.csv`, `competition_context.md`.

Configuration	Public LB
Baseline, focal audio only	0.881
+ weak self-generated pseudo-labels	0.864
+ strong public pseudo-labels, 20 epochs	0.930
+ strong public pseudo-labels, 30 epochs	0.934
Mixing-scheme variants (strong pseudo-labels):	
ConcatDataset (used)	0.934
0.25-ratio pseudo sampling (2025 1st-place style)	0.894
Self-blend 2nd iteration (6-model + Perch)	0.926

model reads as a strong, trustworthy out-of-domain signal, 0.987 for one 10-second model. The signal is partly memorized. Soundscape-native validation is the right idea. It needs file-level disjointness between the pseudo-label source and the validation set. The constraint is easy to break, and breaking the constraint inflates the metric in silence.

5.4. What Did Correlate: a Family-Specific Calibration Proxy

Not everything failed, but the win is narrow and instructive. One proxy combines prediction sharpness with a contrast term. Sharpness is the fraction of windows where some species passes probability 0.5. Contrast is a rank-based AUROC of a candidate against a "good" diverse pole minus a "bad" CNN-ensemble pole (`corr_proxy/proxy_score.py`).

The proxy is family-specific (Figure 4). We score a balanced panel of 8 CNN and 8 SED models. Within the CNN family the proxy ranks models by leaderboard score almost perfectly, Spearman $+0.98$. Within the SED family it carries no ranking signal, Spearman -0.02. Across all 16 models the two families wash out to Spearman $+0.63$. A proxy built to pick CNNs does not pick SEDs. We cross-checked the within-CNN value with the standalone verifier in `corr_proxy/proxy_score.py`, which returns $+0.976$ on the 8 CNN models.

The cause sits in the poles. We built them from CNN soundscape behavior, with the "bad" pole a CNN ensemble. CNN scores climb cleanly with LB. SED models cluster near the top of the proxy range with no internal order, so the proxy neither ranks them nor separates the best SED from the worst.

A second test confirms the limit. No single distribution feature generalizes across the full panel either (Figure 5). The six individual proxy variants reach only Pearson 0.02 to 0.51 against real LB, with most near zero.

The methodological point is direct. A calibration proxy ranks models well, but only inside the family it was tuned on. Build it on CNN soundscape behavior and it transfers to CNNs, not to SED heads. Treat any offline proxy as family-local until a second family proves otherwise.

6. What Moves the Score: the Pseudo-Label Quality Study

If offline AUC does not move the leaderboard, what does? The largest lever we found is the quality of the pseudo-labels used for semi-supervised learning. Figure 6 and Table 5 tell the story with public-LB numbers.

Three findings stand out. **(1) Teacher quality decides the outcome.** A weak self-teacher, our own 0.881 model, generated pseudo-labels and lowered the leaderboard to 0.864. Pseudo-labels from a strong public $\sim$0.95-level teacher [21] raised a single model to 0.930, and to 0.934 at 30 epochs. The $+0.05$ swing comes from label quality, not from quantity or the algorithm. The strong pseudo-labels are also the only training signal for the 28 zero-focal-audio taxa. **(2) With strong labels, the simplest scheme wins.** We concatenated the $\sim$92,700 pseudo-labeled windows to the focal audio and sampled

