Advancements in Semi-Supervised Learning

Rahul Jaisy

Electronics & Telecommunications Engineering

Assam Engineering College, India

OUTLINE

- Abstract
- Introduction
- Related Works
- Objectives
- Methodology
- Algorithm
- Results
- Discussion
- Conclusion
- Acknowledgements
- References

ABSTRACT

- **SimPLE:** Similar Pseudo Label Exploitation for Semi-Supervised Classification introduces a method that leverages similar pseudo- labels to enhance performance on imbalanced datasets.

- **PEFAT:** Boosting Semi-Supervised Medical Image Classification via Pseudo-Loss Estimation and Feature Adversarial Training focuses on improving classification performance by integrating pseudo-loss estimation and feature adversarial training.

- **Rethinking Semi-Supervised Imbalanced Node Classification:** Using bias-variance decomposition and graph augmentation approaches, reevaluates Semi-Supervised Imbalanced Node Classification addressing the bias-variance trade-off in semi-supervised learning.

INTRODUCTION TO SEMI-SUPERVISED LEARNING

- **Context**: In many machine learning applications, obtaining labeled data is costly and time-consuming. Semi-supervised learning (SSL) leverages both labeled and a large amount of unlabeled data to improve model performance.

- **Goal:** This presentation explores three techniques in Semi-supervised learning.

- **Significance:** These approaches aim to address the scarcity of labeled data and improve the generalization of models across various domains.

- **CHALLENGES**

- Data Imbalance: Nowadays, datasets often contain a lot of unlabeled data that we're not fully utilizing with traditional methods.

- Labeling Bottleneck: It’s really expensive and time-consuming to accurately label huge amounts of data, which limits how much we can scale up supervised learning.

- Model Robustness: Without enough labeled examples, it’s hard to build models that can handle a wide range of real-world situations effectively.

- **BENEFITS**

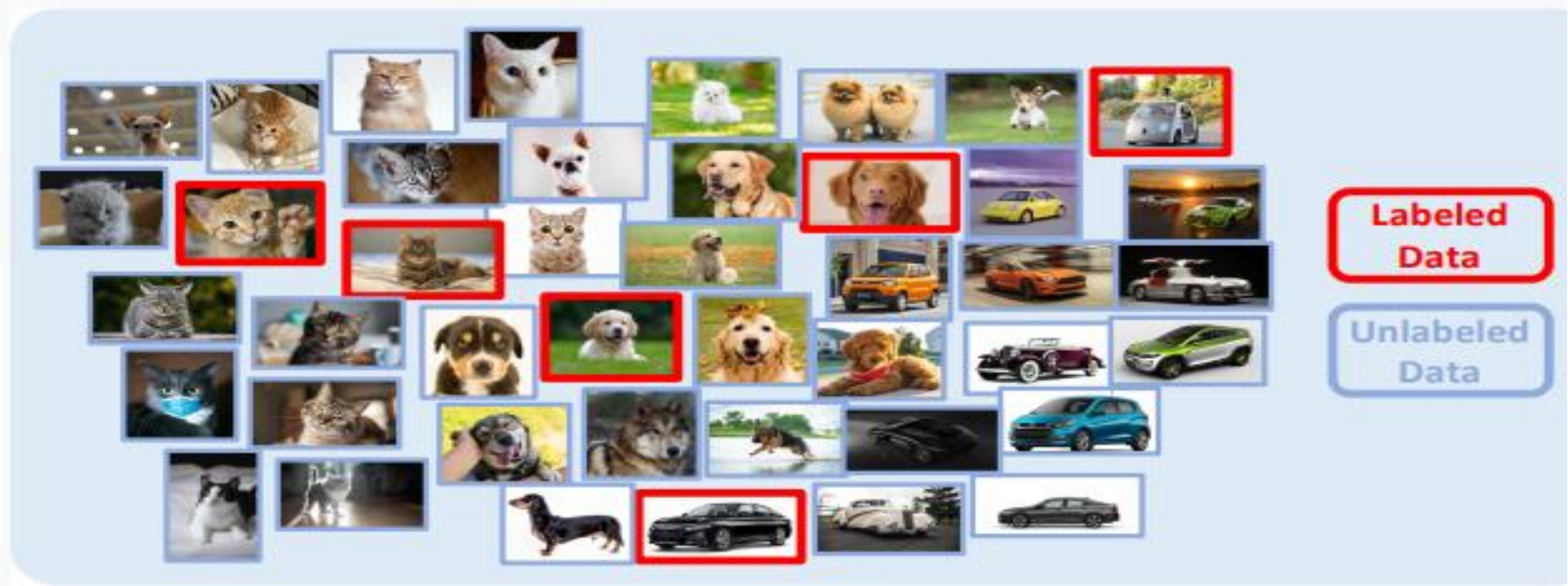
- Optimal Data Use: SSL lets us combine both labeled and unlabeled data, making the most out of all the raw data available to improve how well our models perform.

- Cost Efficiency: By needing fewer fully labeled datasets, SSL cuts down on the high costs linked to manually labeling data.

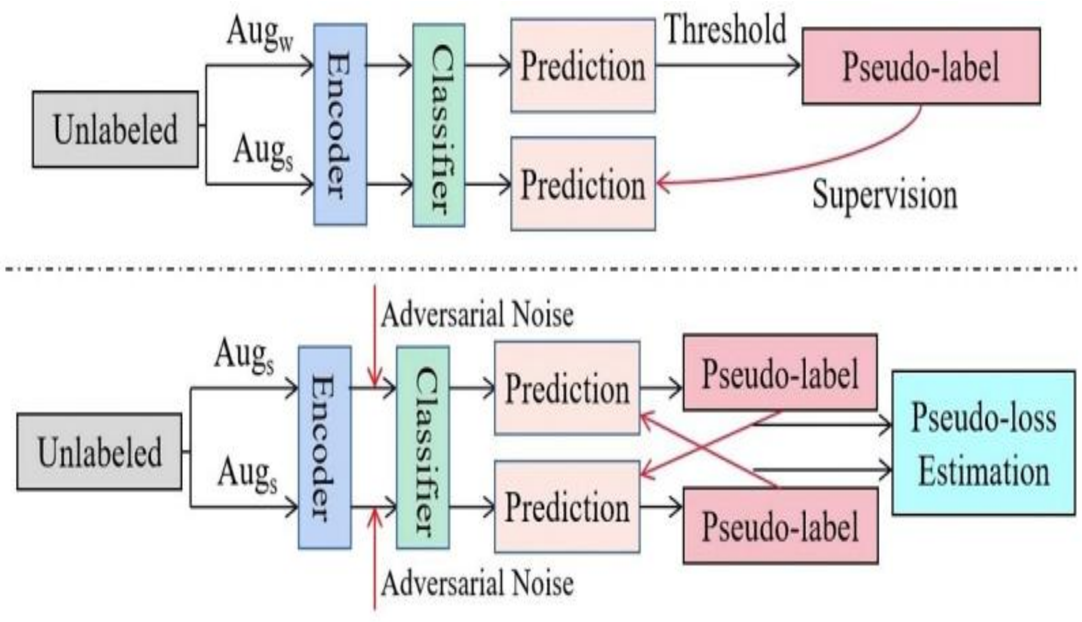
- Better Generalization: Models trained with SSL tend to generalize better, as they learn from a wider range of patterns found in unlabeled data and can apply that knowledge to new data.

