Team 14
Justin-Marian Popescu
Alfred-Andrei Pietraru

Bug fixes

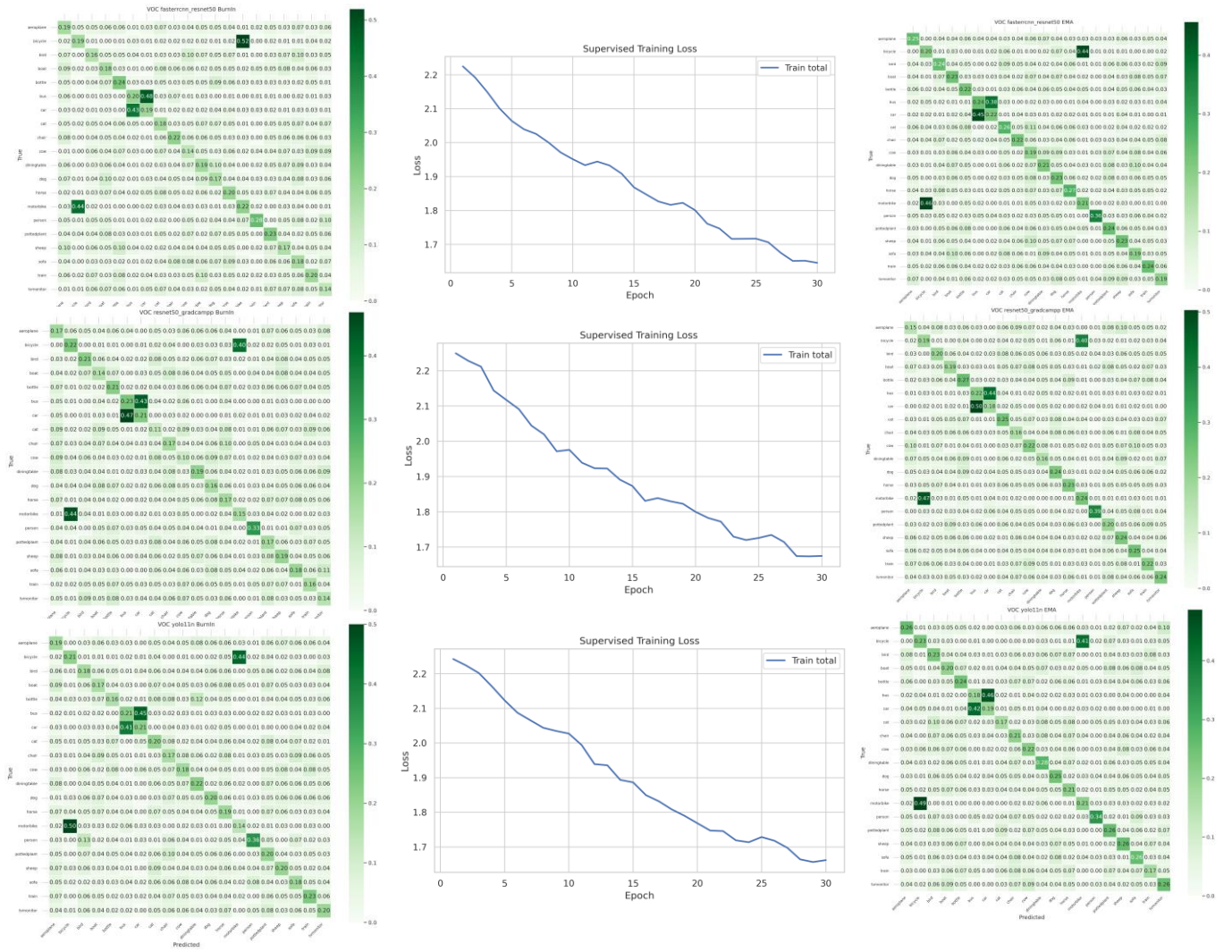
Training phases

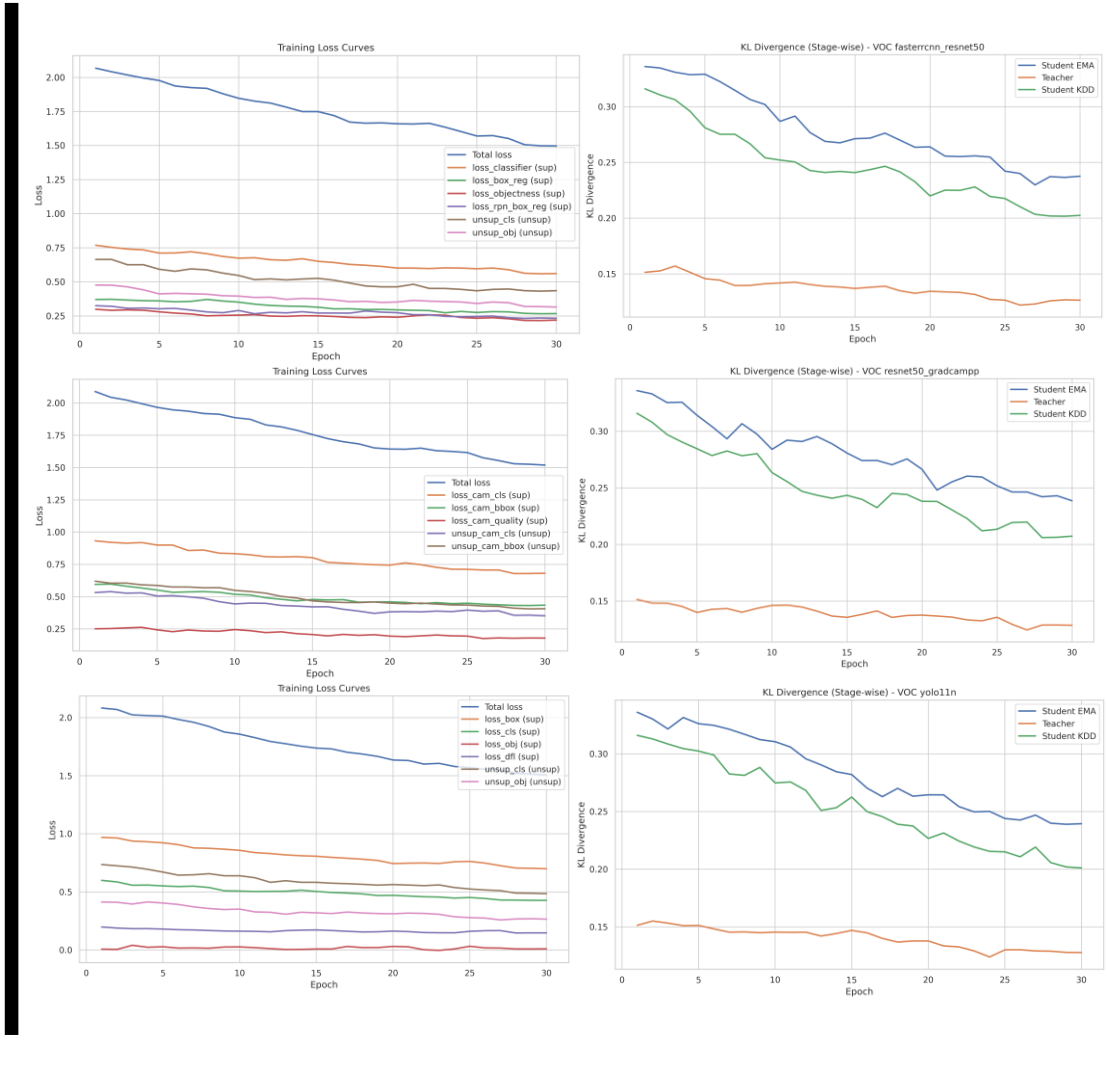
- Burn-in: component losses (**cls / box / obj**) did not converge synchronously → required separate monitoring
- EMA-SSL: EMA accumulation hidden temporary degradation of the student → direct student vs EMA evaluation
- EMA-SSL: teacher-student agreement unstable under class-imbalanced distributions