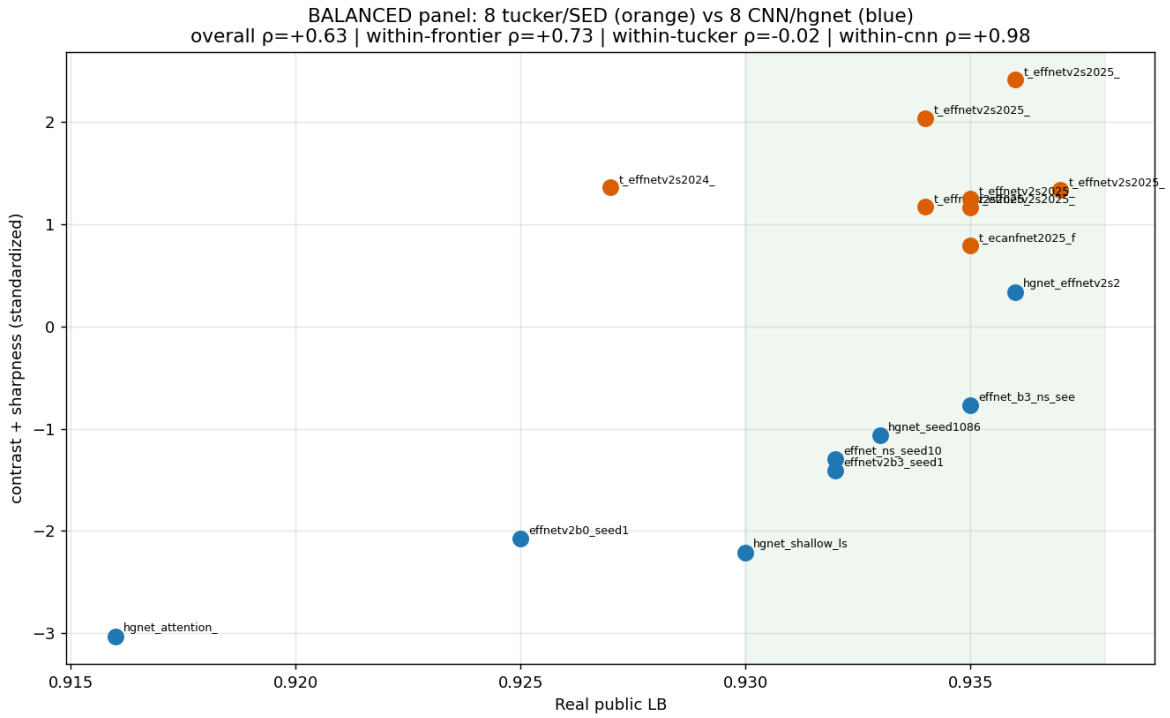

Figure 4: The contrast-plus-sharpness proxy versus public leaderboard score on a balanced panel of 8 CNN models (blue) and 8 SED models (orange). The proxy ranks the CNN family almost perfectly (within-CNN Spearman $+0.98$) and has no ranking power on the SED family (within-SED Spearman -0.02). Overall Spearman is $+0.63$. The shaded band marks the competitive LB region. Source: `corr_proxy/balanced_combo_vs_lb.png`.

uniformly, a `ConcatDataset`. Two alternatives lowered the score. Fixed-amplitude waveform mixing of focal clips with pseudo-labeled soundscapes hurt. Ratio-tuned sampling in the style of the 2025 1st-place solution [13] hurt more, with 0.25-ratio sampling dropping an otherwise-strong model to 0.894. The reason is direct. Those schemes suppress noisy teacher labels. When the labels are near ground truth, dilution destroys signal. **(3) A second self-distillation iteration did not help.** We re-labeled the soundscapes with our own six-model ensemble, optionally blended with Perch logits, and retrained. The result reached only 0.926, below the single-model strong-pseudo result. We kept the original public pseudo-labels.

7. Ensemble Diversity and Final Results

7.1. Diversity Pays, Quantifiably

We measured the offline focal-audio OOF AUC benefit of adding each candidate to a base model through a $0.5/0.5$ blend (`reports/leaderboard_proxy_20260525_blends.csv`; Table 6). The most different candidates help most. Adding the 10-second HGNet to an EfficientNetV2-B3 base gives $+0.0072$. The ECA-NFNet gives $+0.0041$. Same-family additions give less. Weak or pathological candidates lower the blend. The raw-waveform model costs -0.0051. A broken control costs -0.027. Screen diversity, do not assume it. This offline signal behaved sensibly because it measures relative agreement, not absolute AUC.