INTRODUCTION

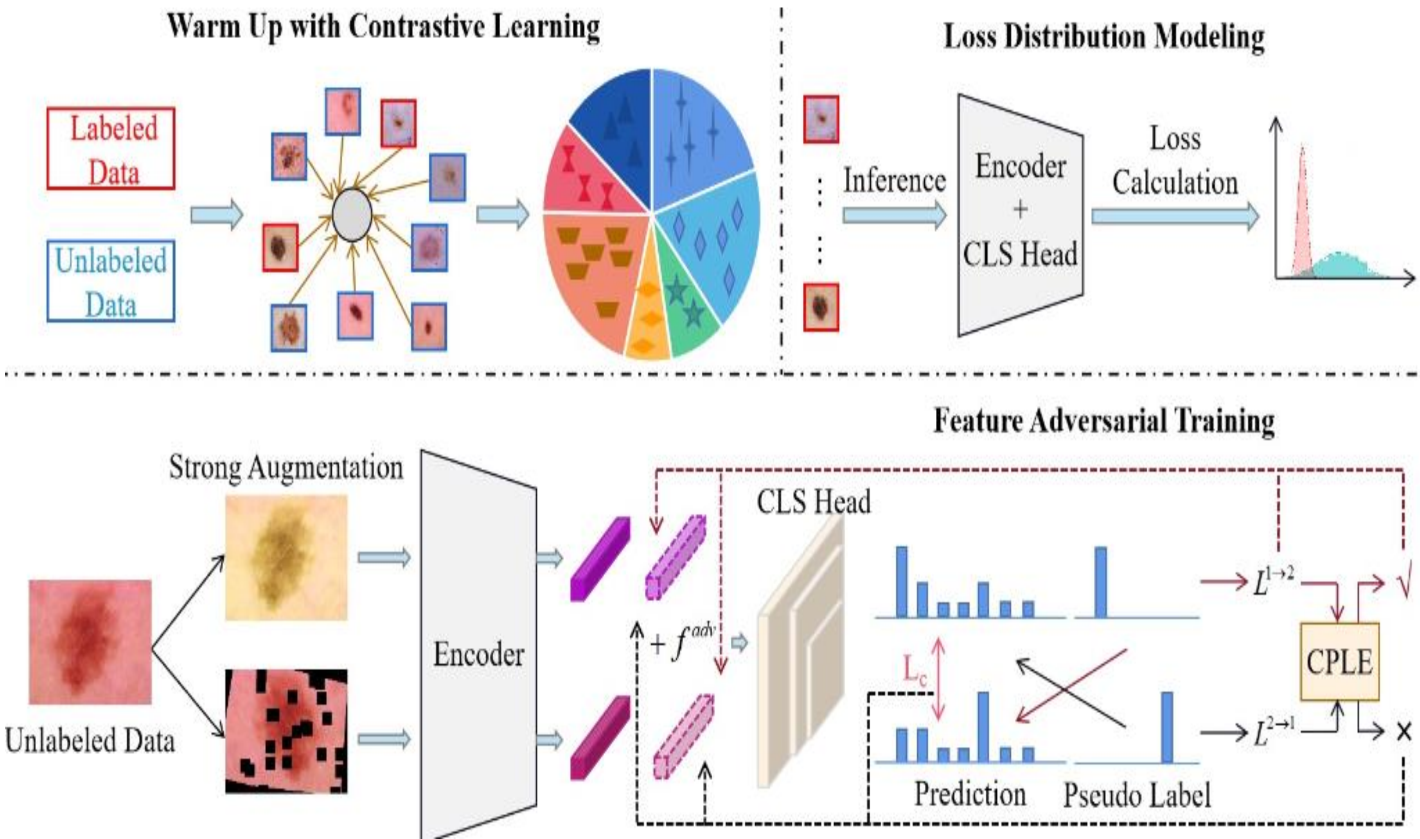
- **<u>SimPLE:</u>** Introduces the use pseudo-labels in semi-supervised learning to improve classification, particularly with unbalanced datasets, by selecting similar pseudo-labels. It integrates MixMatch and Pair Loss methods to achieve significant efficiency improvements.



- **PEFAT:** Focuses on semi-supervised learning in medical imaging, consolidating adversarial training. It utilizes trustworthy pseudo-labeled data and smooths decision boundaries to enhance classification performance.

