Research Logbook

Project: Investigating Dataset Bias with Modern Deep Neural Networks
Period covered: 14 Feb 2025 – 14 Aug 2025

– Part 1. The Ideas

---

14 Feb 2025

Idea / Observation

- All physical systems gravitate toward minimal-energy states. Analogously, DNNs exploit *low-energy* (shortcut) signals—i.e., dataset biases—because they are cheaper to learn.
- Even very large "real-world" datasets still contain abundant shortcuts; absolute removal of bias seems impossible.
- **Question:** If bias is inevitable, can we purposefully harness task-specific biases to save compute, accepting a trade-off in generalization?

Reflection

- Poses a provocative shift: from *removing* bias to *budgeting* it.
- Aligns with literature on energy-based models and implicit regularization of SGD.

Next steps

1. Survey papers on shortcut learning & energy minimisation.
2. Prototype a toy image task (day vs. night) to measure compute savings when the model is encouraged to exploit a known bias.

---

24 Feb 2025

Idea / Analogy

- In wireless communications, controlled *fading* can be exploited. → Hypothesis: controlled *bias* might be similarly exploitable in ML.

Reflection / Implication

- Draws a bridge between communication theory and representation learning; suggests treating bias as a channel property.

Next steps

- Review “fading diversity” techniques; search for existing work on “useful bias” in ML.

---

14 Mar 2025

Concept: Interference Diversity → Bias Diversity

- In CDMA (Code-division multiple access), many weak interferers average out, stabilising SINR.
- Could introducing *multiple, independent biases* average their individual harms, boosting overall generalisation?

Actionables

1. Design synthetic dataset with two orthogonal biases (e.g., colour patch & texture).
2. Measure accuracy on unbiased test set vs. single-bias baseline.

---

27 Mar 2025

Entropy & Complexity Perspective

- Proposes using image (Shannon) entropy as a proxy for bias magnitude.
- **Objective 2:** Bias ≈ low Kolmogorov complexity → networks prioritise compressible (biased) features.
- Draft method:
    - Estimate per-feature entropy via compression metrics.
    - Use Information Bottleneck to track which features survive training

[2 2
1 1] H =1 bits/ pixel

[1 2
3 4] H = 2 bits/pixel

Questions

- Is semantic bias simply *low-entropy structure* in label space?
- Can “semantic entropy” (label histogram entropy) quantify dataset bias?

Planned work

- Implement entropy calculator for image patches & label distributions.
- Compare complexity of shortcut vs. semantic features during training.

---

15 Apr 2025

Transformer Q / K / V Insight

- Values (V) appear to carry semantic content. Could attention weights quantify semantic information retained at different layers?

Next steps

- Instrument ViT to log attention entropies; correlate with bias indicators discovered above.

---

16 Apr 2025

1. **Residual Learning Analogy**
    - Treat *known* bias as the "identity shortcut" and force the network to learn the *residual* (unbiased semantics).
    - Potential method: subtract bias-only prediction from full prediction during training.
2. **Background-free Training Images**
    - Thought experiment: supply images containing *only* the labelled object (pure foreground, plain background).
    - Relates to saliency maps & attention cropping techniques.

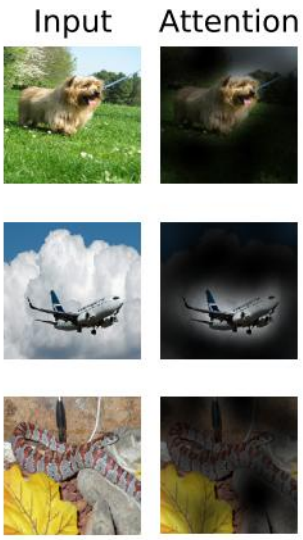


Figure 6: Representative examples of attention from the output token to the input space. See Appendix D.7 for details.

Actionables

- Generate foreground-only CIFAR-100 variant using segmentation masks; benchmark bias metrics.

---

09 May 2025

Nyquist-Style Data Sampling

- Draws analogy: unbiased data collection might need a "sampling rate" ≥ 2 × (max frequency).
- Suggests formalising *bias bandwidth* and deriving a minimal unbiased sampling criterion.

---

24 May 2025 (Entry 1)

Tri-Factor View of Performance

- Performance = f(architecture, training, dataset) → akin to *nature vs. nurture*.
- Mentions "Three-Body Problem" & PID control as metaphors for the complex, coupled dynamics.

Action point

- Sketch a PID-style feedback loop for dataset curation (bias error signal → data augmentation → retrain).

---

24 May 2025 (Entry 2)

Digital vs. Analog Information Capacity

- π requires infinite bits to store exactly in digital form; an ideal analogue system may could store infinite information.
- Speculative leap: Perhaps an *analogue* representation could hold an un-biasable dataset with infinite granularity.

Open questions

1. Is analogue storage viable for storing bias-free data?
2. How would learning algorithms interface with such media?

---

4 Jun 2025

Idea / Observation

- Treat visual *artefacts* (ringing, aliasing, compression blocks) as explicit *objects*.
- Under this view, **bias = systematic relationships between real objects and artefact-objects**.
- If bias is object-like, information-theoretic tools for object discovery (e.g. mutual information, compression) might be reused.

Additional thought

- Could overall model performance itself serve as a *scalar proxy* for bias magnitude? (More bias learned ⇒ higher in-domain accuracy but poorer cross-domain accuracy.)

Next steps

1. Prototype artefact segmentation masks; train a classifier that predicts "artefact class" vs. true class.
2. Correlate artefact-object frequency with cross-domain drop-off to test the performance-as-bias metric.

---

11 Jun 2025

Token Compression Hypothesis

- Low-level tokens ≈ binary patterns; high-level tokens ≈ semantic concepts.
- **Claim:** If high-level (semantic) activations are easily compressible, they must contain *less* semantic information; the converse implies richer semantics.
- Open problem: How to *compress* non-binary token sequences?

Actionables

- Experiment with vector quantisation (VQ-VAE) on intermediate transformer layers to measure code-book usage as a proxy for semantic entropy.
- Compare compression ratios of early vs. late layers under supervised and self-supervised regimes.

---

14 Jun 2025

Distribution-Difference Perspective

- Bias can be reframed as the **distribution gap between the dataset and the real world**.
- Question decomposition:
    1. What fraction of that gap is *semantic* (missing classes, missing contexts)?
    2. What fraction is *non-semantic* (lighting, artefacts, camera pipeline)?

Plan

- Use feature attribution maps to separate class-discriminative (semantic) vs. background (non-semantic) cues.
- Quantify each portion via entropy or compression length.

---

29 Jun 2025

Loss-Level Bias Suppression

- We can't enumerate every bias, but a model *must* learn the bias to "Name-that-Dataset".

- **Proposed loss:**

$$L = -\sum_i P(y_i \mid x_i) + \lambda, P(D_i \mid x_i)$$

  Penalise any image that allows accurate prediction of its source dataset *D*.

Implementation sketch

1. Dual-head architecture: *label head* + *dataset head*.
2. Add gradient reversal layer (GRL) before dataset head (as in DANN).
3. Tune λ to trade off classification accuracy vs. dataset invariance.

Immediate experiments

- Use CIFAR-100 ↔ Tiny-ImageNet joint corpus to verify that raising λ reduces Name-the-Dataset accuracy while maintaining test-set accuracy on held-out natural images.

---

14 Jul 2025

Label-Aware Training Schedule for CIFAR-100 + Tiny-ImageNet (Dataset 1806)
Dataset composition

- 27 labels appear in **both** datasets (overlap).
- 73 labels unique to **CIFAR-100**.
- 173 labels unique to **Tiny-ImageNet**.

Modified DANN training policy

Label group	Update Backbone	Update Main-Task Head (label)	Update Domain Head (dataset)
0 – 26 (overlap)	✓	✓	✓
27 – 99 (CIFAR-only)	✓	✓	✗
100 – 273 (Tiny-only)	✓	✓	✗

- **Rationale:** For unique labels, dataset identity is trivially recoverable through the label itself—training the domain head is pointless and may over-fit.
- **Epoch ordering:** Iterate *backwards*—first train on unique-label batches (domain head frozen), then finish with overlap-label batches (domain head active).

Planned code changes

1. Add new_label_id column to metadata to tag each sample.
2. Implement a conditional optimizer step that skips domain-head gradients for non-overlap IDs.
3. Validate that the modified schedule lowers dataset-predictability without hurting label accuracy.

– Part 2. Progress Meeting Records

---

23 Jan 2025

Meeting Agenda

- **Urgent Task:** Mission Statement (Due Date: 21/02/2025, 12:00)
    - Write a paragraph summarizing project understanding.
    - Compile a list of relevant papers (categories: Green, Orange, Black).
    - Provide a list of datasets (identify missing allocations).
- **Graded Task:** Presentation in May (Due Date: May 2025)
    - Train a deep learning model (e.g., a classifier) and present experimental results.

Key Discussion Points

- **Steps to accomplish project:**
    1. Understand the scope and requirements of the tasks.
    2. Identify and collect relevant papers for reference.
    3. Read and analyze the selected papers thoroughly.
    4. Prepare for 10 weeks of intensive work, focusing on:
        - Python programming
        - Deep learning
        - Dataset manipulation
        - Model training on the Shannon GPU server
- **Project Levels:**
    - **Fundamental Level:** Build a basic classifier using PyTorch (e.g., CNN on MNIST).
    - **Middle Level:** Utilize pre-trained models to enhance efficiency and results.
    - **Upper Level:** (To be confirmed)
        - Investigate dataset bias in ML
        - Understand how biases influence model performance
        - Develop methods to mitigate or leverage bias
        - Explore multi-layer fully connected networks

Action Items & Next Steps

- Draft the mission-statement paragraph by **21/02/2025**.
- Compile the categorized paper list by **30/01/2025**.

- Locate and document all required datasets.
- Begin fundamental-level tasks: implement and train a basic MLP on MNIST.

---

22 Apr 2025

Meeting Agenda

- Review overfitting observed in recent fine-tuning (validation loss – training loss gap).
- Plan next series of experiments:
    1. Reconstruct a combined dataset from overlapping labels in CIFAR-100 and TinyImageNet.
    2. Train ResNet-50 from scratch for image-label classification.
    3. Fine-tune the pre-trained ResNet-50 to predict dataset ID (CIFAR-100 vs. TinyImageNet).

Key Discussion Points

- **Overfitting Analysis:**
  The model began to overfit after a few epochs, as indicated by an increasing delta between validation loss and training loss.
- **Dataset Reconstruction Strategy:**
  For each overlapping label, extract all images from both CIFAR-100 and TinyImageNet to form a new dataset. Annotate each entry with:
    - Original image
    - Semantic label (e.g., “cat,” “dog,” “human”)
    - Dataset ID (CIFAR-100 or TinyImageNet)
- **ResNet-50 Training Plan:**
  Train a ResNet-50 model from scratch using (image, label) pairs—ignoring dataset ID—for standard classification.
- **Fine-Tuning for Dataset ID Classification:**
  Freeze all layers of the newly trained ResNet-50 except the final fully connected layer. Replace this head with a 2-class classifier to predict the dataset source.

Action Items & Next Steps

1. **Implement Dataset Reconstruction**
    - Develop a data loader to extract and annotate overlapping-label images.
    - Target completion: by 06 May 2025.
2. **Configure ResNet-50 Training Pipeline**
    - Set up scripts to train from scratch on the new dataset.
    - Monitor train/validation loss and accuracy.

3. **Design Fine-Tuning Procedure**
    - Freeze backbone, create new FC layer for dataset ID.
    - Validate on held-out split and record performance.
4. **Schedule Next Meeting**
    - Review preliminary results of dataset reconstruction and ResNet-50 label training.
    - Proposed date: late May 2025.

---

03 Jun 2025

Meeting Agenda

- **Identify Potential Sources of Dataset Bias:**
  Acknowledge that numerous unknown or hard-to-pinpoint factors may introduce bias in the combined dataset, beyond those already identified.
- **Mitigation Strategies:**
  Explore methods to reduce or control bias at two levels:
    1. **Data Level:** balancing and filtering of samples.
    2. **Preprocessing Stage:** enforcing consistent image-resizing pipelines.
- **Next Steps & General Quantification:**
  Develop a universal "bias metric" by measuring, for each class, the exact counts of CIFAR-100 vs. TinyImageNet images; use this per-class distribution as a baseline.
- **Note:** Weekly meetings continue through June; no meetings scheduled in July—only the final report.

Key Discussion Points

- **Unknown Bias Sources:**
  Recognize that current identification covers only some sources; further investigation needed into latent biases affecting model performance.
- **Data-Level Controls:**
  Agree on exploring class-wise balancing or selective filtering to mitigate dataset composition biases.
- **Preprocessing Consistency:**
  Highlight resolution artifacts arising from naive upsampling (32×32 vs. 64×64); propose a uniform resize-and-crop pipeline (e.g., shorter side→256 px, then center-crop to 224×224) to minimize such biases.

Action Items & Next Steps

1. **Finalize Splits:**
   Complete merging and re-splitting of train/validation/test sets per the reconstruction plan.

2. **Compute & Visualize:**
   Generate per-class source-distribution statistics and create bar plots showing CIFAR-100 vs. TinyImageNet fractions.
3. **Handle Semantically Mismatched Labels:**
   Decide whether to exclude or separately analyze pairs like “cattle” vs. “ox” and “bus” vs. “school_bus.”
4. **Update Preprocessing Pipeline:**
   Implement a uniform resize-and-crop strategy to ensure identical transformation effects across sources.

---

11 Jun 2025

Meeting Agenda

- Review progress on **Task 1: Constructing New Mixed Dataset**
- Update on **Task 2: Training ResNet-18 from Scratch**
- Overview of **Task 3: Fine-tuning ResNet-18 for Dataset-Source Prediction**
- Confirm hyperparameters, data logging, and evaluation protocols

Key Discussion Points

- **Dataset Construction**
  - Data sources: TinyImageNet (200 classes, 64×64 images) and CIFAR-100 (100 classes, 32×32 images).
  - Category mapping via *ChoosenO3.txt* produced three maps: tiny_to_new, cifar_to_new, new_to_id.
  - Splitting strategy: combine all samples, shuffle (seed 42), train/test ratio 10:1; resize all images to 128×128 (bilinear), save as JPEG quality 95.
  - Metadata schema (metadata.csv):
    image_id, filepath, new_label_id, new_label, tiny_synset_id, tiny_synset_name, cifar_fine_id, cifar_fine_name, dataset_id, split.
- **ResNet-18 Label-Prediction Training**
  - Model: resnet18(weights=None) with final nn.Linear(in_features, 27) head.
  - Loss: CrossEntropyLoss; Optimizer: Adam(lr=1e-3); LR scheduler: ReduceLROnPlateau(mode='max', patience=3, factor=0.1); Early stopping after 10 epochs without improvement.
  - Logging per epoch: train_loss, train_top1, val_loss, val_top1, val_top5 → training_log.csv; save best_resnet18.pth, and intermediate checkpoints every 10 epochs.
- **ResNet-18 Dataset-Source Fine-Tuning**

- Freeze all layers except final fully-connected layer; replace model.fc → nn.Linear(in_features, 2).
- Loss: CrossEntropyLoss; Optimizer: Adam(lr=1e-3) (only new head); Scheduler: ReduceLROnPlateau(mode='max', patience=2, factor=0.1); Early stopping after 5 epochs.
- Evaluation: compute confusion matrix and full classification report on validation and test splits; save per-class accuracies.

Action Items & Next Steps

1. **Validate Dataset Pipeline**
   - Verify shuffling, splitting ratios, and JPEG quality in generated dataset.
   - Load and inspect metadata.csv within custom DataLoader.
2. **Launch ResNet-18 Training**
   - Execute training script with specified hyperparameters.
   - Monitor logs and ensure checkpointing works as intended.
3. **Implement Fine-Tuning Workflow**
   - Develop script to freeze backbone and train new dataset-ID head.
   - Automate generation of confusion matrix and classification reports.
4. **Schedule Follow-Up**
   - Plan next meeting for mid–June to review training curves, bias metrics, and initial fine-tuning results.

---

19 Jun 2025

Meeting Agenda

- **Early-Stopping Criterion**: Switch from monitoring validation accuracy to monitoring validation loss for more stable overfitting detection.
- **Slide Results**: Add interpretive commentary and clear conclusions on each chart and table in the slide deck.
- **GitHub Workflow**:
  1. Set up a central repository for code and slides.
  2. Define a branching strategy (e.g., main for final, feature branches for experiments).
  3. Document pull-request and review process.
- **Experiment 2: Non-Semantic Dataset-Bias Test**:
  1. **Label Selection**

- Reserve the 27 semantically-overlapping classes present in both Tiny-ImageNet and CIFAR-100.
- Randomly select 50 non-overlapping labels from each dataset (100 total).

2. **Test Set Construction**
   - Sample 50 images per non-overlapping label (Tiny-ImageNet validation split; CIFAR-100 test split).
   - Combine all 5,000 images into a "non-semantic" test set.
3. **Evaluation**
   - Use the pre-trained dataset-ID classifier (fine-tuned on CIFAR-100 vs. Tiny-ImageNet) to measure accuracy on this new test set.

Key Discussion Points

- Monitoring validation loss enables earlier detection of overfitting compared to accuracy thresholds.
- Slide annotations must include narrative commentary to guide the audience through each result.
- A standardized GitHub workflow will improve collaboration and reproducibility.
- The non-semantic test assesses whether dataset bias extends beyond shared semantics:
  - **High accuracy** (> 70%) implies transferable non-semantic artifacts.
  - **Chance-level accuracy** (~ 50%) suggests bias depends primarily on semantic overlap.

Action Items & Next Steps

1. **Implement Early-Stopping Update**
   - Modify training scripts to monitor val_loss and trigger early stopping.
2. **Enhance Slide Deck**
   - Integrate interpretive text and conclusions for each figure and table.
3. **Initialize GitHub Repository**
   - Create main branch, set up feature branches, and draft CONTRIBUTING guidelines.
4. **Construct Non-Semantic Test Set**
   - Write data-loading code to sample and combine 5,000 non-overlapping images.
5. **Run Evaluation**
   - Apply the dataset-ID classifier to the non-semantic test set and record accuracy.
6. **Analyze Results**
   - Compute overall and per-label accuracy; quantify bias proportion and transferability strength.

7. **Prepare Thesis Slide**
    - Draft a concise slide summarizing Experiment 2 design, results, and interpretations.

Next meeting: review implementation progress, evaluation results, and finalized slide narrative.

---

25 Jun 2025

Meeting Agenda

- Review and plan **Dataset Construction** for the full 273-class dataset.
- Define **Model Training** procedure on the merged dataset.
- Outline **Dataset Classification** fine-tuning for source prediction (CIFAR-100 vs. TinyImageNet).
- Establish **Evaluation** protocol for label-classification and dataset-naming accuracy.
- Consider theoretical formalization of bias distributions and loss-based mitigation.

Key Discussion Points

- **Dataset Construction:** Merge TinyImageNet and CIFAR-100 into one dataset with 27 semantically overlapping classes, 73 CIFAR-only classes, and 173 Tiny-only classes (total 273).
- **Model Training Plan:** Train ResNet-18 from scratch on the 273-class dataset to predict object labels.
- **Fine-Tuning Strategy:** Freeze the trained ResNet-18 backbone and replace its head with a 2-class classifier to predict dataset source.
- **Evaluation Protocol:** Test label-classification performance on CIFAR-100's test set and TinyImageNet's validation set; compare dataset-naming accuracy against the 27-class baseline to assess semantic representation gains.
- **Optional Extensions:** Time permitting, experiment with ResNet-50 and Vision Transformer (ViT) architectures.
- **Theoretical Consideration:** Formalize dataset bias as divergence between $Pr_{(}CIFAR\text{-}100_{)}(x)$ and $Pr_{(}TinyImageNet_{)}(x)$ and propose adding a penalty term $\lambda \cdot P(D_k|x_i)$ to the loss to discourage bias learning.

Action Items & Next Steps

1. **Implement Full Dataset Merge**
    - Develop and validate scripts for merging into 273 folders with correct counts.
2. **Train ResNet-18 on 273 Classes**
    - Configure hyperparameters, start training, and monitor performance.
3. **Fine-Tune for Source Classification**
    - Freeze backbone, create binary head, and run fine-tuning.

4. **Conduct Evaluation**
    - Compute and compare label accuracy and dataset-naming metrics against baseline.
5. **Explore Architectural Extensions**
    - If feasible, repeat steps 2-4 with ResNet-50 and ViT.
6. **Draft Mathematical Formalism**
    - Write up divergence definitions and integrate the $\lambda \cdot P(D_k|x_i)$ penalty in the loss description.

Next meeting will review merge results, training curves, and initial evaluation metrics.

---

09 Jul 2025

Meeting Agenda

- **Augment "Name-the-Label" Training Details**
  For every experiment, explicitly report both the "name-the-label" and "name-the-dataset" performance in the training logs.
- **Per-Class Accuracy on the Non-Semantic ("Name-the-Bias") Task**
  Include a breakdown of "name-the-label" accuracy for each class/dataset label in the non-semantic test.
- **Figure Annotations**
  Ensure all plots and example images clearly label whether they correspond to the "name-the-label" task or the "name-the-bias" task.
- **Correlation Observation**
  Highlight the strong correlation between linear-probe accuracies on semantically overlapping (27 classes) and non-overlapping (243 classes) sets, using a correlation coefficient or scatterplot in the Results section.
- **Supplementary Log Book Contents**
  Plan to include handwritten notes, weekly meeting slide decks, idea sketches ("idea book" pages), list of all Python scripts, and references to all reviewed papers.
- **(Optional) Probabilistic Interpretation of Dataset Bias**
  Draft a formalism where x denotes images, y the true label, and b the unobserved dataset-specific bias; discuss how b affects distributions and the challenge of inferring y given x in presence of b.

Action Items & Next Steps

1. **Update Training Logs**
    - Modify logging scripts to record both label-prediction and dataset-prediction metrics for every experiment.
2. **Compute & Report Per-Class Non-Semantic Accuracy**
    - Extract and tabulate accuracy per class for the non-semantic test; integrate into results.

3. **Annotate All Figures**
    - Revise figures and example images to include explicit task labels (“Label” vs. “Bias”).
4. **Correlation Analysis**
    - Calculate the correlation coefficient between the two probe-accuracy sets and generate a scatterplot for the manuscript.
5. **Assemble Supplementary Materials**
    - Gather handwritten notes, slide decks, idea sketches, script listings, and paper references for inclusion in the supplementary logbook.
6. **Draft Optional Formalism**
    - If time permits, write up the probabilistic interpretation section and integrate it as an optional appendix.

Outlook

Re: [MSc][BIAS][Project Understanding Submission] Week 6-8

From Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Date Mon 2024-11-04 5:01 PM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>

1 attachment (29 KB)
Understand of the project.docx;

Hi Junhao,

Thank you for summarizing your understanding of the project as of now. Good job!

I have made some comments on the Word document (see attached).

Overall, I would like to emphasize the following:

- The general term "dataset bias" refers to the contents of a dataset that prevent the classifier from achieving strong generalization performance. These biases are unknown to us, and we also do not know any methods to quantify those biases. We can only speculate that these biases can be attributed to an imperfect data selection process.
- If you want to refer to the features that are useful for the task at hand (e.g., classification), try to mention them as "useful features" or "task-relevant features" etc. This way you will clearly separate "useful features" from "biased features" — both can be used to solve the task, but ideally we would like to learn only what is useful!

You already picked up the important aspects of the problem we are interested in. Now, I would suggest that you go through each of the two suggested papers and try to read them more thoroughly. For example, you can focus on the experiments they ran and even compare the two papers directly (note that the Torralba and Efros paper was released more than a decade ago, therefore many datasets and networks that are now popular did not exist back then). Reading the papers again will help you understand the problem and their proposed methodology much better!

Let us know if you have any questions in the meantime.

Also, for future reference, please only include Sotos, Jingshuai and me as recipients of your emails.

Thanks again and keep up the good work!

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 04 November 2024 08:00

To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>; Yuning Du <yuning.du@ed.ac.uk>; Jun Yan <Junyu.Yan@ed.ac.uk>; Zhuofei Lu <Z.Lu-45@sms.ed.ac.uk>
Subject: [MSc][BIAS][Project Understanding Submission] Week 6-8

Dear All,

I hope this message finds you well. I have reviewed the two provided papers and completed a document summarizing my understanding of the project. The document is attached to this email, and I look forward to any feedback or suggestions from anyone.

Thank you for your guidance, and I'm eager to continue progressing on this work.

Best regards,
Junhao

Outlook

[MSc][BIAS][Project Follow-Up & Next Steps] Week 1 Sem2

From Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Date Wed 2025-01-15 8:00 AM
To Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>

Dear All,

I hope this message finds you well and that you had a good slack during the Christmas.

Over the winter break, I spent time revisiting the Python basics using the "Learn Python" website recommended by the MSc Project Slides. This has given me a foundational understanding of the syntax.

As we move forward, I would appreciate your guidance on the next steps. Are there specific tasks or objectives you would like me to prioritize at this stage?

To make our discussion more focused and productive, I suggest including the following points in the meeting agenda:

1. An overview of the project's current status.
   Literature Review : Data collection, Google Scholar, More graphs
   Python: use more to improve proficiency.
2. Specific tasks or deliverables for the next phase.
   Looking forward to your guidance
3. Key skills or concepts from the "Machine Learning in Signal Processing" module that could align with the project goals.
   Looking forward to your guidance
4. A brief review of the Anaconda, PyCharm tools relevant to the project.
   Looking forward to your guidance

Following the meeting, I will prepare a summary of the discussion, including key takeaways, agreed-upon actions, and any questions that arise. This will ensure we remain aligned and provide a chance to address any misunderstandings or follow-ups asynchronously.

Thank you all for your continuous support and guidance. I look forward to hearing your thoughts and moving ahead with the next steps.

Best regards,
Junhao Sun

Outlook

Re: [MSc][BIAS][Meeting Summary & Questions] Week 1 Sem2

From Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Date Fri 2025-01-24 1:08 PM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>

Hi Junhao,

Thank you for providing the meeting summary. I think you have included all the important points that we discussed yesterday. My comments are the following:

- Regarding Project Level 3, the goal of your MSc project will be clearer to you once you become familiar with neural networks and the concept of dataset bias in practice.
- I've posted the PyTorch tutorial in the **MSc SPC 2024-25** Teams channel. Could you please leave a comment on that post to confirm that you have seen it?
- Since you decided to read the Deep Learning book, if you have access to the version with those chapters https://www.deeplearningbook.org/contents/TOC.html, I believe that Chapter 15 (from Part III) will help you understand some aspects of your project. That being said, you will also have to go through some of the earlier chapters (Part I-II) to learn about modern neural networks. However, keep in mind that practical experience is really important for your project (and deep learning research in general), so I would suggest that your primary focus now will be on developing your coding skills. Reading about deep learning theory can be a parallel task.
- My apologies for the misunderstanding about the dataset (I decided last minute to exclude it from the project description). Since there is no official benchmark dataset for this task, we will have to come up with a "clever" data collection process (note that it will also be a contribution of your thesis). In short, based on the definition of dataset bias, it would be interesting to combine two different datasets (collected from different sources) that share a few classes — this will help us conduct interesting experiments for the dataset prediction task (i.e., "*Name that Dataset*"). These two datasets will be **Tiny ImageNet** (https://huggingface.co/datasets/zh-plus/tiny-imagenet) and **CIFAR-100** (https://www.cs.toronto.edu/~kriz/cifar.html). We will help you with data collection, so you don't have to worry about it.

Let me know if you have any questions.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 24 January 2025 08:00
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc][BIAS][Meeting Summary & Questions] Week 1 Sem2

Dear All,

I hope this email finds you well.

Following our recent meeting, I wanted to let you know that I have completed the meeting summary as discussed. Please let me know if there's anything you would like me to add or clarify.

Additionally, I have a few follow-up requests that I would appreciate your guidance on:

1. Could you please share the PyTorch tutorial link mentioned during the session? I would like to review and practice it further.

2. Regarding the book Deep Learning by Ian Goodfellow, Yoshua Bengio, and Aaron Courville, could you advise on how best to utilize it for this project? Are there specific chapters or sections that align with our goals?

3. Lastly, I seem to have missed the dataset link referenced in the project allocation list. My project number is 29. Could you kindly confirm where I can access it?

Thank you for your continuous support and guidance. I look forward to your reply and am eager to make further progress on the project.

Best regards,
Junhao Sun

Outlook

Re: [MSc][BIAS][Meeting2 Task Sumbmission] Week 3 Sem2

From Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Date Tue 2025-01-28 8:41 AM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>

Hello,
Which one articles do you think you should read first such that it will improve what you will write in the mission statement?

In my opinion I will split this week 50% mission statement by reading papers and 50% improving your computational skills and the following weeks reduce the effort on computational skills as needed so you can spend time on the mission statement.
After the mission statement deadline focus 80% on computational skills and 20% paper reading.

best
Sotos

++++++++++++++++++++++++++
Professor Sotirios A. Tsaftaris
Chair in Machine Learning and Computer Vision
Canon Medical/Royal Academy of Engineering Research Chair in Healthcare AI
Director EPSRC AI Hub for Causality in Healthcare AI with Real Data (CHAI)
ELLIS Fellow
School of Engineering, University of Edinburgh,
Rm 2.06, Alexander Graham Bell, King's Buildings, EH9 3FG, Edinburgh, UK
+44 (0)131 650 5796 // S.Tsaftaris@ed.ac.uk // https://vios.science
++++++++++++++++++++++++++

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Date: Tuesday, 28 January 2025 at 08:00
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>, Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>, Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc][BIAS][Meeting2 Task Sumbmission] Week 3 Sem2

Dear All,

I hope this email finds you well.

I'm writing to provide an update on the tasks assigned to me this week:

1. New understanding of the project.
2. Dataset list.
3. Paper list.

I'd appreciate your guidance on the next steps. Should I focus on drafting the Mission Statement first, or would it be more valuable to start by training a small neural network on

the FashionMNIST dataset (a grayscale image dataset)?

Thank you for your support, and I look forward to your advice on the best course of action.

Best regards,
Junhao Sun

Outlook

Re: [MSc][BIAS][Initial Draft of Mission Statement] Week 6 Sem2

From Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Date Fri 2025-02-21 12:13 PM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>

Hi Junhao,

Thank you for sharing the draft. Good job!

Overall, I'm pleased with the Mission Statement. It's great to see that you expanded your literature review to include topics such as spurious correlations and OOD generalization. Also, the "Scope for extension" section refers to interesting real-world problems.

As for the next steps, I suggest that you:

- Keep practicing on your Python coding skills (complete the Pytorch tutorial and get familiar with popular pre-trained vision models for classifications found in https://pytorch.org/vision/0.20/models.html#classification)
- Analyse the "Name that Dataset" task a bit more in detail. You can go through the "Experiments" sections of the corresponding papers ("Unbiased Look at Dataset Bias" and "A Decade's Battle on Dataset Bias") to better understand the goal of this proposed experiment and how they execute this experiment in practice. Also keep in mind that there are subtle differences between the terms *dataset bias* and *spurious correlations*. We can discuss this further in our next meeting.

In the meantime, let us know if you have any questions or issues.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 20 February 2025 08:00
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc][BIAS][Initial Draft of Mission Statement] Week 6 Sem2

Dear All,

I hope you are doing well.

I am writing to inform you that I have completed the initial draft of the Mission Statement.

Given that this is an early version, I expect it to evolve as the project progresses.

I would greatly appreciate your feedback and suggestions for improvement. Please let me know your thoughts at your convenience.

Thank you for your time and support.

Best regards,
Junhao Sun

Outlook

Re: [MSc][Bias] Schedule a supervision meeting

From Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Date Thu 2025-04-24 4:56 PM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>

Hi Junhao,

Thank you for the meeting summary!

The only comment I have is that you can start your experiments with an even smaller model at first, for example ResNet-18
https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html

Best of luck with your exams!

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 22 April 2025 15:25
To: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Dear all,

I've completed the meeting summary as discussed (see attachment). Let me know if you'd like any adjustments before I share it with the team.

Our next meeting will be in June—I'll coordinate schedules and send an invite closer to the date.

Looking forward to continuing our progress then.