7.2. Final Leaderboard Results

Table 7 reports the headline numbers. Our six-model CNN and SED ensemble scores 0.943 public and 0.953 private. Blending it $0.5/0.5$ with the public Perch/ProtoSSM pipeline raises the public score to 0.957. The public-private gap returns at the last step. On the private split our CNN and SED ensemble

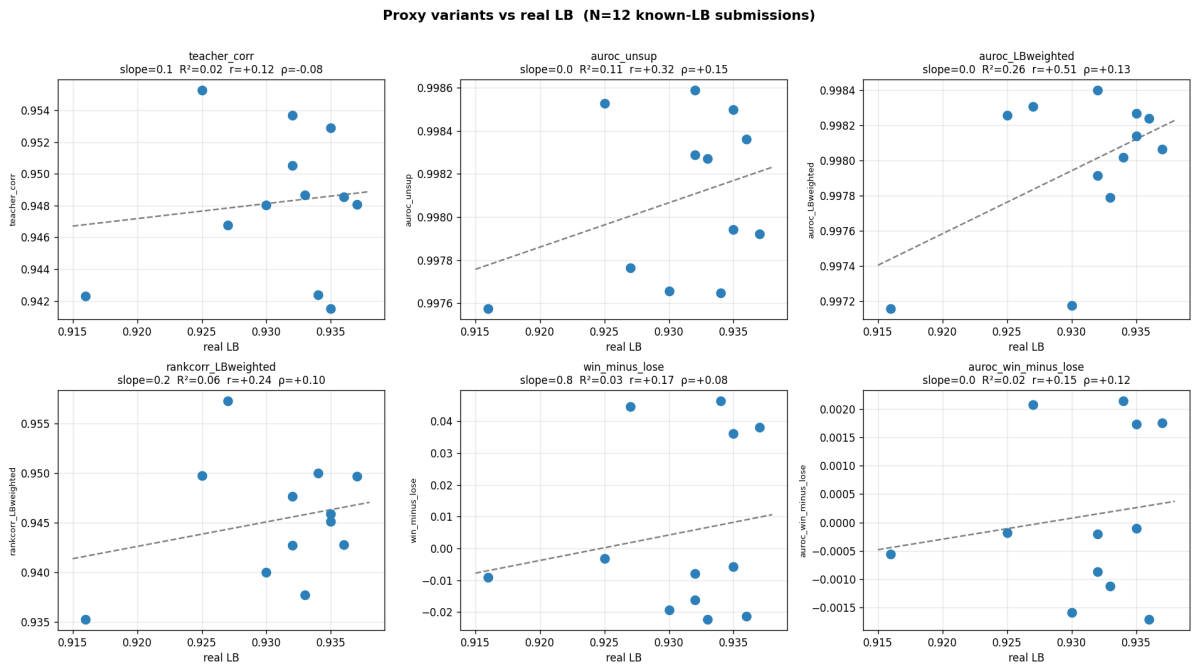

Figure 5: Six single proxy variants versus public leaderboard score over a 12-model panel spanning both families. Each panel lists its own slope and R^2. No variant generalizes. Correlations sit near zero, from Pearson 0.02 to 0.51. Source: `corr_proxy/all_variants_vs_lb.png`.

