
Supplementary Materials

Anonymous Author(s)

Affiliation

Address

email

1 Chapter 1 Understanding The Noisy Labels In Pre-Training Data

We provide additional experiment details for the motivating example of ResNet-50 in this section. We also present the detailed results on each downstream dataset for noisy pre-trained models on both ImageNet-1K and YFCC15M.

1.1 Pre-training datasets and Hyper-parameters

For analysis in Section 2, we conduct pre-training of ResNet-50 on ImageNet-1K and YFCC15M. For ImageNet-1K pre-training, we follow the training recipe in Wightman et al. (2021). To introduce noise in ImageNet-1K, we use function cleanlab (Northcutt et al., 2021) to introduce symmetric noise in each class. For YFCC15M CLIP pre-training, we follow the training recipe in Cherti et al. (2023). To introduce noise in YFCC15M, we swap the text description between two randomly sampled image-text pairs until the noise ratio is achieved. We show the validation accuracy on ImageNet-1K of the noisy ResNet-50 models pre-trained on ImageNet-1K and zero-shot accuracy on ImageNet-1K of the noisy ResNet-50 models pre-trained on YFCC15M in Table 3. The results show that our pre-training achieves the state-of-the-art results (Wightman et al., 2021; Cherti et al., 2023), as a basis for our further analysis.

1.2 Downstream Vision Datasets and Hyper-parameters

We present the details of the in-domain (ID) vision datasets in Table 4 and out-of-domain vision datasets Table 5. For ID, we conduct training on the training set and test on the validation set of the downstream dataset. For OOD on DomainNet (Peng et al., 2019), we conduct training on the training set of DomainNet Real or DomainNet Sketch, and test on all the other three DomainNet datasets not used in training. For OOD on ImageNet (Russakovsky et al., 2015), we conduct training on ImageNet training split and test on its variants. To transfer a pre-trained model, we use linear probing (LP) for analysis as shown in Section 2. We train the linear classifier for 30 epochs on each downstream dataset, using AdamW (Kingma Ba, 2014) optimizer with a cosine scheduler. We do not use weight decay for linear probing and set the learning rate to 0.1 for all tasks.

Table 1: ImageNet-1K validation and zero-shot accuracy of ImageNet-1K pre-trained and YFCC15M CLIP pre-trained noisy ResNet-50 models.

Noise Ratio	ImageNet-1K Pre-train Validation Accuracy	YFCC15M CLIP Pre-train Zero-shot Accuracy
0%	79.96	32.64
5%	79.18	30.86
10%	78.61	29.54
20%	76.27	27.72
30%	73.11	26.53

Table 2: Details of the 3 medical vision datasets used to evaluate transfer performance.

Dataset	Classes	Train Size	Test Size	Evaluation Metric
Camelyon17	2	302,436	33,501	accuracy
HAM10000	7	8,138	2,000	accuracy
NIH ChestX-ray	14	89,322	25,596	accuracy