Best regards,
Junhao Sun

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Sent: Saturday, April 19, 2025 3:14 PM
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Hi Junhao,

Let's keep the meeting as it is then. I'll see you on Tuesday.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 19 April 2025 14:59
To: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Dear All,

Konstantinos, Thank you for your suggestion.

The reason I initially moved the meeting to April 22nd and changed it to an in-person meeting at Usher Building is that I was planning to inquire with PhD students, who previously enrolled in the MSc SPC programme and had applied for the SPC scholarship, about the SPC scholarship interview suggestions, as I know some of them have been through it. The application deadline for the SPC scholarship is April 25th, so I wanted to get their insights as soon as possible.

If we are unable to meet at Usher Building on the 22nd at 12:00(noon), we can certainly move the meeting back to the 23rd at 3p.m. at the Usher Building as originally planned.

Looking forward to your response.

Best regards,
Junhao

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Sent: Friday, April 18, 2025 5:24 PM
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Hi everyone,

I just realized that I have to be at King's Buildings this Tuesday.

Would it perhaps be more suitable for you to meet at King's Buildings instead of the Usher institute? I think AGB Room 1.01 is still empty, so we can use it for a meeting.

If it does not work for any of you, we will leave the meeting as it is.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 18 April 2025 08:00
To: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Dear All,

Jingshai, thank you for confirming for availability at noon on the 22nd. **Let's meet, 12:00 (noon) on Tuesday, April 22nd, at the Usher Building.**

The agenda proposed for our meeting:

1. The performance of the two recently trained models (ResNet and Vision Transformer).
2. The Transformer architecture from *Attention is All You Need.*
3. (Optional) My PhD research proposal **attached in the email,** if you are interested.
4. A concept inspired by my image processing course work: **Semantic entropy**, which can be used to quantify image semantic information and might relate to dataset bias.

Please let me know if you are happy with the agenda. If so, I will look forward to our discussion on Tuesday.

Best regards,
Junhao Sun

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Sent: Wednesday, April 16, 2025 1:03 PM
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Hi Junhao,

Unfortunately, I will not be available at these times. I could meet earlier on Tuesday though, for example around 12pm?

If that does not suit you or Jingshuai, feel free to meet without me.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 16 April 2025 12:48
To: Jingshuai Liu <jliu11@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Dear all,

Thanks for your emails.

1. **GitHub:** I've invited you as collaborators on GitHub and uploaded the ResNet and Transformer code/results. Unfortunately, due to GitHub's 100MB file size limit, I couldn't upload the model files.https://github.com/BoShao-Edin/Bias/invitations

2. **Transformer :**Since I hadn't worked with Transformer before, I spent more time than expected. While reading "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", I traced back to "Attention is All You Need", but I find it hard for me to explore the underlying principles (e.g., multi-head attention).

3. **Meeting**: As I still need more understanding of Transformer, would it be possible to have a face-to-face meeting for our next discussion? Could we possibly meet on April 21st Afternoon or 22nd Afternoon Or earlier Afternoon in Usher Building.?

Looking forward to hearing from you!

Best regards,
Junhao

From: Jingshuai Liu <jliu11@ed.ac.uk>
Sent: Tuesday, April 15, 2025 1:54 PM
To: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Hi, Junhao,

Hope everything is going well with you.

I can make it on 23rd April online. If you want to chat in-person, please come and find us in Usher Building.

My Github is https://github.com/JLiuED.

Best regards,
Jingshuai Liu

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Sent: 15 April 2025 12:15
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Hi Junhao,

Thank you for your reply.

I suggest that we meet **online** on **Wednesday 23rd of April at 3pm** if that suits @Jingshuai Liu too. Let me know if that works and I will send a meeting invite later.

My Github profile is: https://github.com/kvilouras

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 14 April 2025 13:30
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: Re: [MSc][Bias] Schedule a supervision meeting

Dear All,

I hope you are all doing well.

1. **Meeting Availability:**
   I am available for a meeting between April 16th and April 27th, from 2:00 p.m. to 6:00 p.m. each day. Please let me know a time that suits you. I am open to either an in-person or online meeting.

2. **GitHub Usernames:**
   Could you kindly share your GitHub username, as well as Jingshuai Liu's, so I can add you both as collaborators?

3. **ResNet Model:**
   I have just finished reading "7 Deep Residual Learning for Image Recognition" and have retrained the final fully connected layer of ResNet-101 using the CIFAR-100 dataset.

4. **Vision Transformer:**
   I am currently reading the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" and plan to use the Tiny-ImageNet dataset to train a Vision Transformer pre-trained on ImageNet. I expect to complete this by 10:00 p.m. on April 15th.

5. **MSc preliminary assessment (In person or online presentation with Q&A): Week 26/5-1/6, dates to be agreed with supervisor :**
   @Sotos Tsaftaris, could you please let me know when you are available to discuss this and the timing for the assessment?

   - The preliminary assessment for my project will be part of my practical work (10% of the grade) and will involve a 10-15 slide presentation. The presentation will cover the following:
   - A clear statement of the project's problem and objectives
   - A summary of the most relevant state-of-the-art literature
   - Mention of any progress made so far (e.g., software tools learned, demo code run)
   - A work plan for the project's implementation phase.

Looking forward to your response.

Best regards,
Junhao

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Sent: Monday, April 14, 2025 11:27 AM
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Cc: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc][Bias] Schedule a supervision meeting

Hi Junhao,

I hope you are well.

Would you like to schedule a meeting to discuss about your MSc project? I think it would be helpful to discuss any progress you have had, or any issues that you encountered, since the last time we discussed, and also plan ahead for the following weeks.

Could you please let us know when you are available? I'm aware that the exams' period is approaching, so please feel free to suggest the date and time that suits you best. Also let us know if you prefer an in-person or an online meeting.

Last, could you please remind me if you have any upcoming deadlines related to your MSc project?

Thanks in advance!

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

Outlook

Re: [MSc] [BIAS] [Deadline:01/06/2025] [Scheduling Preliminary Project Assessment]

From Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Date Mon 2025-05-26 8:04 AM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Cc Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>

Hello Junhao,

This is in my radar. I want to finish grading the exams before we do this presentation.
I will get back to you after also liaising with the other student as we tend to do this one after the other.

Best,
Sotos

++++++++++++++++++++++++++
Professor Sotirios A. Tsaftaris
Canon Medical/Royal Academy of Engineering Research Chair in Healthcare AI
Chair in Machine Learning and Computer Vision
Director EPSRC AI Hub for Causality in Healthcare AI with Real Data (CHAI)
ELLIS Fellow
School of Engineering, University of Edinburgh,
Rm 2.06, Alexander Graham Bell, King's Buildings, EH9 3FG, Edinburgh, UK
+44 (0)131 650 5796 // S.Tsaftaris@ed.ac.uk // https://vios.science
++++++++++++++++++++++++++

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: Monday, May 26, 2025 8:00:00 AM
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Cc: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc] [BIAS] [Deadline:01/06/2025] [Scheduling Preliminary Project Assessment]

Dear Sotos,

I hope this message finds you well.

As per the recent announcement from Dr. Nick Polydorides, the MSc preliminary project assessment is scheduled to take place during the week of 26th May to 1st June. I would like to kindly ask when you might be available during that period for my presentation and Q&A session.

Please let me know your availability so that we can agree on a suitable date and time, my calendar has been attached.

Looking forward to your reply.

Best regards,
Junhao

From: Nick Polydorides - npolydor@ed.ac.uk <learn@ed.ac.uk>
Sent: Friday, March 7, 2025 9:49 AM
Subject: Signal Processing & Communications: Project and Thesis (2024-2025)[SB5+]: Presentation about your projects

PGEE110102024-5SV1SB5plus

Signal Processing & Communications: Project and Thesis (2024-2025)[SB5+]

New announcement from Nick Polydorides

Presentation about your projects
Dear SPC MSc students,

I hope you are making progress with your courses and also find some time to prepare for your upcoming dissertation projects. As you will now be able to see under the "Assessment Information" page on Learn, the next deadline is coming in a couple of months time. In particular, note that

10% out of the Practical work will be on your preliminary project assessment to take place at the end of semester 2. Exact dates to be agreed with supervisors around your exam dates. This preliminary assessment will have the form of a 10-15 slide presentation to include:
i) A clear statement of the project's problem and objectives
ii) A succinct summary of most pertinent state-of-art literature. Your supervisor will guide you on which literature you have to focus on to get an in-depth understanding and of course you can also rely on the literature review you have done for your Engineering Research methods course in semester 1, and
iii) Mention of any progress achieved so far, e.g. learning a software tool, running some external demo code to get some insights etc
iv) A work plan for the project implementation phase.

Deadline: MSc preliminary assessment (In person or online presentation with Q&A): Week 26/5-1/6, dates to be agreed with supervisor

If you have any questions please let me know by emailing me.

Regards,
Nick Polydorides

View Announcement

Any multimedia items must be viewed online

Outlook

Re: [MSc][BIAS][Meeting to Discuss ResNet-18 Results on 3rd June, Bias]

From Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Date Mon 2025-06-09 12:44 PM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Cc Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>

Hi Junhao,

Thank you for your update!

The plan of actions seems reasonable to me. It would be great if you could show us some data visualizations on Wednesday.

Also, you raise a good point in terms of data preprocessing: torchvision's resnet18 model requires that the images used for training/inference have a consistent size of 224x224 (this transform is part of the model class, see here https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html). We can discuss this further on Wednesday.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 04 June 2025 08:00
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: Re: [MSc][BIAS][Meeting to Discuss ResNet-18 Results on 3rd June, Bias]

Dear all,

I have finished the summary of our recent meeting and attached it here. Please take a look and let me know if any revisions are needed.

Looking forward to your suggestion!

Best regards,
Junhao

From: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Sent: Wednesday, May 28, 2025 5:17 PM
To: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Jingshuai

Liu <jliu11@ed.ac.uk>
Subject: Re: [MSc][BIAS][Meeting to Discuss ResNet-18 Results on 3rd June, Bias]

Can you please send me an invite for this meeting?

Best
Sotos

++++++++++++++++++++++++++
Professor Sotirios A. Tsaftaris
Director Causality in Healthcare AI Hub (CHAI)
Chair in Machine Learning and Computer Vision
Canon Medical/Royal Academy of Engineering Research Chair in Healthcare AI
ELLIS Fellow
School of Engineering, The University of Edinburgh
Usher Building, 5-7 Little France Road, Edinburgh BioQuarter - Gate 3
EDINBURGH, EH16 4UX
+44 (0)131 650 5796 // S.Tsaftaris@ed.ac.uk // https://vios.science
++++++++++++++++++++++++++

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Date: Wednesday, 28 May 2025 at 12:41
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>, Jingshuai Liu <jliu11@ed.ac.uk>
Cc: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: Re: [MSc][BIAS][Meeting to Discuss ResNet-18 Results on 3rd June, Bias]

Hi Junhao,

Thank you for your email. Good job!

May I ask that you also take a look at Step 1 "Reconstruct a new dataset from the original datasets" before our meeting? Specifically, it would be great if you can identify the overlapping classes in both datasets. Also keep in mind that some of the overlapping classes might not share the exact same word indicating the label (we will discuss that during the meeting).

Also, Tuesday 3rd of June at 3pm works well for me (Jingshuai will not attend the meeting). I'll send you a meeting invite shortly.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 28 May 2025 11:07
To: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Cc: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: [MSc][BIAS][Meeting to Discuss ResNet-18 Results on 3rd June, Bias]

Dear All,

I hope you are all doing well.

I wanted to let you know that I am currently training the ResNet-18 model as per Konstantinos's suggestion (see attachment), and it is estimated to be finished by 01/06/2025. @Konstantinos Vilouras,@Jingshuai Liu Shall we schedule a short in-person meeting to go over the results and plan the next steps?

Would you be available on Tuesday, 3rd June 2025 at 3 p.m. in the Usher Building (room 1.03 or another convenient room)? Please let me know if this time and location work for you, or suggest an alternative slot in the same week.

Looking forward to our discussion!

Best regards,
Junhao Sun

Outlook

Re: [MSc][BIAS][Enquiry On Dissertaion Format]

From Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Date Fri 2025-06-27 1:50 PM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>

Hello a latex template should be available within the course information on learn. If not, please ask the course coordinator.

Best,
Sotos

++++++++++++++++++++++++++++
Professor Sotirios A. Tsaftaris
Director Causality in Healthcare AI Hub (CHAI)
Chair in Machine Learning and Computer Vision
Canon Medical/Royal Academy of Engineering Research Chair in Healthcare AI
ELLIS Fellow
School of Engineering, The University of Edinburgh
Usher Building, 5-7 Little France Road, Edinburgh BioQuarter - Gate 3
EDINBURGH, EH16 4UX
S.Tsaftaris@ed.ac.uk // https://vios.science
++++++++++++++++++++++++++++

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: Friday, June 27, 2025 1:48:25 PM
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>
Subject: [MSc][BIAS][Enquiry On Dissertaion Format]

Dear Sotos,

I hope you're well.

Could you please let me know if you have any **specific formatting requirements** or a preferred template for the dissertation (e.g. LaTeX style file versus Word)? If there is a template or style file you recommend, would you be able to share it with me? Otherwise, should I simply follow the general structure and formatting guidelines posted on Learn (Document)?

Thank you very much for your advice.

Best regards,
Junhao

The following is the general structure and formatting guidelines posted on Learn:

MSc Dissertation report - structure and content

Here is some useful information about the dissertation report structure, format and content. This serves as a guidance for good practice in academic writing, but you are expected to consult your project supervisor before you finalize your report to ensure that it meets their specific requirements.

1) Please use the LaTeX style file for typesetting your report. It is recommended that you use this, but you may also use MicroSoft Word if your supervisor agrees to it.

2) Font sizes, margins and line spacing should match or be very similar to those of the pre-dissertation report, e.g. 12pt regular Arial, Times or Calibri, default margins in word, 1.5 line spacing, and page numbering should be as {1,2,3,...}. You should also follow the conventional section, subsection, subsubsection format with proper enumeration and headings beginning with a capital letter. We anticipate that the length of these reports will be around 50 pages, but this is only indicative, if you have many numerical experiments and images you may need to add more pages.

3) Ensure that the report is well structured to chapters without significant text repetition, perhaps only a small critical component to ensure the flow of the information from the one chapter to the other.

4) Do not copy or re-write material from other publications but rather explain how does that relate to your own work and perhaps how you extend it.

5) The report should begin with a front page showing your name, S number, and project title. You should also include, in a separate page, a declaration of originality stating that the report contains your own personal work and any external information used is appropriately cited. Use something like "I declare that the contents of this report is my own work and any external information used is cited in the bibliography section".

6) Contents: In terms of contents, I suggest the following generic template but do check with your supervisor to make sure that these suffice.

An introduction with a concise statement of the technical problem to remind the reader why you are doing this project. This could fit in a page or two at most, but if possible do not copy the relevant section of the pre-dissertation report. By now you should have a good idea about this.

A Literature review section covering the state of the art in the problem under consideration.

A section dedicated on developing your methodology on addressing this problem. This should include the necessary "theory" or the "scientific foundation" on which your methodology is based. You may want to include a separate section to include all your proofs if you have many of those in your methodology, but that's optional and you need to discuss this with your supervisor.

A section on results, describing the numerical or otherwise experiments you have done in demonstrating the methodology with adequate discussion and analysis is necessary. This is likely to contain several figures and tables so please make sure that these are sufficiently captioned. Try to avoid using captions like "This is the graph of f against x" or "this is the error plot" as these are obvious by inspection and do not offer any scientific insight about what is important in the plotted curve. Another advice is to make sure you discuss adequately the results that confirm the hypothesis of the methodology as well as those that (maybe) do not.

A conclusions and further work section: These are two distinct elements sharing the same section of the report. Please give and substantiate the conclusions of your work and explain their potential impact or implications to the targeted application(s). Please make sure to understand what "conclusions" are before you do that. Hint: Conclusions are not the summary of what you have done or set out to do in your project. There is no official

requirement for having an impacts section in the report but it is usually a sign of deeper understanding if you can explain “how the world is a better place for what you have done in your project” and how could this work find its way into implementation into the various application sectors. In terms of the future work, we expect to see if you have a broader vision of the area of your topic to see what extensions would be necessary to extend this work further.

Outlook

Re: [MSc][BIAS] [Meeting Summary and Proposed Next Steps]

From Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Date Fri 2025-06-13 12:21 PM
To Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Cc Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>

Dear Konstantinos ,

I will write down the steps to conduct this experiment.

Thursday 19th of June 3pm works for me.

Best,
Junhao

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Sent: Friday, June 13, 2025 12:00 PM
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>; Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: Re: [MSc][BIAS] [Meeting Summary and Proposed Next Steps]

Hi Junhao,

Thank you for the update.

For your "Experiment 2" idea, please make sure to write down the steps you will follow to conduct this experiment before the next meeting so we can discuss it a bit more.

Also, I will be available on Thursday 19th of June if that works. Should we say at 3pm?

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 12 June 2025 16:32
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc][BIAS] [Meeting Summary and Proposed Next Steps]

Dear all,

I hope you're well.

I've finished writing up our meeting summary for **11.06.2025(see attachment)** – please feel free to have a look when you get a chance.

Looking ahead, I would like to suggest two follow-up experiments (details in the word document for meeting summary)that we could consider:

Experiment 1: *Which layer of ResNet-18 is more sensitive to dataset bias?*

- We will add a linear probe after **each block** of ResNet-18
- The aim is to analyze which stage of the network is more sensitive to dataset-specific bias

---

Experiment 2: *Dataset bias – non-semantic information test*
(Inspired by: "A Decade's Battle on Dataset Bias")

- Prior work shows that dataset bias often aligns with **semantic features** useful for classification
- This experiment is simply to test if the dataset bias discovered by neural network models are relevant to **non-semantic features** that are meanless for image classification.

Also, regarding our next group meeting – would it be possible to move it to **Friday (June 21st, 1 p.m.)** or **Thursday (June 20th)** instead of Wednesday? I have my SPC scholarship interview on Monday, and performing well in it is my top priority at the moment and I plan to devote most of my time into it before Monday.

Looking forward to your response

Best regards,
Junhao

From: Junhao Sun
Sent: Wednesday, June 11, 2025 4:36 PM
To: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Subject: Accepted: MSc Meeting
When: Occurs every Wednesday from 3:00 PM to 3:30 PM effective 6/11/2025 until 7/2/2025.
Where: ENG MR Usher 1.04 D3 Meeting Room (GB)

Outlook

Re: [MSc][BIAS] [Meeting Summary6.25]

From Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Date Fri 2025-06-27 8:59 AM
To Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>

Dear Konstantinos,

Thank you for your email and for the helpful clarification regarding the probabilistic interpretation of dataset bias — I agree that defining the variables explicitly will be useful for structuring the thesis.

Regarding the meeting, Wednesday 9/7 works better for me. In the meantime, if I encounter any issues or questions, I will reach out via email or Teams as suggested.

Best regards,
Junhao

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Sent: Thursday, June 26, 2025 11:20 AM
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Subject: Re: [MSc][BIAS] [Meeting Summary6.25]

Hi Junhao,

Thank you for the summary.

Regarding the last question (probabilistic interpretation of dataset bias): It will be helpful to first define the variables of interest, i.e. images **x**, labels **y** that overlap between dataset sources (here, CIFAR and TinyImagenet), and dataset-specific biases **b** which are not useful for the actual classification task (predicting y from x). Also note that these biases **b** exist in the given images **x**, but we do not observe **b** during training!
I think it will be nice to describe the problem of dataset bias in such way in your thesis. However, this is **not** mandatory!

Also, based on our discussion yesterday, I suggested that we should meet one more time in the following 2 weeks in person to answer any remaining questions. For any other issues that might occur, we can talk either via email or via Teams. Therefore, can you please let me know if you want to meet on Wednesday 2/7 or Wednesday 9/7 in Usher?

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 26 June 2025 08:00

To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc][BIAS] [Meeting Summary6.25]

Dear all,

I’ve finished drafting the meeting summary 6.25.

Please take a look and share any feedback or suggested edits directly in the file or via email.

Looking forward to your response.

Best regards,
Junhao

Outlook

Re: [MSc][Bias][An Idea to Mitigate Dataset Bias]

From Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Date Tue 2025-07-01 11:22 AM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>

Hi Junhao,

You mentioned that Wednesday 9/7 works best for you, so we will meet then in Usher.

I will update the meeting invite in a bit.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 01 July 2025 08:00
To: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Subject: Re: [MSc][Bias][An Idea to Mitigate Dataset Bias]

Dear Konstantinos,

Thank you for your email. To make sure I prepare the document for the correct date, could you please confirm whether you mean Wednesday **2 July 2025** or Wednesday **9 July 2025**?

Best regards,
Junhao

From: Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Sent: Monday, June 30, 2025 10:49 AM
To: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Subject: Re: [MSc][Bias][An Idea to Mitigate Dataset Bias]

Hi Junhao,

Thank you for your email.

Please write down in a Word document (or a PPT presentation) and we can discuss it on Wednesday.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 30 June 2025 08:00
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc][Bias][An Idea to Mitigate Dataset Bias]

Dear all,

I hope you are well.

An Idea to Mitigate Dataset Bias:

In fact, there are infinite varieties of dataset bias, some of which are even unknown to us humans, and it is impossible to name all the dataset biases. But dataset bias requires at least two datasets if it is to be captured by the model. **We don't necessarily need to completely eliminate dataset bias in our dataset, we just need to make sure that the model doesn't capture (learn) dataset bias during training.**

To minimize the dataset bias learned by the model, the idea is that we directly add a term λ P ($D_k|\ x_i$) to the loss function and **apply a penalty if the model can successfully predict the dataset.**

Loss Func1 = sum_i [-log (P ($y_j|\ x_i$))+ λ P ($D_k|\ x_i$)]

$$\text{Loss Func} = \sum_{i=0}^{Batch\ Size} \left[-log_2\left(P\left(y_j|x_i\right)\right) + \lambda\ P\ \left(D_k|x_i\right)\ \right]$$

Loss Func2 = sum_i [-log (P (y_j| x_i))+ λ (P (D_k| x_i) - Name the dataset Accuracy of Guess by Chance)]

$$\text{Loss Func2} = \sum_{i=0}^{Batch\ Size} \left[-log_2\left(P\left(y_j|x_i\right)\right) + \lambda\left(P(D_k|x_i) - Name\ the\ dataset\ Accuracy\ of\ Guess\ by\ Chance\right) \right]$$

y_j: *Image label*

x_i: *Image*

D_k: *Dataset ID (which dataset the image x_i comes from)*

λ: *Coefficient indicating how much punishment imposed on the term P (D_k| x_i)*

Shall we discuss the idea and give me some guidance to design some experiments to implement it?

Best,
Junhao

Outlook

Re: [MSc][BIAS][Meeting Summary 7.9]

From Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>
Date Thu 2025-07-10 10:45 AM
To Junhao Sun <J.Sun-90@sms.ed.ac.uk>

Hi Junhao,

Thank you for your email.

For point "5. Log book Contents", the last 2 bullets "List of all Python scripts and References to all reviewed papers" are not required. The references will be part of your thesis document, and you will also upload your code on a Github repository.

Please take a look at my post on the Teams channel, which I hope would make things clear.

Best,

Konstantinos Vilouras
PhD student
School of Engineering, University of Edinburgh
https://vios.science/

From: Junhao Sun <J.Sun-90@sms.ed.ac.uk>
Sent: 10 July 2025 08:00
To: Sotos Tsaftaris <S.Tsaftaris@ed.ac.uk>; Konstantinos Vilouras <konstantinos.vilouras@ed.ac.uk>; Jingshuai Liu <jliu11@ed.ac.uk>
Subject: [MSc][BIAS][Meeting Summary 7.9]

Dear all,

I’ve finished drafting the meeting summary 7.9.

Please take a look and share any feedback or suggested edits directly in the file or via email.

Looking forward to your response.

Best regards,
Junhao

Investigate Dataset Bias with Modern Deep Neural Networks

MSc SPC Project

06/06/2025

Exam No. B273213

Dataset Introduction

Dataset02, Bicubic

Tiny64 × 64 → 32 × 32 (BICUBIC) → 128 × 128 (BICUBIC)
CIFAR32 × 32 → 128 × 128 (BICUBIC)

Data Split
test = (CIFAR-100 test + TinyImageNet validation);
(No TinyImageNet test annotation access;
CIFAR-100 do not have a validation set)

train/validation = 90%/10% random split
(CIFAR-100 Train+ TinyImageNet Train)

Dataset03, Bicubic

Tiny64×64 → 256×256 (BICUBIC) → 224×224 (Random Crop)
CIFAR32 × 32 → 256 × 256 (BICUBIC) → 224×224 (Random Crop)

Data Split
test = (CIFAR-100 test + TinyImageNet validation);
(No TinyImageNet test annotation access;
CIFAR-100 do not have a validation set)

train/validation = 90%/10% random split
(CIFAR-100 Train+ TinyImageNet Train)

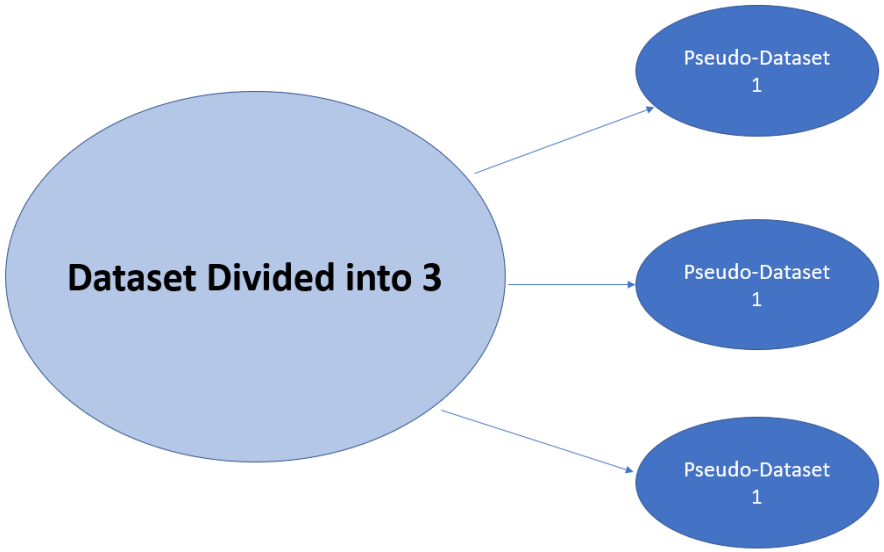
Dataset04, Bicubic

Tiny64 × 64 → 32 × 32 (BICUBIC 256 × 256 (BICUBIC) → 224 × 224 (Random Crop)
CIFAR32×32 → 256×256 (BICUBIC) → 224×224 (Random Crop)

Data Split
test = (CIFAR-100 test + TinyImageNet validation);
(No TinyImageNet test annotation access;
CIFAR-100 do not have a validation set)

train/validation = 90%/10% random split
(CIFAR-100 Train+ TinyImageNet Train)

Pseudo-dataset Experiment



If the samples of different datasets were **unbiasedly drawn from the same distribution**, the model should not discover any dataset-specific bias. To check this, we study a pseudo-dataset classification task, in which the different "datasets" are uniformly sampled from a single dataset.

imgs per set	w/o aug	w/ aug
100	100.0	100.0
1K	100.0	100.0
10K	100.0	fail
100K	fail	fail

Table 6: *Training* accuracy on a pseudo-dataset classification task. Here we create 3 pseudo-datasets, all of which are sampled without replacement from the same source dataset (YFCC). This *training* task becomes more difficult for the network to solve if given more training images and/or stronger data augmentation. Validation accuracy is ~33% as no transferrable pattern is learned.

——from *A Decade's Battle*

Dataset Bias is some kinds of distribution difference between the sampled data(dataset) and the real world, which may make a model's output systematically deviates its expectation(target).

Conclusion

I. Conclusion: Dataset Bias captured by DNN is, in essence, a kind of **distribution difference** (pseudo-dataset)

II. Conclusion: Dataset Bias contain **semantic information** and/or **non-semantic information** (simultaneously) (leverage bias; Non-Semantic Test)

III. Conclusion: To sum up, name the dataset result is determined **primarily by non-semantic information**, and semantic **information** have only a subordinate influence

(model primarily use non-semantic information to name the dataset)

I. Guess: Deep models gradually transform non-semantic inputs into semantic concepts—yet every layer may still retain useful low-level detail. (Is it true? I Guess)

PhD Direction Plan – Leverage Bias (Semantic)

Pre-train to name *Labels* ; Fine-tune to name *Dataset*

Pre-train to name *Dataset* ; fine-tune to name *Labels*

It has been discovered that such **bias** may contain **some generalizable and transferrable patterns**, carrying some semantic information that is transferrable to image classification tasks.

The dataset bias discovered by neural network models are relevant to **semantic features** that are useful for image classification.

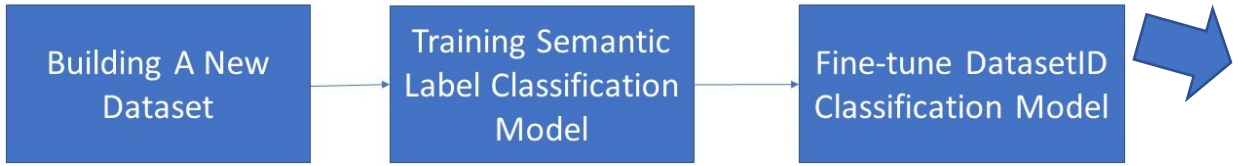
——from *Unbiased_look & A Decade's Battle*

case	transfer acc
random weights	6.7
Y+C+D	27.7
Y+C+D+W	34.2
Y+C+D+W+L	34.2
Y+C+D+W+L+I	34.8
MAE [19]	68.0
MoCo v3 [5]	76.7

Table 8: Features learned by classifying datasets can achieve nontrivial results under the linear probing protocol. Transfer learning (linear probing) accuracy is reported on ImageNet-1K, using ViT-B as the backbone in all entries. The acronyms follow the first letter of each dataset in Table 2.