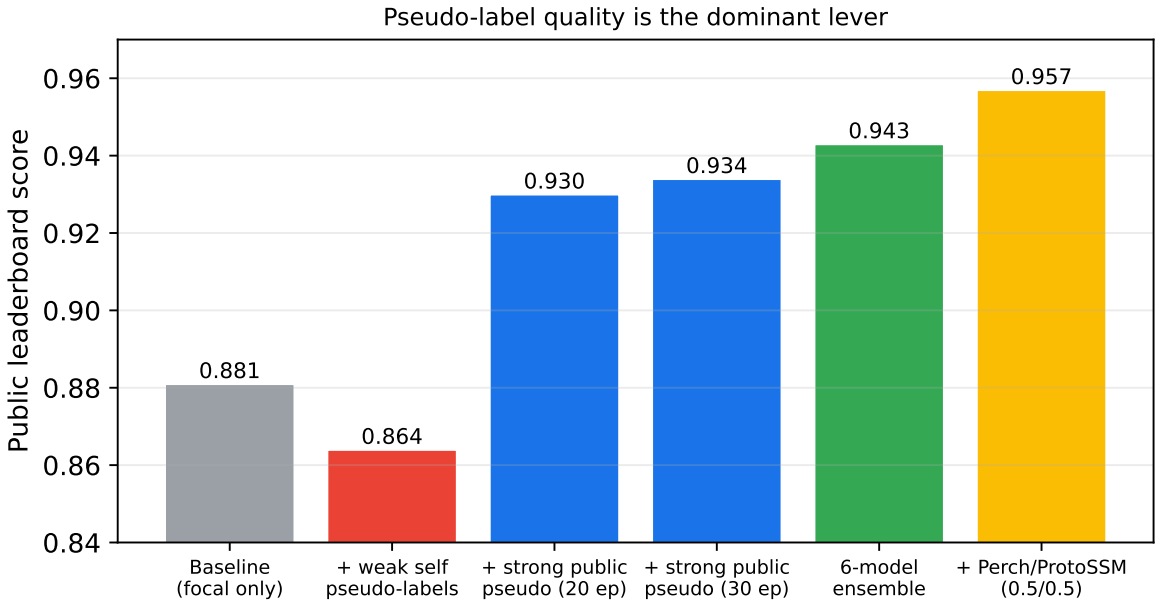

Figure 6: Public leaderboard score across the project. The decisive step is strong public pseudo-labels (blue). A weak self-teacher (red) lowers the score. Diversity and the Perch/ProtoSSM blend add the final increments. Sources: `corr_proxy/lb_log.csv`, `competition_context.md`, and the team writeup.

alone already scores 0.953. The $0.5/0.5$ Perch blend gives 0.955. A $0.6/0.4$ blend favoring our models would have reached 0.957 private, better than the weight we submitted. We chose the submitted weight to maximize the public score. The last lesson matches the first. Even the blend weight fell to public-private decorrelation.

Table 6
Offline $0.5/0.5$ blend deltas (focal-audio OOF AUC) for adding each candidate to the EfficientNetV2-B3 base (`reports/leaderboard_proxy_20260525_blends.csv`). Diverse architectures help. Weak and pathological models hurt.

Candidate added to base	Δ AUC
HGNet, 10 s (bin-pool)	$+0.0072$
HGNet, 10 s (orig. recipe)	$+0.0067$
ECA-NFNet-L0	$+0.0041$
EfficientNet-B3-NS	$+0.0026$
Raw-waveform HGNet	-0.0051
Broken control (untrained)	-0.0271

Table 7
Final ensemble results. "Ours" is the equal-weight six-model CNN and SED average. "Perch" is the public Perch/ProtoSSM pipeline [23]. The $0.6/0.4$ row is post-hoc.

System	Public LB	Private LB
Ours (6 models, equal weight)	0.943	0.953
Ours 0.5 + Perch 0.5 (submitted)	0.957	0.955
Ours 0.6 + Perch 0.4 (post-hoc)	n/a	0.957

Table 8
ASL versus focal-BCE distance to a four-model pseudo-teacher on three disjoint unlabeled-soundscape slices (`reports/asl_gap_root_cause_20260525.md`). Lower abs. error and higher top-5 agreement are better. ASL is worse on both, across all slices.

Slice	BCE err	ASL err	BCE top5	ASL top5
files 0:64	0.00297	0.00368	0.942	0.925
files 3000:3064	0.00287	0.00334	0.883	0.870
files 7000:7064	0.00255	0.00271	0.893	0.893

8. Failure Analysis

We document failed ideas not as a formality but because, under the weak CV–LB correlation described above, they are as informative as the successes.

Asymmetric loss. A source-aware asymmetric-loss (ASL) [28] SED outscored the matched focal-BCE baseline on the focal OOF proxy (0.9802 vs. 0.9786), yet fell behind on the leaderboard (0.932). Pinning down why was complicated by the fact that our labeled-soundscape validation AUC (0.9945) was contaminated. Those rows had been seen during training. We instead evaluated both models against a four-model pseudo-teacher on three disjoint 64-file slices of unlabeled soundscapes (Table 8). ASL sat farther from the teacher and agreed less on the top-5 predictions in every slice, despite the two models correlating above 0.98 with each other. The most plausible reading is that ASL shifted the model's calibration and class ranking in a direction the focal OOF happened to reward but the soundscape distribution did not.

Other failures. Several other directions were explored and abandoned. Plain cross-entropy (0.843 LB) and a differentiable AUC surrogate (0.851 LB) both fell well short of focal BCE, which we treated as settled after those comparisons (`corr_proxy/lb_log.csv`). A 1D raw-waveform model reached only 0.930 focal-OOF AUC and reduced any blend it entered by roughly 0.005, so it was excluded from the ensemble. A second attempt at waveform-level mixup, combining real and pseudo-labeled audio at $\lambda{=}0.5$ for the SED, peaked at 0.903 offline AUC, far below the $\sim$0.979 of the concatenation recipe.