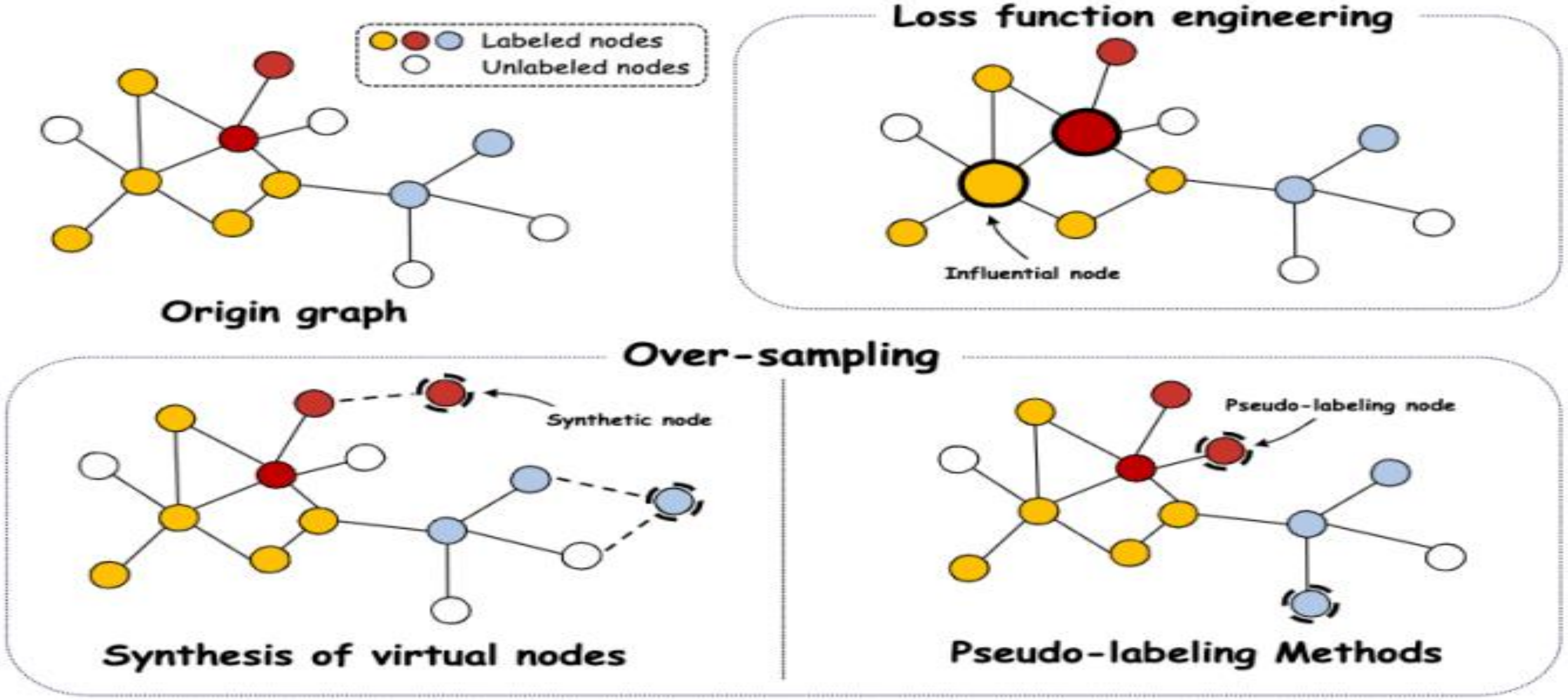


(a) Illustration of traditional SSL methods (top) and our method (bottom).



• **Rethinking Semi-Supervised Imbalanced Node Classification:**

- Examining semi-supervised imbalanced node classification helps to solve the problem of unbalanced node classification in graph neural networks. Understanding and solving model performance problems using bias-variance decomposition is essential. Regularizing methods and graph augmentation strategies can be used to properly manage data imbalances.



RELATED WORKS

- **<u>Consistency Regularization:</u>**

- Relevance: Guarantees strong resistance against disturbances and ensures model stability.

- ReMixMatch: Designed to match model responses to various perturbation levels, it introduces augmentation.

- FixMatch: Confirmed simplified pseudo-labeling with confidence criteria increases resilience.

- **<u>Pseudo-Labeling:</u>**

- Relevance: It generates and improves labels for unlabeled data, thereby enhancing learning.

- MixMatch: For improved generalization, it combines mix-up with pseudo-labeling.

- FixMatch: It filters pseudo-labels using confidence thresholds, hence improving label accuracy.

- **<u>Label Propagation:</u>**

- Relevance: Utilizes graph structures to propagate label information efficiently.

- Label Propagation: Builds graphs based on sample similarities and spreads labels through graph nodes.

- GraphSMOTE: Uses the SMOTE algorithm to handle imbalances in graph nodes by synthesizing minority nodes.

- **<u>Additional Techniques:</u>**

- Entropy Minimization: Reduces uncertainty in model predictions by focusing on confident samples.

- Generic Regularization: Applies various regularization techniques to improve model performance.

- Relevance: Improves model performance by reducing prediction uncertainty and overfitting.

OBJECTIVES

- **PURPOSE:** To provide a detailed overview and synthesis of recent advancements in semi-supervised learning through the integration of different methodologies.
- **SimPLE:** Introduce and evaluate a method to exploit similar pseudo-labels for enhanced semi-supervised classification on imbalanced datasets.
- **PEFAT:** Using adversarial training and pseudo-loss estimates to enhance understanding in semi-supervised medical image categorization.
- **Rethinking Semi-Supervised Imbalanced Node Classification:-** Explore the effects of variation and bias on semi-supervised unbalanced node classification, and develop strategies to mitigate these impacts.

METHODOLOGY

- **<u>SimPLE:</u>** Pseudo-Label Generation: Generate pseudo-labels depending on model predictions in a similar fashion.
- Similarity-Based Selection: Select similar pseudo-labels by measuring feature representation similarity.
- Model Update: Select pseudo-labels and update the model.
- Augmentation Strategy: Apply both weak and strong augmentations.
- Pair Loss Method: Add a loss function that matches strongly augmented samples with their pseudo-labels.