Name the Dataset – Build New DataSet

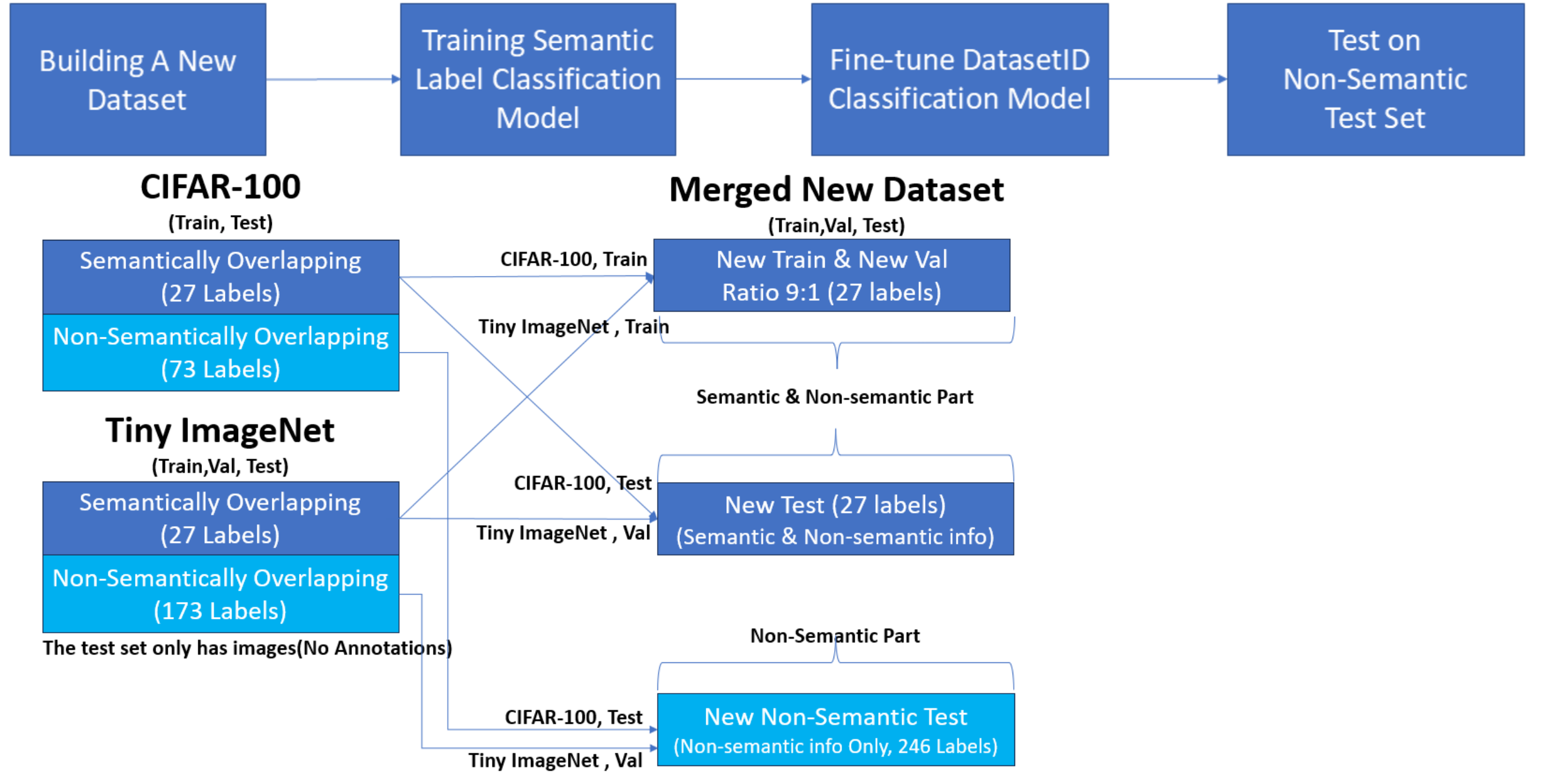
Overall Task Workflow



Sematically Overlapping Labels NewDataset

	A	B	C	D	E	F
1	new_label_id	new_label	tiny_synset_id	tiny_synset_name	cifar_fine_id	cifar_fine_name
2	0	goldfish	n01443537	goldfish, Carassius auratus	1	aquarium_fish
3	1	brown bear	n02132136	brown bear, bruin, Ursus arctos	3	bear
4	2	bee	n02206856	bee	6	bee
5	3	lady beetle	n02165456	ladybug, ladybeetle, lady beetle, ladybird, ladybird beetle	7	beetle
6	4	pop bottle	n03983396	pop bottle, soda bottle	9	bottle
7	5	suspension bridge	n04366367	suspension bridge	12	bridge
8	6	school bus	n04146614	school bus	13	bus
9	7	monarch butterfly	n02279972	monarch, monarch butterfly, milkweed butterfly, Danaus plexippus	14	butterfly
10	8	Arabian camel	n02437312	Arabian camel, dromedary, Camelus dromedarius	15	camel
11	9	ox	n02403003	ox	19	cattle
12	10	rocking chair	n04099969	rocking chair, rocker	20	chair
13	11	chimpanzee	n02481823	chimpanzee, chimp, Pan troglodytes	21	chimpanzee
14	12	cockroach	n02233338	cockroach, roach	24	cockroach
15	13	African elephant	n02504458	African elephant, Loxodonta africana	31	elephant
16	14	computer keyboard	n03085013	computer keyboard, keypad	39	keyboard
17	15	lawn mower	n03649909	lawn mower, mower	41	lawn_mower
18	16	lion	n02129165	lion, king of beasts, Panthera leo	43	lion
19	17	American lobster	n01983481	American lobster, Northern lobster, Maine lobster, Homarus americ	45	lobster
20	18	mushroom	n07734744	mushroom	51	mushroom
21	19	orange	n07747607	orange	53	orange
22	20	plate	n07579787	plate	61	plate
23	21	snail	n01944390	snail	77	snail

Experiment 2: *Dataset bias – Non-semantic Information Test*



Dataset Introduction

Dataset03, Bicubic

Tiny64×64 → 256×256（BICUBIC）→ 224×224 (Random Crop)

CIFAR32 × 32 → 256 × 256（BICUBIC）→ 224×224 (Random Crop)

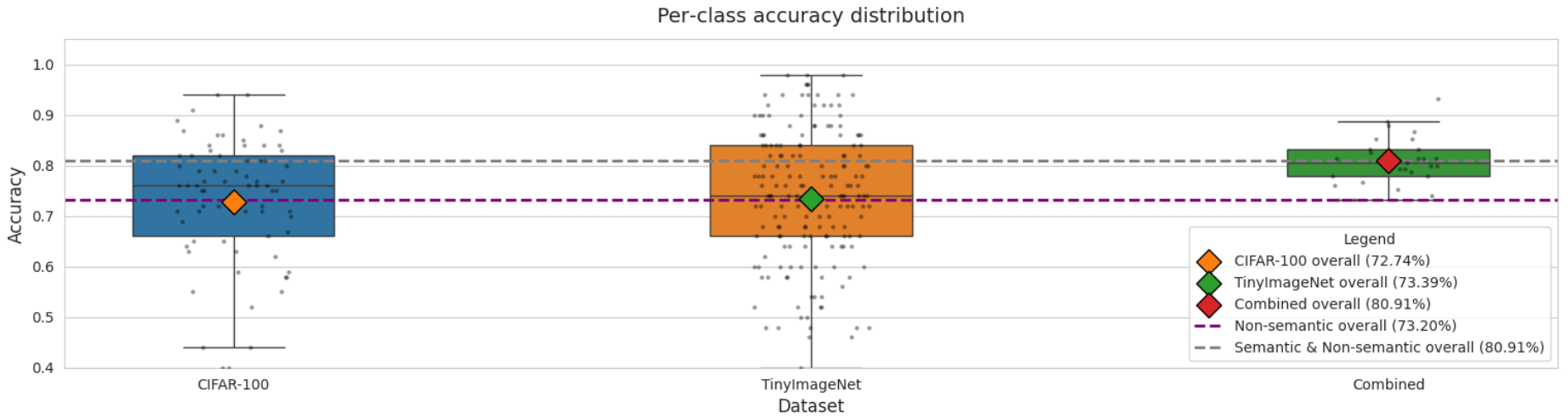
Dataset04, Bicubic

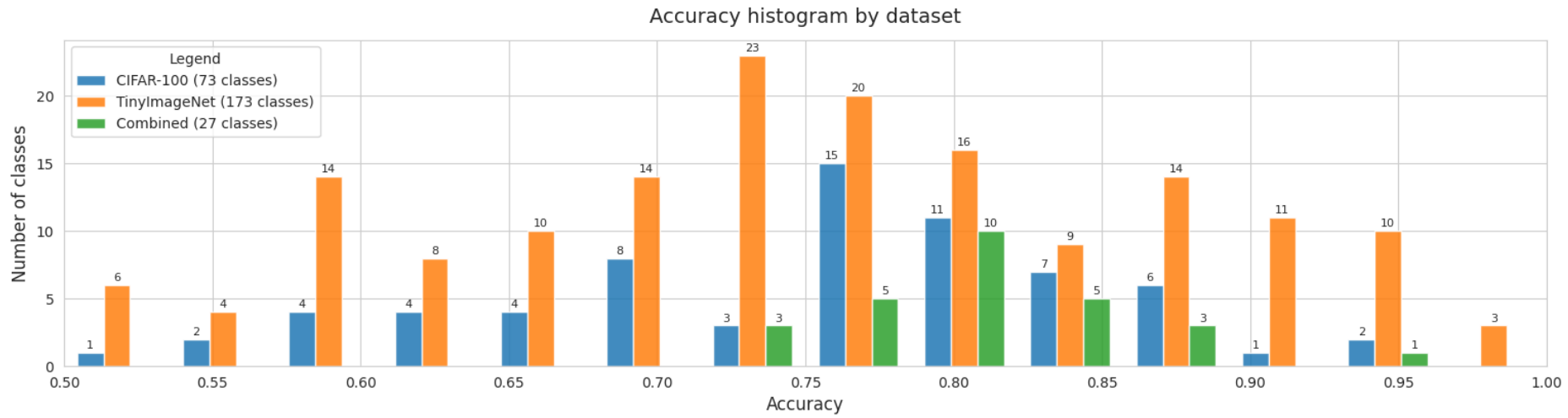
Tiny64 × 64 → 32 × 32（BICUBIC 256 × 256（BICUBIC）→ 224 × 224 (Random Crop)

CIFAR32×32 → 256×256（BICUBIC）→ 224×224 (Random Crop)

Experiment 2: *Non-semantic Information Test* ***for 1804***

TINY and CIFAR contain non-semantic information only; Combined contains both semantic and non-semantic information.

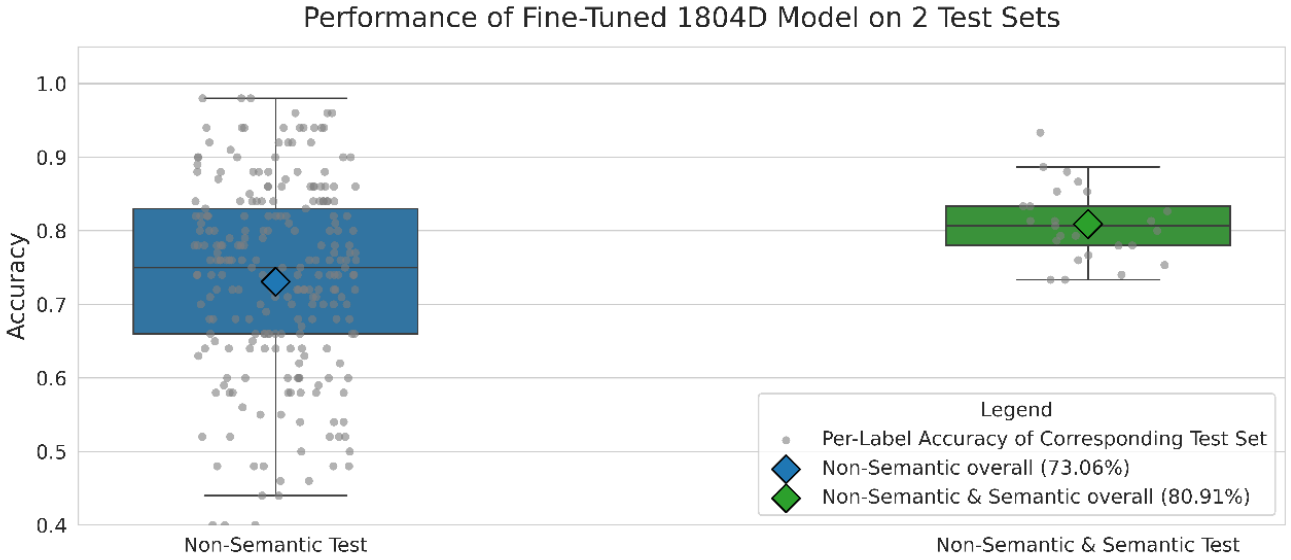


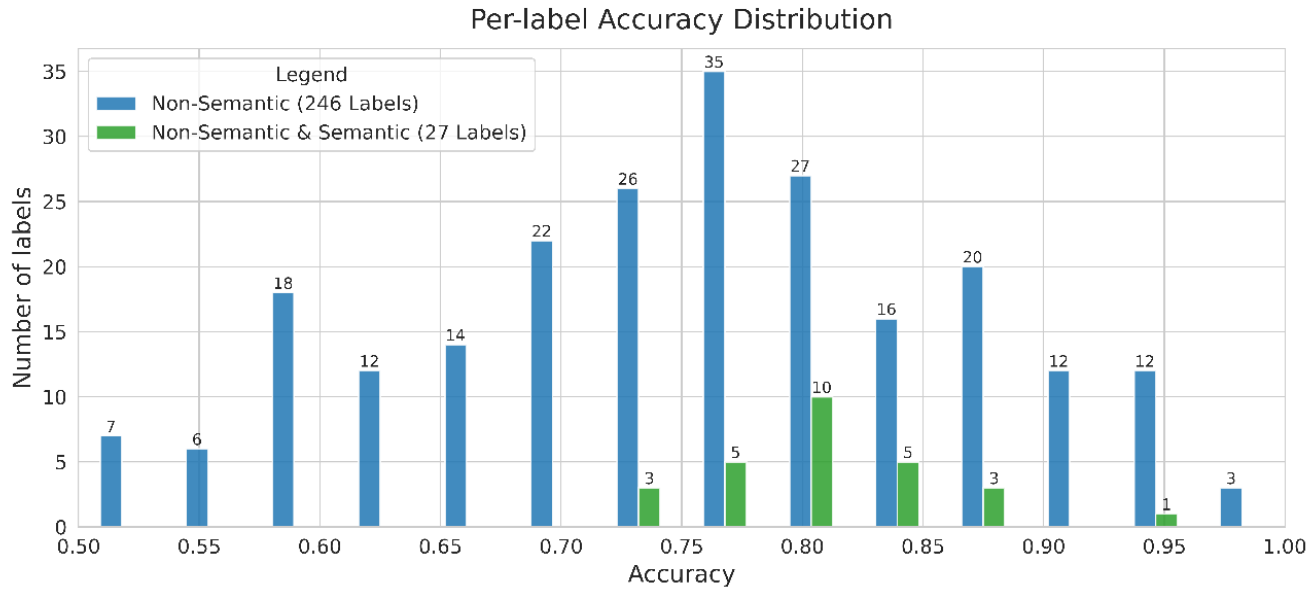


Total classes – CIFAR-100: 73, TinyImageNet: 173, Combined: 27
Overall accuracy – CIFAR-100: 0.7274 | TinyImageNet: 0.7339 | Combined: 0.8091
Non-semantic overall (avg classes CIFAR-100 & TinyImageNet): 0.7320 | Semantic & Non-semantic overall (Combined): 0.8091

Combining Semantic info and non-semantic info achieves a higher accuracy rate than using non-semantic information only

Experiment 2: *Non-semantic Information Test* ***for 1804***

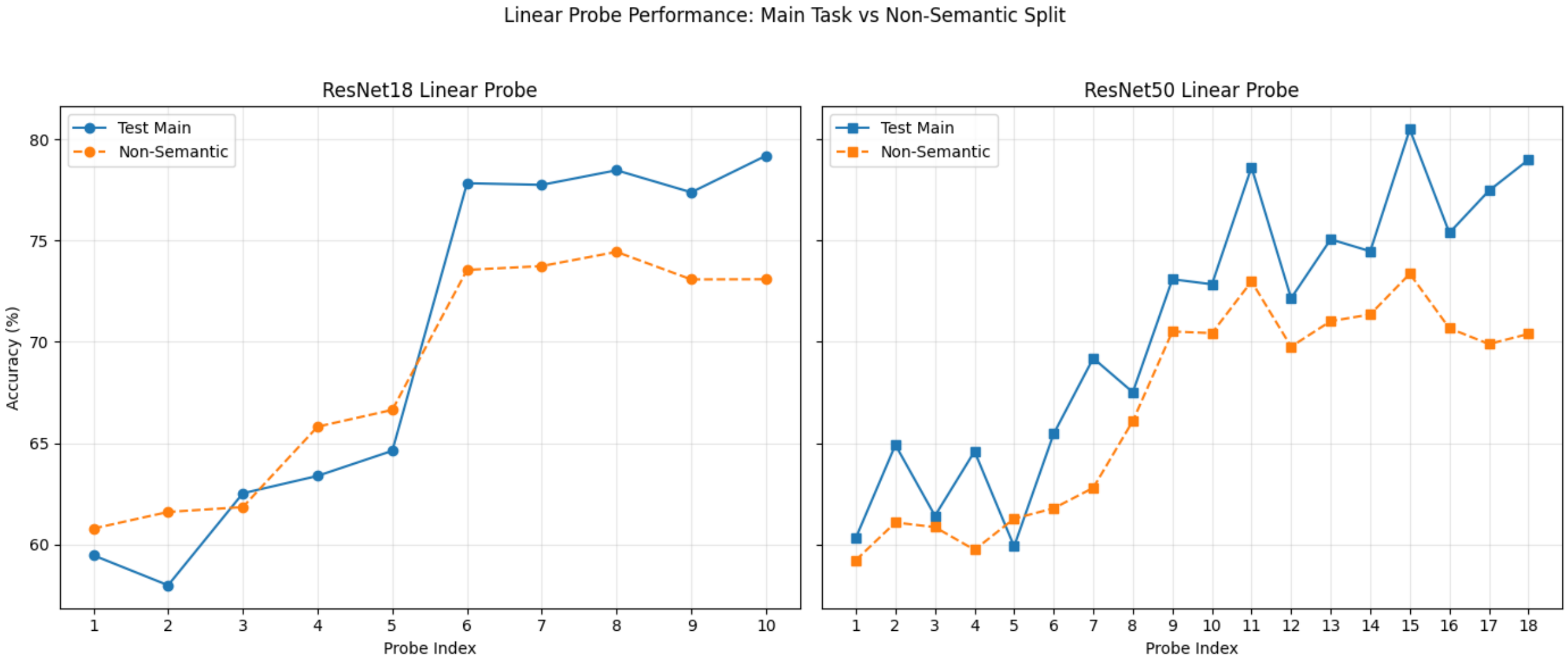




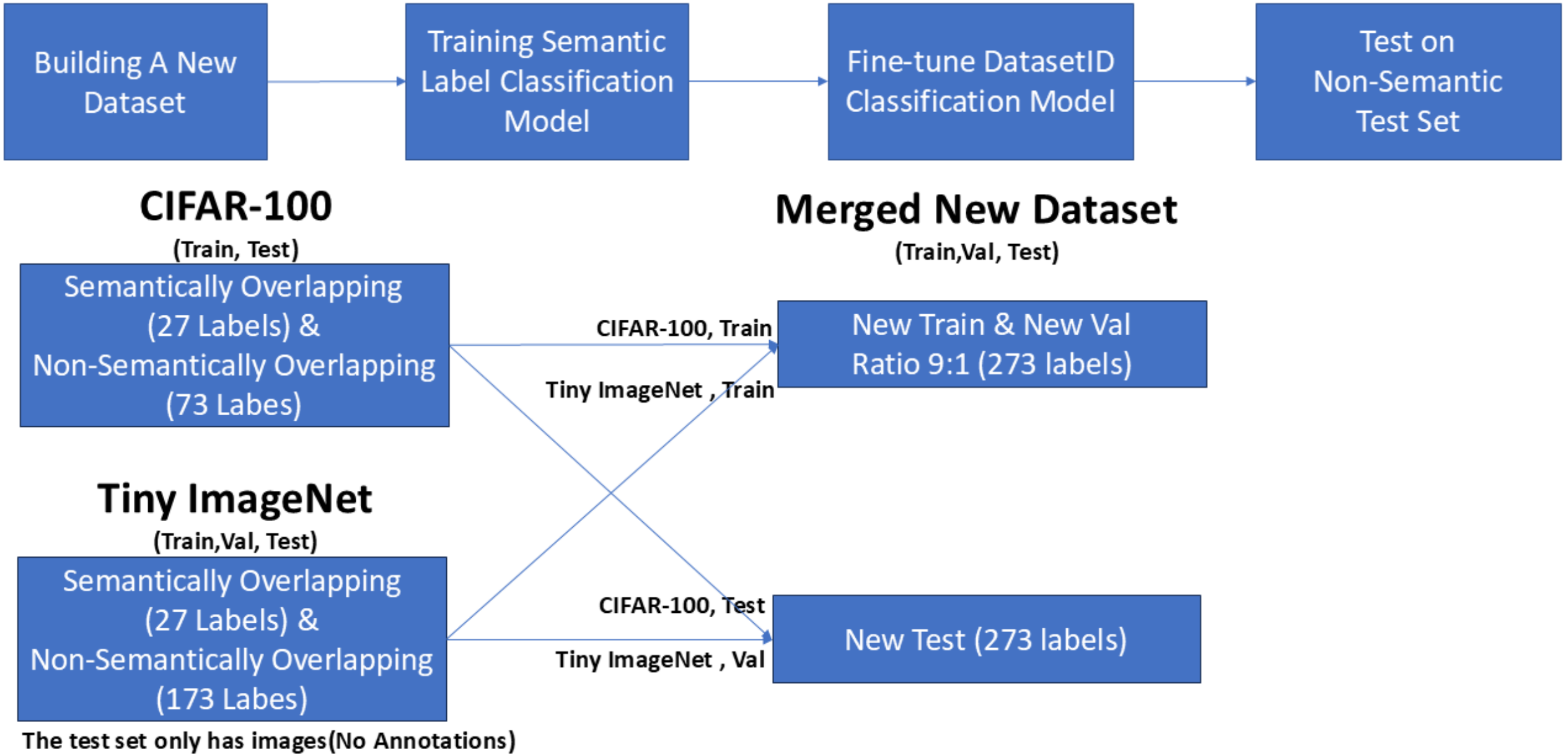
Total labels – Non-Semantic: 246, Full: 27. Overall accuracy – NS: 73.06%, Full: 80.91%

Combining Semantic info and non-semantic info achieves a higher accuracy rate than using non-semantic information only

Experiment 2: *Non-semantic Information Test* ***for 1804***



Experiment 3: *2 Full Dataset Dataset05,06*



Dataset Introduction

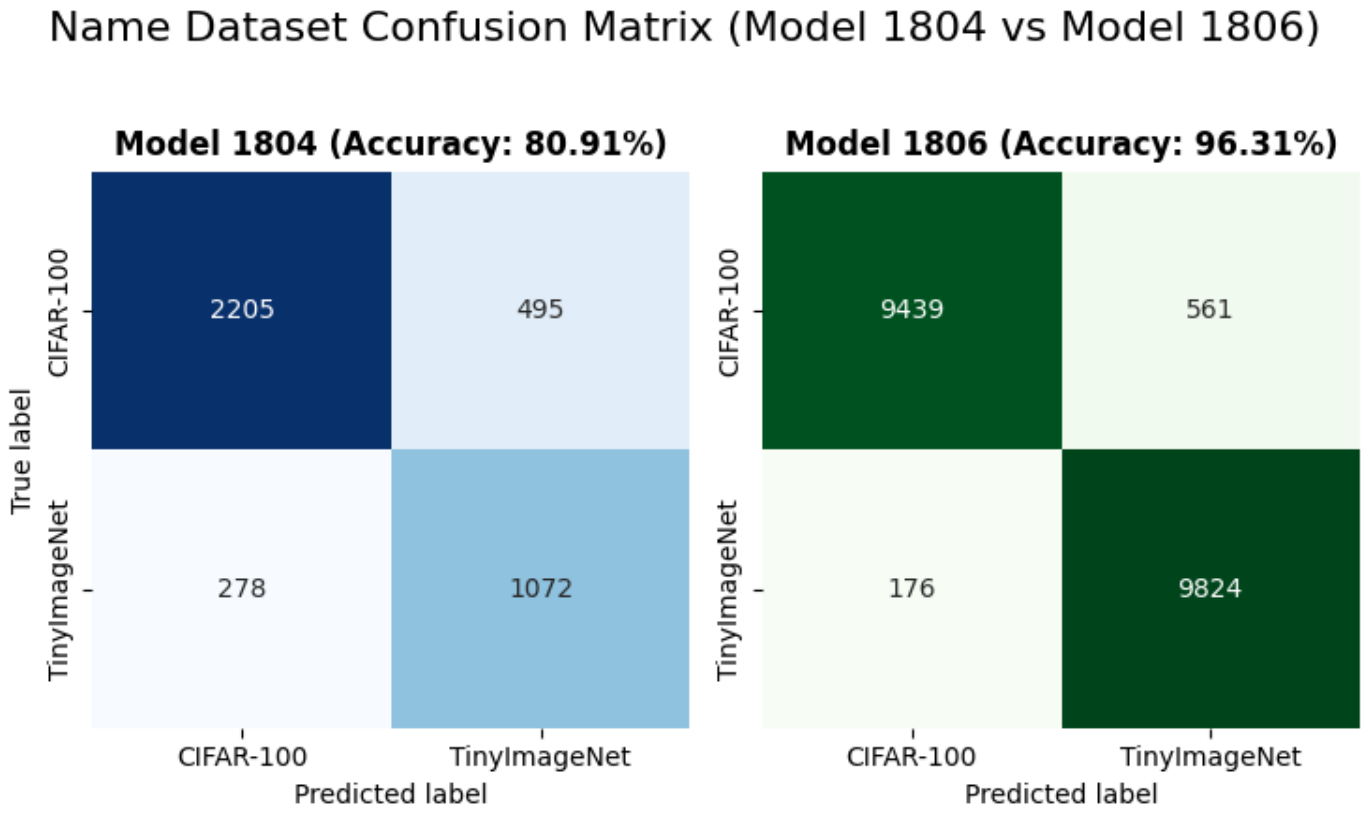
Dataset05, Bicubic

Dataset05, Bicubic
Tiny64×64 → 256×256 （BICUBIC） → 224×224 (Random Crop)
CIFAR32×32 → 256×256 （BICUBIC） → 224×224 (Random Crop)

Dataset06, Bicubic

Tiny64 × 64 → 32 × 32 （BICUBIC
256 × 256 （BICUBIC） → 224 × 224 (Random Crop)
CIFAR32×32 → 256×256 （BICUBIC） → 224×224 (Random Crop)

Experiment 2: *Full dataset* ***for 18041806***



as the sample size increases, the model learns more patterns of dataset bias, causing accuracy to rise dramatically

1804

[Test Classification Report]

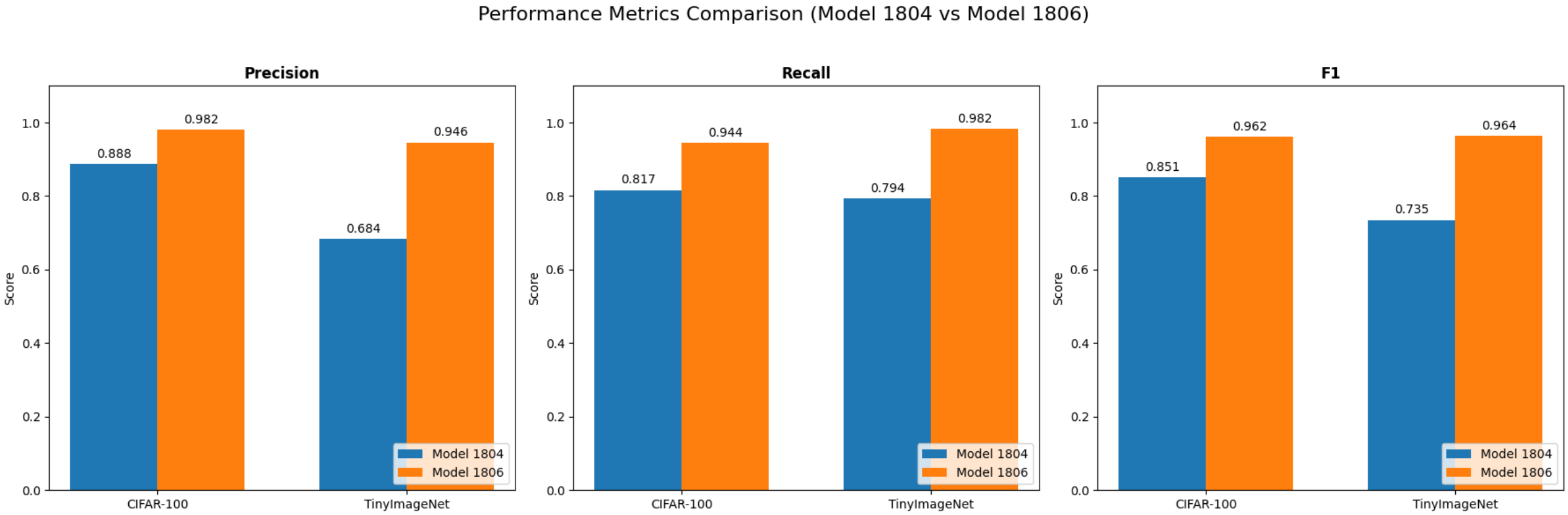
	precision	recall	f1-score	support
CIFAR-100	0.8880	0.8167	0.8509	2700
TinyImageNet	0.6841	0.7941	0.7350	1350
accuracy			0.8091	4050
macro avg	0.7861	0.8054	0.7929	4050
weighted avg	0.8201	0.8091	0.8122	4050

1806

[Test Classification Report]

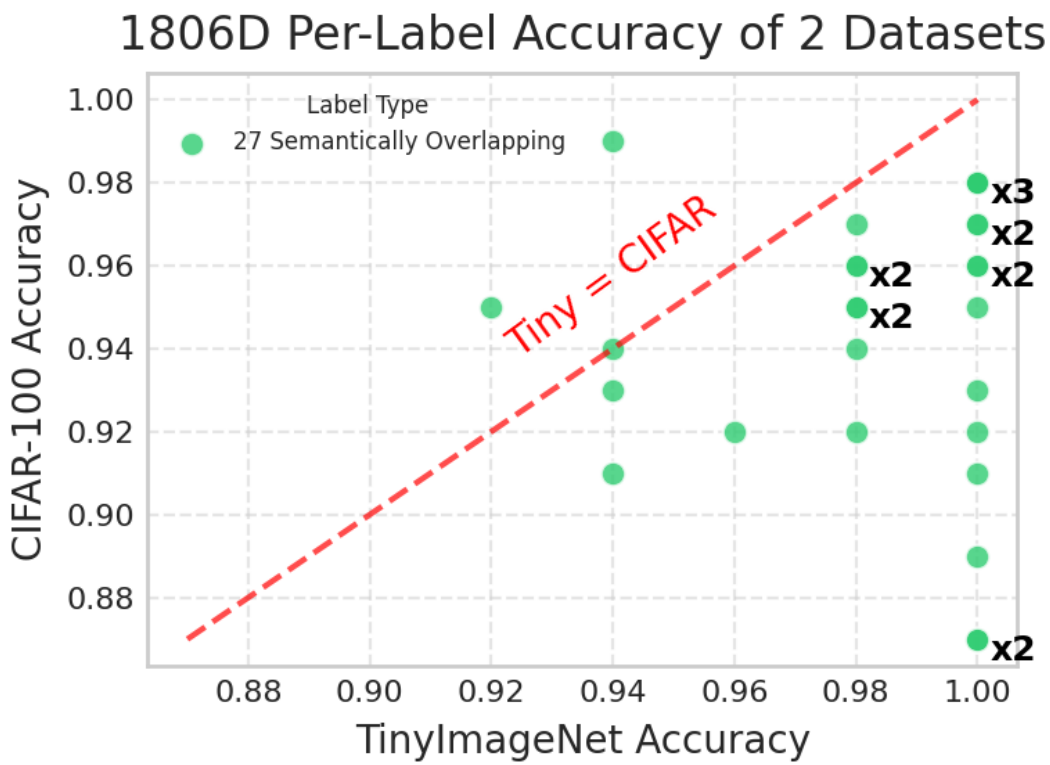
	precision	recall	f1-score	support
CIFAR-100	0.9817	0.9439	0.9624	10000
TinyImageNet	0.9460	0.9824	0.9638	10000
accuracy			0.9631	20000
macro avg	0.9638	0.9631	0.9631	20000
weighted avg	0.9638	0.9631	0.9631	20000

Experiment 2: *Full dataset* ***for 18041806***

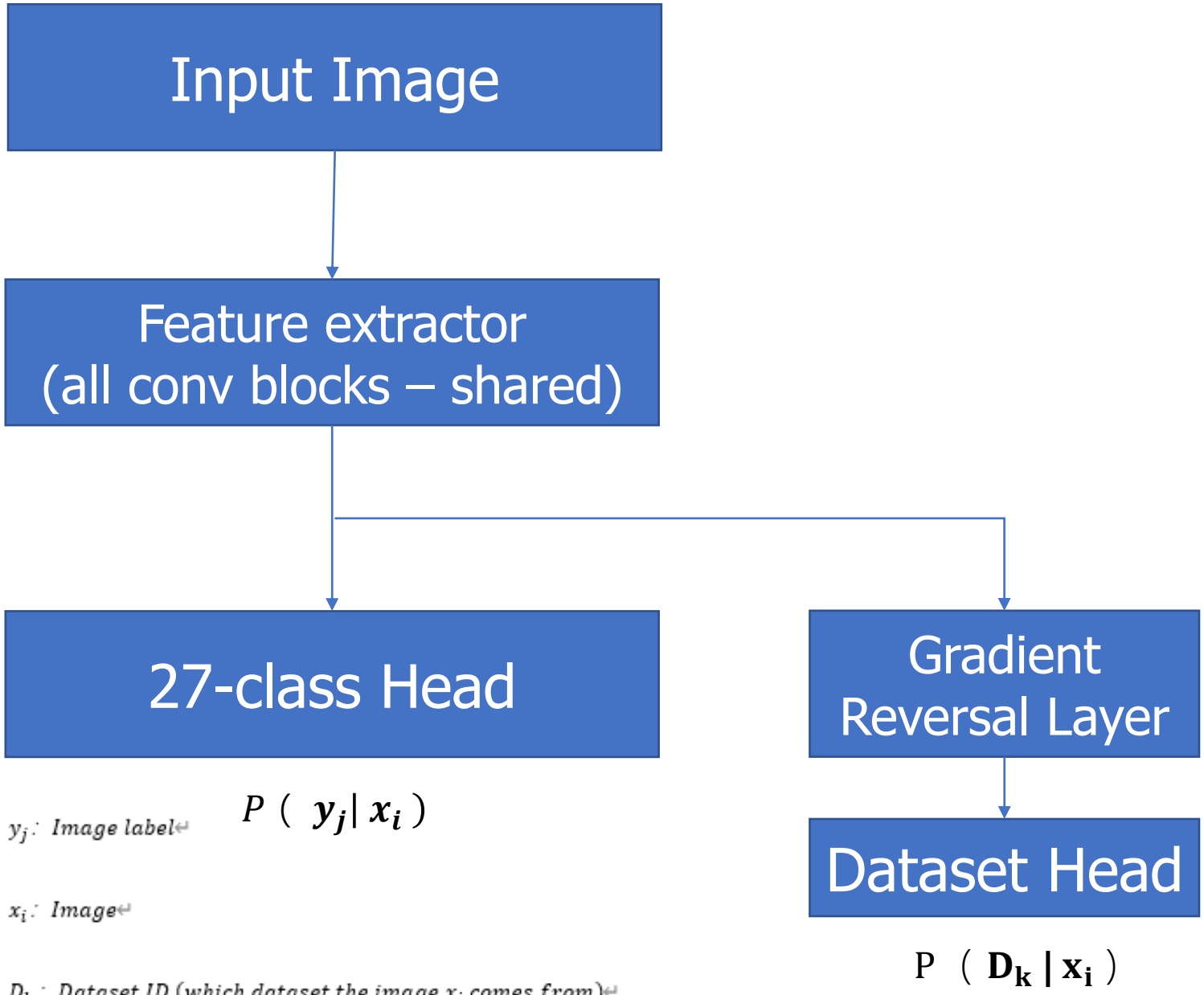


as the sample size increases, the model learns more patterns of dataset bias, causing accuracy to rise dramatically

Experiment 2: *Full dataset* ***for 18041806***



It can be observed that TINY has a higher acc than CIFAR because we downsampled Tiny to produce distinguishable artifacts, whereas CIFAR does not.



y_j: *Image label*

x_i: *Image*

D_k: *Dataset ID (which dataset the image x_i comes from)*

λ: *Coefficient indicating how much punishment imposed on the term $P\ (D_k|\ x_i\ )$*

GRL multiplies incoming gradients by $-\lambda$. During the forward pass it is the identity; during back-prop it reverses the sign, so *minimising* the dataset-head loss forces the shared features to *maximise* it.