1.3 Detailed results

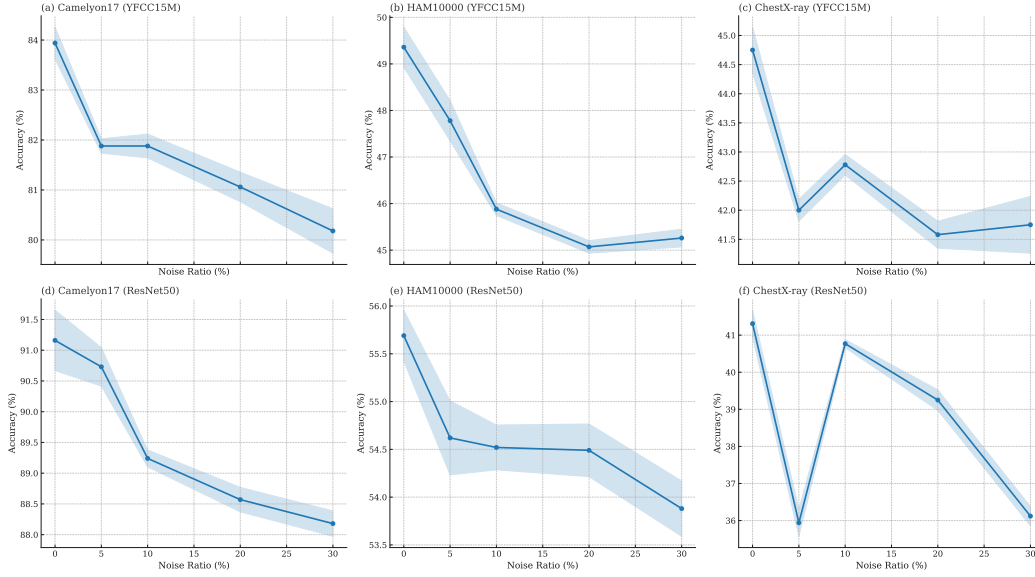


Figure 1: Average evaluation results of ImageNet-1K (IN-1K) fully supervised pre-training on downstream tasks with various percentages of data using ResNet-50. The robustness performance constantly decreases once noise is introduced in pre-training.

Figure 1 reports the detailed accuracy trends across different levels of synthetic label noise for YFCC15M and ResNet-50 pre-training. Across all settings, SKD demonstrates consistent robustness compared to LP and NML. On Camelyon17, accuracy under YFCC15M pre-training (Figure 1a) remains above 83% up to 30% noise, with SKD maintaining a clear margin over both baselines. For HAM10000 (Figure 1b), the performance is more sensitive to noise, but SKD still preserves $\sim 2\text{--}3\%$ improvement under high-noise regimes. A similar trend is observed on ChestX-ray (Figure 1c), where accuracy degrades steadily, but SKD slows the collapse. Under supervised ResNet-50 pre-training (Figures 1d–f), SKD consistently achieves top performance across all datasets. For example, on Camelyon17 (Figure 1d), accuracy stays above 90% even at 30% label noise, whereas LP and NML drop significantly. These trends confirm the effectiveness of SKD in preserving feature integrity under both contrastive and supervised pre-training paradigms, particularly in safety-critical medical tasks.

1.4 Detailed Feature and Logit Results

Skewness and kurtosis degradation under noise. We analyze the feature distributional changes of ResNet-50 models pre-trained with varying noise ratios by reporting the mean and variance of skewness and kurtosis on Camelyon17, HAM10000, and NIH ChestX-ray datasets. As shown in Figure 2, both skewness and kurtosis consistently degrade with increasing label noise. On Camelyon17, the skewness mean drops from 6.00 to 3.61, and kurtosis mean plummets from 110.02 to 57.28 as noise increases from 0% to 30%, accompanied by a sharp decline in kurtosis variance from 232.15 to 93.66. Similar trends are observed on HAM10000 (skewness mean: 5.49 \rightarrow 3.28; kurtosis mean: 78.05 \rightarrow 43.71) and NIH ChestXray (skewness mean: 5.11 \rightarrow 3.12; kurtosis mean:

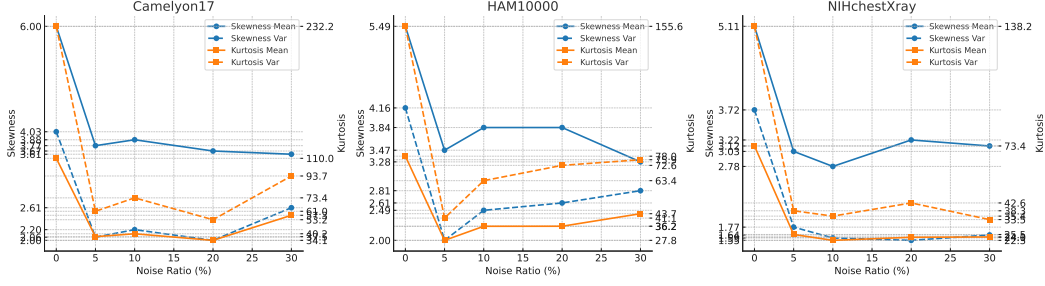


Figure 2: **Degradation of skewness and kurtosis under label noise.** We visualize the mean and variance of feature-wise skewness and kurtosis for ResNet-50 pre-trained with different levels of label noise. All three datasets (Camelyon17, HAM10000, NIH ChestXray) exhibit a consistent downward trend in both metrics as noise increases, indicating a flattening of the representation space. Skewness becomes closer to zero (more symmetric), and kurtosis drops significantly (less peaky), reflecting reduced expressiveness and discriminative structure.