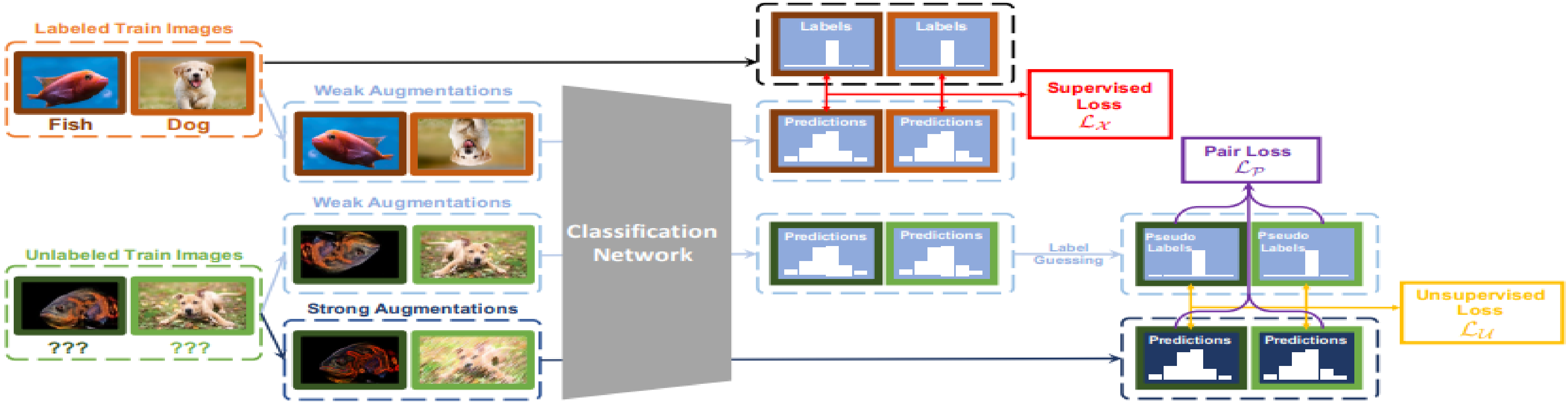


Figure 2: An overview of the proposed SimPLE algorithm. SimPLE optimizes the classification network with three training objectives: 1) supervised loss $\mathcal{L}_{\mathcal{X}}$ for augmented labeled data; 2) unsupervised loss $\mathcal{L}_{\mathcal{U}}$ that aligns the strongly augmented unlabeled data with pseudo labels generated from weakly augmented data; 3) Pair Loss $\mathcal{L}_{\mathcal{P}}$ that minimizes the statistical distance between predictions of strongly augmented data, based on the similarity and confidence of their pseudo labels.

- **PEFAT:** Pseudo-Loss Estimation: Generate pseudo-losses for unlabeled data to guide training.
- Feature Adversarial Training: Add adversarial training to the feature space.
- Data Augmentation: Apply both weak and strong augmentations to increase data diversity.
- Loss Calculation: For model optimization, combine adversarial losses with both supervised and unsupervised losses.

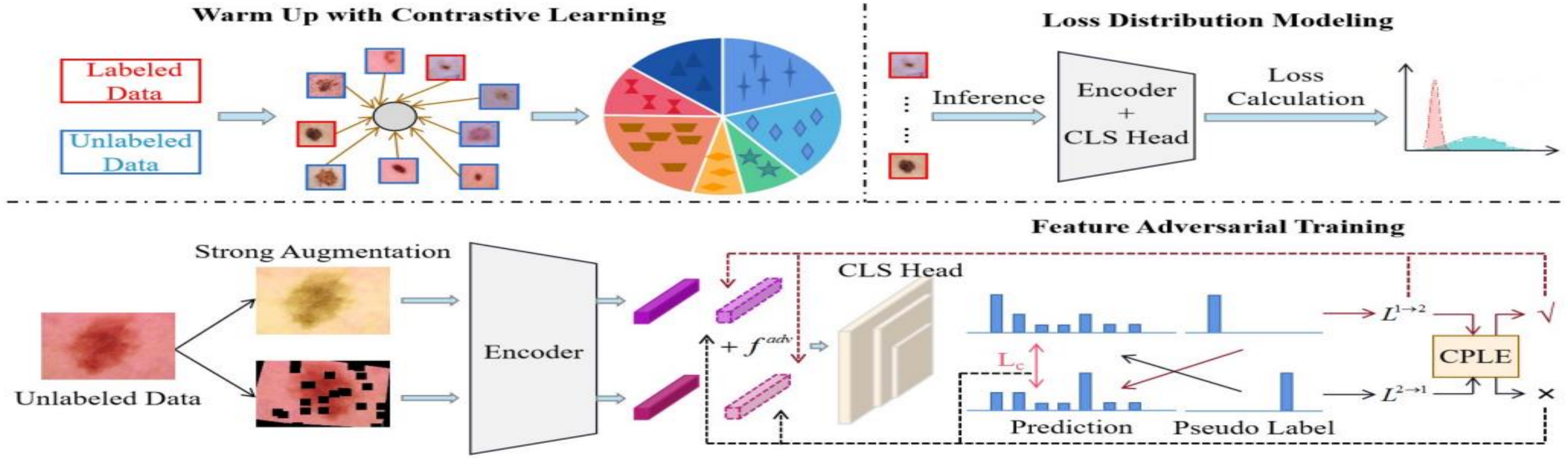


Figure 2. Illustration of our proposed PEFAT. We first warm up the model with contrastive learning on training data to learn unbiased representation. Then we set up a two-component GMM to construct the loss distribution calculated on labeled data. As for the unlabeled data utilization, we use the cross pseudo-loss estimation (CPLE) for trustworthy pseudo-labeled data exploration. Beyond that, adversarial noises are injected in the feature-level for better unlabeled data mining.

Pair Loss: Complete Form

$$\mathcal{L}_{\mathcal{P}} = \frac{1}{\binom{KB}{2}} \sum_{\mathcal{U}'} \mathbb{1}_{\max(q_l)>\tau_c} \cdot \mathbb{1}_{f_{\text{sim}}(q_l,q_r)>\tau_s} \cdot f_{\text{dist}}\left(q_l, p_\theta\left(y|v_r\right)\right)$$