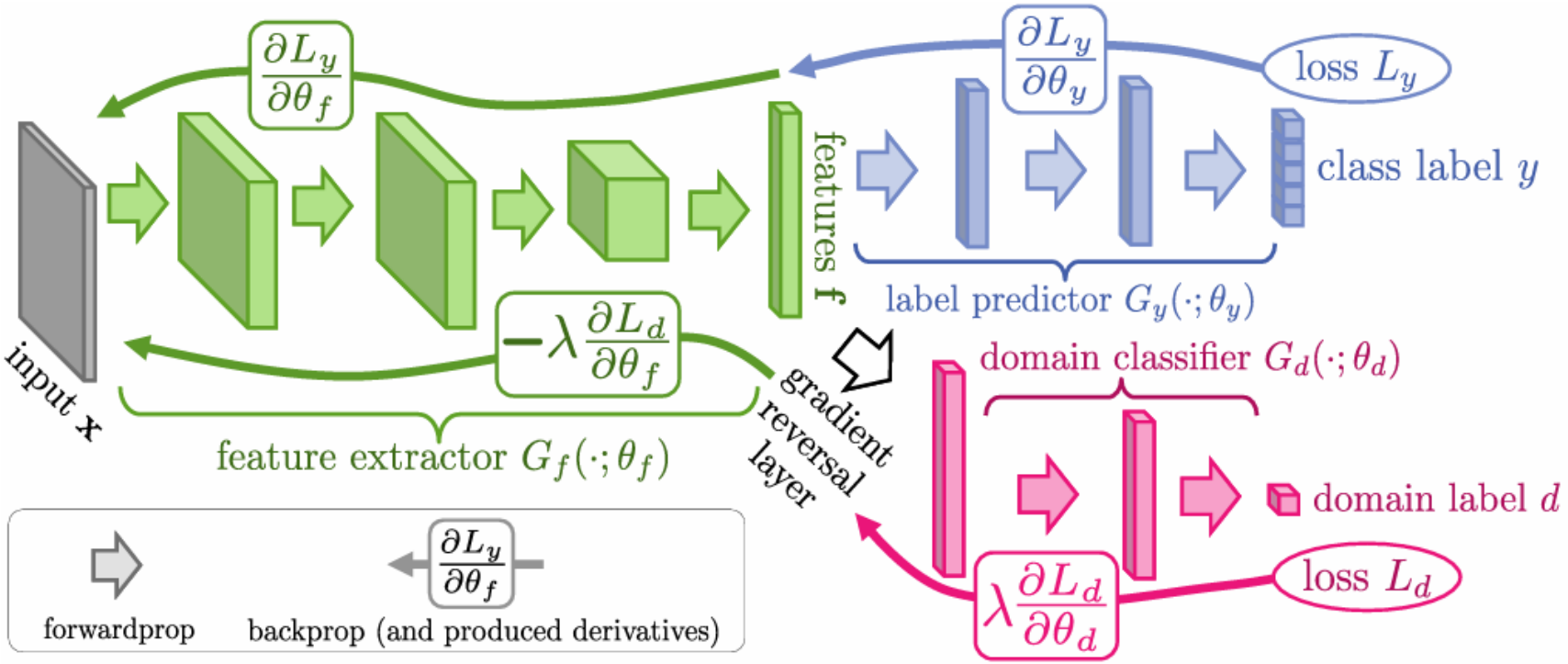


Figure 1: The **proposed architecture** includes a deep *feature extractor* (green) and a deep *label predictor* (blue), which together form a standard feed-forward architecture. Unsupervised domain adaptation is achieved by adding a *domain classifier* (red) connected to the feature extractor via a *gradient reversal layer* that multiplies the gradient by a certain negative constant during the backpropagation-based training. Otherwise, the training proceeds standardly and minimizes the label prediction loss (for source examples) and the domain classification loss (for all samples). Gradient reversal ensures that the feature distributions over the two domains are made similar (as indistinguishable as possible for the domain classifier), thus resulting in the domain-invariant features.

Val-loss as Early Stop Indicator

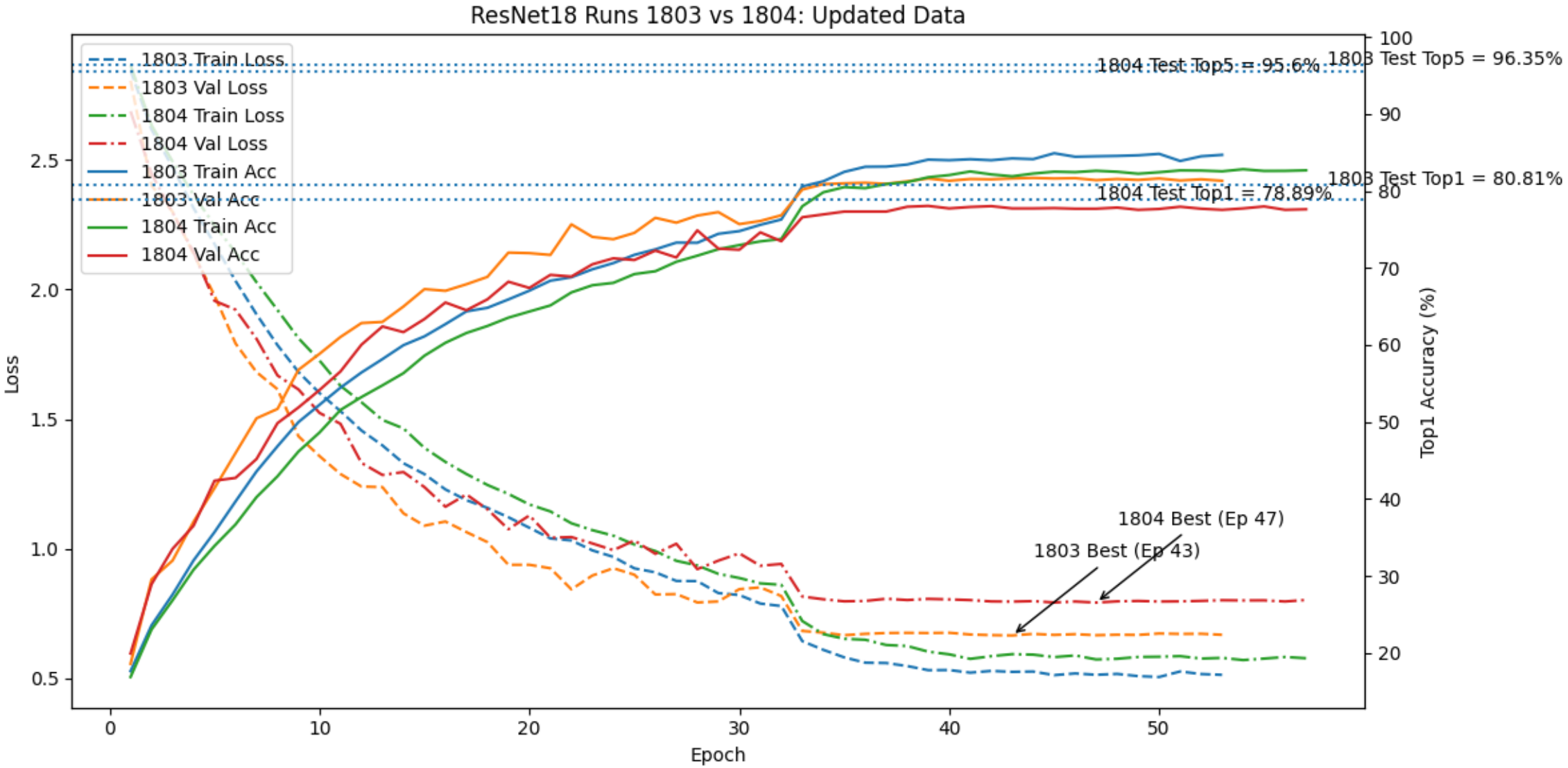
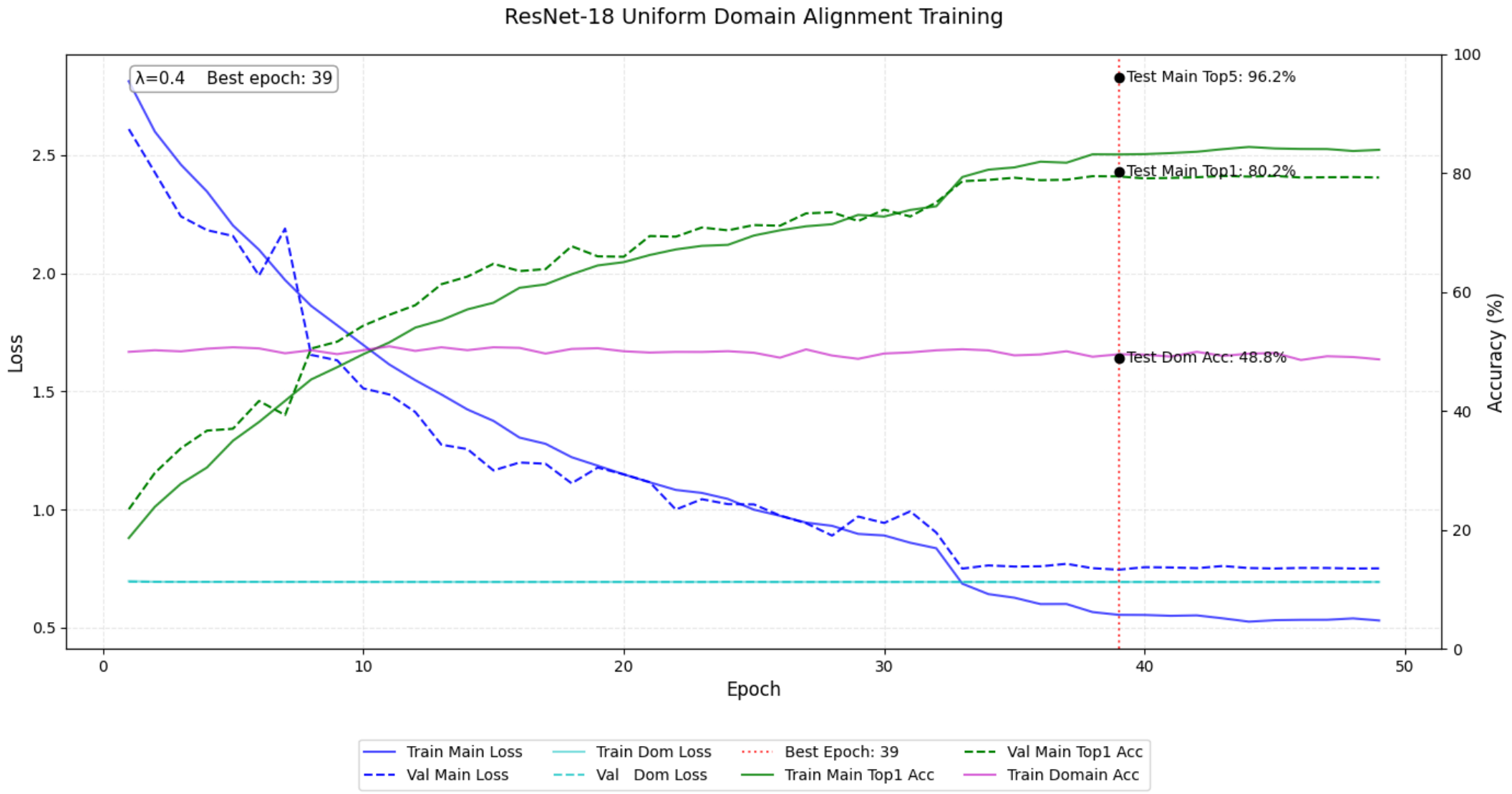


Figure for Semantic Label Classifier Training

Bias Mitigation Method : Uniform Double Min



Bias Mitigation Method : Uniform Double Min

1804D (D:Name Dataset)

[Test Confusion Matrix]

	CIFAR-100	TinyImageNet
CIFAR-100	2205	495
TinyImageNet	278	1072

[Test Classification Report]

	precision	recall	f1-score	support
CIFAR-100	0.8880	0.8167	0.8509	2700
TinyImageNet	0.6841	0.7941	0.7350	1350
accuracy			0.8091	4050
macro avg	0.7861	0.8054	0.7929	4050
weighted avg	0.8201	0.8091	0.8122	4050

1804D Uni Double Min (D:Name Dataset) lambda =0.4

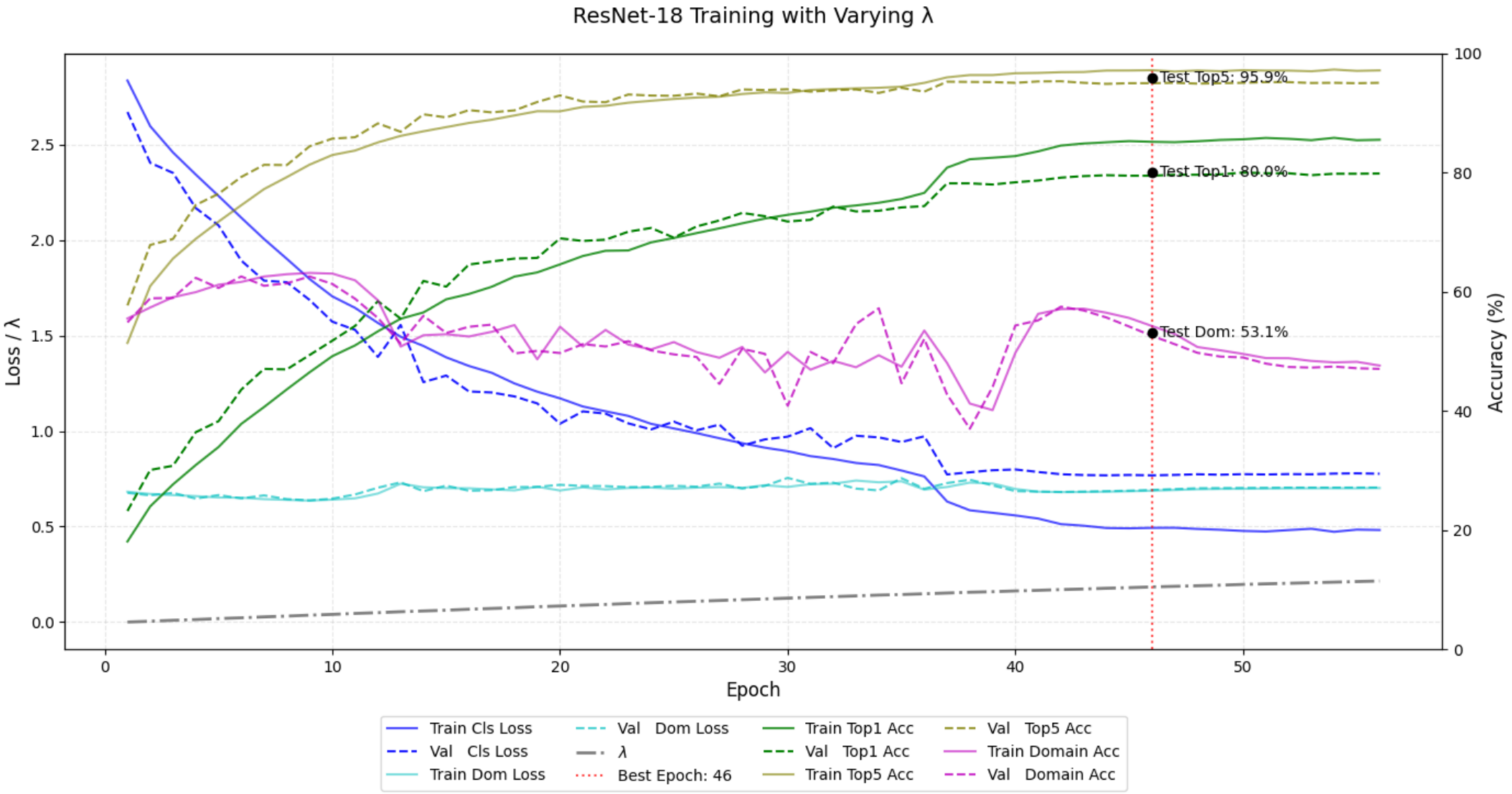
[Test Confusion Matrix]

	CIFAR-100	TinyImageNet
CIFAR-100	1754	946
TinyImageNet	484	866

[Test Classification Report]

	precision	recall	f1-score	support
CIFAR-100	0.7837	0.6496	0.7104	2700
TinyImageNet	0.4779	0.6415	0.5478	1350
accuracy			0.6469	4050
macro avg	0.6308	0.6456	0.6291	4050
weighted avg	0.6818	0.6469	0.6562	4050

Bias Mitigation Method : One-Hot Min-Max



Bias Mitigation Method : One-Hot Min-Max

1804D (D:Name Dataset)

[Test Confusion Matrix]

	CIFAR-100	TinyImageNet
CIFAR-100	2205	495
TinyImageNet	278	1072

[Test Classification Report]

	precision	recall	f1-score	support
CIFAR-100	0.8880	0.8167	0.8509	2700
TinyImageNet	0.6841	0.7941	0.7350	1350
accuracy			0.8091	4050
macro avg	0.7861	0.8054	0.7929	4050
weighted avg	0.8201	0.8091	0.8122	4050

1804D One-Hot Min-Max (D:Name Dataset)

--lambda-max=0.36

--lambda-gamma = 2

[Test Confusion Matrix]

	CIFAR-100	TinyImageNet
CIFAR-100	1743	957
TinyImageNet	445	905

[Test Classification Report]

	precision	recall	f1-score	support
CIFAR-100	0.7966	0.6456	0.7132	2700
TinyImageNet	0.4860	0.6704	0.5635	1350
accuracy			0.6538	4050
macro avg	0.6413	0.6580	0.6383	4050
weighted avg	0.6931	0.6538	0.6633	4050

Bias Mitigation Method : One-Hot Min-Max

1804D (D:Name Dataset)

[Test Confusion Matrix]

	CIFAR-100	TinyImageNet
CIFAR-100	2205	495
TinyImageNet	278	1072

[Test Classification Report]

	precision	recall	f1-score	support
CIFAR-100	0.8880	0.8167	0.8509	2700
TinyImageNet	0.6841	0.7941	0.7350	1350
accuracy			0.8091	4050
macro avg	0.7861	0.8054	0.7929	4050
weighted avg	0.8201	0.8091	0.8122	4050

1804D One-Hot Min-Max (D:Name Dataset)

--lambda-max=0.36

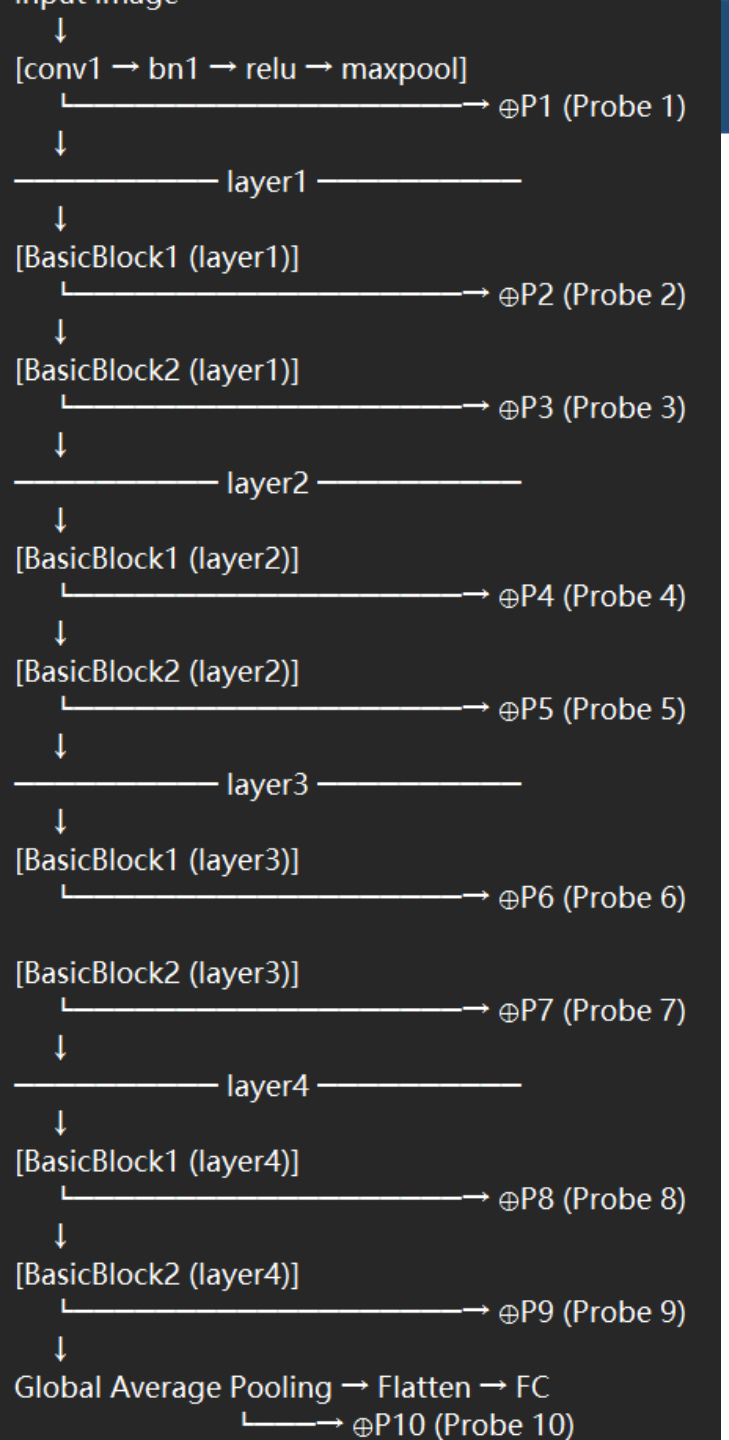
--lambda-gamma = 2

[Test Confusion Matrix]

	CIFAR-100	TinyImageNet
CIFAR-100	1743	957
TinyImageNet	445	905

[Test Classification Report]

	precision	recall	f1-score	support
CIFAR-100	0.7966	0.6456	0.7132	2700
TinyImageNet	0.4860	0.6704	0.5635	1350
accuracy			0.6538	4050
macro avg	0.6413	0.6580	0.6383	4050
weighted avg	0.6931	0.6538	0.6633	4050



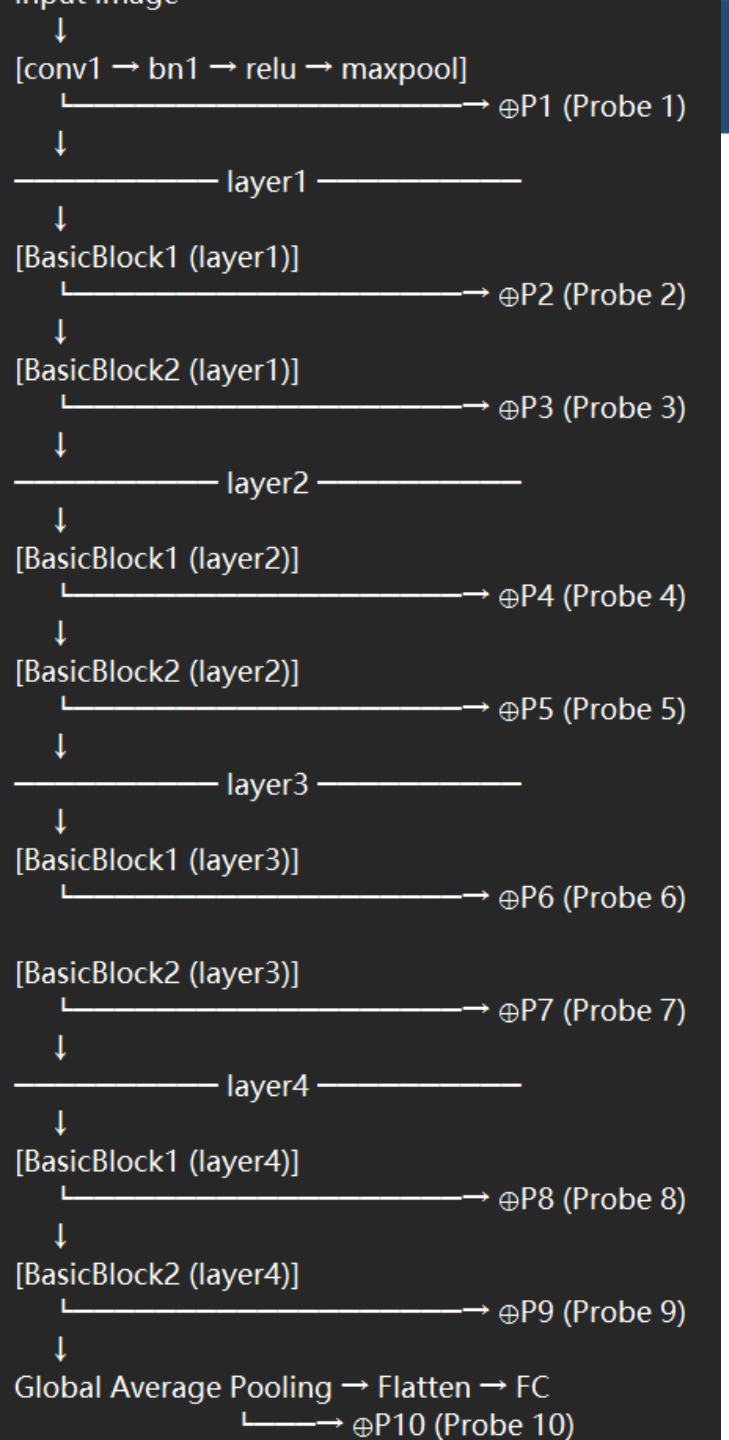
Experiment 1: Which layer of ResNet-18 is more sensitive to dataset bias?

10 Linear Probe Task Workflow for ResNet1802, ResNet1803, ResNet1804 (18Layer Dataset02)

I. Same Hyperparameters for ResNet1802,3,4 Training

II. Train until find the best（ Highest Val Accuracy ）,
for All models(1802,3,4), All Probes(1802,3,4;0-9)

III. Different Dataset(02,03,04)



Experiment 1: Which layer of ResNet-18 is more sensitive to dataset bias?