On the distillation side, the SED notebook's default Perch distillation [19] degraded performance once pseudo-labels were added, and was disabled. External background-noise mixing similarly offered no benefit, likely because the pseudo-labeled soundscapes already provide realistic in-domain backgrounds. Self-generated in-domain pretraining did not transfer usefully; the public 2025 checkpoints [20] were both better and cheaper as a starting point. Finally, layering our own post-processing on top of the already heavily post-processed Perch pipeline hurt rather than helped.

9. Conclusion

We used a 13th-place BirdCLEF+ 2026 system to answer the question frustrating work on this benchmark. In the competitive region, offline validation does not predict the leaderboard. We measure the correlation falling from Pearson 0.68 overall to about 0 above LB 0.92. We trace this to a saturated labeled-soundscape surface, with an AUC spread of 0.00089, and to pseudo-label leakage, which inflates the otherwise-correct soundscape-native validation in silence. The levers moving the score are few and identifiable. Pseudo-label quality ranks first. A strong public teacher buys $+0.05$ where a weak self-teacher loses ground. Screened architectural diversity and a complementary embedding pipeline supply the rest. Our failures, once diagnosed, all fit one pattern: offline-rewarded, leaderboard-penalized calibration shifts. The highest-value direction for future editions is leakage-free, soundscape-native validation. You need a held-out soundscape set with file-level disjointness from the pseudo-label source. This is the only surface to separate our models where ranking mattered.

Acknowledgments

We thank the competition hosts and the wider Kaggle community. Our solution builds directly on publicly shared work: Tawara's HGNetV2 training notebook [18], Tucker Arrants' SED notebook [19], the BirdCLEF 2023 2nd-place augmentation recipe [17], Babych's BirdCLEF 2025 1st-place inference strategy [13], the in-domain pretrained checkpoints [20], Ali Ozan Memetoglu's public pseudo-labels [21], and the public Perch/ProtoSSM pipeline [23]. This work would not have been possible without them.

Declaration on Generative AI

During the preparation of this work, the author(s) used a generative AI assistant in order to generate figures from the authors' own experiment logs and to check grammar and spelling. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: Ai challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2026.

[2] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, BirdNET: A deep learning solution for avian diversity monitoring, Ecological Informatics 61 (2021) 101236. URL: https://doi.org/10.1016/j.ecoinf.2021.101236. doi:10.1016/j.ecoinf.2021.101236.

[3] S. Kahl, T. Denton, H. Klinck, V. Ramesh, V. Joshi, A. Srivathsa, V. V. R. Anand, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF 2024: Acoustic identification of understudied bird species in the Western Ghats, in: Working Notes of CLEF 2024 – Conference

and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3740/paper-184.pdf.

[4] S. Kahl, T. Denton, H. Klinck, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the Middle Magdalena, Colombia, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[5] S. Kahl, T. Denton, L. Sugai, L. Piatti, W. H. Arruda Moreno, M. M. Barbosa, M. Eibl, C. Z. Fieker, C. M. Garcia, D. L. Hokama Sousa, J. E. d. A. Júnior, R. C. Kridler, M. Lasseck, A. V. d. Melo, H. Müller, M. d. O. Neves, M. G. d. Reis, L. K. Sampaio Teles, K.-L. Schuchmann, J. L. M. M. Sugai, K. T. P. Azevedo, P. d. N. Lopes, M. I. Marques, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2026: Acoustic species identification in the Pantanal, South America, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.

[6] G. Lu, fastlight2007, noobanot, BirdCLEF+ 2026 13th place solution: Code, proxy logs, and diagnostics, https://www.kaggle.com/competitions/birdclef-2026/writeups/13th-solution, 2026.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, PANNs: Large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2880–2894. URL: https://doi.org/10.1109/TASLP.2020.3030497. doi:10.1109/TASLP.2020.3030497.

[8] M. Tan, Q. V. Le, EfficientNetV2: Smaller models and faster training, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 10096–10106. URL: https://proceedings.mlr.press/v139/tan21a.html.