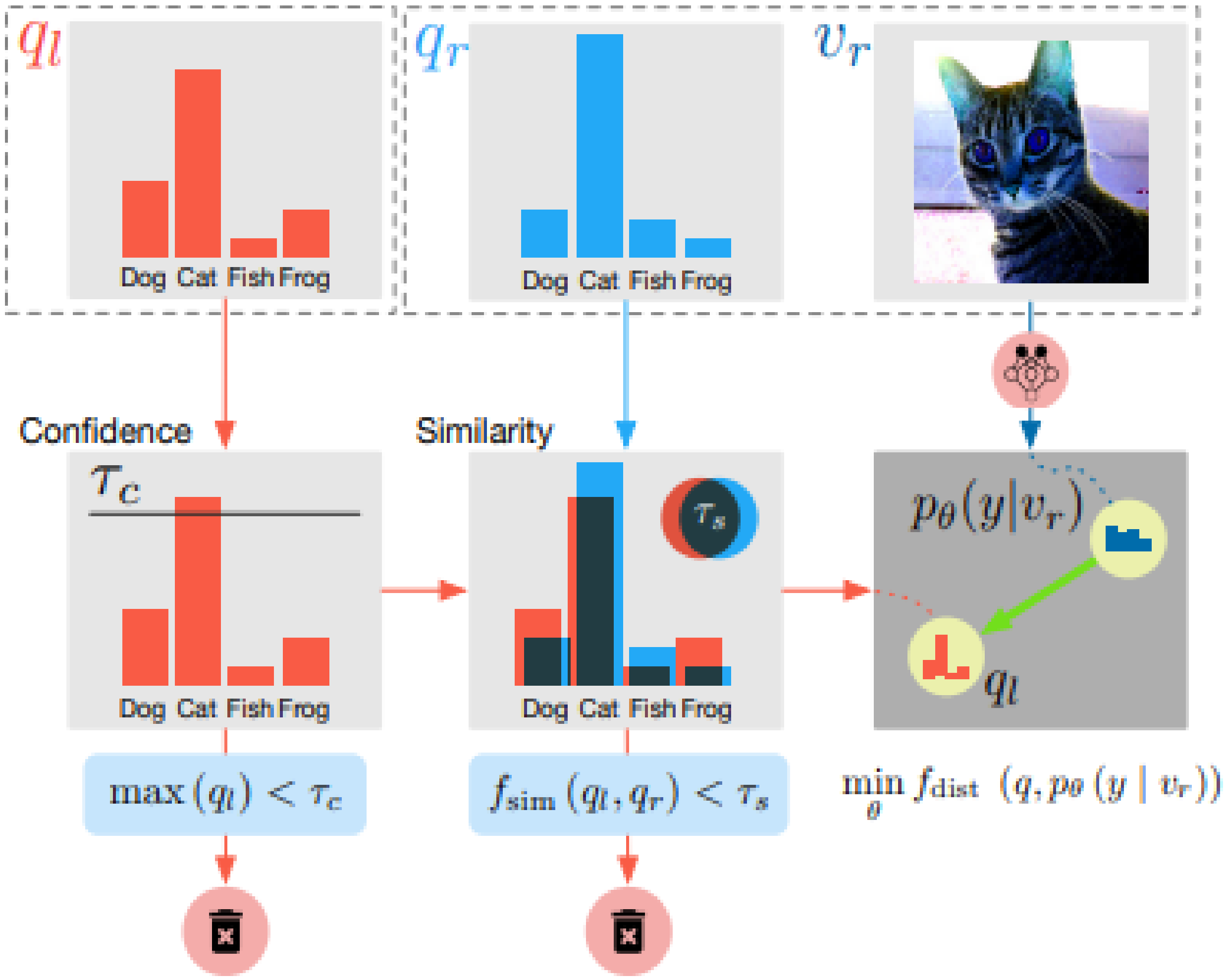
K: number of augmentations

B: mini-batch size

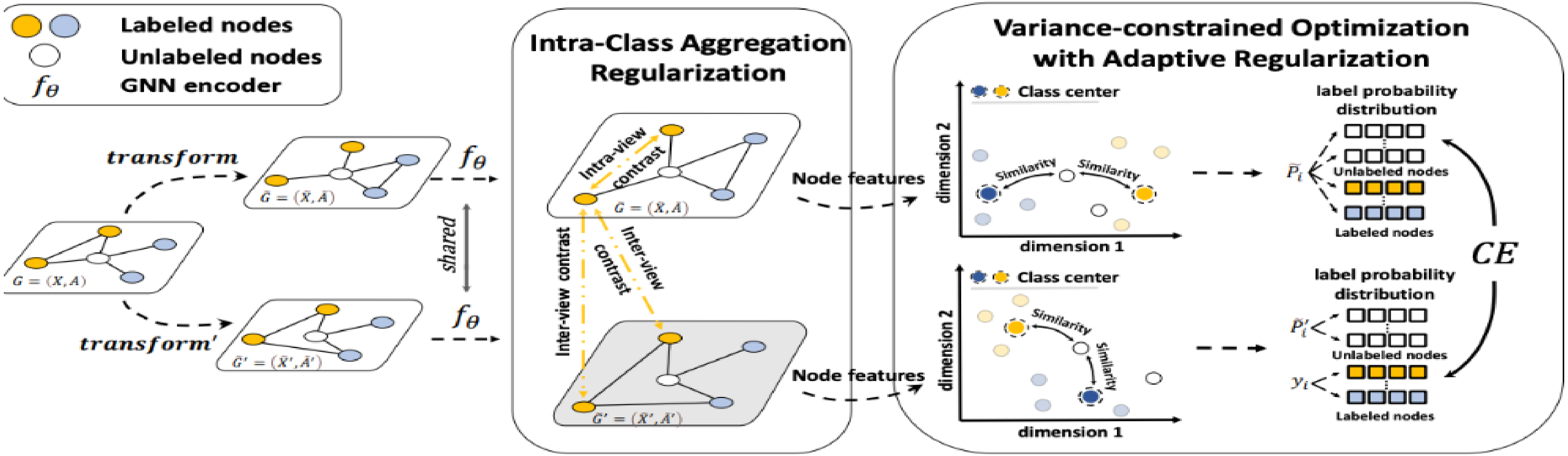
$\mathcal{U}'$: unlabeled mini-batch

$f_{\text{sim}}(\cdot,\cdot)$: similarity function

$f_{\text{dist}}(\cdot,\cdot)$: distance function



- **Rethinking Semi-Supervised Imbalanced Node Classification:-** Decomposition: Break down the bias and variance components in the classification process.
- Semi-Supervised Techniques: Apply approaches utilizing unlabeled data.
- Evaluation Metrics: To assess performance, apply the F1 score and accuracy.
- Optimization: Balance bias and variance to enhance overall performance.



(a) Graph augmentations (b) Intra-view and inter-view regularization (Corresponding to $\mathcal{L}_{IR}$) (c) Variance regularization (Corresponding to $\mathcal{L}_{VR}$)