Hyperparameter for 10 Linear Probe Training

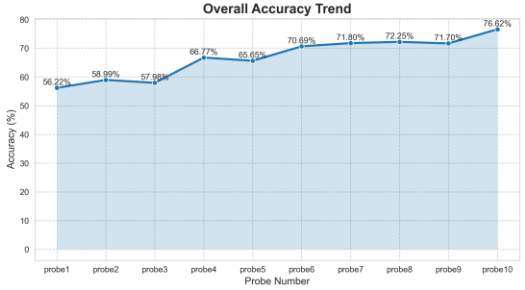
epochs	1
batch size	64
num_workers	4
early stopping patience	5 (--early-stop)
optimizer	Adam (optim.Adam(model.cls.parameters(), lr=1e-3))
initial learning rate (lr)	1e-3
LR scheduler	ReduceLROnPlateau (mode='max', patience=2, factor=0.1)
loss function	CrossEntropyLoss
input size	224×224
train data augmentation	RandomResizedCrop(224) + RandomHorizontalFlip()
val/test preprocessing	CenterCrop(224)
normalization mean/std	mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]
freeze backbone parameters	Yes (requires_grad=False for all backbone parameters)
probe tap points	10 taps: maxpool, layer1.0, layer1.1, layer2.0, layer2.1, layer3.0, layer3.1, layer4.0, layer4.1, avgpool

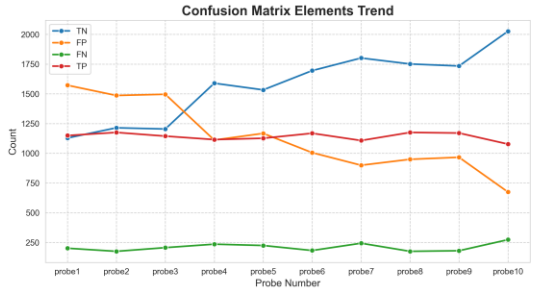
Semantic Information & Non-Semantic Information

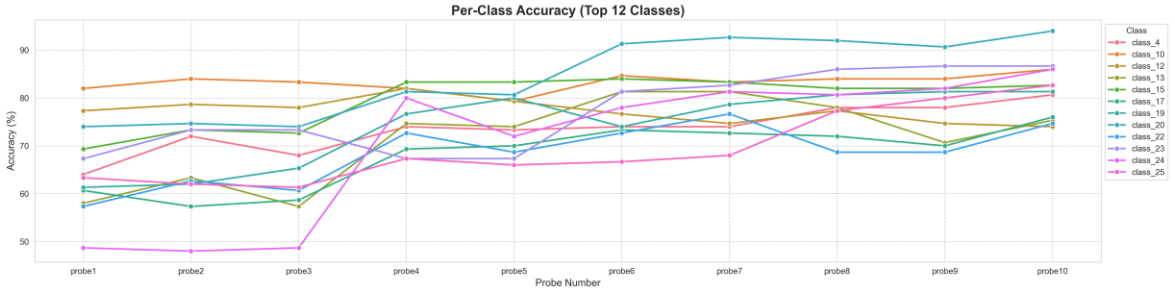
	Semantic Information	Non-Semantic Information
Definition	Carries meaning/concepts ("what it is")	Raw signal/attributes without direct meaning
Level	High-level abstraction	Low-level, detailed
Purpose	Conveys categories, labels, intent	Basis for feature extraction
Examples	"Cat" label in vision; word meaning in NLP; "Hello" text	RGB pixel values; audio waveform; embedding vector
Distribution	Dominant in higher layers but present (to a lesser extent) throughout the network	Dominant in lower layers but residuals persist up to the top layers

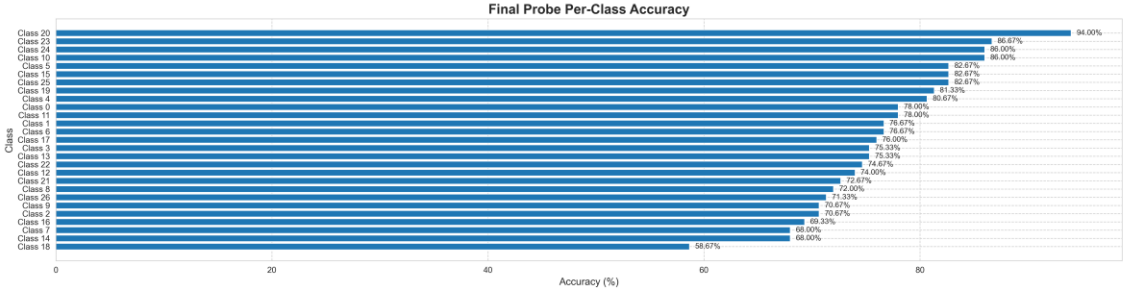
Deep models gradually transform non-semantic inputs into semantic concepts—yet every layer may still retain useful low-level detail. (Is it true? I Guess)

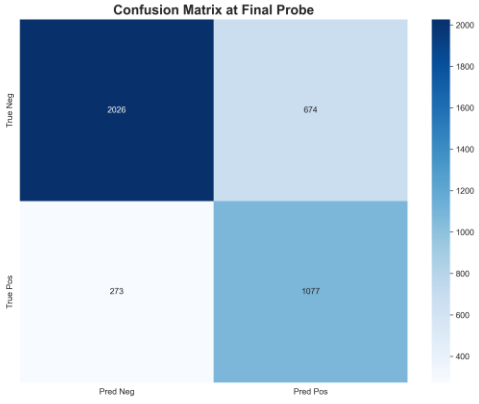
ResNet-1802 10 Probe Performance Analysis

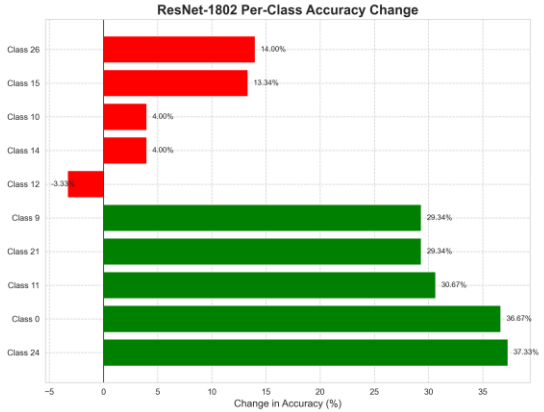




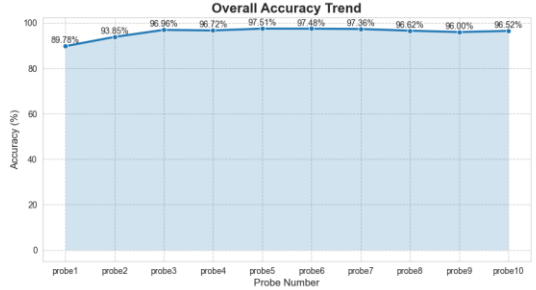


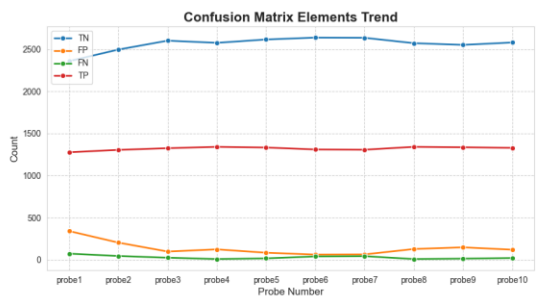


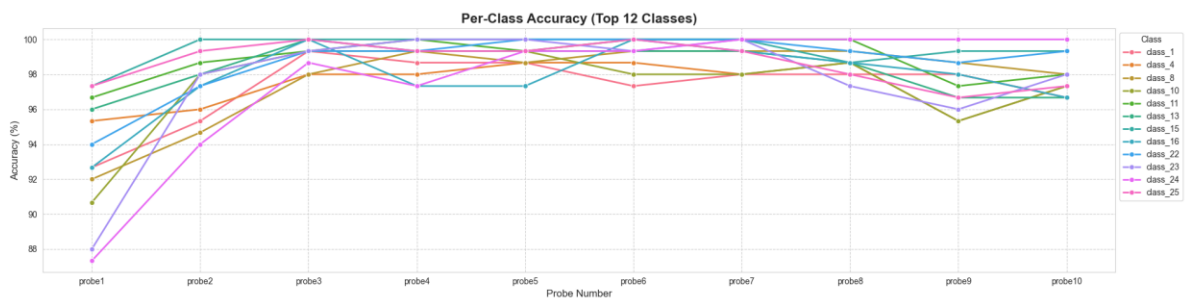


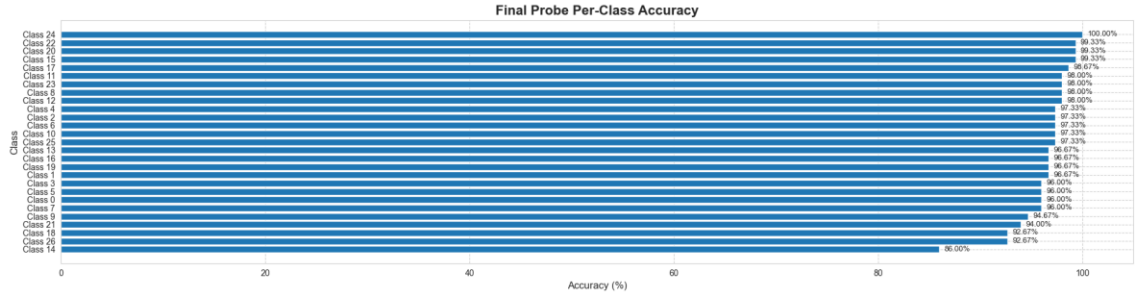


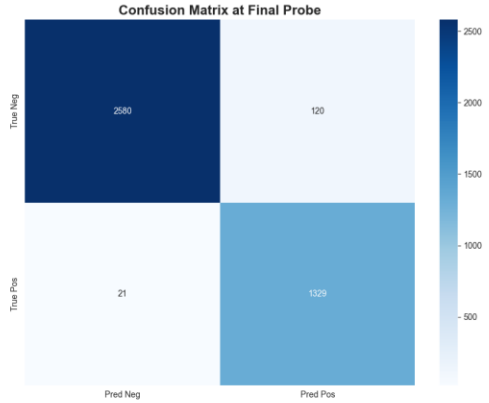
ResNet-1803 10 Probe Performance Analysis

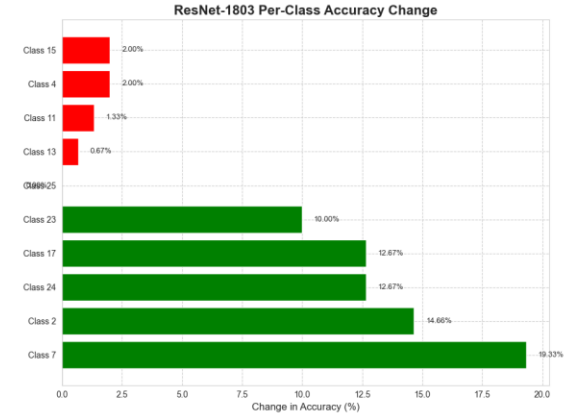












Semantic Information & Non-Semantic Information

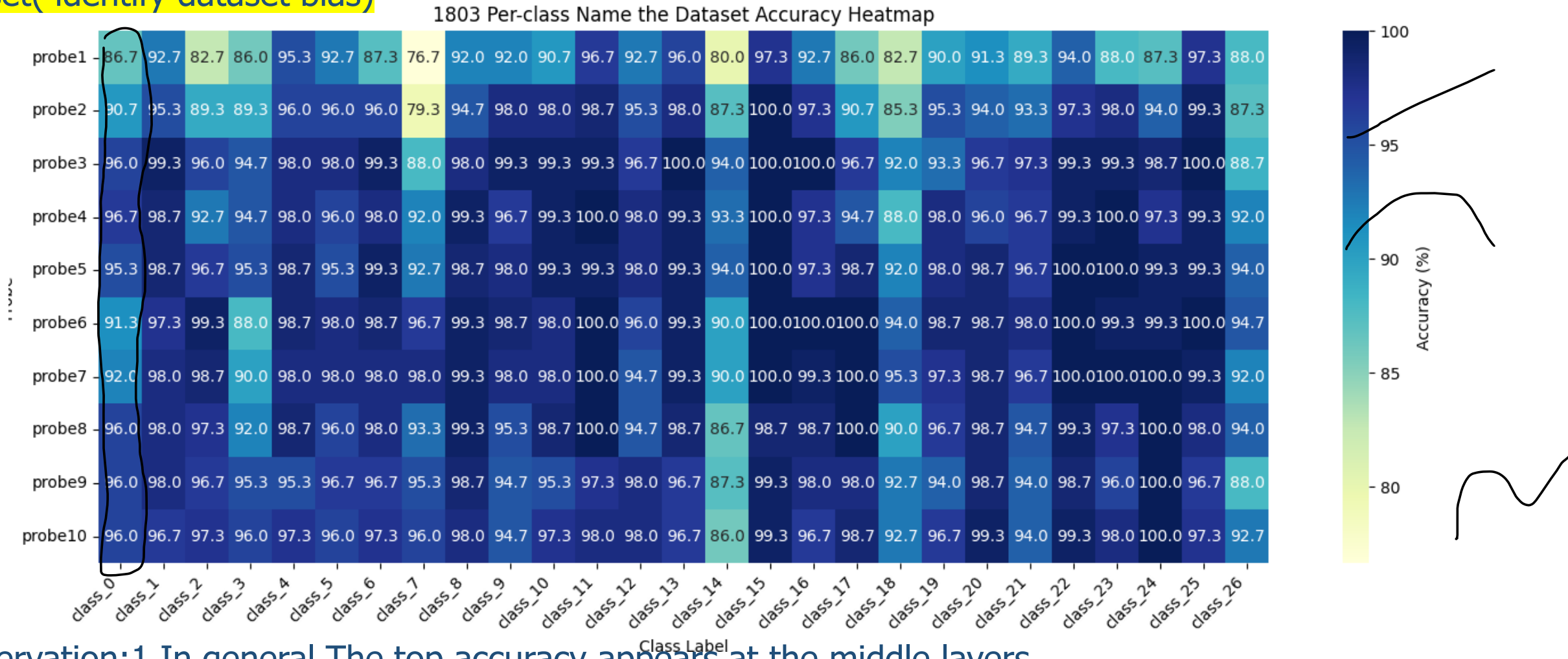
- Can we see dataset bias as a combination of Semantic info and non-semantic info?

- For different Probes (0-9), different percentage of reliance on Semantic info and non-Semantic info for the model

- E.g.
- Probe 1 10% seman 90% non-seman
- Probe 1 20% seman 80% non-seman

Assumption: Low layer(more non-semantic info) and high layer (more semantic info) contain semantic and non semantic information simultaneously

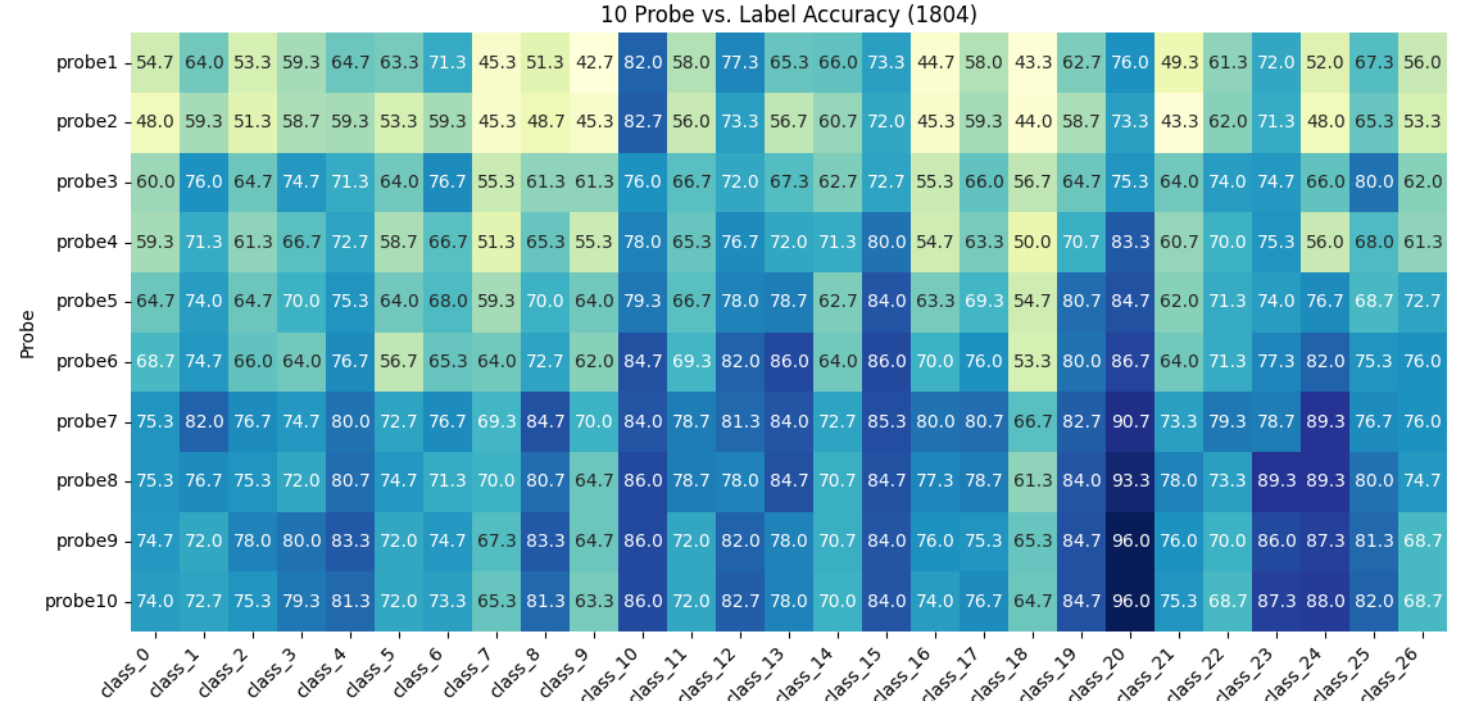
Conclusion: the model use semantic and non semantic information simultaneously to name the dataset(identify dataset bias)

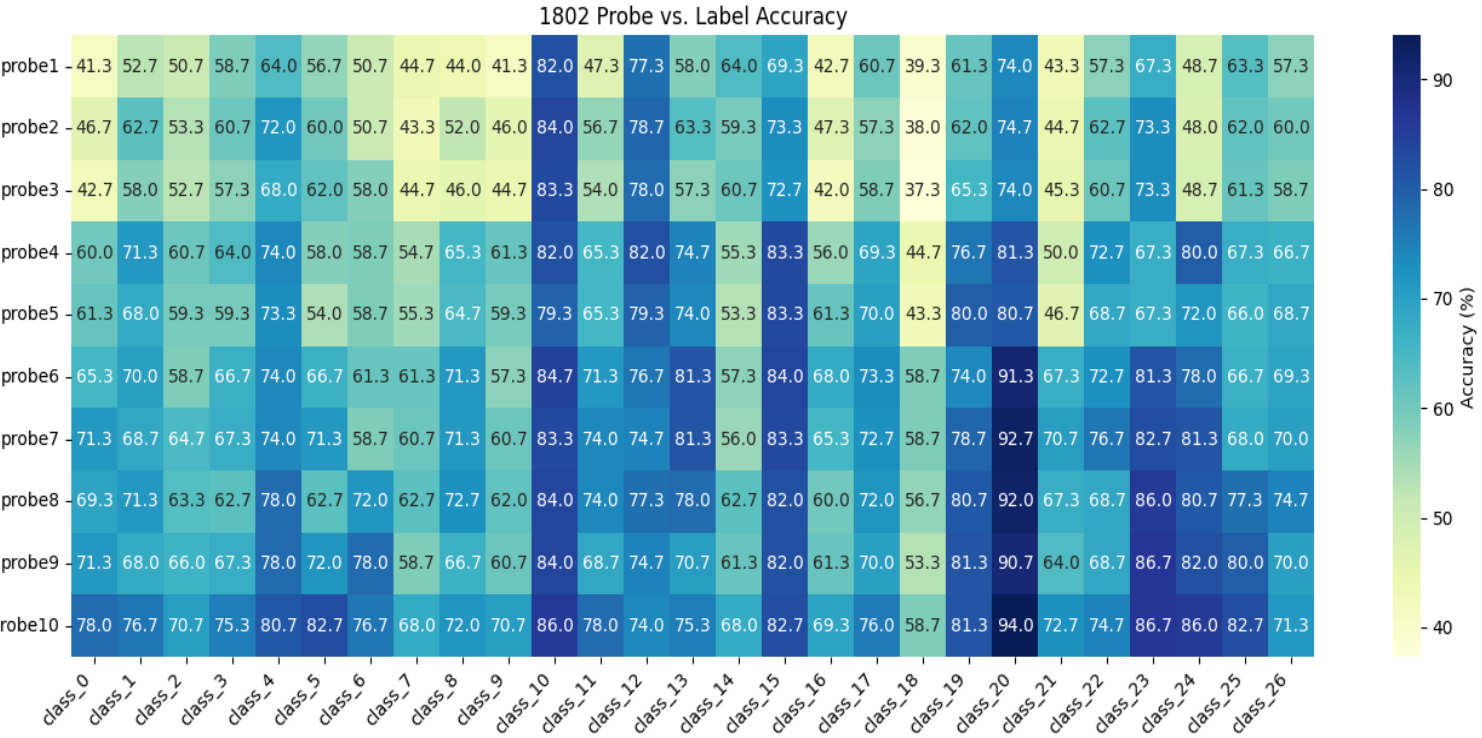


Observation:1.In general The top accuracy appears at the middle layers

2. Accuracy fluctuates

3. Accuracy goes up and then drops down slightly





[TEST] Loading best checkpoint and evaluating on test split…

[Test Confusion Matrix]1802

	pred CIFAR-100	pred TinyImageNet
True CIFAR-100	1795	905
True TinyImageNet	193	1157

The model likes to mistake **CIFAR-100** as **TinyImageNet**

[Test Classification Report]

	precision	recall	f1-score	support
CIFAR-100	0.9029	0.6648	0.7658	2700
TinyImageNet	0.5611	0.8570	0.6782	1350
accuracy			0.7289	4050
macro avg	0.7320	0.7609	0.7220	4050
weighted avg	0.7890	0.7289	0.7366	4050

→ why?

DownSampling (TinyImageNet) seems to make the dataset specific feature much harder for the model to capture

Dataset04, Bicubic

Tiny64×64 → 32×32 (BICUBIC 256×256 (BICUBIC) → 224×224 (Random Crop)

CIFAR32×32 → 256×256 (BICUBIC) → 224×224 (Random Crop)

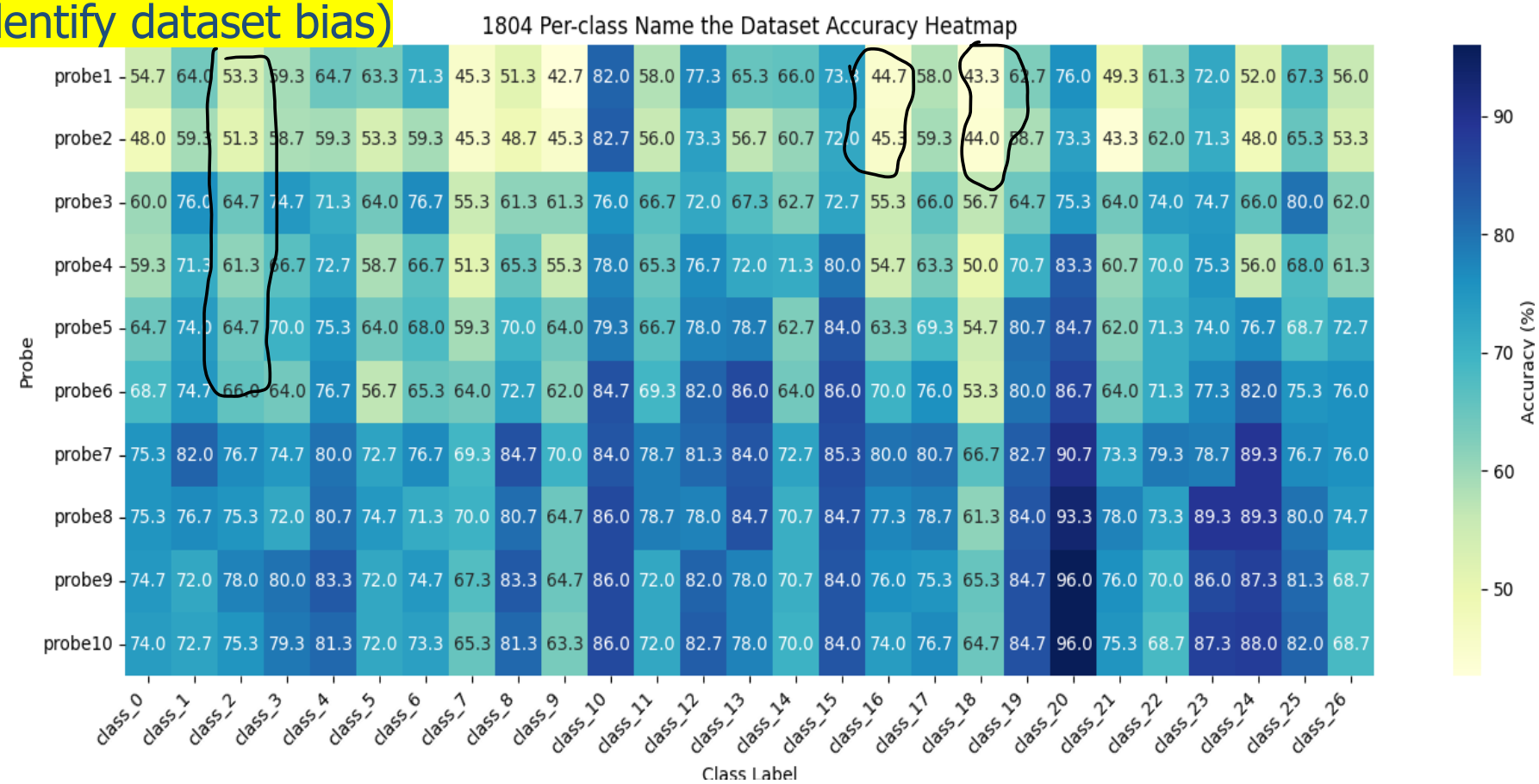
Dataset02, Bicubic

Tiny64×64 → 32×32 (BICUBIC) → 128×128 (BICUBIC)

CIFAR32×32 → 128×128 (BICUBIC)

Assumption: Low layer(more non-semantic info) and high layer (more semantic info) contain semantic and non semantic information simultaneously

Conclusion: the model use semantic and non semantic information simultaneously to name the dataset(identify dataset bias)

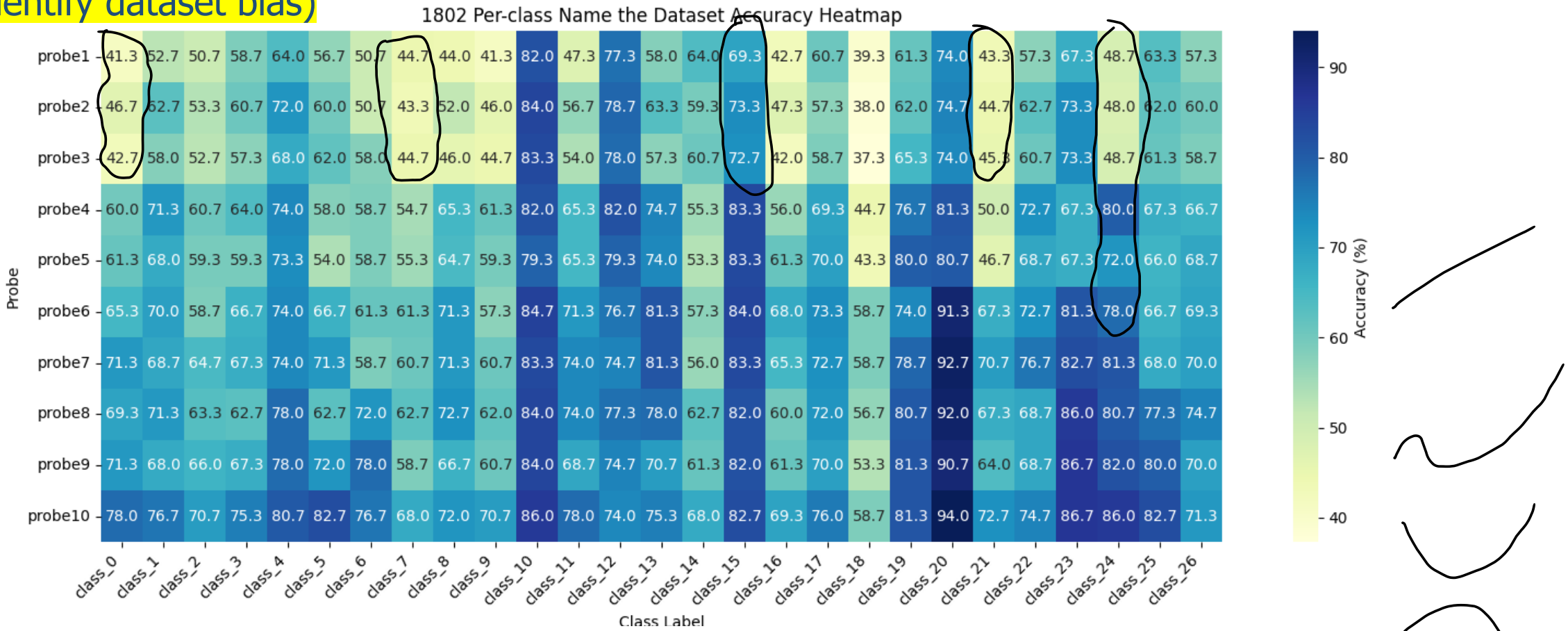


Observation:1.In general, the top accuracy appears at the middle layers

2. Accuracy fluctuates
3. Accuracy goes up and then drops down slightly
4. Accuracy BELOW 50%

Assumption: Low layer(more non-semantic info) and high layer (more semantic info) contain semantic and non semantic information simultaneously

Conclusion: the model use semantic and non semantic information simultaneously to name the dataset(identify dataset bias)



Observation:1.In general, the top accuracy appears at the middle layers

2. Accuracy fluctuates
3. Accuracy goes up and then drops down slightly
4. Accuracy BELOW 50%

Intro/ Background/ Story 01

As machine learning advances at breakneck speed, our deep neural networks (DNNs) are constantly evolving to achieve stronger generalization—a core challenge in the field that directly determines a model's performance in real-world scenarios. Much like the age-old "nature versus nurture" debate in biology (are our behaviors driven by innate genetics or shaped by environment?), a model's outputs hinge on two complementary factors: architecture and training. If the architecture is underpowered, even the highest-quality data won't reveal complex patterns; conversely, if the training data are sparse or noisy, even the most sophisticated architecture will be held back.

Yet today, research overwhelmingly emphasizes novel network architectures and training methodologies (e.g., supervised and self-supervised learning), while systematic examination of datasets themselves remains surprisingly rare. However, as the field has advanced, an inherent problem has been brought to light: **dataset bias**—the systematic distribution discrepancies between different datasets, or between a dataset and real-world distributions—that, much like the nature versus nurture debate, profoundly influences a model's performance.

Semantic Information & Non-Semantic Information

	Semantic Information	Non-Semantic Information
Definition	Carries meaning/concepts ("what it is")	Raw signal/attributes without direct meaning
Level	High-level abstraction	Low-level, detailed
Purpose	Conveys categories, labels, intent	Basis for feature extraction
Examples	"Cat" label in vision; word meaning in NLP; "Hello" text	RGB pixel values; audio waveform; embedding vector
Distribution	Dominant in higher layers but present (to a lesser extent) throughout the network	Dominant in lower layers but residuals persist up to the top layers

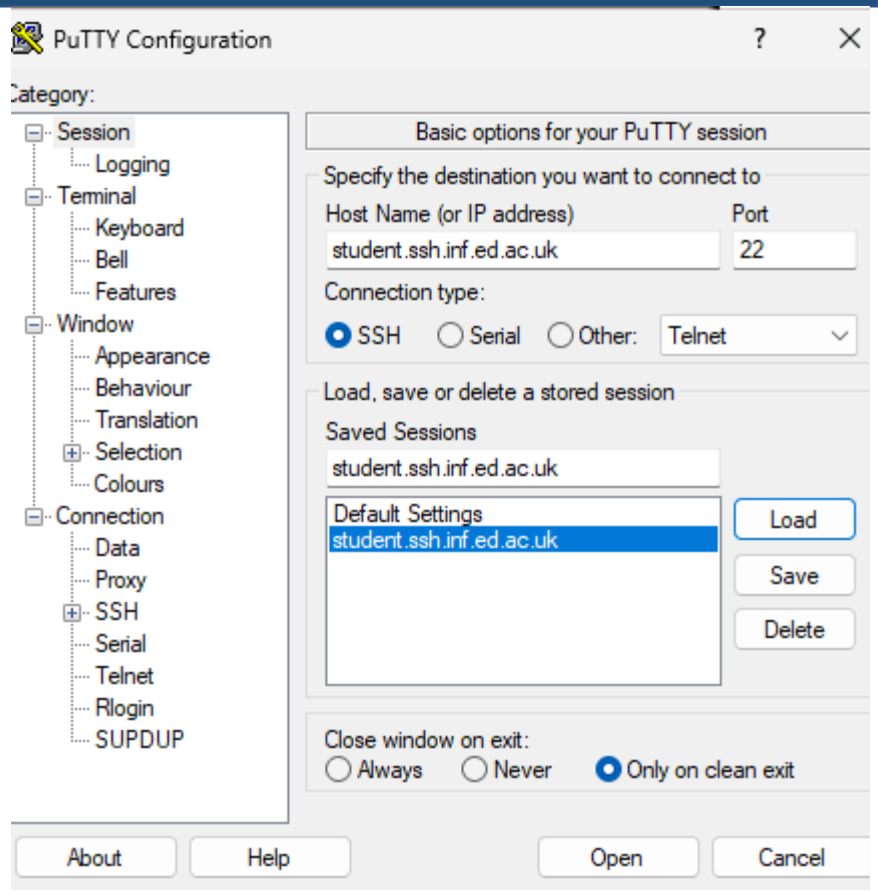
Deep models gradually transform non-semantic inputs into semantic concepts—yet every layer may still retain useful low-level detail.

Do we hope the model to learn semantic information only?

Is training a process to maximize semantic info and minimize non-semantic info?

Can we write a loss function to maximize semantic info and minimize non-semantic info?

Shannon



Must USE Anaconda ?

Next Steps?