Validation phases

- Burn-in: validation metrics increased slower than training loss → early overconfidence indicator
- Burn-in: high confusion between visually similar classes, stabilized only after detection head convergence
- EMA-SSL: train–validation gap reduced via filtered pseudo-labeling (**threshold + NMS + agg class boxes**)
- EMA-SSL: clear recall improvements, with increased risk of false positives
- EMA-SSL: EMA validation is more stable than raw student evaluation

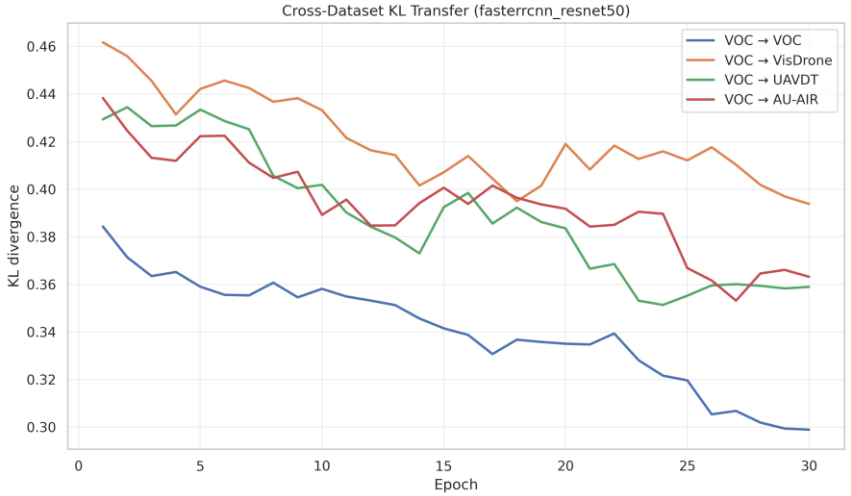
Improved evaluation on an already selected dataset