ALGORITHM

SimPle

Algorithm 1 SimPLE algorithm

Input: Batch of labeled examples and their one-hot labels $\mathcal{X} = ((x_b, y_b); b \in (1, \ldots, B))$, batch of unlabeled examples $\mathcal{U} = (u_b; b \in (1, \ldots, B))$, sharpening temperature T, number of weak augmentations K, number of strong augmentations K_{strong}, confidence threshold τ_c, similarity threshold τ_s.
for $b = 1$ to B **do**
  $\tilde{x}_b = A_{\text{weak}}(x_b)$ ▷ Apply weak data augmentation to x_b
  for $k = 1$ to K **do**
    $\tilde{u}_{b,k} = A_{\text{weak}}(u_b)$ ▷ Apply k^{th} round of weak data augmentation to u_b
  end for
  for $k = 1$ to K_{strong} **do**
    $\hat{u}_{b,k} = A_{\text{strong}}(u_b)$ ▷ Apply k^{th} round of strong augmentation to u_b
  end for
  $\bar{q}_b = \frac{1}{K}\sum_{k=1}^{K} \mathrm{p}_{\text{model}'}(\tilde{y} \mid \tilde{u}_{b,k}; \theta)$ ▷ Compute average predictions across all **weakly** augmented u_b using EMA
  $q_b = \text{Sharpen}(\bar{q}_b, T)$ ▷ Apply temperature sharpening to the average prediction
end for
$\hat{\mathcal{X}} = ((\tilde{x}_b, y_b); b \in (1, \ldots, B))$ ▷ Weakly augmented labeled examples and their labels
$\hat{\mathcal{U}} = ((\hat{u}_{b,k}, q_b); b \in (1, \ldots, B), k \in (1, \ldots, K_{\text{strong}}))$ ▷ Strongly augmented unlabeled examples, guessed labels
$\mathcal{L}_{\mathcal{X}} = \frac{1}{|\mathcal{X}'|}\sum_{x,y \in \hat{\mathcal{X}}} H(y, \mathrm{p}_{\text{model}}(\tilde{y} \mid x; \theta))$ ▷ Compute supervised loss
$\mathcal{L}_{\mathcal{U}} = \frac{1}{L|\hat{\mathcal{U}}|}\sum_{u,q \in \hat{\mathcal{U}}} \mathbb{1}_{(\max(q) > \tau_c)} \| q - \mathrm{p}_{\text{model}}(\tilde{y} \mid u; \theta)\|_2^2$ ▷ Compute thresholded unsupervised loss
$\mathcal{L}_{\mathcal{P}} = \text{PairLoss}\left(\hat{\mathcal{U}}, \tau_c, \tau_s\right)$ ▷ Compute Pair Loss
return $\mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}}\mathcal{L}_{\mathcal{U}} + \lambda_{\mathcal{P}}\mathcal{L}_{\mathcal{P}}$ ▷ Compute loss $\mathcal{L}$ from $\hat{\mathcal{X}}$ and $\hat{\mathcal{U}}$

PEFAT

Algorithm 1: PEFAT Algorithm

Input: Labeled dataset D_l; unlabeled dataset D_u; initialized model h_θ.
Initialize a two-componet GMM;
Warm up h_θ with Eq. (1) and Eq. (2);
for $(x_i, y_i) \in D_l$ **do**
  Calculate loss l_{x_i} according to Eq. (3);
end
Fit GMM with $\{l_{x_i}\}_{i=1}^{N_l}$ with Eq. (4) and Eq. (5);
for $u_i \in D_u$ **do**
  Make cross prediction by Eq. (6) and Eq. (7);
  Get cross pseudo-loss by Eq. (8) and Eq. (9);
  Obtain p_{gmm} according to Eq. (10);
  if $p_{gmm}^{k=0} > p_{gmm}^{k=1}$ **then**
    Calculate L_{FAT} and L_{ce} with pseudo-label;
  else
    Calculate L_{FAT} with Eq. (13);
  end
end
Return h_θ;

- **Rethinking Semi-Supervised Imbalanced Node Classification:-**

1. $$\mathcal{L}_{\mathrm{VR}} = \frac{1}{|V_{\mathrm{conf}}|}\sum_{i\in V_{\mathrm{conf}}} CE\left(\tilde{\pi}'_i, \tilde{\pi}_i\right) + \frac{1}{|V_L|}\sum_{i\in V_L} CE\left(y_i, \tilde{\pi}_i\right)$$

2. $$\mathcal{L}_{\mathrm{IR}} = -\frac{1}{|V_U|}\sum_{\mathrm{h}_i,\mathrm{h}'_i\in V_U} \mathrm{sim}\left(\mathrm{h}_i \cdot \mathrm{h}'_i\right) - \frac{1}{N_{all}}\left(\sum_{l=1}^{K}\sum_{\mathrm{h}_i,\mathrm{h}'_j\in \mathbf{C}_l} \mathrm{sim}\left(\mathrm{h}_i \cdot \mathrm{h}'_j\right) + \sum_{l=1}^{K}\sum_{\substack{\mathrm{h}_i,\mathrm{h}_j\in \mathbf{C}_l \\ i\neq j}} \mathrm{sim}\left(\mathrm{h}_i \cdot \mathrm{h}_j\right)\right)$$

3. $$\mathcal{L}_{\mathrm{composite}} = \lambda_1\mathcal{L}_{\mathrm{VR}} + \lambda_2\mathcal{L}_{\mathrm{IR}} + \mathcal{L}_{\mathrm{sup}}$$

RESULTS

SimPLE:

Datasets: CIFAR-10, SVHN, CIFAR-100

Metrics: Test accuracy (%)

Results:

- Competitive performance compared to FixMatch and ReMixMatch with a difference of less than 1%.
- Slightly underperformance on CIFAR-100 brought on by demanding samples.
- Convergence speed Advantage: Converges in 4.7 hours on Mini-ImageNet against FixMatch's 8 hours with convergence speed advantage.
- Consistency: SimPLE consistently achieved the highest Top-1 Accuracy across all datasets.
- Superiority: SimPLE's performance surpasses that of MixMatch, ReMixMatch, and FixMatch approaches.
- Robustness: Strong results in transfer contexts and several dataset scales show robustness.

Experiment Result

Dataset	# Labels	Method	Backbone	Top-1 Accuracy
CIFAR-100	10000	MixMatch	WRN 28-8	71.69%
		ReMixMatch	WRN 28-8	76.97%
		FixMatch	WRN 28-8	77.40%
		SimPLE	**WRN 28-8**	**78.11%**
Mini-ImageNet	4000	MixMatch	WRN 28-2	55.47%
		MixMatch Enhanced	WRN 28-2	60.50%
		SimPLE	**WRN 28-2**	**66.55%**
ImageNet to DomainNet-Real	3795	MixMatch	ResNet-50	35.34%
		MixMatch Enhanced	ResNet-50	35.16%
		SimPLE	**ResNet-50**	**50.90%**
DomainNet-Real to Mini-ImageNet	4000	MixMatch	WRN 28-2	53.39%
		MixMatch Enhanced	WRN 28-2	55.75%
		SimPLE	**WRN 28-2**	**58.73%**

*** Shaded rows are in transfer setting**

PEFAT

- **Datasets:** CT-CRC-HE, ISIC2018, Chest X-Ray
- **Metrics:** AUC, Sensitivity, Specificity, Accuracy, F1-score
- **Results:** NCT-CRC-HE (200 labeled images)