1. Other models(e.g. ViT? Bigger model)?(bigger model needs to use shannon)
2. Other Dataset? (Without using up sampling and downsampling, use randomcrop/centercrop to achieve the same size)
3. Design the experiment to prove :

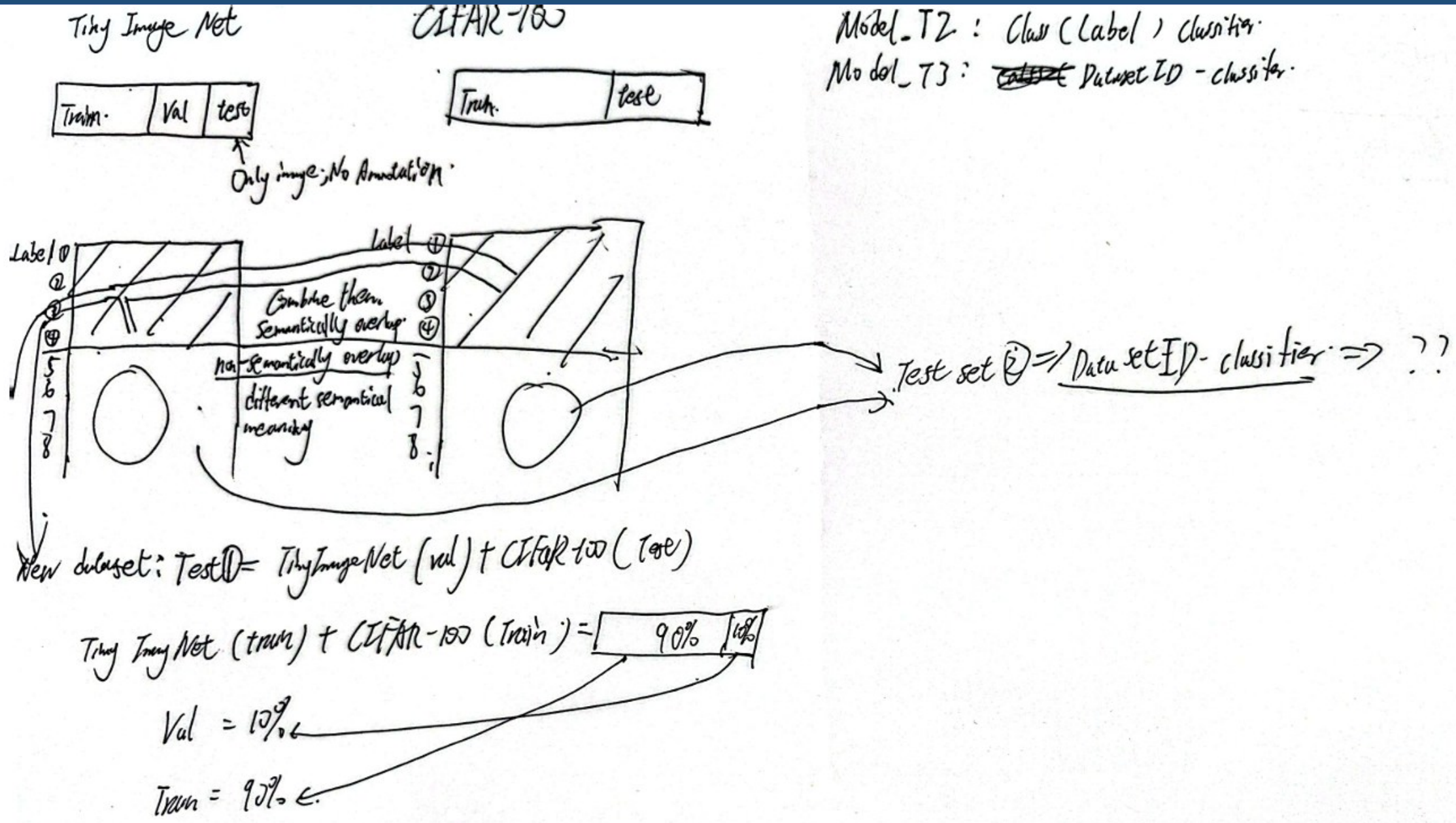
Guess: Dataset Bias captured by DNN is, in essence, a kind of **distribution difference** (pseudo-dataset)

4. Design other interesting experiments?

Or?

Start Writing and Publish as soon as possible?

Experiment 2: *Dataset bias – Non-semantic Information Test*



ResNet1803

probe	overall_cm_00	overall_cm_01	overall_cm	overall_cm_11	overall_acc	class_0	class_1	class_2	class_3	class_4	class_5	class_6	class_7	class_8	class_9	class_10	class_11	class_12	class_13	class_14
probe1	2359	341	73	1277	89.78	86.67	92.67	82.67	86	95.33	92.67	87.33	76.67	92	92	90.67	96.67	92.67	96	80
probe2	2496	204	45	1305	93.85	90.67	95.33	89.33	89.33	96	96	96	79.33	94.67	98	98	98.67	95.33	98	87.33
probe3	2602	98	25	1325	96.96	96	99.33	96	94.67	98	98	99.33	88	98	99.33	99.33	99.33	96.67	100	94
probe4	2576	124	9	1341	96.72	96.67	98.67	92.67	94.67	98	96	98	92	99.33	96.67	99.33	100	98	99.33	93.33
probe5	2616	84	17	1333	97.51	95.33	98.67	96.67	95.33	98.67	95.33	99.33	92.67	98.67	98	99.33	99.33	98	99.33	94
probe6	2638	62	40	1310	97.48	91.33	97.33	99.33	88	98.67	98	98.67	96.67	99.33	98.67	98	100	96	99.33	90
probe7	2636	64	43	1307	97.36	92	98	98.67	90	98	98	98	98	99.33	98	98	100	94.67	99.33	90
probe8	2572	128	9	1341	96.62	96	98	97.33	92	98.67	96	98	93.33	99.33	95.33	98.67	100	94.67	98.67	86.67
probe9	2552	148	14	1336	96	96	98	96.67	95.33	95.33	96.67	96.67	95.33	98.67	94.67	95.33	97.33	98	96.67	87.33
probe10	2580	120	21	1329	96.52	96	96.67	97.33	96	97.33	96	97.33	96	98	94.67	97.33	98	98	96.67	86

class_15	class_16	class_17	class_18	class_19	class_20	class_21	class_22	class_23	class_24	class_25	class_26
97.33	92.67	86	82.67	90	91.33	89.33	94	88	87.33	97.33	88
100	97.33	90.67	85.33	95.33	94	93.33	97.33	98	94	99.33	87.33
100	100	96.67	92	93.33	96.67	97.33	99.33	99.33	98.67	100	88.67
100	97.33	94.67	88	98	96	96.67	99.33	100	97.33	99.33	92
100	97.33	98.67	92	98	98.67	96.67	100	100	99.33	99.33	94
100	100	100	94	98.67	98.67	98	100	99.33	99.33	100	94.67
100	99.33	100	95.33	97.33	98.67	96.67	100	100	100	99.33	92
98.67	98.67	100	90	96.67	98.67	94.67	99.33	97.33	100	98	94
99.33	98	98	92.67	94	98.67	94	98.67	96	100	96.67	88
99.33	96.67	98.67	92.67	96.67	99.33	94	99.33	98	100	97.33	92.67

ResNet1802

probe	overall_cm_00	overall_cm_01	overall_cm	overall_cm	overall_acc	class_0	class_1	class_2	class_3	class_4	class_5	class_6	class_7	class_8	class_9	class_10	class_11	class_12	class_13	class_14
probe1	1128	1572	201	1149	56.22	41.33	52.67	50.67	58.67	64	56.67	50.67	44.67	44	41.33	82	47.33	77.33	58	64
probe2	1214	1486	175	1175	58.99	46.67	62.67	53.33	60.67	72	60	50.67	43.33	52	46	84	56.67	78.67	63.33	59.33
probe3	1204	1496	206	1144	57.98	42.67	58	52.67	57.33	68	62	58	44.67	46	44.67	83.33	54	78	57.33	60.67
probe4	1589	1111	235	1115	66.77	60	71.33	60.67	64	74	58	58.67	54.67	65.33	61.33	82	65.33	82	74.67	55.33
probe5	1533	1167	224	1126	65.65	61.33	68	59.33	59.33	73.33	54	58.67	55.33	64.67	59.33	79.33	65.33	79.33	74	53.33
probe6	1695	1005	182	1168	70.69	65.33	70	58.67	66.67	74	66.67	61.33	61.33	71.33	57.33	84.67	71.33	76.67	81.33	57.33
probe7	1801	899	243	1107	71.8	71.33	68.67	64.67	67.33	74	71.33	58.67	60.67	71.33	60.67	83.33	74	74.67	81.33	56
probe8	1751	949	175	1175	72.25	69.33	71.33	63.33	62.67	78	62.67	72	62.67	72.67	62	84	74	77.33	78	62.67
probe9	1734	966	180	1170	71.7	71.33	68	66	67.33	78	72	78	58.67	66.67	60.67	84	68.67	74.67	70.67	61.33
probe10	2026	674	273	1077	76.62	78	76.67	70.67	75.33	80.67	82.67	76.67	68	72	70.67	86	78	74	75.33	68

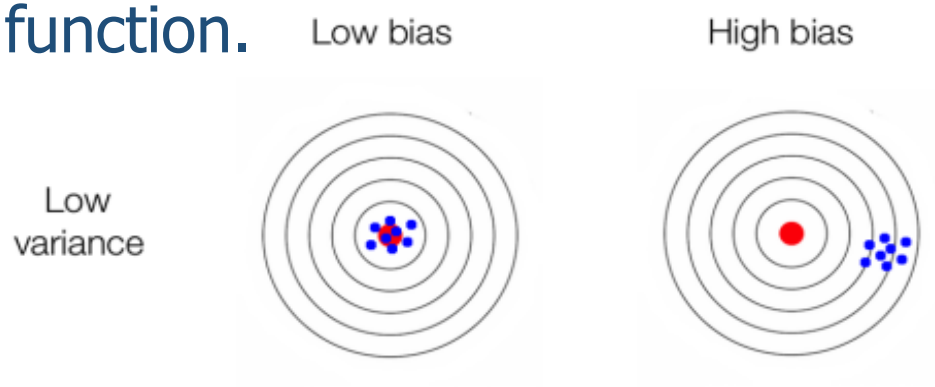
class_15	class_16	class_17	class_18	class_19	class_20	class_21	class_22	class_23	class_24	class_25	class_26
69.33	42.67	60.67	39.33	61.33	74	43.33	57.33	67.33	48.67	63.33	57.33
73.33	47.33	57.33	38	62	74.67	44.67	62.67	73.33	48	62	60
72.67	42	58.67	37.33	65.33	74	45.33	60.67	73.33	48.67	61.33	58.67
83.33	56	69.33	44.67	76.67	81.33	50	72.67	67.33	80	67.33	66.67
83.33	61.33	70	43.33	80	80.67	46.67	68.67	67.33	72	66	68.67
84	68	73.33	58.67	74	91.33	67.33	72.67	81.33	78	66.67	69.33
83.33	65.33	72.67	58.67	78.67	92.67	70.67	76.67	82.67	81.33	68	70
82	60	72	56.67	80.67	92	67.33	68.67	86	80.67	77.33	74.67
82	61.33	70	53.33	81.33	90.67	64	68.67	86.67	82	80	70
82.67	69.33	76	58.67	81.33	94	72.67	74.67	86.67	86	82.67	71.33

Definition of BIAS ———(Adopted from PGEE11164, Probabilty)

- In statistics and machine learning, **bias** typically refers to the **systematic deviation** between the expected value of a model's output (prediction) and the true value.
- In other words, bias measures how far the model's average prediction deviates from the actual target value — it represents a form of systematic error. Mathematically, the bias of a model for a given input x can be expressed as:

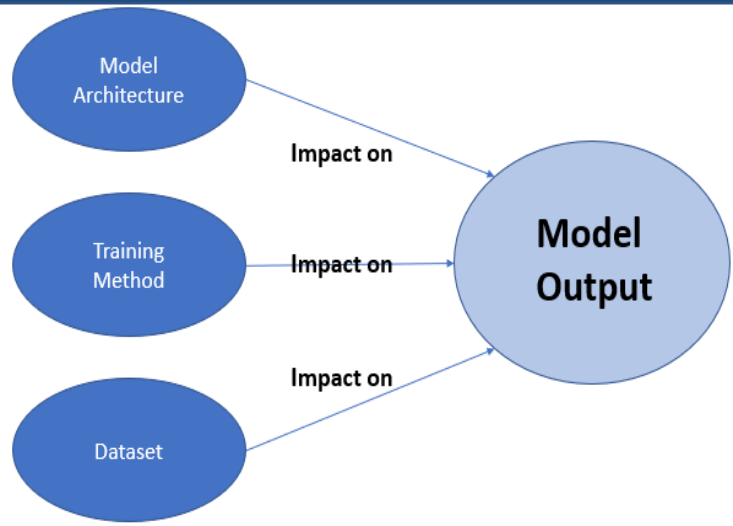
$$\text{Bias}(x) = E\left[\hat{f}(x)\right] - f(x)$$

- where $E\left[\hat{f}(x)\right]$ denotes the expected prediction of the model over different training datasets, and $f(x)$ is the true underlying function.

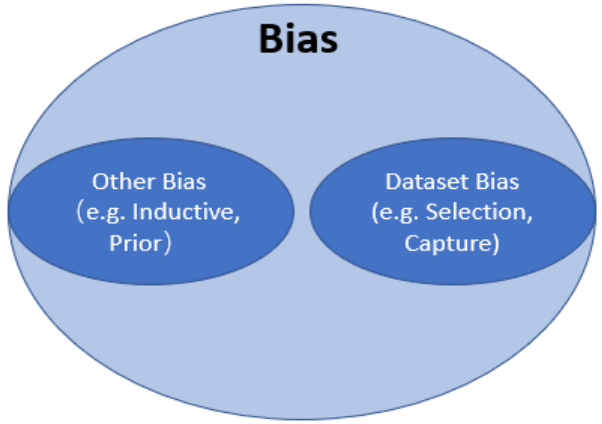


Definition of Dataset BIAS ——*(Discussed with gpt o3)*

- In practice, a **model's output** and its prediction error are jointly determined by the model architecture, training method, and dataset.



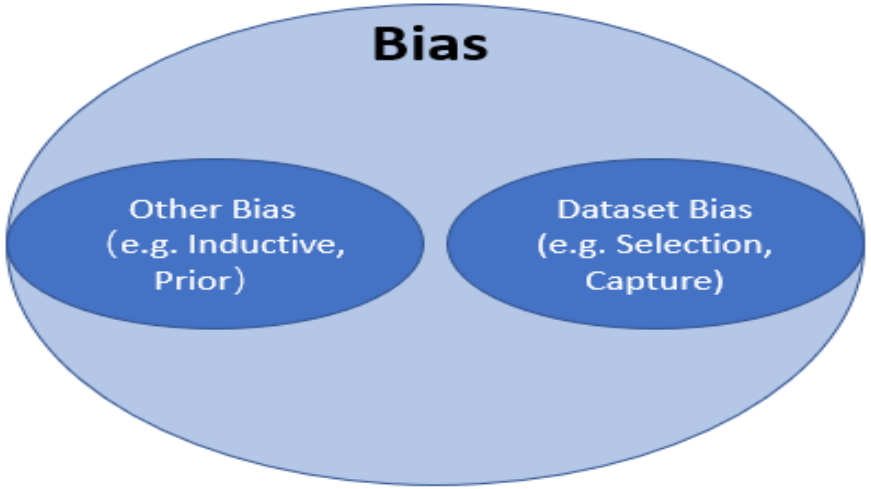
- **Dataset bias** refers specifically to distributional differences, which is **currently hard to identify and predict**, between the dataset and the real world (or between datasets) caused by systematic—rather than random—factors during data collection, filtering, annotation, or pre-processing (whether technical, content-based, or social). Models can hardly detect intra-dataset bias —— pseudo-dataset experiment from 2A Decade's Battle on Dataset Bias
- **Such bias** induces systematic errors (i.e., bias irrespective of variance) between the **model's outputs** and the expected targets, typically manifesting as poor generalization performance.

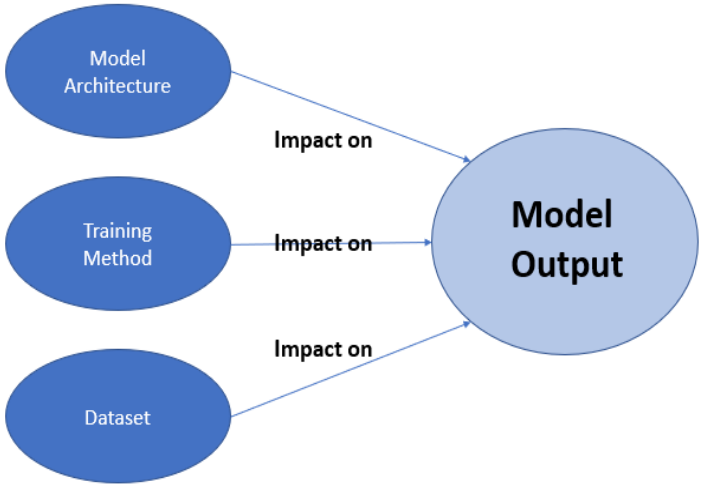


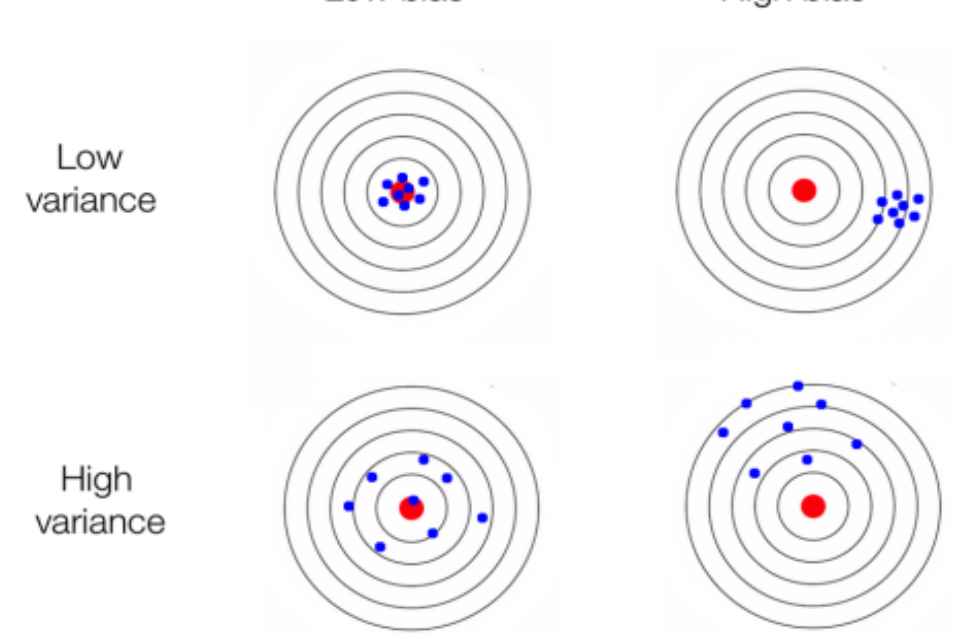
———(Discussed with gpt o3)

Relationship between bias and dataset bias

- In summary, **dataset bias is the concretization of the concept of bias at the data level**, and both exert the same directional effect on a model's generalization performance: any systematic imbalance or deviation at the data level (dataset bias) will ultimately manifest as systematic prediction error (induced by bias) in the model.
- Dataset bias is part of total bias.







Problem

- **Core Problem**: Dataset bias arises when systematic (by design,non-random) factors during data collection, filtering, annotation, or preprocessing cause **deviation of data distribution** between a dataset's distribution and the real world (or between datasets).
- As a result:

1. Models **generalize poorly** to unseen scenarios (e.g. low light, rare classes, cross-domain), making systematic errors. ——from *Unbiased_look & A Decade's Battle*
2. Simply enlarging or merging datasets often **fails to eliminate bias** and may even **amplify** it. ——from *A survey on bias in visual datasets & short cut learning*
3. There is **no unbiased (benchmark) dataset**, and it is hard to identify or predict dataset bias, as some of types of the bias are difficult for human to perceive ——from *A survey on bias in visual datasets*

Aim & Objectives

Aim: To **improve the model's generalization performance** by constructing a better (potentially less biased) **dataset**, or/and by using specific model **architectures** or/and **training methods** to suppress dataset bias.

Objectives:

1. **Literature Review on Dataset Bias and Mitigation Strategies;**
2. **Dataset Familiarisation and Their Analysis:**
   – Load, pre-process and augment popular image datasets.
3. **Model Implementation**
   – Build and train various deep architectures (CNNs, ViTs, etc.).
4. **Study of Learning Paradigms:**

- Through transfer learning, generative models (e.g., GANs / VAEs) and other training setups, deeply observe how dataset bias emerges, is amplified, or is mitigated in practice.

5. **(If Possible) Bias Detection & Quantification**
   1. Analyze internal biases (class imbalance, annotation rules) and external biases (domain shifts) in datasets like ImageNet, PASCAL VOC and Cityscapes.
   2. Develop interpretable metrics (feature-space clustering scores, cross-dataset accuracy, information theory) to quantify bias.
6. **(If Possible) Bias Impact Analysis**
   1. Run controlled experiments to demonstrate how dataset bias causes large performance drops in specific scenarios (e.g. low-light, rare classes).
   2. Evaluate bias-correction methods (domain adaptation, GAN-based augmentation, synthetic data).

Paper Read So far

	Title	Authors	Year
2	Unbiased Look at Dataset Bias	Antonio Torralba, Alexei A. Efros	2011
3	A Decade's Battle on Dataset Bias: Are We There Yet?	Zhuang Liu, Kaiming He	2024
4	Bi-Mix: Bidirectional Mixing for Domain Adaptation Between Day and Night	Z. Yang, J. Tang, J. Xu	2021
5	A Deeper Look at Dataset Bias	T. Tommasi, N. Patricia, B. Caputo, T. Tuytelaars	2017
6	Shortcut Learning in Deep Neural Networks	R. Geirhos et al.	2020
7	Undoing the Damage of Dataset Bias	Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Ant	2012
8	A Deeper Look at Dataset Bias	Tatiana Tommasi, Tinne Tuytelaars, Barbara Caputo	2017
9	A Survey on Bias in Visual Datasets	Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntout	2021
10	Attention Is All You Need	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob U	2017
11	Deep Residual Learning for Image Recognition	Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun	2015
12	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale	Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesniko	2021
13	Lable Me		
14	Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper		
15	The Need for Biases in Learning Generalizations		
16	Chen_An_Empirical_Study_of_Training_Self-Supervised_Vision_Transformers_ICCV_2021_paper		
17	Domain Adaptation? transfer learning and generative modelling?Synthetic Data Generation?GAN?		
18	Mitigating Dataset Bias via Image Translation	J. An, Y. Kwak, J. Kim	2022
19	Performance of Machine Learning Algorithms and Diversity in Data	H. Sug	2018
20	Diversity in Machine Learning	Z. Gong, P. Zhong, W. Hu	2018
21	Improving Generalisation of AutoML Systems with Dynamic Fitness Evaluations	B. P. Evans, B. Xue, M. Zhang	2020
22	Learning Transferable Features with Deep Adaptation Networks	M. Long et al.	2015
23	Domain Invariant Transfer Kernel Learning	M. Long, J. Wang, J. Sun, P. S. Yu	2015
24	A Unified Causal View of Domain Invariant Representation Learning	Z. Wang, V. Veitch	2022
25	A General Methodology to Quantify Biases in Natural Language Data	J. Chen et al.	2020
26	An Approach to Identifying and Quantifying Bias in Biomedical Data	M. C. De Paolis Kaluza, S. Jain, P. Radivojac	2023
27	OccamNets: Mitigating Dataset Bias by Favoring Simpler Hypotheses	R. Shrestha et al.	2022
28	AVATAR - Machine Learning Pipeline Evaluation Using Surrogate Model	T.-D. Nguyen et al.	2020
29	Harnessing Synthetic Datasets: The Role of Shape Bias in Deep Neural Network Generalization	E. Benarous et al.	2023
30	In Rain or Shine: Understanding and Overcoming Dataset Bias for Improving Robustness Against Weather Corr	A. R. Marathe et al.	2022

GREEN: READ ALL AT LAST ONCE
ORGRANGE: READ ABSTRACT ONLY
BLACK: NOT YET READ

The most pertinent state-of-art literature summary

1. Datasets bias can **still be easily captured** by modern neural networks.
2. This **phenomenon** is robust **across model architectures**, **dataset combinations**, and **many other settings(hyperparameters)**.
3. It is worth pointing out that the **concrete forms** of the bias captured by **neural networks** remain **largely unclear**.
4. It has been discovered that such **bias** may contain **some generalizable and transferrable patterns**, carrying some semantic information that is transferrable to image classification tasks.

——from *Unbiased_look & A Decade's Battle*

Literature-bias as a Learning Short-cut

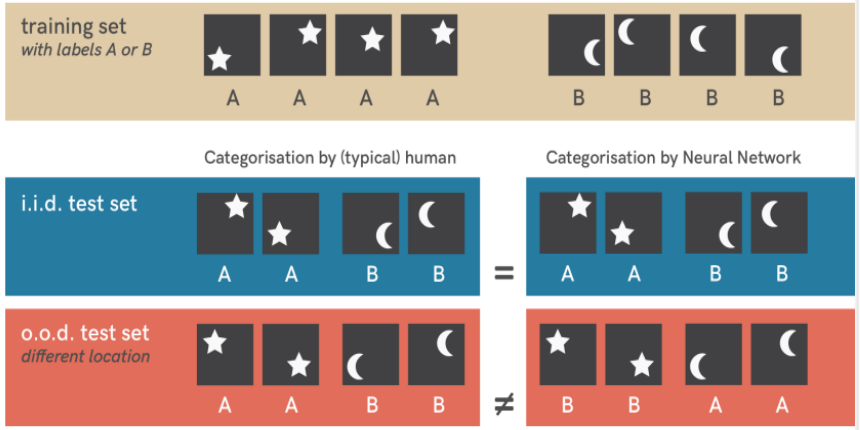


Figure 2. Toy example of shortcut learning in neural networks. When trained on a simple dataset of stars and moons (top row), a standard neural network (three layers, fully connected) can easily categorise novel similar exemplars (mathematically termed i.i.d. test set, defined later in Section 3). However, testing it on a slightly different dataset (o.o.d. test set, bottom row) reveals a shortcut strategy: The network has learned to associate object location with a category. During training, stars were always shown in the top right or bottom left of an image; moons in the top left or bottom right. This pattern is still present in samples from the i.i.d. test set (middle row) but not in o.o.d. test images (bottom row), exposing the shortcut.

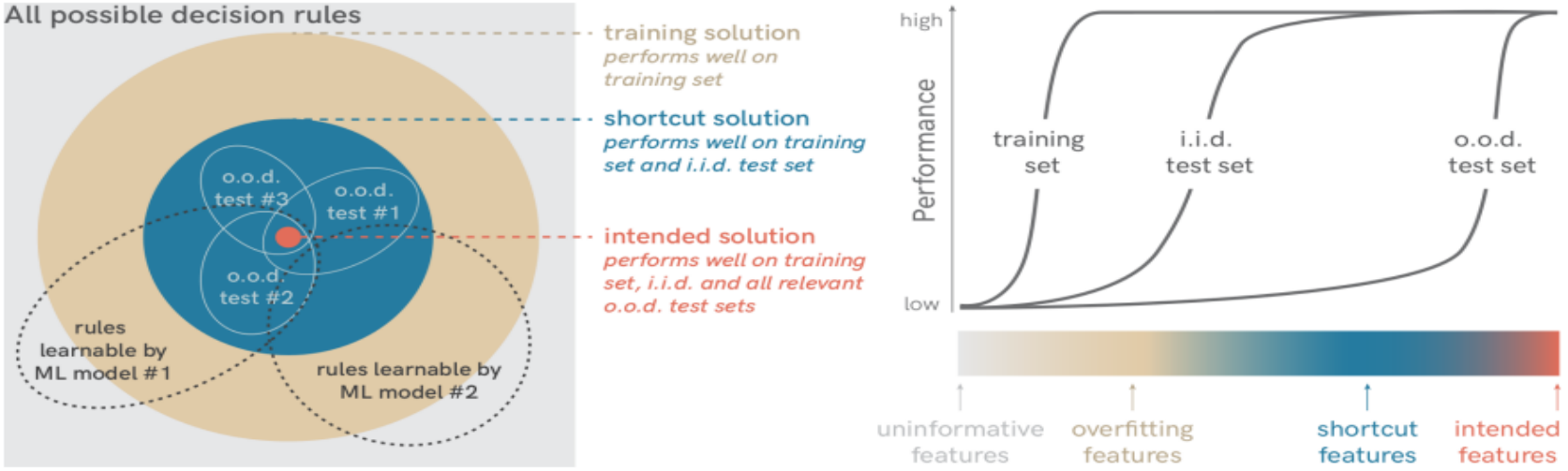


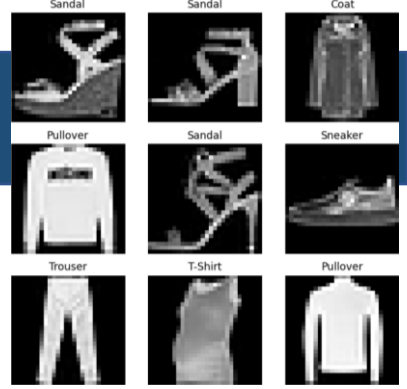
Figure 3. Taxonomy of decision rules. Among the set of all possible rules, only some solve the training data. Among the solutions that solve the training data, only some generalise to an i.i.d. test set. Among those solutions, shortcuts fail to generalise to different data (o.o.d. test sets), but the intended solution does generalise.

Testing O.O.D. Generalisation

——from *Short-Cut Learning Shortcut Learning in Deep Neural Networks*

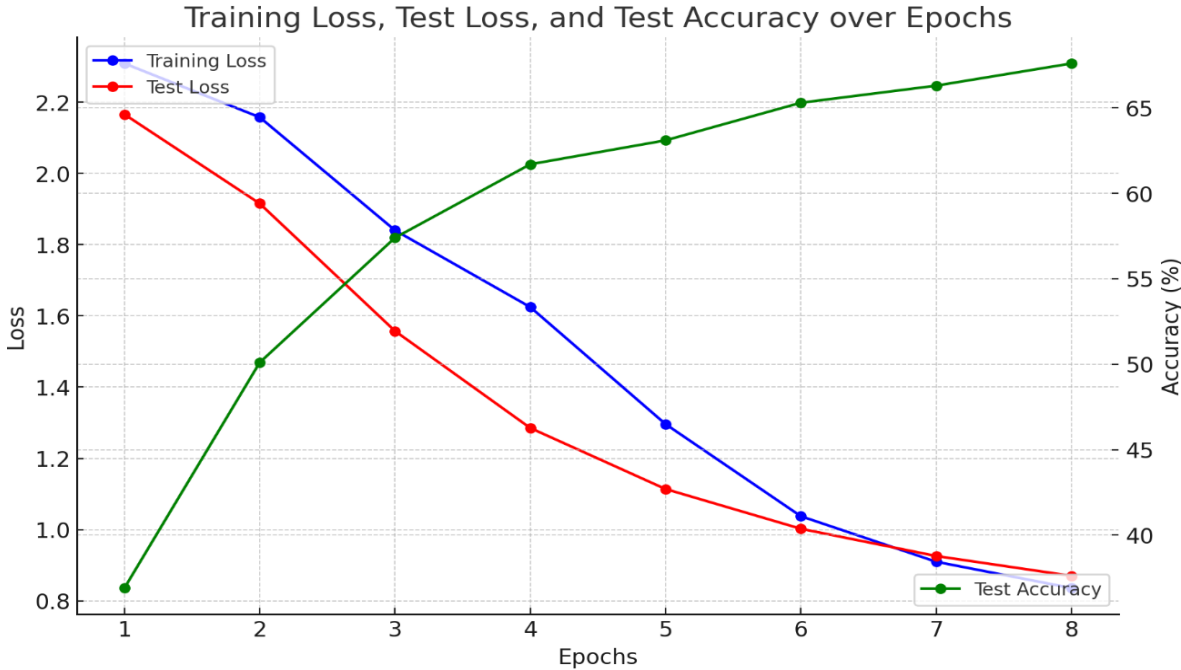
1. Evaluating only on i.i.d. test data is insufficient: high performance may exploit dataset-specific shortcuts rather than true task understanding.
2. O.O.D. tests (on data from a different distribution) reveal whether the model has learned the intended solution or merely over-relied on spurious correlations.
3. Incorporate O.O.D. evaluation as a standard practice to ensure robust performance in real-world scenarios.

