

Peer Review

Review of: "Enhancing Dataset Distillation via Label Inconsistency Elimination and Learning Pattern Refinement"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

This is a technical report on the solution that took first place in Track 1 of the ECCV-2024 Dataset Distillation Challenge. The authors start from DATM, a trajectory-matching method that controls the difficulty of what the synthetic set learns by restricting which segment of the expert training trajectories is matched. They make two changes and call the result M-DATM. First (M1), they drop DATM's learned soft labels and train the synthetic images with plain cross-entropy against labels in the default order (Eq. 5), because, as Fig. 2(b) shows, the soft labels learned by DATM do not line up one-to-one with the labels the official evaluation script assigns. Second (M2), they shrink the matching range to $(T^-, T^+) = (0, 20)$ (§2.2, "Adjust Matching Range"), on the argument that without soft labels a 10-IPC synthetic set cannot absorb the "hard patterns" in late trajectories, especially on Tiny ImageNet.

The headline result is 0.4061 on CIFAR-100 and 0.1831 on Tiny ImageNet (Abstract), which match the 40.61 and 18.31 in Table 2, row "DATM+M1+M2 (M-DATM)". The authors also claim that early-stage matching is best on both datasets (Fig. 3) and that medium- and late-stage matching leave the synthetic images nearly unchanged (Fig. 4, §3.4).

2. Whether the conclusion holds on the data

I liked several things here. The report is honest about where the gain comes from: it is a practical diagnosis of a mismatch between a research codebase and a competition's evaluation protocol, and the authors say so rather than dressing it up as a new theory. The ablation in Table 2 is organised the right

way, adding one change at a time. All numbers are stated to come from the official evaluation script (§3.3), which is exactly what a challenge report should do. And the code is released (footnote 1). For practitioners who will run trajectory matching under someone else’s evaluator, the label-order warning alone is useful. Now the evidence itself. Table 2 gives three rows. On CIFAR-100: DATM 31.11, DATM+M1 39.90, M-DATM 40.61. On Tiny ImageNet: 6.90, 7.10, 18.31. I checked the “+11.21%” quoted in §3.3 (Ablation Study): $18.31 - 7.10 = 11.21$. The arithmetic is right, but it is an absolute gain in percentage points; in relative terms it is $11.21 / 7.10 \approx 158\%$. I would write “+11.21 points” to avoid confusion.

The CIFAR-100 gain from M1 (31.11 to 39.90, i.e. +8.79 points) is the largest single effect in the paper, and here I think the claim slightly exceeds what was tested. Fig. 2(b) and §2.2 describe the problem as an *ordering* mismatch: some synthetic images end up paired with the wrong class index under the official script. If that is the cause, there is a cheaper control than removing soft labels altogether: keep DATM’s soft-label training, then at export time sort or relabel the synthetic images by the argmax of their learned logits so they agree with the default order. If that control recovers most of the 8.79 points, the lesson is “export in the right order”; if it does not, then soft labels themselves are harmful under hard-label evaluation, which is a different and more interesting finding. As written, M1 bundles two things (fixing the order and switching from soft to hard targets) and the paper cannot tell them apart. I would also like to see how many of the 1,000 CIFAR-100 images (100 classes $\times$ IPC 10) or 2,000 Tiny ImageNet images (200 $\times$ 10) were actually mismatched; Fig. 2(b) shows red squares but I could not read a count from the text.

For Tiny ImageNet, the story is that M1 barely helps (6.90 to 7.10) and M2 does the work. The explanation offered is reduced “information capacity” after soft labels are removed. That is plausible. It is also untested. Table 2 has no DATM+M2 row, i.e. the shorter matching range with soft labels kept. That single row would show whether M2 helps only because M1 removed capacity, as the hypothesis predicts, or whether the original DATM range was simply a poor choice for 10 IPC under this protocol regardless of labels.

Each number in Table 2 appears to be a single run. I could not find seeds, the number of repeated evaluations, or standard deviations anywhere in §3. The official evaluator trains a ConvNet from scratch on the synthetic set, and at 10 IPC that step alone has non-trivial variance. The 0.71-point difference between DATM+M1 and M-DATM on CIFAR-100 (40.61 – 39.90) could well be within that variance, so I would not read it as evidence that M2 helps on CIFAR-100 until spreads are reported.

Fig. 3 is where the matching range is chosen. The stages are defined in §3.3 as early ($T^- = 0$, $T = 15$, $T^+ = 20$), medium (30, 45, 60), and late (60, 75, 80). Two points here. The stages are coarse and non-contiguous (nothing between 20 and 30), so “optimal” really means “best of three”. And the values plotted in Fig. 3 are not given in the text, so as rendered on Qeios I cannot check them. More importantly, if the stage was selected using the same official test-set evaluation that produces Table 2, the final number carries some selection on test. Tiny ImageNet has a validation split (§3.1 mentions 50 validation images per class), and CIFAR-100 can be split; picking the range on held-out data would make the result cleaner.

