A Visual Denoising Model Based on Vision Transformer and Image Groups

Dianshi Moses Li[1], Yanyi Deborah Wang[2], Qiuyi Gao[3]

[1] Centre for Empirical Legal Studies, Faculty of Law, University of Macau; Taipa, Macau SAR, China, 999078.
[2] School of Accounting, Dongbei University of Finance and Economics; Dalian, Liaoning, China, 116025
[3] Department of Nephrology, The First Affiliated Hospital of Dalian Medical University, Dalian, China, 116011

Keywords

Computer Vision, Noisy Labels, Image Classification, Vision Transformer, Instance-Dependent Noise

Abstract

With the substantial advancements of deep learning in computer vision and natural language processing, the issue of data hunger, which is prevalent in real-world datasets, has garnered increasing attention. Learning with Noisy Labels (LNL), as a crucial method to address data scarcity, has attracted significant interest from researchers. However, most existing works are based on the Class-Conditional Noise (CCN) assumption rather than the Instance-Dependent Noise (IDN) assumption. Since the CCN assumption deviates considerably from real-world scenarios, models that perform well under CCN may exhibit poor performance in practical settings. Moreover, IDN noise possesses memory effects and instance-level memorization characteristics, making it more challenging than CCN. To tackle this problem, this paper proposes a novel model to overcome IDN noise: leveraging the backbone structure of Vision Transformer (ViT) and image groups to mitigate memory effects and instance-level memorization, thereby enabling deep neural networks to learn superior representations under IDN noise. Comprehensive experiments on the CIFAR-10 and MNIST datasets, as well as the real-world Clothing1M dataset, demonstrate the effectiveness of the proposed method. Notably, the model outperforms nearly all existing models on the MNIST dataset.

Research Article
Submitted: 1 February 2025 / Accepted: 7 February 2025 / Published online: 16 February 2025
Trans. Soc. 2025. 1(1): 1–16
https://doi.org/10.63336/TransSoc.18
Online ISSN: 3079-8310

1. Introduction

Processing computer vision tasks with deep learning necessitates large amounts of labeled data. However, in real-world scenarios, acquiring a substantial volume of high-quality data is often challenging due to issues such as data privacy and the high cost of annotation, especially in some high-stake academic and practice fields, like health care (Gao et al., 2024; Herm et al., 2023) and criminal justice (Berk et al., 2009). To address this practical problem, researchers have proposed various approaches, including Learning with Noisy Labels (LNL) (P. Chen et al., 2021; Han et al., 2019; Ma et al., 2020; Zheltonozhskii et al., 2022), Semi-Supervised Learning (SSL) (Berthelot et al., 2019; J. Li et al., 2020; Liang et al., 2020), Domain Adaptation (DA) (Kundu et al., 2020; R. Li et al., 2020; Liang et al., 2020), Data Augmentation (Cubuk et al., 2019; Nishi et al., 2021), and Self-Supervised Learning (SSL*) (Caron et al., 2020; Grill et al., 2020), aiming to mitigate the data hunger problem.

Among these, LNL seeks to leverage a large quantity of readily available but low-quality labeled data (P. Chen et al., 2021) (e.g., numerous images scraped from Amazon using keywords) to learn useful representations. SSL models utilize a subset of labeled data alongside a vast amount of unlabeled data (Berthelot et al., 2019; J. Li et al., 2020) to model decision boundaries between different classes. DA adopts the principles of transfer learning to transfer knowledge from a source domain to a target domain (R. Li et al., 2020; Saito et al., 2019; Yang et al., 2023), thereby significantly reducing the number of samples required for training in the target domain. Data Augmentation is the most straightforward approach, employing various augmentation techniques (such as cropping, rotation, grayscale adjustment, Mixup, etc.) to generate multiple images from a single image, thereby substantially increasing the sample size. SSL* (Caron et al., 2020; X. Chen et al., 2021; Grill et al., 2020) primarily relies on the concept of contrastive learning, comparing the similarity between different representations generated from a single image and those generated from other images to learn effective features.

All the aforementioned approaches can alleviate or even eliminate data hunger to varying degrees. However, even the current state-of-the-art (SOTA) models within these tasks exhibit numerous drawbacks and limitations, which severely constrain their industrial application. Consequently, the models currently adopted in industry remain large-parameter models renowned for their brute-force aesthetics and simplicity, such as GPT-3 (Brown et al., 2020), DALL-E (Reddy et al., 2021), and M6 (A Chinese Multimodal Pretrainer) (Lin et al., 2021). These large-parameter models are extremely time-consuming and resource-intensive to train, and presently, only well-capitalized companies possess the capability to deploy them in production and commercial settings, leaving small and medium-sized enterprises at a disadvantage.

As an economically viable task to address the data hunger problem, LNL has seen the proposal of numerous models in recent years, such as AugDesc-WS-SAW (Nishi et al., 2021) and DivideMix (J. Li et al., 2020). These models are predominantly based on the Class-Conditional Noise (CCN) (P. Chen et al., 2021) assumption, which posits that label noise is independent of the images themselves. However, recent studies have indicated that noise based on the CCN assumption does not align with real-world scenarios. Models that perform well under the CCN assumption may not necessarily be more suitable for practical applications, nor do they necessarily better address the data hunger problem (P. Chen et al., 2021).