FashionMNIST (MLP)




```
Epoch 8
-------------------------------
loss: 0.835908  [    64/60000]
loss: 1.098426  [ 6464/60000]
loss: 0.918809  [12864/60000]
loss: 0.886673  [19264/60000]
loss: 0.823191  [25664/60000]
loss: 0.764925  [32064/60000]
loss: 0.934324  [38464/60000]
loss: 0.845133  [44864/60000]
loss: 0.853848  [51264/60000]
loss: 0.970004  [57664/60000]
Test Error:
 Accuracy: 67.6%, Avg loss: 0.870076
```

- **Data Preprocessing**: Loaded the FashionMNIST dataset and applied ToTensor() to convert images into tensors.
- **Model Architecture**: A simple MLP withInput: 28×28 images flattened into a 784-dim vectorHidden
- **Layers**: Two layers of 512 units each, with ReLU activations
- **Output**: 10 units for the FashionMNIST classes
- **Training & Optimization:** Trained over multiple epochs using SGD and cross-entropy loss; final model weights have been saved.



ResNet

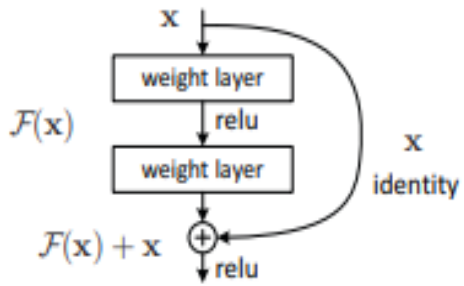
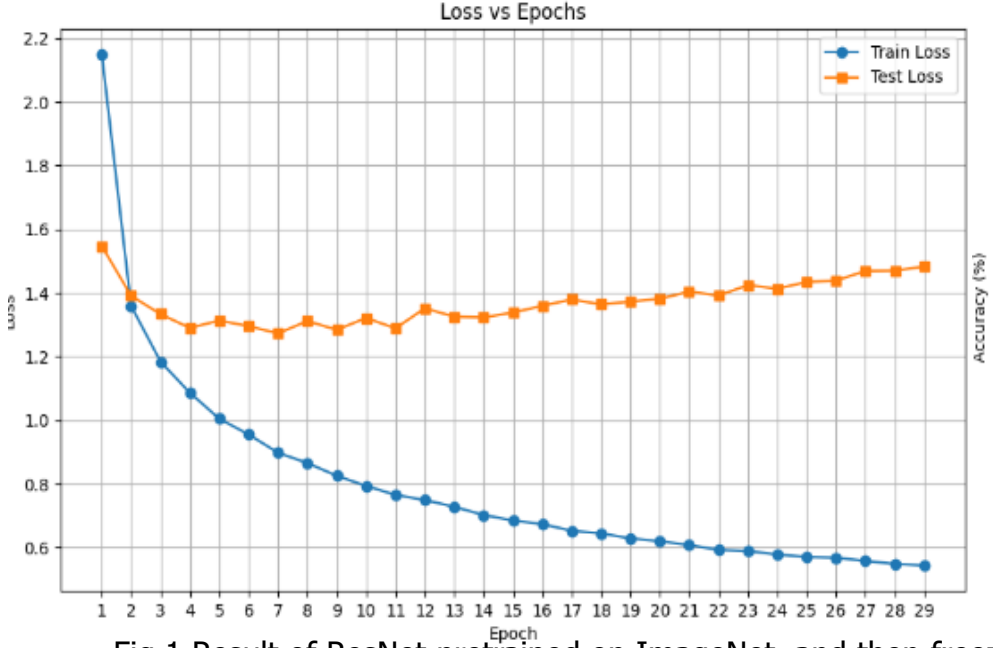


Figure 2. Residual learning: a building block.



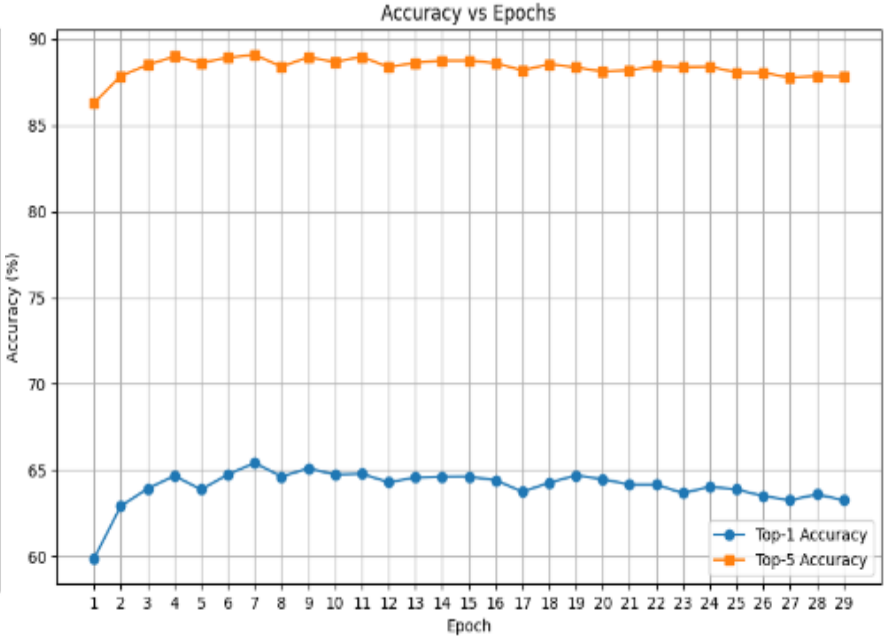


Fig.1 Result of ResNet pretrained on ImageNet, and then freeze the feature layer parameters then, modify the last layer (FC)(train the last C layer) on CIFAR-100

- **Fine-tuning ResNet-101 on CIFAR-100**
- **Data prep & augmentation**
  – Resize images to 224×224
  – Random horizontal flips on training set
- **Model setup**
  – Pretrained ResNet-101 (ImageNet)
  – Freeze all conv layers; replace final FC with 100-way classifier
- **Training**
  – Optimizer: Adam on last FC only
  – Loss: cross-entropy
  – Trained for 30 epochs

- **Results**
  – **Loss:** train loss steadily ↓; test loss bottoms at ~ep 5 then slowly ↑ (mild overfitting)
  – **Accuracy:** Top-1 peaks ~65 %; Top-5 ~89 % around ep 6
- **Next steps**
  – Introduce stronger augmentations (e.g. random crop, color jitter)
  – Tune learning rate schedule & regularization
  – Optionally unfreeze deeper blocks for full fine-tuning

Vision Transformer

Token ←→bit

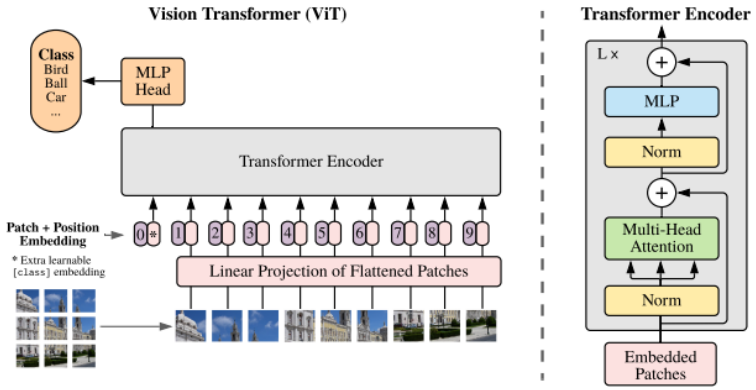


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

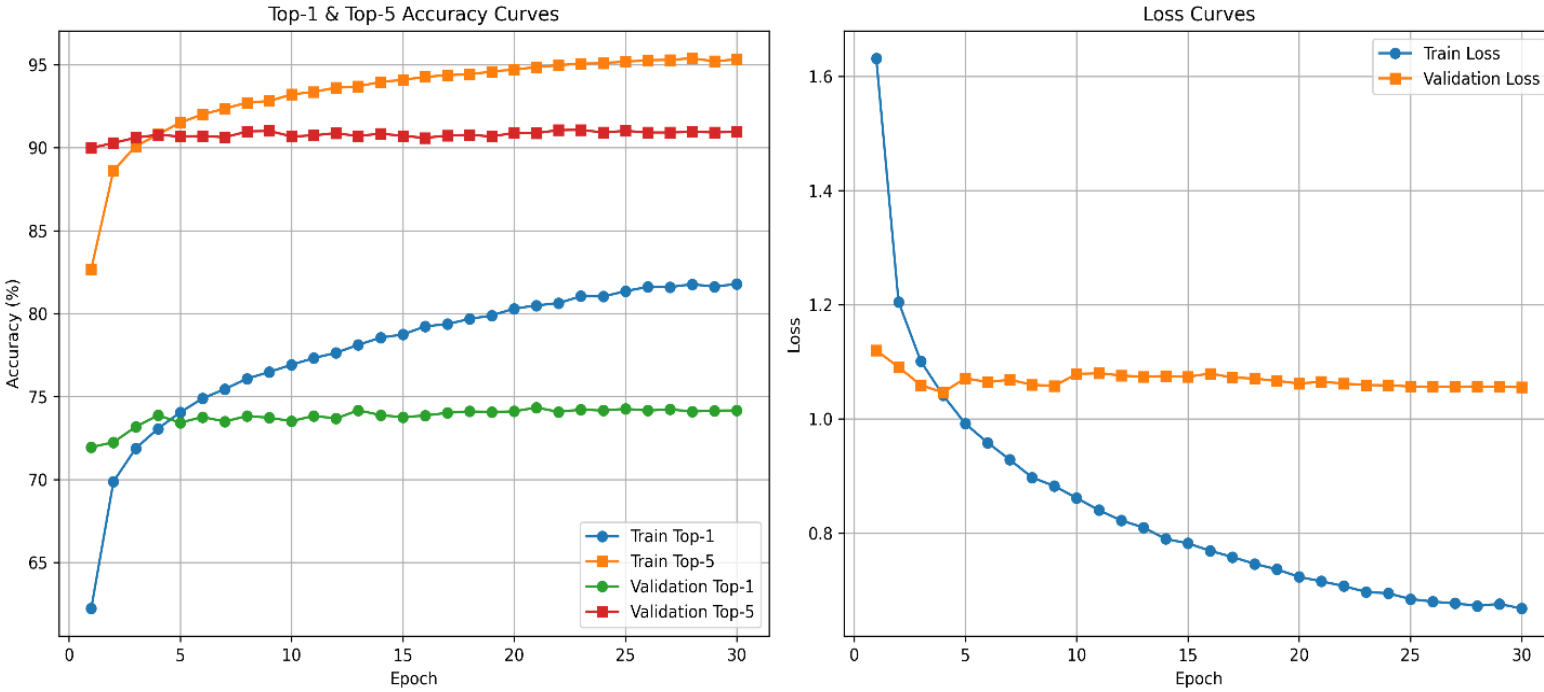


Fig.2 Results of pre-training the visual transformer on ImageNet and then training the last FC layer parameters on Tiny ImageNet

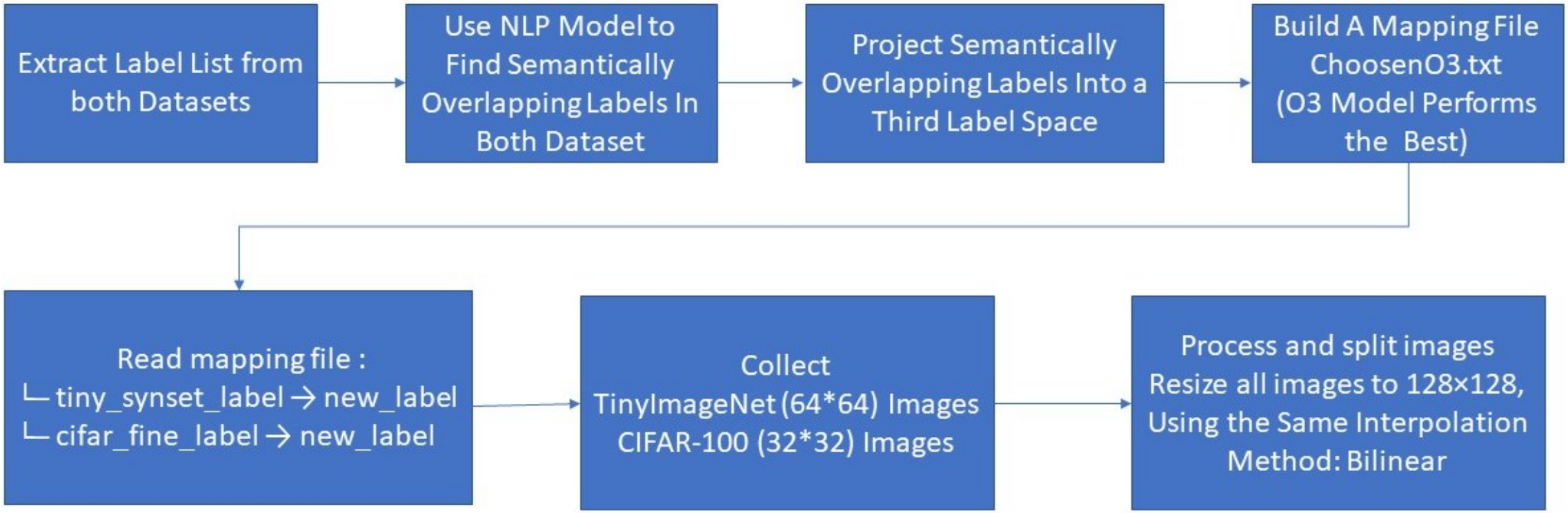
- **Fine-tuning ViT on** ImageNet
- **1. Experimental Setup Recap**
- **Pre-training**: ViT-B/16 pre-trained on ImageNet
- **Fine-tuning**: Only the final FC layer unfrozen, trained on Tiny-ImageNet (200 classes)
- **Hyperparameters**: 30 epochs, LR=0.001, CosineAnnealingLR scheduler, early-stop patience=26

- **3. Key Observations**
- **Overfitting Trend**
  - Training loss steadily decreases to ~0.65, while validation loss plateaus around ~1.05
  - The gap (Δ = val loss – train loss) widens as epochs progress
- **Model Saturation**
  - Validation Top-1/Top-5 accuracy levels off after epochs 4
  - Further training does not yield significant generalization gains

Name the Dataset – Build New Dataset

Overall Task Workflow

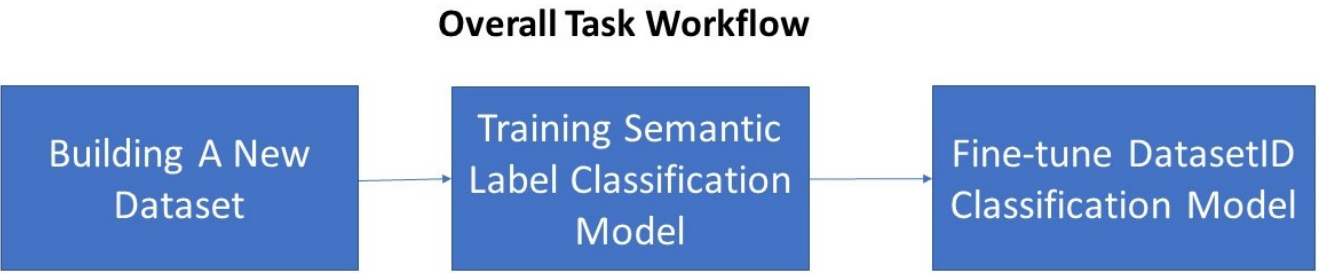
Sematically Overlapping Labels NewDataset

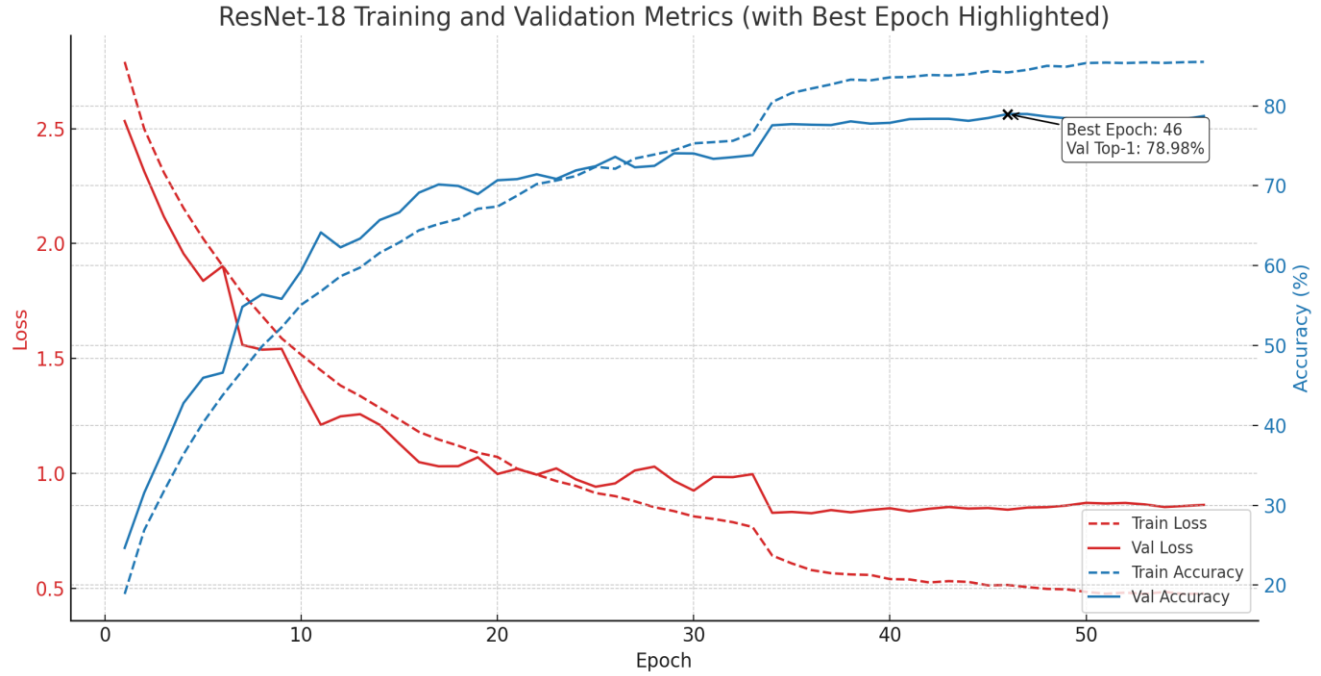


The same up-sampling method "does not equal" the same effect for the two datasets, but if we don't do it there exist size short-cut.

Upsampling :Ringing effect, artifacts
DownSampling: Aliasing, Information loss

Name the Dataset – ResNet Sematic Classifier





Input: (image, image label) (ignoring the dataset ID)

Task 2: Train ResNet-18 from Scratch

•Read metadata.csv

• Load the CSV generated in Task 1

• Split records into training and test sets according to the split column

•Create PyTorch Dataset & DataLoader

• Define CustomImageDataset that:

– Reads each image path and its new_label_id

– Applies transforms (data augmentation + normalization)

– Returns a Tensor + integer label

• Instantiate a DataLoader for the train split (shuffle=True) and the test split (shuffle=False)

•Initialize ResNet-18 (randomly)

• model = resnet18(weights=None)

• Replace model.fc with a new nn.Linear(in_features, 27) layer (27 = number of classes)

• Move model to the chosen device ("cuda" or "cpu")

•Training Loop (with Early Stopping & Checkpoints)

• Loss = CrossEntropyLoss; Optimizer = Adam

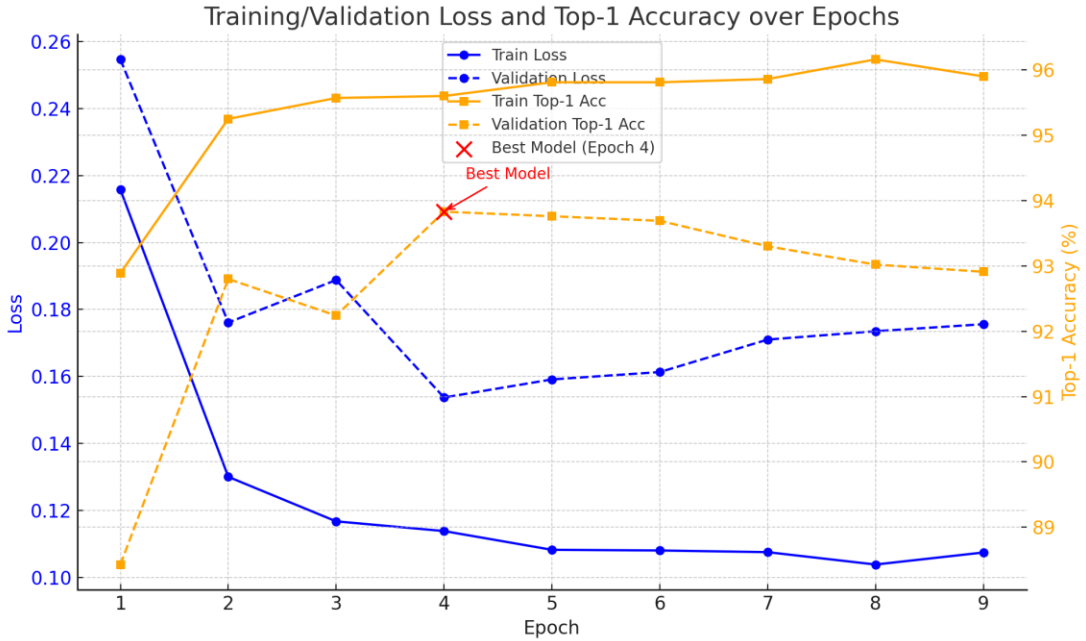
• For each epoch:

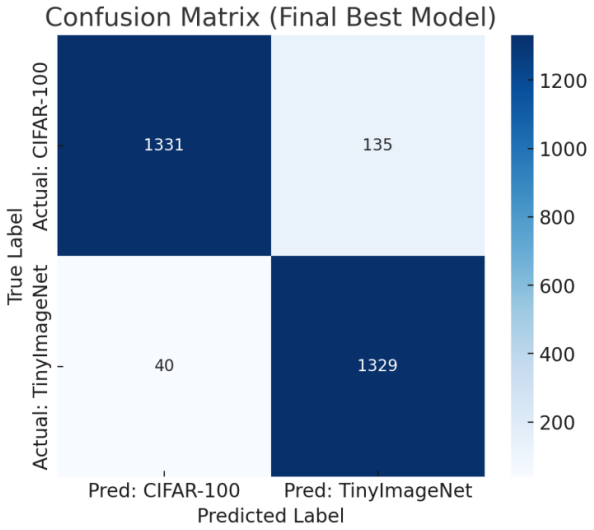
1. **train_one_epoch** → compute train_loss & train_top1
2. **validate** → compute val_loss, val_top1 & val_top5
3. If val_top1 improved → save best_resnet18.pth and reset patience
4. Else increment "stale" counter; if it reaches Early Stopping patience, break

• Every 10 epochs, also save a checkpoint: resnet18_epoch_<n>.pth

DatasetID Classifier

Input: (image, dataset ID (CIFAR-100 or TinyImageNet))





Best epoch: 4 with val_top1=93.83%

Final Validation Set Performance:

Classification Report:

	precision	recall	f1-score	support
CIFAR-100	0.97	0.91	0.94	1466
TinyImageNet	0.91	0.97	0.94	1369
accuracy			0.94	2835
macro avg	0.94	0.94	0.94	2835
weighted avg	0.94	0.94	0.94	2835

Overall Task Workflow

Building A New Dataset → Training Semantic Label Classification Model → Fine-tune DatasetID Classification Model

Task 3: Fine-Tune for Dataset ID Prediction

•Read metadata.csv

• Similar to Task 2, but use only filepath, dataset_id (CIFAR-100 = 0, TinyImageNet = 1) and split

•Create Dataset for "Dataset ID" Prediction

• Define DatasetIDDataset that:
  – Reads each image from filepath and converts to Tensor (with transforms)
  – Maps dataset_id text to integer (0 or 1)
• Instantiate a DataLoader for train (shuffle=True) and validation (shuffle=False)

•Load Pretrained ResNet-18 from Task 2

• model = resnet18(weights=None) + original fc layer (27 outputs)
• Load weights from best_resnet18.pth

•Freeze All Layers Except the Last

• Iterate model.named_parameters(), set requires_grad = False for all except parameters in fc

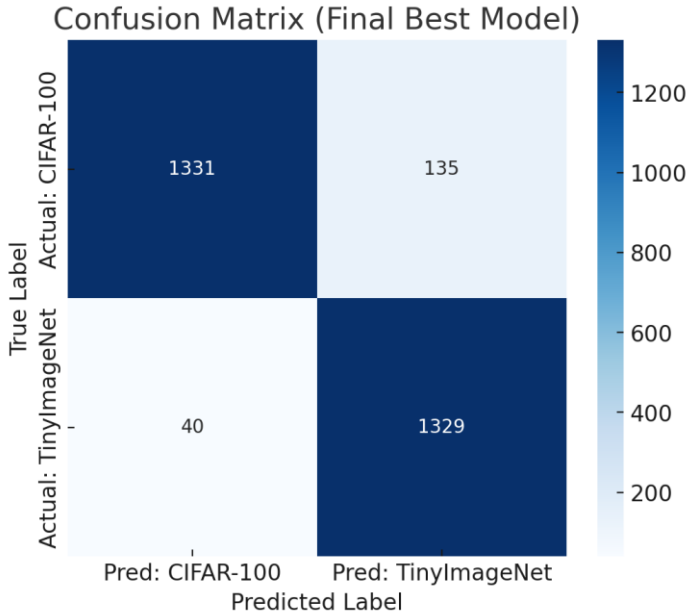
•Replace Last Fully-Connected Layer

• Set model.fc = nn.Linear(in_features, 2) to output [CIFAR-100 vs TinyImageNet]

•Training & Evaluation

• Loss = CrossEntropyLoss; Optimizer = Adam; use Early Stopping
• For each epoch:
  1. Train → compute train_loss, train_top1 (and collect preds/labels)
  2. Validate → compute val_loss, val_top1 (and collect preds/labels)
  3. If val_top1 improves → save best_resnet18_datasetid.pth and print confusion matrix + classification report; reset patience
  4. Else increment "stale" counter; if it hits patience, stop training
• After training, load best_resnet18_datasetid.pth and run a final validation on the test split; print the confusion matrix and classification report

DatasetID Classifier



Best epoch: 4 with val_top1=93.83%

Final Validation Set Performance:

Classification Report:

	precision	recall	f1-score	support
CIFAR-100	0.97	0.91	0.94	1466
TinyImageNet	0.91	0.97	0.94	1369
accuracy			0.94	2835
macro avg	0.94	0.94	0.94	2835
weighted avg	0.94	0.94	0.94	2835

Precision (per class)

Formula:

precision = TP / (TP + FP)

CIFAR-100 precision = 1331 / (1331 + 40) ≈ 0.97

TinyImageNet precision = 1329 / (1329 + 135) ≈ 0.91

Recall (per class)

Formula:

recall = TP / (TP + FN)

CIFAR-100 recall = 1331 / (1331 + 135) ≈ 0.91

TinyImageNet recall = 1329 / (1329 + 40) ≈ 0.97

F1-score (per class)

Formula:

F1 = 2 × (precision × recall) / (precision + recall)

Both classes yield F1 ≈ 0.94

(e.g., for CIFAR-100: 2 × (0.97 × 0.91) / (0.97 + 0.91))

Support

The number of true samples of each class:

CIFAR-100: 1466

TinyImageNet: 1369

Accuracy (overall)

Formula:

accuracy = (TP_total + TN_total) / (total samples)

= (1331 + 1329) / 2835 ≈ 0.94

Average Metrics

Macro Average (macro avg)

Arithmetic mean of each class's precision, recall, and F1:

(0.97 + 0.91) / 2 = 0.94

Weighted Average (weighted avg)

Sample-weighted mean of each class's precision, recall, and F1:

(0.97 × 1466 + 0.91 × 1369) / (1466 + 1369) ≈ 0.94

Models trained for classifying semantic labeling still retain enough information to recognize the dataset source; (even if it was trained only to classify semantic labels in the first place).

Future Plan – Quantization & Mitigation Paper

4. Study of Learning Paradigms:

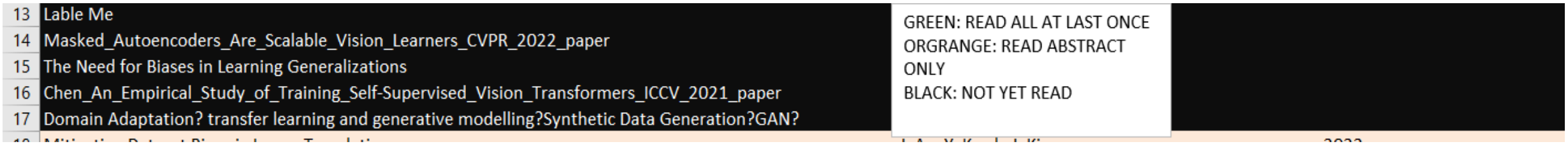
- Through transfer learning, generative models (e.g., GANs / VAEs) and other training setups, deeply observe how dataset bias emerges, is amplified, or is mitigated in practice.

5. (If Possible) Bias Detection & <u>Quantification</u>

1. Analyze internal biases (class imbalance, annotation rules) and external biases (domain shifts) in datasets like ImageNet, PASCAL VOC and Cityscapes.
2. Develop interpretable metrics (feature-space clustering scores, cross-dataset accuracy, information theory) to quantify bias.

6. (If Possible) Bias Impact Analysis & Mitigation Strategies

1. Run controlled experiments to demonstrate how dataset bias causes large performance drops in specific scenarios (e.g. low-light, rare classes).
2. Evaluate bias-correction methods (domain adaptation, GAN-based augmentation, synthetic data).