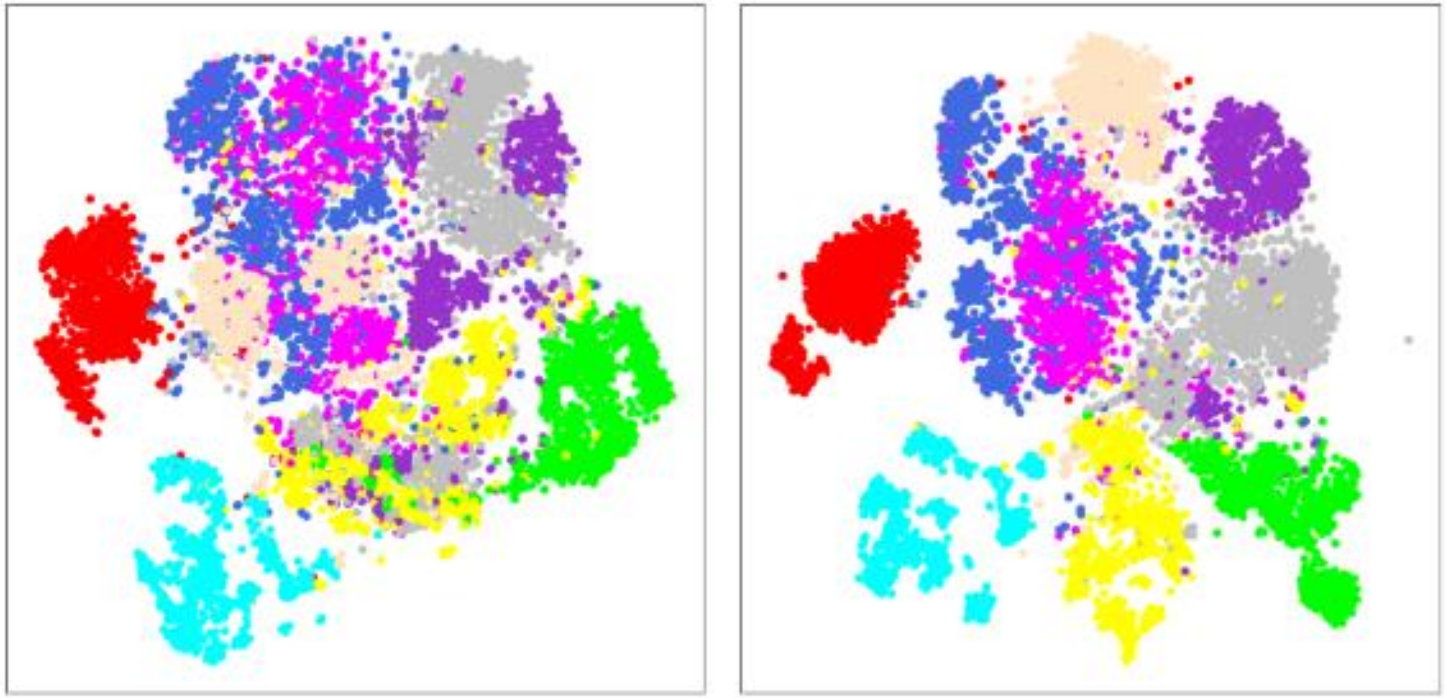
 - Accuracy: 90.29% (PEFAT) vs. 81.63% (baseline), improvement of 8.66%.
 - Sensitivity: 89.68% (PEFAT) vs. 78.12% (baseline), improvement of 11.56%.
 - Specificity: 91.18% (PEFAT) vs. 83.06% (baseline), improvement of 8.12%.
 - PEFAT demonstrates significant improvements across all metrics and datasets.

CPLE-FAT

- **Datasets:** NCT-CRC-HE, Skin Cancer Image Classification.
- **Metrics:** AUC, Sensitivity, Precision, Accuracy, F1-score.
- **Results:** NCT-CRC-HE (200 labeled images)

 - Accuracy: 86.33% (CPLE-FAT) vs. 73.29% (baseline), improvement of 13.04%.
 - Precision: 85.76% (CPLE-FAT) vs. 76.25% (baseline), improvement of 9.51%.
 - F1-score: 85.73% (CPLE-FAT) vs. 73.48% (baseline), improvement of 12.25%.

VISUALIZATION



(a) (b)

Figure 4. The t-SNE visualization on NCT-CRC-HE validation set. (a) is the result when using VAT; (b) shows the feature embedding when using FAT.

Rethinking Semi-Supervised Imbalanced Node Classification from Bias-Variance Decomposition:

- **ReVar's Performance:**
- **Datasets:** CiteSeer-Semi, Computers-Semi, Pubmed-Semi
- **Models:** GCN, GAT, GraphSAGE
- **Metrics:** Balanced accuracy (bAcc%), F1-score (%)
- **Results:**

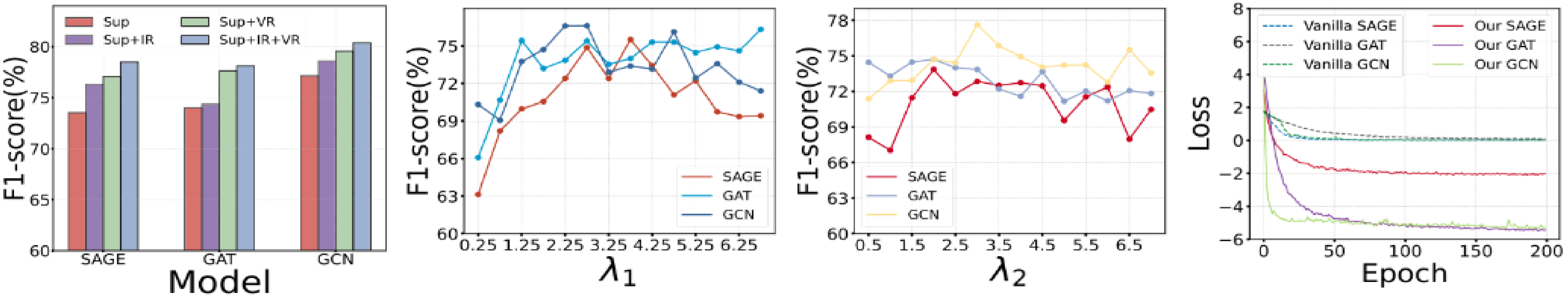
- CiteSeer-Semi (GCN): ReVar achieves 65.28% bAcc and 79.29% F1-score vs. baseline's 38.72% bAcc and 28.74% F1-score. Improvement: 26.56% (bAcc) and 50.55% (F1-score).

- ReVar consistently outperforms across all datasets and models, with significant statistical improvements.

- For example, on CS-Random with the GNN model: ReVar achieves 82.14% bAcc vs. baseline's 68.43% (13.71% improvement).

- Convergence: ReVar shows faster loss convergence compared to the vanilla model using only cross-entropy loss.