The baseline is not quite the published DATM either. §3.2 says ZCA whitening was removed and the data were pre-normalised with the evaluator’s statistics, and Table 2’s “DATM” row is defined as this modified version. That is a sensible engineering choice for the challenge, but it is a third change that is not ablated. A row for “DATM with ZCA, own evaluator” would let readers connect these numbers to the original DATM paper.

Finally, the report positions M-DATM as a “strong baseline for future works” (contribution 3), but it does not compare against any other method under the same script: not MTT, not FTD, not the other challenge entries. Ranking first is informative. A small table with the scores of, say, the top three entries or two classical baselines would make the “strong baseline” claim concrete. Fig. 4’s claim that late-stage images remain “nearly unchanged” is visual only; a simple mean pixel distance between initial and final synthetic images per stage would turn it into a measurement.

Overall, the main conclusion (these two changes lift DATM substantially under the challenge protocol and gave the winning entry) holds on the data shown. The causal explanations attached to each change go somewhat beyond what the experiments isolate.

3. Reproducibility

This is better than many challenge reports. The code is public (footnote 1, github.com/ChuhaoZhou99/M-DATM), it builds on the official DATM repository (footnote 2), expert trajectories are produced as in FTD, the distillation runs for 10,000 iterations, the network is the challenge’s default ConvNet, and the hardware is stated (4 NVIDIA A100 80 GB, §3.2). Eq. 5 and the worked label example ($\{0,0,1,1,2,2\}$ for 3 classes at $\text{IPC} = 2$) make the hard-label scheme unambiguous.

The gaps I found:

- **Table 1 is not legible on Qeios.** Only its caption, “Hyper-parameters for different datasets”, is rendered; the values of N, M, the learning rates, batch sizes, and the per-dataset matching ranges are missing from the text I reviewed. Since the whole contribution is about settings, this table matters most. I would re-upload it as an HTML table rather than an image.
- It is not stated whether CIFAR-100 also uses (0, 20) in the final submission, or only Tiny ImageNet. §2.2 introduces (0, 20) in the Tiny ImageNet paragraph; Fig. 3 suggests early is best for both. One sentence would settle it.
- No seeds, no number of expert trajectories, no number of evaluation runs, no variance.
- No wall-clock time or GPU-hours for expert training and for distillation; “4 A100” does not tell a reader what a reproduction will cost.
- The version or commit of the official evaluation script is not given. Since the paper’s first finding is precisely that results depend on evaluator details, pinning that script is worth doing.

4. *What specifically to fix*

1. §2.2, Eq. 5, definition of y_i . As rendered, the label reads “ $y_i = i \mid \text{IPC}$ ”, with a vertical bar (the “divides” symbol). The example {0,0,1,1,2,2} requires floor division, $\lfloor i / \text{IPC} \rfloor$ with zero-based i . Please correct the notation.
2. **Table 2.** Add a DATM+M2 row (shorter range, soft labels kept) so that the two modifications are actually separated, and add mean $\pm$ standard deviation over at least three seeds for every row.
3. Add a control to §2.2 “Remove Soft Labels”: DATM with soft labels but images reordered by argmax before submission. Report the count of mismatched images behind Fig. 2(b).
4. **§3.3, Ablation Study:** change “+11.21%” to “+11.21 points (7.10 $\rightarrow$ 18.31)”.
5. **Fig. 3.** Print the numeric accuracies (a small table is enough), say which data split was used for choosing the stage, and if possible add one intermediate range between 20 and 30.
6. **Table 1** needs to be readable on Qeios, with an explicit row for the final (T^- , T , T^+) of each dataset.
7. §3.2 removes ZCA whitening. I would ablate that change too, or at least report the original DATM number under its own evaluator next to Table 2’s “DATM” row.
8. Contribution 3 and §4 call M-DATM a “strong baseline”. Back this with the scores of other challenge entries or of MTT/FTD under the official script, or soften the wording.
9. **Fig. 2(c)** is the only evidence that late-range DATM “could not be effectively optimized” on Tiny ImageNet; please state the range and number of iterations shown in that plot.

10. Quantify Fig. 4 (mean change from initialisation per stage).
11. **Fig. 1 caption** gives the synthetic set sizes as 500 for CIFAR-100 and 1,000 for Tiny ImageNet. With $IPC = 10$ (§3.1) the sizes are $100 \times 10 = 1,000$ and $200 \times 10 = 2,000$. One of the two statements is wrong; the caption reads as if IPC were 5.
12. Small text points in §3.1 and §4: “The DD challenge utilize” should be “utilizes”, and “Eventually, We thank” has a stray capital.

5. Recommendation

Minor revisions. This is an honest, useful challenge report with a working solution, released code, and a sensible one-step-at-a-time ablation. What it needs is not new methodology but cleaner attribution: seeds and spreads, the missing DATM+M2 row, a reorder-only control for the soft-label issue, a legible Table 1, and the numbers behind Fig. 3. With those, the two explanations in §2.2 would rest on direct evidence rather than on plausibility.

Declarations

Potential competing interests: No potential competing interests to declare.