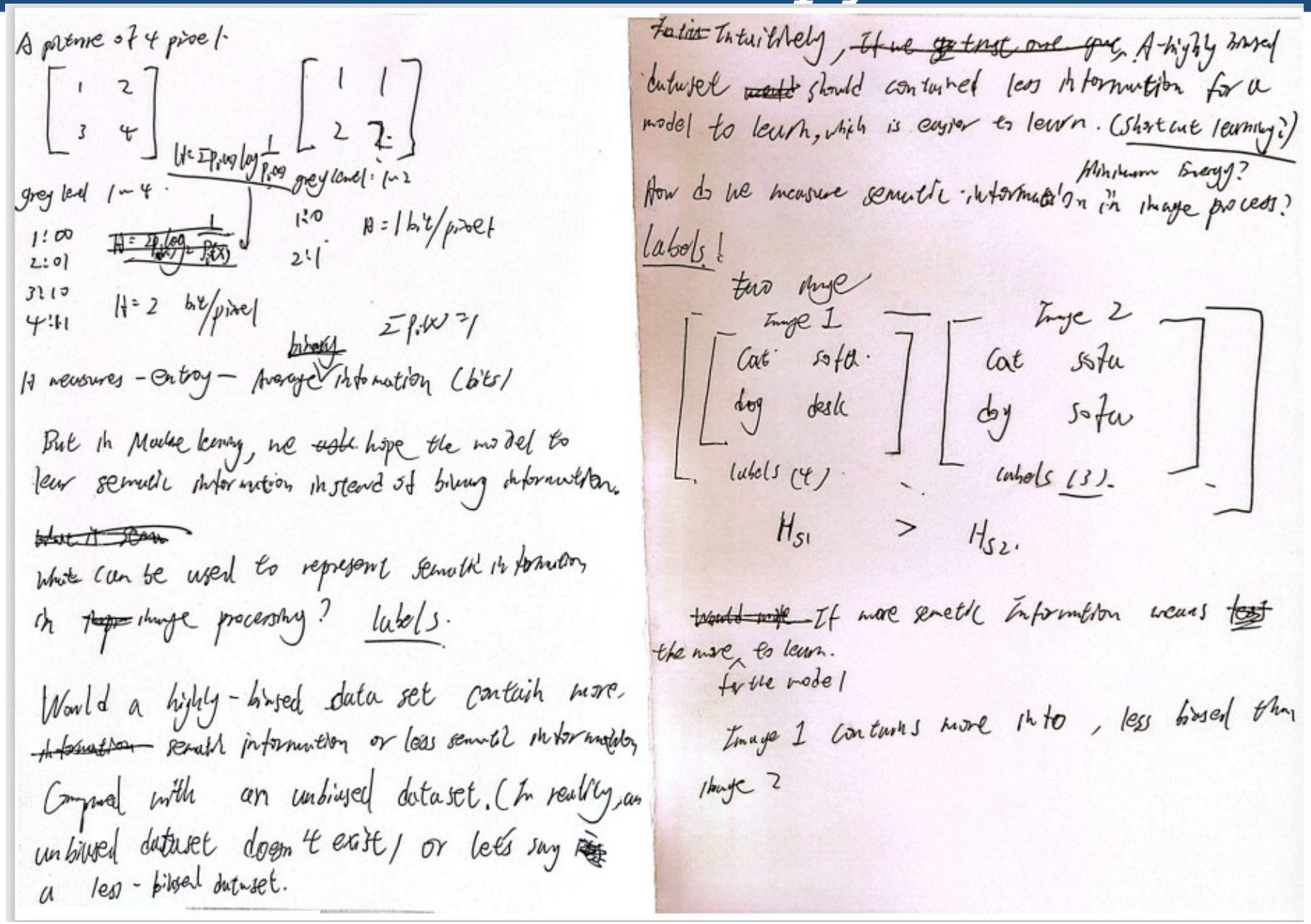


Future Work Plan – Sematic Entropy

Will a highly-biased dataset contain more semantic information or less semantic information compared with a less-biased dataset?

Sematic Entropy :

We need a method to quantify how much sematic information is contained in an image.



Future Work Plan – Sematic Entropy

A picture of 4 pixel:

$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ → $\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$

grey level: 1 ~ 4

1: 00
2: 01
3: 10
4: 11

$H = \sum_i p(x_i) \log \frac{1}{p(x_i)}$

grey level: 1 ~ 2

1: 0
2: 1

$H = 1$ bit/pixel

$H = 2$ bit/pixel

$\sum p_i(x) = 1$

binary

H measures - entropy - Average information (bits)

But in Machine learning, we ~~only~~ hope the model to learn semantic information instead of binary information.

~~what it learn~~

What can be used to represent semantic information in ~~image~~ image processing? labels.

World a highly-biased data set contain more ~~information~~ semantic information or less semantic information Compared with an unbiased dataset. (In reality, an unbiased dataset doesn't exist / or let's say a less-biased dataset.

~~Factor~~ Intuitively, ~~If we get trust one guy~~, A highly biased dataset ~~would~~ should contained less information for a model to learn, which is easier to learn. (Shortcut learning?)

Minimum energy?

How do we measure semantic information in image process? labels!

two image

Image 1

[[Cat sofa
dog desk]]

labels (4)

Image 2

[[Cat sofa
dog sofa]]

labels (3)

$H_{S1} > H_{S2}$.

~~would make~~ If more semantic information means ~~less~~ the more to learn. for the model

Image 1 contains more info, less biased than image 2

Analogy to ~~binary~~ ~~cross-entropy~~ entropy, lets define the same sematic entropy. for a certain label i:

$P(x_i)$ = How many ~~label~~ images contain the label (n_i).
(S-label: number of images in a dataset (N))

How likely a label will appear in a image

~~$P_{S\text{-label}_i}(x) \leq 1$~~

what would be $\sum_i P_{S\text{-label}_i}$?

How can we normalize it to be 1.

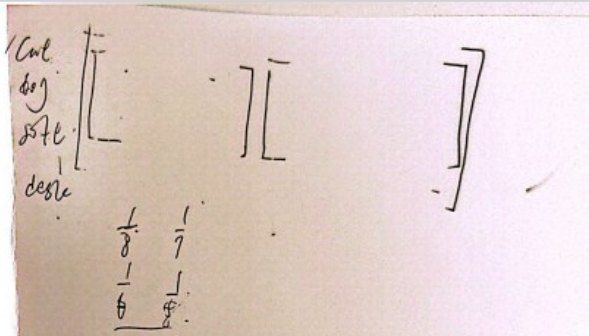
What is $H_S = \sum_i P_{S\text{-label}_i} \log \frac{1}{P_{S\text{-labels}}}$? ← Can we use this to quantify bias?

bits/pixel → labels/image ?

label / token?

MLP.

Semantic / ~~label~~ token



token?

being info → Transformer middle layer → semantic info

low layer

lose layer

↑

cluster tokens

$P_{S\text{-label}_i} = \frac{\text{How many images contain one label}_i(n_i)}{\text{Number of images in a dataset } (N)}$

$H_{S\text{-label}_i} = (1 - P_{S\text{-label}_i}) \log (1 - P_{S\text{-label}_i}) + P_{S\text{-label}_i} \log P_{S\text{-label}_i}$

$H_S = \sum_i \frac{1}{\text{num of } i} \cdot H_{S\text{-label}_i}$

[[dog sofa
cat desk] [dog desk
cat desk]]

1

Future Work Plan – Quantize Dataset Bias

Why Purely Statistical Measures Are Not Enough to Quantify Dataset Bias

•Bias comes from many subtle sources, it's hard to name them all.
•Truly "unbiased" datasets don't exist
•The goal of bias analysis is to improve **generalization performance**

Model Generalization Should Guide Bias Quantification

Don't just compare data distributions, we also need to evaluate model **generalization performance** across them

Suggested approaches:

- Cross-domain/Cross-distribution Validation: Compare performance on data from different domains or conditions
- Sensitivity Analysis: Modify features (e.g., color, background) and measure performance drops
- Distribution Alignment Metrics: Compute distribution discrepancies (e.g., MMD) between training data and real-world data in a latent space

4..

It has been discovered that such **bias** may contain **some generalizable and transferrable patterns**, carrying some semantic information that is transferrable to image classification tasks

Reference

•A. Torralba and A. A. Efros, “Unbiased Look at Dataset Bias,” in *CVPR Workshop on Dataset Bias*, 2011.

•F. Wang, X. Xue, and J. Huang, “A Decade’s Battle on Dataset Bias,” *International Journal of Computer Vision*, vol. 128, no. 4, pp. 780–795, 2020.

•A. Khosla, B. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, “Undoing the Damage of Dataset Bias,” in *ECCV*, 2012, pp. 158–171.

•X. Peng, B. Xiao, Z. Zhao, and Q. Huang, “A Deeper Look at Dataset Bias,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1132–1145, 2020.

•H. Wang, A. Gupta, and L. Davis, “A Survey on Bias in Visual Datasets,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 13, no. 1–2, pp. 1–151, 2022.

•F. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. Wichmann, “Shortcut Learning in Deep Neural Networks,” *Nature Machine Intelligence*, vol. 2, pp. 665–673, 2020.

•K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *CVPR*, 2016, pp. 770–778.

•A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner et al., “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in *ICLR*, 2021.

•A. Vaswani et al., “Attention Is All You Need,” in *NeurIPS*, 2017, pp. 5998–6008.

Future Work Plan- Dataset List

	A	B	C
1	**Dataset Name**	**Description**	**Link**
2	ImageNet	Large-scale dataset for classification and transfer learning.	https://www.image-net.org/
3	CIFAR-10	Small image dataset with 10 classes for quick experiments.	https://www.cs.toronto.edu/~kriz/cifar.html
4	CIFAR-100	Extension of CIFAR-10 with 100 classes.	https://www.cs.toronto.edu/~kriz/cifar.html
5	Tiny ImageNet	Subset of ImageNet for smaller-scale experiments.	https://huggingface.co/datasets/zh-plus/tiny-imagenet
6	MNIST	Handwritten digit classification dataset.	http://yann.lecun.com/exdb/mnist/
7	Fashion-MNIST	Classification dataset of fashion products.	https://github.com/zalandoresearch/fashion-mnist
8	COCO	Dataset for object detection, segmentation, and captioning.	https://cocodataset.org/
9	PACS	Dataset for domain generalization with different styles.	https://domaingeneralization.github.io/
10	Office-31	Dataset for domain adaptation in office environments.	https://www.cc.gatech.edu/~judy/domainadapt/
11	DomainNet	Large-scale dataset for domain adaptation with diverse styles.	http://ai.bu.edu/M3SDA/
12	CelebA	Dataset of celebrity faces with attribute labels.	https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
13	Stanford Cars	Fine-grained car classification dataset.	https://ai.stanford.edu/~jkrause/cars/car_dataset.html
14	ChestX-ray14	Medical image dataset for thorax disease classification.	https://nihcc.app.box.com/v/ChestXray-NIHCC
15	ADE20K	Dataset for semantic segmentation tasks.	https://groups.csail.mit.edu/vision/datasets/ADE20K/
16	VISDA-17	Dataset for domain adaptation tasks.	https://github.com/VisionLearningGroup/taskcv-2017-public/tree/master/classification
17	Camelyon16	Pathology image dataset for small sample analysis.	https://camelyon16.grand-challenge.org/
18	YFCC [49]		
19	CC [4]		
20	DataComp [15]	YFCC+CC+ImageNet	
21	WIT [45]	YFCC100M [49], CC12M [4], and DataComp-1B [15	
22	LAION [42]		

Literature Summary

single-object-in-the-center mentality
——from *Unbiased_ook at dataset bias*

Figure 3. Dataset Look-alikes: Above, ImageNet is trying to impersonate three different datasets. Here, the samples from ImageNet that are closest to the decision boundaries of the three datasets are displayed. Look-alikes using PASCAL VOC are shown below.

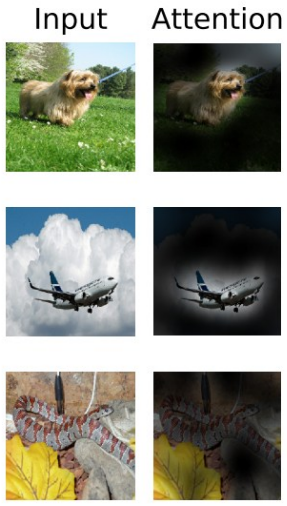


Figure 6: Representative examples of attention from the output token to the input space. See Appendix D.7 for details.

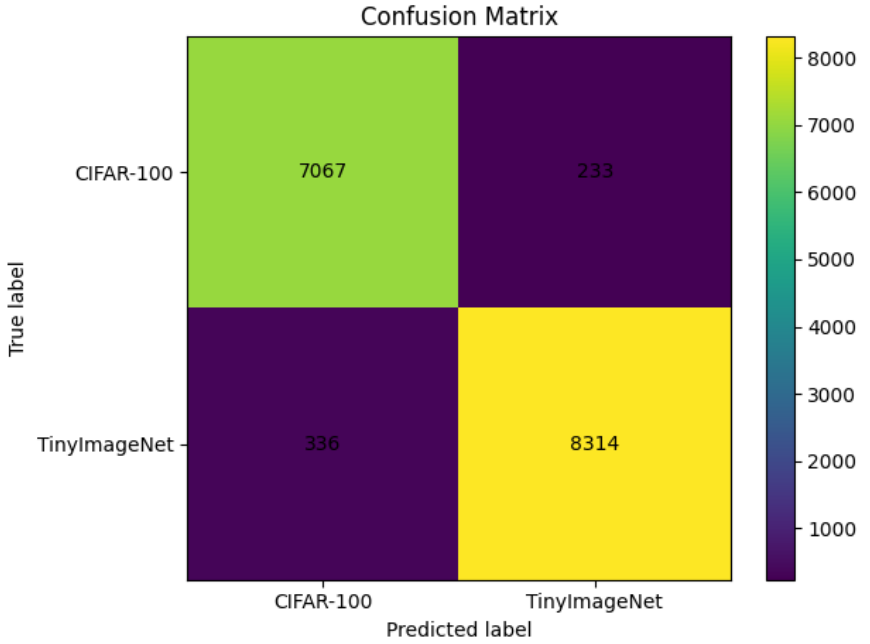
ViT self-attention

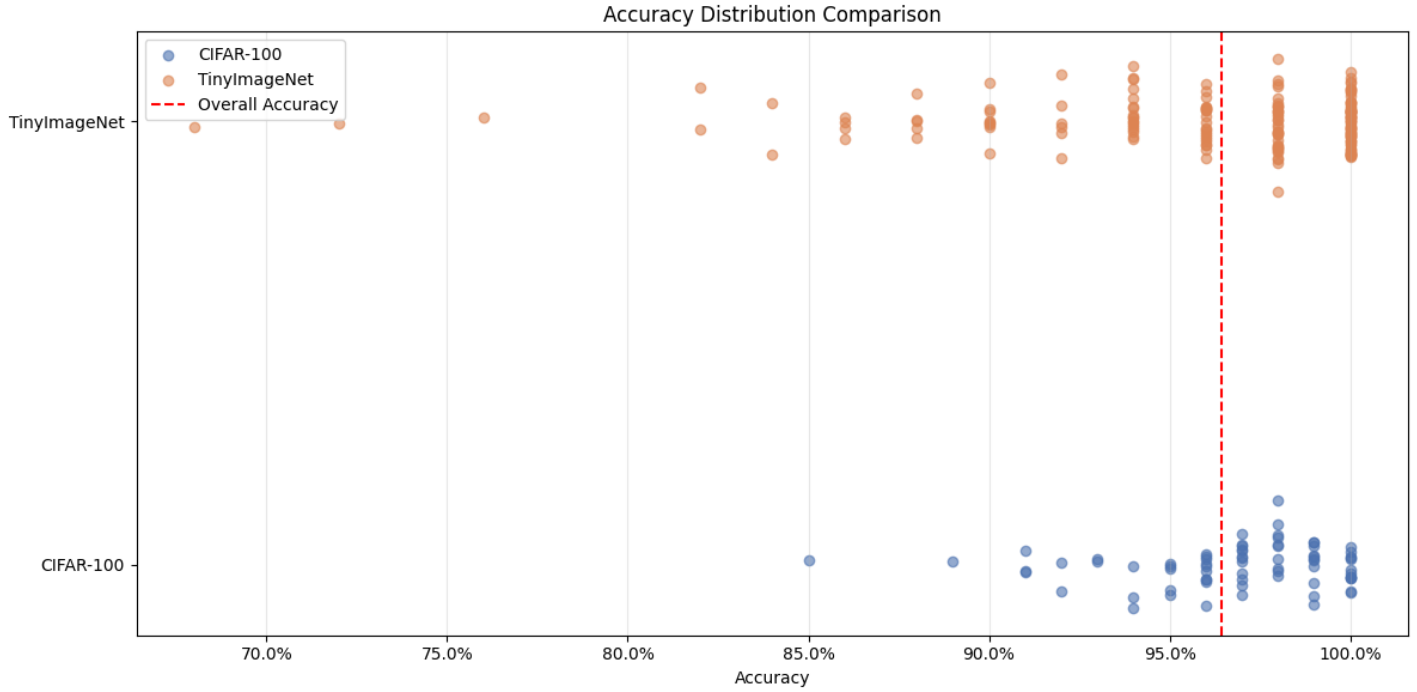
– Sometimes naturally focuses on object patches and downweights background clutter
– Reduces need for "object-on-white" or perfectly centered datasets
– Still benefits from diverse poses, lighting and contexts for true robustness.

Experiment 2: *Non-semantic Information Test* ***for 1803***

[1803 Classification Report]

	precision	recall	f1-score	support
CIFAR-100	0.9489	0.9659	0.9573	7300
TinyImageNet	0.9708	0.9561	0.9634	8650
accuracy			0.9606	15950
macro avg	0.9598	0.9610	0.9603	15950
weighted avg	0.9607	0.9606	0.9606	15950





A picture of 4 pixels.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$$

grey level 1 ~ 4.

1: 00
2: 01
3: 10
4: 11

Use $\sum p_i \log \frac{1}{p_i}$ to ~~$H = \sum p_i \log \frac{1}{p(x)}$~~

$H = 2$ bit/pixel

grey level: 1 ~ 2

1: 0
2: 1

$H = 1$ bit/pixel

$\sum p(x) = 1$

Entropy measures - entropy - average information (bits)

But in Machine learning, we ~~also~~ hope the model to learn semantic information instead of biasing information.

~~What I mean~~

What can be used to represent semantic information in ~~image~~ image processing? labels.

Would a highly-biased data set contain more ~~information~~ semantic information or less semantic information compared with an unbiased dataset. (In reality, an unbiased dataset doesn't exist, or let's say ~~a~~ a less-biased dataset.

~~To put it~~ Intuitively, ~~If we get just one guy~~, A highly biased dataset ~~could~~ should contained less information for a model to learn, which is easier to learn. (Shortcut learning?)

Minimum Energy?

How do we measure semantic information in image process?

labels!

two image

$$\begin{bmatrix} \begin{bmatrix} \text{cat} & \text{sofa} \\ \text{dog} & \text{desk} \end{bmatrix} & \begin{bmatrix} \text{cat} & \text{sofa} \\ \text{dog} & \text{sofa} \end{bmatrix} \end{bmatrix}$$

Image 1: labels (4) — Image 2: labels (3)

$$H_{S1} > H_{S2}.$$

~~would make~~ If more semantic information means ~~text~~ the more ~~to~~ to learn.

for the model

Image 1 contains more info, less biased than Image 2

Analogy to ~~entropy~~ ~~entropy~~ entropy, lets define the same semantic entropy. for a certain label i:

$P_{s\text{-}label_i}(x) = \frac{\text{How many images contain the label (m)}}{\text{number of images in a data set (N)}}$

How likely a label will appear in a image

$P_{s\text{-}label_i}(x) \leq 1$

what would be $\sum_i P_{s\text{-}label_i}(x)$?

How can we normalize it to be 1.

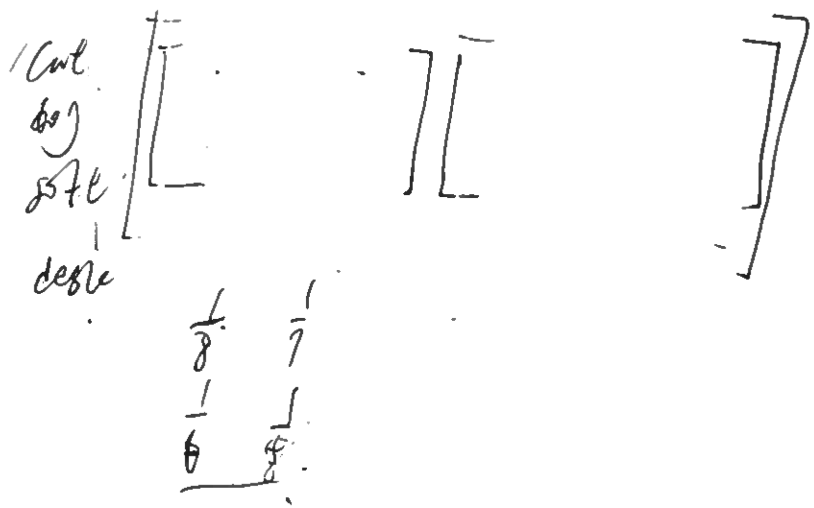
what is $H_s = \sum_i P_{s\text{-}label_i} \log_2 \frac{1}{P_{s\text{-}label_i}}$? ← Can we use this to quantify bias?

bits/pixel → labels/image ?

label / token?

NLP · something / ~~label~~ token

token?   binary info —Transformer / middle layer→ semantic info
~~low~~ low layer   last layer ↑ classifier token



$P_{s\text{-}label_i} = \frac{\text{How many images contain the label}(i)}{\text{Number of images in a dataset } (N)}$

$H_{s\text{-}label_i} = (1 - P_{s\text{-}label_i}) \log_2 (1 - P_{s\text{-}label_i}) + P_{s\text{-}label_i} \log_2 P_{s\text{-}label_i}$

$H_s = \sum_i \frac{1}{\text{num of } i} \cdot H_{s\text{-}label_i}$

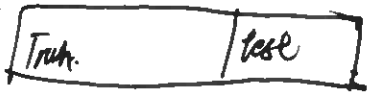
[[dog sofa] [dog desk]]
[[cat desk] [cat desk]]

1

Tiny Image Net

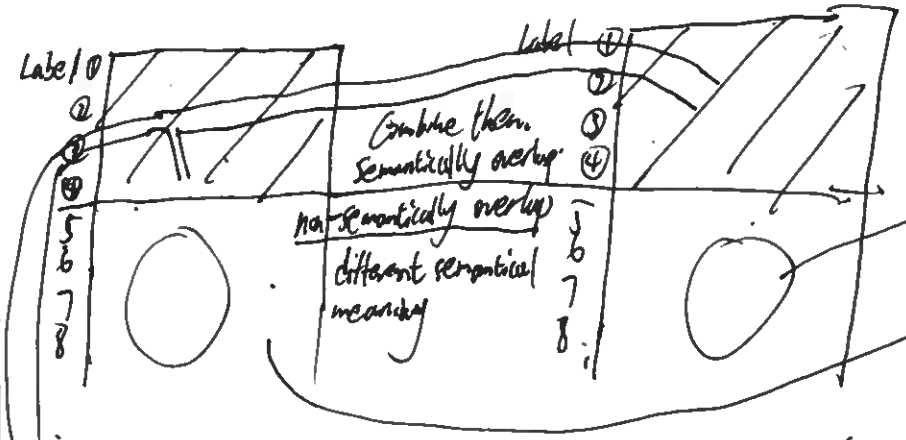
Only image; No Annotation.

CIFAR-100



Model_T2 : Class (Label) classifier.

Model_T3 : ~~Class~~ Dataset ID - classifier.



Test set ② => Data set ID - classifier => ??

New dataset: Test ① = Tiny Image Net (val) + CIFAR-100 (test)

Tiny Image Net (train) + CIFAR-100 (Train) = 90% | 10%

Val = 10%

Train = 90%

① model 多大的参数 可见的 ~~[illegible]~~ Alone the dataset；即使是很小的模型 也能达到 提高 76.62 Accuracy
ResNet 50 层

② model - Architecture
- size
- different layers

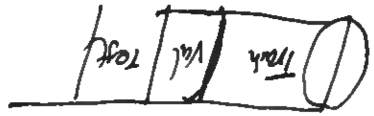
③ Dataset — Augmentation
preprocessing < up sampling / Down sampling

Dataset 本身 里面有多少是 Semantic noise
多少是 non-Semantic noise

MAE — self-supervised
supervised 呢？

~~还是说~~ 仅训练一层???　~~（不是说 仅仅一层 就能够分的）~~
二层???　~~（很好）~~
三层???

我明明是用 MAE 在 Imagenet-1k 上训练，为什么会学到
区分 CIFAR 的特征。



Tiny ImageNet

CIFAR-100

最小充分统计量: minimal sufficient statistic

$X \xrightarrow{T} T(X)$　　$p(y|X=x) = p(y|T(x)), \forall x, y$

一旦知道了 $T(x)$，额外知道 X 对于预测 Y 不再提供任何更多信息；

$\hat{X} = T(X)$ 即是 关于 X with respect to Y 的 ~~suf~~ sufficient statistic

当 $T(X)$ 是 compressed 时为 min sufficient statistic

取最大的信息最少的 statistic　min $I(X; T(X))$

最小互信息

$Y \rightarrow X \rightarrow \hat{X}$　(cat → cat photo → compressed photo)

Information Bottleneck:

$$\mathcal{L}[p(\hat{x}|x)] = \underbrace{I(X;\hat{X})}_{\min} - \beta \underbrace{I(\hat{X};Y)}_{\max}$$

$I(\hat{X};Y) \le I(X;Y)$; take "=" 当 $\hat{X}$ 是 Y 的 ~~min sufficient~~ 最小充分统计量 min sufficient statistic

$\beta \to 0$ 极度压缩；$I(X;\hat{X})$ 主导

$\beta \to \infty$ 极力保留相关信息；$I(\hat{X};Y)$ 主导（无失真）

$d_{IB}(x,\hat{x}) = D_{KL}[p(y|x) \,||\, p(y|\hat{x})]$

KL divergence.

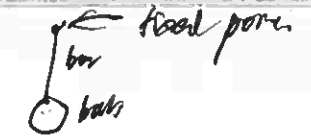
$D_{KL}[P||Q] = \sum_y P(y) \log \frac{P(y)}{Q(y)}$　　$P(y) \leftarrow$ 真实分布，$Q(y) \leftarrow$ 近似分布

① 非负性：$D_{KL}[P||Q] \ge 0$，当仅当 $P=Q$ 时等号

② 不对称性：$D_{KL}[P||Q] \ne D_{KL}(Q||P)$

③ KL 散度，量化了用 $\hat{X}$ 代替 x 对预测 Y 的信息损失

Lyapunov stability



$\vec{x}(t) = \begin{bmatrix} \theta(t) \\ \dot{\theta}(t) \end{bmatrix}$

$\vec{x}_{eq} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ ← equilibrium point.

Build

$V(\vec{x}(t)) \geq 0$ 始终大于 it is stable $x=0$

↑
energy or error.

① positive definite: $V(\vec{x}) > 0$, $\forall x \neq 0$

AND $V(\vec{0}) = 0$

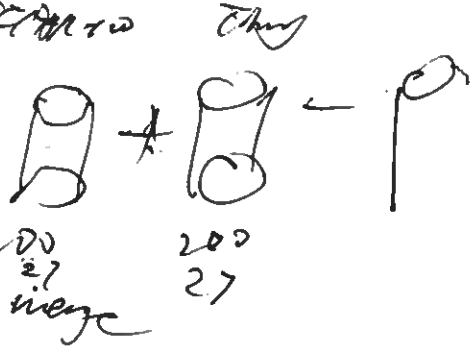
This means that $V(\vec{x})$ must be positive definite

② $\dot{V}(\vec{x}) = \frac{\partial V}{\partial x} \cdot \dot{\vec{x}} < 0$ $\forall x \neq 0$

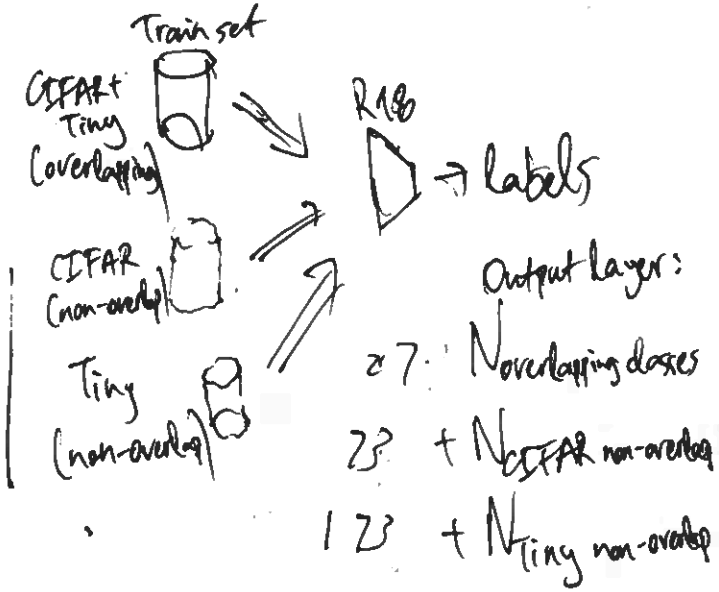
$\dot{\vec{x}} = \frac{d\vec{x}}{dt}$

系统能够自随着时间推移逐渐减小

Legend.



① less info



① Train to name the label

② fine-tune to name the dataset.

③ Test on CIFAR-100 + Tiny test set. 273 classes (labels)

Assumption: During training the model will learn more semantic info to name the classes (labels) instead of non-semantic info

semantic info can by human.

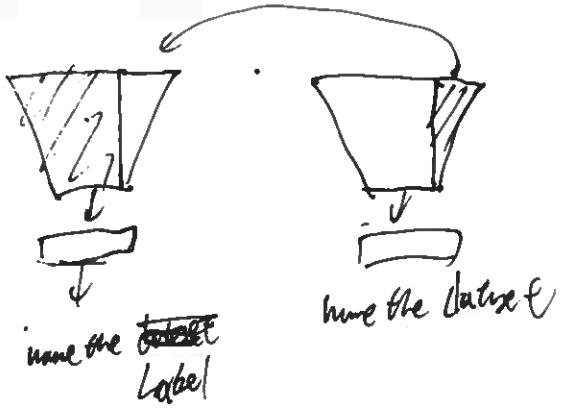
273

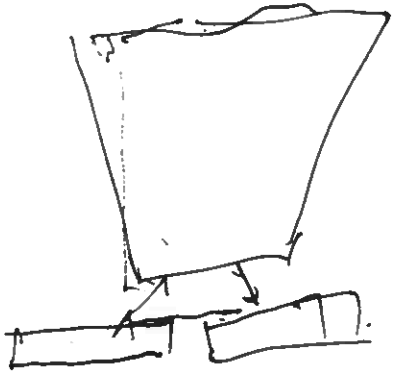
$\vec{y}$ useful info (e.g., Labels)

$\vec{x}$ image

$\vec{b}$ meaningless bias (e.g., artifacts, spurious correlation)

$p(\vec{x}, \vec{y}, \vec{b})$





$p(y_j | \vec{x}_i)$ $p(b_k | \vec{x}_i)$

Trade off between the two.

Dataset distribution:

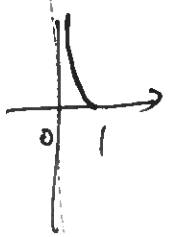
$Pr(\vec{D})$

CIFAR-100

$\vec{y}$ useful semantic info for specific task (e.g. classification)

$\vec{b}$

$Pr(\vec{D})$ Thyroid cell



$Loss = P(y_{\vec{y}} | \vec{x}_i) + \lambda P(D_{\vec{j}} | \vec{x}_i)$

~~$= -\log P(x_i)$~~

$= -\log P(y_{\vec{y}} | \vec{x}_i) + \lambda P(D_{\vec{j}} | \vec{x}_i)$

Minimize this term

label image dataset id image

cross entropy loss

cross entropy loss

~~human uses~~

human uses

$\vec{D}$ — $\vec{y}$ useful semantic info to ~~use~~ to finish specific task (e.g., classification)

$\vec{D}$ — $\vec{b}$ Meaningless non-semantic info to finish specific task.

$\vec{D} = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \end{pmatrix}$ different images. $P(y_{\vec{y}} | \vec{x}_i)$

$P(D_{\vec{j}} | \vec{x}_i)$

$P(D_j | \vec{b}_i)$

$Loss = P(y_{\vec{y}} | \vec{x}_i) + \lambda P(D_{\vec{j}} | \vec{x}_i)$

Cross-entropy

$\vec{y} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}_{c \times 1}$ ← $y_k = 1$，只有是正确类，其它均为0，one-hot 向量

$y_k = 1$

$\hat{\vec{p}} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_c \end{pmatrix}_{c \times 1}$ $\quad \hat{p} = \begin{pmatrix} 0.3 \\ 0.6 \\ 0.1 \end{pmatrix} \quad \begin{pmatrix} 0.1 \\ 0.8 \\ 0.1 \end{pmatrix}$

$\sum \hat{p}_i = 1, \ \hat{p}_i > 0$

$CE(\vec{y}, \hat{\vec{p}}) = -\sum_{i=1}^{c} y_i \log \hat{p}_i = -y_k \log \hat{p}_k = -\log \hat{p}_k$

（其中 $\hat{p}_k$，k 是真实类别。）

$\vec{y}$：真实分布（标签）

$\hat{\vec{p}}$：model 预测出来的分布

比如 $\vec{y} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ 真实 label 为 2.

$\hat{p} = \begin{pmatrix} 0.3 \\ 0.6 \\ 0.1 \end{pmatrix}$

$CE = -(0 \cdot \log 0.3 + 1 \cdot \log 0.6 + 0 \cdot \log 0.1) = -\log 0.6$