73.44 \rightarrow 24.14). These shifts indicate a progressive flattening of the representation space under noise—feature dimensions become more symmetric (lower skewness) and less heavy-tailed (lower kurtosis), pointing to a collapse of expressive capacity. Such structural degradation motivates our SKD design, which explicitly regularizes these higher-order moments to preserve discriminative geometry.

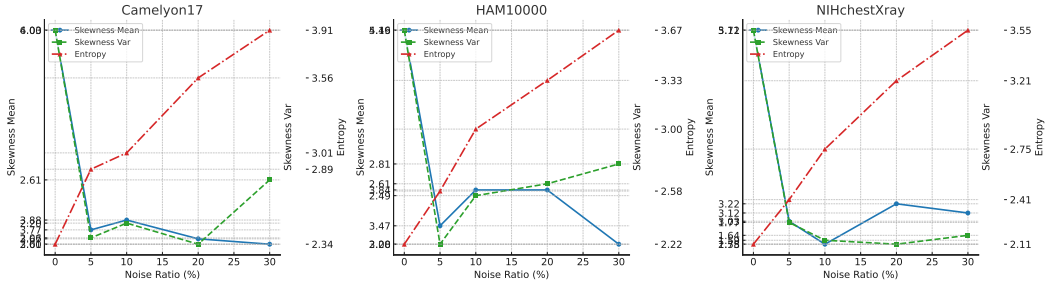


Figure 3: **Logit-level changes under pre-training noise.** We report logit entropy, energy, and maximum softmax probability (MSP) for CLIP models pre-trained with different noise levels on Camelyon17, HAM10000, and NIH ChestXray. As noise increases, entropy rises while both energy and MSP decrease, indicating higher prediction uncertainty and lower confidence. These trends mirror the flattening of feature-level skewness and kurtosis, suggesting that representational degradation propagates through to the output space.

Logit-level analysis. We examine how pre-training noise affects the distribution of logit outputs across different datasets. As shown in Figure 3, increasing noise levels lead to a consistent rise in logit entropy (entropy_mean), reflecting increased prediction uncertainty. For example, in NIH ChestXray, entropy increases from 1.5710 at 0% noise to 1.6830 at 30% noise. Simultaneously, both logit energy (energy_mean) and maximum softmax probability (msp_mean) decrease, indicating lower confidence and greater dispersion in the predictions. On Camelyon17, energy drops from 5.1151 to 4.2888, while msp_mean falls from 0.9301 to 0.9006.

These patterns are tightly connected to the structural degradation observed in feature-level statistics—specifically, reduced skewness and kurtosis under noisy supervision. Lower skewness suggests more symmetric and less distinctive feature distributions, whereas reduced kurtosis indicates a lack of peakedness and diminished confidence concentration. Together, these shifts in the representation space translate into softer and more ambiguous logit-level predictions, underscoring the downstream impact of representational collapse. Our results support the view that structural noise inherited during

pre-training propagates through to the output layer, impairing model reliability in high-stakes clinical settings.

2 Chapter 2 Experiment

More details of experiments in Section 4 are shown here.

Table 3: Details of the biomedical NER datasets used for evaluation. Each dataset provides token-level annotations for domain-specific entity types (e.g., diseases, chemicals, genes). Following prior work, we report both F1 score and accuracy.