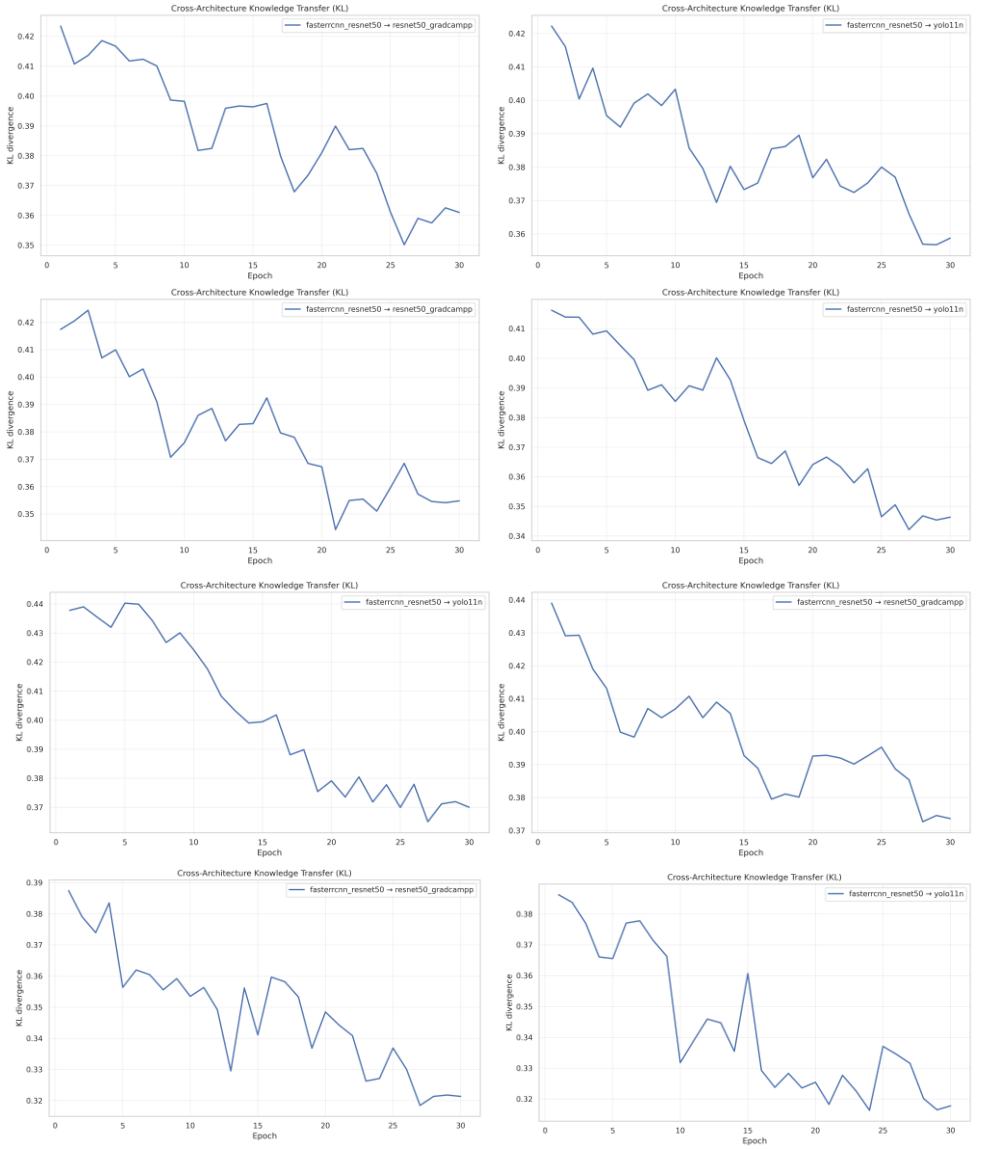




- Stable loss convergence across Burn-in, EMA-SSL, and KDD stages
- EMA-SSL reduces training noise and improves validation stability
- KDD consistently lowers KL divergence, indicating stronger teacher-student alignment
- Confusion matrices show reduced class confusion for dominant VOC classes
- Overall performance improves progressively from Burn-in to EMA and KDD

🧪 Evaluate on a new dataset (generalize)





- **Transfer setting**
  - trained on VOC, evaluated on VisDrone / UAVDT / AU-AIR
- **Motivation**
  - assess robustness under domain shift (scale, density, viewpoint)
- **Burn-in**
  - provides usable baseline but high confusion for domain-specific classes
- **EMA-SSL**
  - improves precision-recall balance and stabilizes predictions
- **KDD**
  - lowest KL divergence and strongest teacher–student consistency
- **Takeaway**
  - staged training improves cross-dataset transfer despite expected performance drop

🔍 New step added to the experiment results

What it includes

- loss curves per stage (Burn-In, Unbiased-Teacher, KL Models comparations)
- evaluation curves per stage
- confusion matrices + detection grids
- agreement matrices + KL curves (EMA vs KDD)

[Does the student predict the same class as the teacher and where the confusions come from?]

Result

- faster iteration and easier comparison between architectures/dataset (one file per run).

Practical win

- reduced debugging time;
- issues become visually detectable (class imbalance, collapse, unstable agreement).

EMA stabilize training and reduce noise in pseudo-label supervision.
KDD transfer teacher knowledge by aligning student output distributions via KL regularization.

🔁 Changes in proposed experiment setup and motivation

- **Decision timing:** weakly supervised approach abandoned **before Project Status 1**
- **Reason:** intermediate experiments showed poor inference behavior and unclear architectural gains
- **Issue:** proposed weak-supervision architectures were hard to control and evaluate consistently
- **Action:** reverted to a simpler and stable staged learning pipeline

---

- **New direction:** test student-teacher learning with **different model architectures**
- **EMA stage (2):** use heterogeneous teacher-student pairs to stabilize learning
- **KDD stage (*NEW 3*):** evaluate knowledge distillation as a quality refinement step
- **Benchmarking setup:** Faster R-CNN (ResNet-50) kept as the main benchmark
- **Comparison models:** GradCAM++-ResNet50 and YOLOv11 evaluated under the same training pipeline
- **Goal:** ensure fair, controlled comparison within a unified learning pipeline