Dataset(*CS-Random*)	GCN		GAT		SAGE	
Imbalance Ratio($\rho = \mathbf{41.00}$)	bAcc.	F1	bAcc.	F1	bAcc.	F1
Vanilla	84.85 $\pm$ 0.16	87.12 $\pm$ 0.14	82.47 $\pm$ 0.36	84.21 $\pm$ 0.31	83.76 $\pm$ 0.27	86.22 $\pm$ 0.19
Re-Weight	87.42 $\pm$ 0.17	88.70 $\pm$ 0.10	83.55 $\pm$ 0.39	84.73 $\pm$ 0.32	85.76 $\pm$ 0.24	87.32 $\pm$ 0.16
PC Softmax	88.36 $\pm$ 0.12	88.94 $\pm$ 0.04	85.22 $\pm$ 0.31	85.54 $\pm$ 0.33	87.18 $\pm$ 0.14	88.00 $\pm$ 0.19
GraphSMOTE	85.76 $\pm$ 1.73	87.31 $\pm$ 1.32	84.65 $\pm$ 1.32	85.63 $\pm$ 1.01	85.76 $\pm$ 1.98	87.34 $\pm$ 0.98
BalancedSoftmax	87.72 $\pm$ 0.07	88.67 $\pm$ 0.07	84.38 $\pm$ 0.20	84.53 $\pm$ 0.41	86.78 $\pm$ 0.10	88.05 $\pm$ 0.09
+ TAM	88.22 $\pm$ 0.11	89.22 $\pm$ 0.08	85.48 $\pm$ 0.24	85.77 $\pm$ 0.50	87.83 $\pm$ 0.13	88.77 $\pm$ 0.07
Renode	87.53 $\pm$ 0.11	88.91 $\pm$ 0.06	85.98 $\pm$ 0.19	86.97 $\pm$ 0.09	86.13 $\pm$ 0.10	87.89 $\pm$ 0.09
+ TAM	87.55 $\pm$ 0.06	89.03 $\pm$ 0.05	86.61 $\pm$ 0.30	87.42 $\pm$ 0.24	85.21 $\pm$ 0.33	87.01 $\pm$ 0.31
GraphENS	85.97 $\pm$ 0.29	86.68 $\pm$ 0.20	85.86 $\pm$ 0.19	86.51 $\pm$ 0.32	85.39 $\pm$ 0.26	86.41 $\pm$ 0.24
+ TAM	86.34 $\pm$ 0.12	87.36 $\pm$ 0.08	86.29 $\pm$ 0.20	87.28 $\pm$ 0.13	85.99 $\pm$ 0.13	87.25 $\pm$ 0.07
ReVar	**88.44** $\pm$ 0.16	**89.54** $\pm$ 0.11	**87.33** $\pm$ 0.04	**88.33** $\pm$ 0.06	**90.11** $\pm$ 0.11	**91.18** $\pm$ 0.11
Δ	**+ 0.08**(0.09%)	**+ 0.32**(0.36%)	**+ 0.72**(0.83%)	**+ 0.91**(1.04%)	**+ 2.28**(2.60%)	**+ 2.41**(2.71%)



(a) *Computers-Semi* (b) *Pubmed-λ_1* (c) *PubMed-λ_2* (d) *CiteSeer-Loss*

Figure 3: Analysis of ReVar.

- **Skin Cancer Image Classification:**

  - Highest accuracy (93.68%) compared to other SSL methods with 20% labeled data.

  - Significant improvement over baseline accuracy (78.90%) with limited labeled data.

  - Maintains high accuracy (>93%) across various label percentages (2%, 5%, 10%, 15%, 20%).

DISCUSSION

- **<u>SimPLE:</u>**

- Effectiveness: Use of same pseudo-labels selectively improves classification accuracy by more precisely matching model predictions with actual labels.

- Impact: It effectively addresses class imbalance to produce on a variety of datasets more consistent and fair model performance.

- **<u>PEFAT:</u>**

- Effectiveness: Combining feature adversarial training with pseudo-loss estimation has produced effective categorization of medical images.

- Impact: This approach increases the application of the model in the real world by enhancing its generalization and resilience, especially on unbalanced medical datasets.

- **<u>Rethinking Semi-Supervised Imbalanced Node Classification:</u>**

- Effectiveness: It is shown that the model performs better generally when variance components and bias are broken down.

- Impact: Balance of bias and variance considerably increases the accuracy and efficacy of semi-supervised learning methods.

CONCLUSION

- **SimPLE:** Simples the Pair Loss objective by minimizing the statistical distance between pseudo-labels, hence boosting semi-supervised learning and hence classification accuracy.
- **PEFAT:** Introduces pseudo-loss estimation and adversarial training to improve semi-supervised learning in medical picture categorization.
- **Rethinking Semi-Supervised Imbalanced Node Classification:** Offers theoretical understanding to address imbalanced node classification, directly connecting data imbalance to model variance for optimal performance.
- **Performance Improvements:** Outfits current approaches by significantly lowering bias and variance over several datasets and by improving accuracy.
- **Future Research:** Suggests ongoing investigation on advanced pseudo-labeling strategies, modified adversarial training methodologies, and the construction of robust bias-variance decomposition frameworks for improved semi-supervised learning.

ACKNOWLEDGEMENTS

- **Paper Title:** SimPLE: Similar Pseudo Label Exploitation for Semi-Supervised Classification

- Authors: Zijian Hu, Zhengyu Yang, Xuefeng Hu, Ram Nevatia

- Affiliation: University of Southern California

- Presented at: CVPR, 2021

- **Paper Title:** PEFAT: Boosting Semi-supervised Medical Image Classification via Pseudo-loss Estimation and Feature Adversarial Training

- Authors: Qingjie Zeng, Yutong Xie, Zilin Lu, Yong Xia

- Affiliation: School of Computer Science and Engineering, Northwestern Polytechnical University, China & The University of Adelaide, Australia

- Presented at: CVPR, 2023

- **Paper Title:** Rethinking Semi-Supervised Imbalanced Node Classification from Bias-Variance Decomposition

- Authors: Divin Yan , Gengchen Wei , Chen Yang , Shengzhong Zhang , Zengfeng Huang

- Affiliation: Fudan University

- Presented at: NeurIPS, 2023

REFERENCES

- https://proceedings.neurips.cc/paper_files/paper/2023/hash/5d1233f819202ade06023346df80a6d2-Abstract-Conference.html
- https://ieeexplore.ieee.org/document/9577604
- https://ieeexplore.ieee.org/document/10203490

THANK YOU