Dataset	Entity Type	Annotation Scheme	Evaluation Metric
BC2GM	Gene/protein	IOB	F1, Accuracy
BC4CHEMD	Chemicals	IOB	F1, Accuracy
CRAFT	Multiple (e.g., cell, gene)	IOB	F1, Accuracy
BC5CDR-chem	Chemicals	IOB	F1, Accuracy
BC5CDR-disease	Diseases	IOB	F1, Accuracy
JNLPBA	Biomedical terms	IOB	F1, Accuracy
NCBI-disease	Diseases	IOB	F1, Accuracy
BioNLP11	Multiple event/mention types	IOB	F1, Accuracy
BioNLP13	Multiple entity types (species, cell, etc.)	IOB	F1, Accuracy
Ex-PTM	Protein post-translational modifications	IOB	F1, Accuracy
AnatEM	Anatomical entities	IOB	F1, Accuracy

2.1 Detailed Setup For Language Model Experiment

To assess the generalizability of our method beyond vision, we evaluate its effectiveness in natural language processing (NLP) via biomedical named entity recognition (NER). Specifically, we use the BiomedBERT model as our encoder backbone. This model is a domain-adapted variant of BERT, pre-trained on PubMed abstracts and PMC full-text articles, making it well-suited for biomedical language tasks.

We conduct experiments across 32 standard biomedical NER benchmarks, including BC2GM, BC4CHEMD, NCBI-disease, JNLPBA, and multiple BioNLP and CRAFT datasets. These datasets cover a wide range of biomedical entity types such as genes, proteins, chemicals, diseases, and anatomical structures. A summary of representative datasets is provided in Table 3, while full results across all benchmarks are included in the supplementary materials.

We compare our proposed SKD method against linear probing (Baseline) and a recent baseline NML, reporting both F1 score and accuracy. Results demonstrate that SKD consistently improves upon prior methods across nearly all datasets, highlighting its robustness and transferability in language settings.