In contrast, the Instance-Dependent Noise (IDN) assumption (P. Chen et al., 2021) posits that label noise is related to the images themselves. This is because, during the label generation process, label noise often arises from failures in search engine retrieval or errors in annotator recognition. Therefore, IDN is considered a more realistic assumption.

This paper aims to develop a model robust to Instance-Dependent Noise (IDN). The specific approach involves using point-wise convolution to abstract semantic information (semantic concepts) within images, leveraging the self-attention mechanism in the Vision Transformer (ViT) backbone (Dosovitskiy et al., 2020) to model semantic information within image groups. By taking advantage of the characteristics that noise information is difficult to fit (P. Chen et al., 2021) and constitutes a small proportion of the total information, the model reduces the weight of noisy samples, thereby achieving denoising.

In summary, this study makes the following three contributions:

1. Introduction of the Image Group Concept: The concept of image groups is proposed, enabling the convolutional layers to predominantly extract high-level semantic concepts belonging to the image group categories.

2. Novel Use of Vision Transformer and Self-Attention Mechanism for Noise Handling: For the first time, Vision Transformer and self-attention mechanisms are employed to process noise, thereby partially exploiting the transferability and attention mechanism advantages inherent in Transformer models.

3. Empirical Validation of Effectiveness: Through experiments conducted on multiple datasets, the proposed method demonstrates effectiveness in handling IDN noise, particularly achieving state-of-the-art performance on the MNIST (handwritten digit recognition) dataset.

2. Literature Review

2.1 Research on Learning with Noisy Labels

Label noise refers to errors present in the data labels. Specifically, let $\mathcal{X}$ denote the sample space, $\mathcal{Y} = 1, \dots, \mathrm{c}$ represent the label space, and $(\mathrm{X}, \mathrm{Y}) \in \mathcal{X} \times \mathcal{Y}$ be a random vector following the distribution $\mathcal{D}_{\mathcal{X},\mathcal{Y}} = (x_i, y_i)_{i=1}^{n}$. In noisy scenarios, the true label $\overline{Y}$ is observable, i.e., $(\mathrm{X}, \overline{\mathrm{Y}}) \in \mathcal{X} \times \overline{\mathcal{Y}} \sim \overline{\mathcal{D}}_{\mathcal{X},\overline{\mathcal{Y}}}$, where $\overline{\mathcal{D}}_{\mathcal{X},\overline{\mathcal{Y}}} = \left(x_i, \overline{y}_i\right)_{i=1}^{t}$.

Currently, in Learning with Noisy Labels (LNL) tasks, the mainstream noise generation paradigm is the Class-Conditional Noise (CCN) assumption (Menon et al., 2020; Scott et al., 2013; Zhang & Sabuncu, 2018). Under the CCN assumption, noise is generated using a noise transition matrix, where the flipping rules depend solely on the true label $y$and are independent of the image x itself.

However, recent studies have indicated that CCN noise does not accurately reflect real-world scenarios. In contrast, Instance-Dependent Noise (IDN) (P. Chen et al., 2021) is considered a more realistic form of label noise. This is because, in real-world settings, human annotation (or data scraping) is dependent on the data itself rather than merely on its true category. In other words, the mislabeling of a sample is often due to annotators confusing it with another category because the image is inherently similar to images of that other category. Additionally, during the data scraping process, overly vague keywords or excessively broad search engine association algorithms can result in images that should belong to other categories being erroneously classified into the same category. As illustrated in Figure 2-1, when searching for "dust coat" on Taobao, a portion of the retrieved products are not actually dust coats.

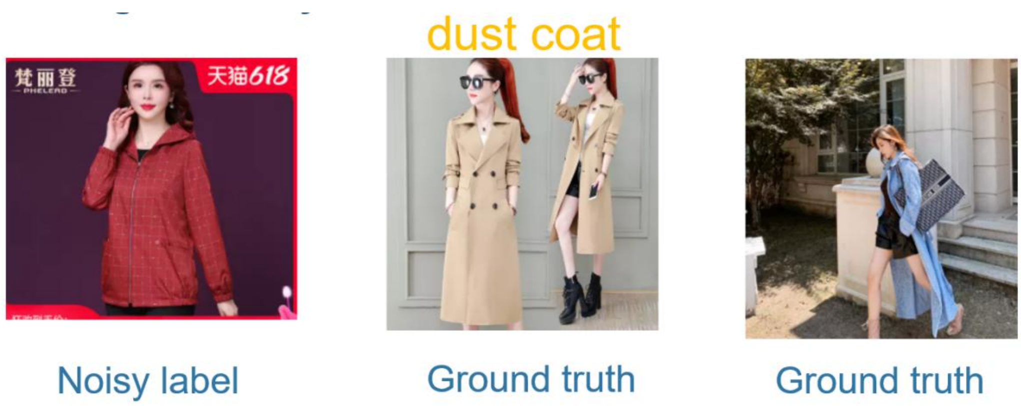

Figure 2-1: When searching for "dust coat" on Taobao, some products that are not dust coats are retrieved due to instance-dependent noise.