[9] A. Brock, S. De, S. L. Smith, K. Simonyan, High-performance large-scale image recognition without normalization, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 1059–1071. URL: https://proceedings.mlr.press/v139/brock21a.html.

[10] PaddlePaddle Authors, PP-HGNetV2: A high-performance GPU-friendly backbone, PaddleClas Technical Report (2023). URL: https://github.com/PaddlePaddle/PaddleClas/blob/release/2.5/docs/en/models/PP-HGNet_en.md.

[11] B. Ghani, T. Denton, S. Kahl, H. Klinck, Global birdsong embeddings enable superior transfer learning for bioacoustic classification, Scientific Reports 13 (2023). URL: https://doi.org/10.1038/s41598-023-49989-z. doi:10.1038/s41598-023-49989-z.

[12] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: ICML Workshop on Challenges in Representation Learning, 2013. URL: https://www.kaggle.com/blobs/download/forum-message-attachment-files/746/pseudo_label_final.pdf.

[13] N. Babych, BirdCLEF 2025: 1st place solution, https://www.kaggle.com/competitions/birdclef-2025/writeups/nikita-babych-1st-place-solution-multi-iterative-n, 2025.

[14] V. Sydorskyi, F. Gonçalves, BirdCLEF 2025: Journey down the rabbit hole of pseudo labels, in: Kaggle BirdCLEF+ 2025 Solution Writeups, 2025. URL: https://www.kaggle.com/competitions/birdclef-2025/writeups/volodymyr-vialactea-2nd-place-journey-down-the-rab.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations (ICLR), 2018. URL: https://openreview.net/forum?id=r1Ddp1-Rb.

[16] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, in: Proceedings of Interspeech, 2019, pp. 2613–2617. URL: https://doi.org/10.21437/Interspeech.2019-2680. doi:10.21437/Interspeech.2019-2680.

[17] RihanPiggy, BirdCLEF 2023: 2nd place solution: SED + CNN with 7 models ensemble, https://www.kaggle.com/competitions/birdclef-2023/writeups/griffith-2nd-place-solution-sed-cnn-with-7-models-, 2023. Original discussion: https://www.kaggle.com/competitions/birdclef-2023/discussion/412707.

[18] Tawara, BirdCLEF 2026 HGNetV2-b0 baseline training, https://www.kaggle.com/code/ttahara/birdclef-2026-hgnetv2-b0-baseline-training, 2026.

[19] T. Arrants, BirdCLEF distilled SED training notebook, https://www.kaggle.com/code/tuckerarrants/bc2026-distilled-sed, 2026.
[20] V. Sydorskyi, BirdCLEF 2025 in-domain pretrained backbone checkpoints, https://www.kaggle.com/datasets/vladimirsydor/bird-clef-2025-all-pretrained-models, 2025.
[21] A. O. Memetoglu, Public soundscape pseudo-labels for BirdCLEF+ 2026, https://www.kaggle.com/datasets/aliozanmemetoglu/raw-pseudos, 2026.
[22] M. Wortsman, G. Ilharco, S. Y. Gadre, et al., Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, in: Proceedings of the 39th International Conference on Machine Learning (ICML), 2022, pp. 23965–23998. URL: https://proceedings.mlr.press/v162/wortsman22a.html.
[23] R. Dey, contributors, BirdCLEF+ 2026 public Perch/ProtoSSM inference pipeline (public notebook), https://www.kaggle.com/code/raunakdey07/birdclef-2026-v6, 2026.
[24] Intel Corporation, OpenVINO toolkit: Open visual inference and neural network optimization, https://github.com/openvinotoolkit/openvino, 2024.
[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. URL: https://doi.org/10.1109/ICCV.2017.324. doi:10.1109/ICCV.2017.324.
[26] L. N. Smith, N. Topin, Super-convergence: Very fast training of neural networks using large learning rates, in: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, 2019. URL: https://arxiv.org/abs/1708.07120.
[27] R. Wightman, PyTorch image models, https://github.com/huggingface/pytorch-image-models, 2019.
[28] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 82–91. URL: https://openaccess.thecvf.com/content/ICCV2021/html/Ridnik_Asymmetric_Loss_for_Multi-Label_Classification_ICCV_2021_paper.html.