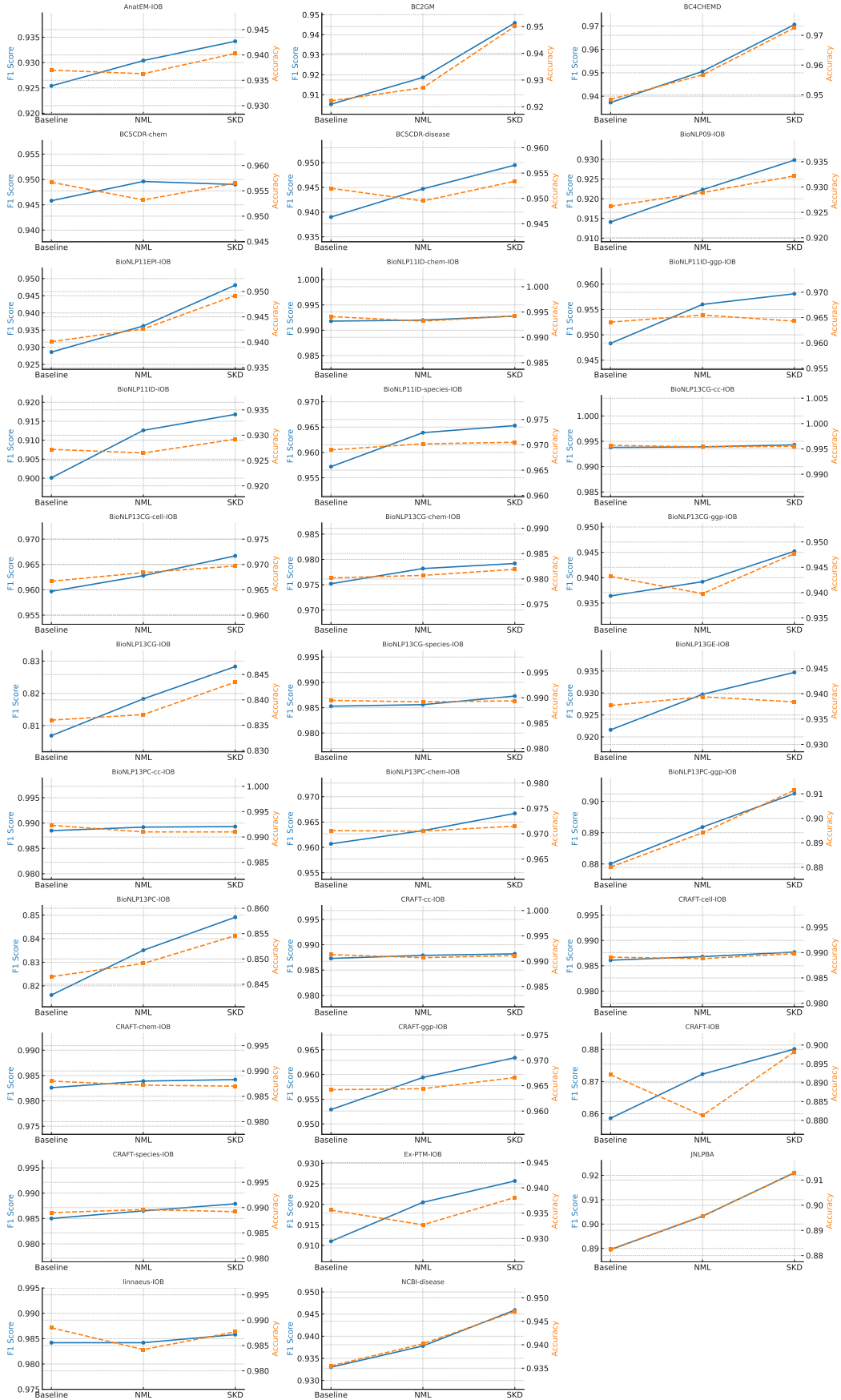


Figure 4: **F1 and Accuracy on Biomedical NER Benchmarks.** We report F1 score and accuracy for 32 biomedical NER datasets using the BioMedBERT backbone under three fine-tuning strategies: linear probing (LP), NML, and our proposed SKD.

Biomedical NER Results. We evaluate the effectiveness of our method SKD on 32 biomedical named entity recognition (NER) datasets spanning various entity types (e.g., chemical, gene/protein, species, disease) from benchmark corpora such as BioNLP, CRAFT, BC2GM, and NCBI-disease. All experiments use the BioMedNLP-BioMedBERT-base-uncased-abstract-fulltext model as the backbone, with fine-tuning conducted using three strategies: LP (Baseline), NML, and our proposed SKD.

As shown in Figure 4, SKD achieves consistent improvements across both F1 and accuracy metrics. For instance, on BC2GM, SKD achieves an F1 score of 0.9459 compared to 0.9053 (LP) and 0.9187 (NML); on BC4CHEMD, SKD further improves the F1 to 0.9706. Across all datasets, SKD outperforms prior methods by an average margin of **1.21%** in F1 and **0.57%** in accuracy. These results highlight SKD’s ability to enhance robustness and representation quality in biomedical NLP tasks under noisy transfer settings.

3 Chapter 3 In-depth Analysis

3.1 Ablation study

Table 4: **Ablation study on SKD components using ResNet-50.** We evaluate variants of SKD by selectively removing components: skewness (S), kurtosis (K), and disagreement (D), across three medical datasets. The full SKD consistently achieves the best performance under varying pre-training noise ratios, confirming the complementary benefits of each regularizer.

Dataset	Method	0	5	10	20	30	Avg
Camelyon17	S	90.10	91.05	88.24	88.77	88.10	89.25
	K	91.41	92.53	88.70	90.57	88.06	90.25
	D	91.02	90.77	89.01	88.27	88.78	89.57
	S&K	91.52	92.11	90.72	88.92	88.79	90.41
	S&D	91.09	91.54	90.22	87.88	87.74	89.69
	K&D	91.28	92.13	90.72	88.64	88.69	90.29
	SKD	91.76	92.50	91.39	89.12	89.02	90.76
HAM10000	S	56.02	55.25	55.55	57.53	55.72	56.01
	K	56.14	55.21	54.74	57.03	56.08	55.84
	D	55.95	53.73	55.62	56.45	55.17	55.38
	S&K	57.34	55.98	55.42	57.90	56.87	56.70
	S&D	56.87	56.45	55.68	57.44	56.15	56.52
	K&D	56.96	56.05	55.75	57.37	56.54	56.53
	SKD	58.25	57.54	56.94	58.37	57.54	57.73
NIHchestXray	S	41.75	39.90	40.66	41.20	41.91	41.08
	K	43.85	41.98	42.52	43.18	43.45	43.00
	D	41.70	38.56	41.02	40.35	38.23	39.97
	S&K	43.80	41.90	42.56	43.23	43.45	42.99
	S&D	43.57	42.08	42.24	42.89	43.58	42.87
	K&D	43.62	42.44	42.36	43.15	44.38	43.19
	SKD	44.81	43.37	44.22	44.25	45.39	44.41

We conduct ablation experiments to disentangle the contributions of the three components in SKD: skewness regularization (S), kurtosis regularization (K), and logit disagreement regularization (D). As shown in Table 4, removing any individual component consistently reduces performance across all noise levels and datasets. Among the single-component variants, kurtosis (K) alone typically performs best, aligning with our hypothesis that high-order statistics play a dominant role in mitigating representation collapse. Combinations of two terms improve upon single terms, while the full SKD achieves the best average performance on Camelyon17 (90.76%), HAM10000 (57.73%), and NIHchestXray (44.41%), highlighting the complementary effects of the three regularizers.

3.2 Hyperparameter sensitivity

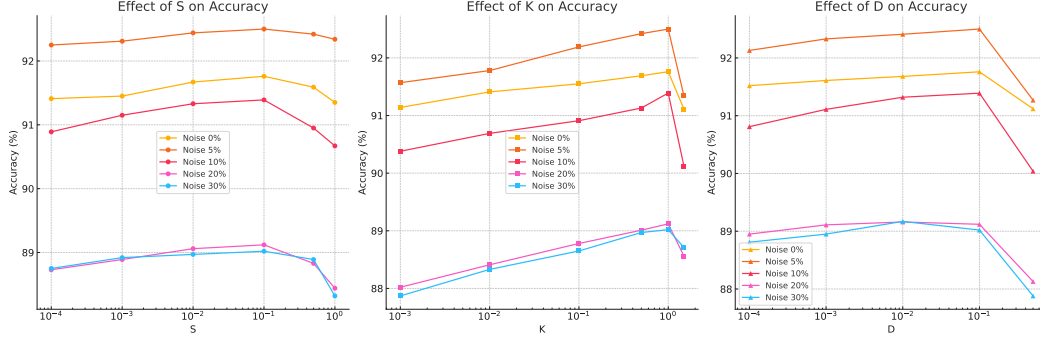


Figure 5: **Hyperparameter sensitivity of SKD components on Camelyon17.**

Hyperparameter Sensitivity. We conduct a systematic sensitivity analysis on the three regularization terms in SKD—skewness ($\mathcal{L}_{\text{skew}}$), kurtosis ($\mathcal{L}_{\text{kurt}}$), and disagreement (\mathcal{L}_{dis})—using ResNet50 on Camelyon17. For each component, we vary its weight over a range of values and report downstream accuracy under different noise ratios. The results show that SKD is robust across a wide range of values. Specifically, $\mathcal{L}_{\text{skew}}$ and $\mathcal{L}_{\text{kurt}}$ exhibit stable improvements, with optimal performance typically reached around 0.1. In contrast, \mathcal{L}_{dis} is more sensitive: when its weight exceeds 0.5, performance degrades substantially, and training may collapse under high noise. For example, accuracy under 30% noise drops from 89.02% at 0.1 to 87.88% at 0.5. Therefore, we do not experiment with larger values of λ_{dis} . Full results are shown in Figure 5.

3.3 Running Time

Table 5: **Training time comparison on PanNuke using PLIP.**

Dataset	LP	NML	SKD
PanNuke	20(s)	100(s)	150(s)

To assess computational efficiency, we measure the wall-clock training time of LP, NML, and SKD on the PanNuke dataset using the PLIP model. As shown in Table 5, LP is the fastest with only 20 seconds per run, while NML takes approximately 100 seconds. SKD introduces additional overhead due to the skewness, kurtosis, and disagreement regularization losses, requiring 150 seconds. Despite the added complexity, SKD remains lightweight and practical for real-world deployment, with only a $2.5\times$ increase over LP and a $1.5\times$ increase over NML. All experiments were conducted on a single NVIDIA A100 40GB GPU.