The Clothing1M dataset includes real-world images collected by Xiao et al. (2015) from several online shopping websites. The annotations of this dataset rely entirely on automatically assigned labels that contain real-world noise (Xiao et al., 2015). Researchers' calculations on the real noise scenario dataset Clothing1M (P. Chen et al., 2021) indicate that the probability of this dataset conforming to the CCN assumption is approximately 10^{-21250}, which implies that noise in real-world scenarios satisfying the CCN assumption is virtually an impossible event.

Currently, methods for handling noise are primarily based on the Class-Conditional Noise (CCN) assumption (Menon et al., 2020; Scott et al., 2013; Xiao et al., 2015; Zhang & Sabuncu, 2018), and these methods can be broadly categorized into four types: label correction methods, loss correction methods, strategies for modifying the training process, and robust loss function-based methods.

2.1.1 Label Correction Methods

The idea behind label correction is to predict the probability that a given label is noisy using certain means and then correct it accordingly. A natural approach is label inference (prediction), and models capable of performing this task include graphical models (Xiao et al., 2015), conditional random forests (Vahdat, 2017), neural networks (Lee et al., 2018; Veit et al., 2017), and even knowledge graphs (Li et al., 2017). However, these models typically require additional correct labels or incur high detection costs, making this class of methods not a research hotspot.

2.1.2 Loss Correction Methods

This category of methods focuses on adjusting the loss function during training by modifying weights based on the labels (Natarajan et al., 2013), or by estimating the noise transition matrix (Han et al., 2018), thereby weighting the loss. For example, Patrini et al. (2017) proposed Backward and Forward, which are two loss correction-based approaches for estimating the noise transition matrix. Additionally, Reed et al. (2014) introduced methods based on bootstrap techniques, and Szegedy et al. (2016) as well as Pereyra et al. (2017) proposed methods based on Label Smoothing Regularization (LSR).

However, methods that estimate the noise transition matrix often imply a form of "overfitting" to the noise task for the sake of publishing papers. In other words, the noise transition matrix is used by researchers to generate noise based on precisely labeled datasets to establish a unified standard (benchmark) for noise tasks. It merely specifies the probability of a particular label being flipped to another label to generate noise (Goldberger & Ben-Reuven, 2017; Scott et al., 2013; Sukhbaatar et al., 2014). However, this artificially generated noise on “clean” datasets significantly differs from noise in real-world applications: real-world noise often depends on the image itself rather than solely on its original category, and the probability of mislabeling images of the same true category into other categories varies. Therefore, the noise transition matrix should not exist in real-world scenarios; in other words, real-world noise is difficult to describe using only a noise transition matrix. This determines that methods based on estimating the noise transition matrix are unsuitable for real-world noise scenarios.

2.1.3 Strategies for Modifying the Training Process

Designing certain training strategies during the training process can often yield better results, leading to the emergence of many classic works (Kumar & Ithapu, 2020; Malach & Shalev-Shwartz, 2017). Examples include MentorNet (Jiang et al., 2018; Yu et al., 2019), the Co-teaching series (Han et al., 2018), and the Curriculum Learning series (TCL) (Shu et al., 2019), among others. Currently, the top three methods in the label noise domain (AugDesc-WS-SAW (Nishi et al., 2021), DivideMix (J. Li et al., 2020), and Nested Co-teaching (Y. Chen et al., 2021), data sourced from Papers with Code) are also based on this approach.

Among these, AugDesc-WS-SAW proposes a data augmentation strategy for noise scenarios (Cubuk et al., 2019). It suggests that augmentation in noisy scenarios should be divided into two types: strong augmentation (Strong), which involves modifying the image itself (Raw) by adjusting grayscale, altering colors, randomly erasing, etc., to generate images that appear similar to the human eye but differ during training; and weak augmentation (Weak), which only employs minor alterations such as random cropping, random flipping, and random rotation to generate new images.

AugDesc-WS-SAW posits that different augmentation strategies should be used at different stages of training and that different augmentation methods should be selected for different datasets. Specifically, in the domain of label noise handling, unsupervised augmentation learning methods (AutoAugmentation, i.e., automatically learning the optimal augmentation strategies for a given dataset and model) (Cubuk et al., 2019) should be employed.

The AugDesc-WS-SAW method in the paper is based on the DivideMix model, which remains among the top three in label noise handling tasks. Its approach is as follows: it attempts to separate noisy data from clean data using a certain threshold and treats noisy data as unlabeled data. Together with clean data, it is trained using the semi-supervised MixMatch (Berthelot et al., 2019) method, which involves generating k different images from the unlabeled data using different augmentation methods and using the average prediction of these k images by the current trained network as their pseudo labels. The model is then trained using Mean Squared Error Loss (MSEL) for pseudo-labeled data and Cross-Entropy Loss for labeled data (Berthelot et al., 2019).

2.1.4 Robust Loss Function-Based Methods

Compared to the above three categories, designing a robust loss function is a simpler approach. However, no models with satisfactory performance have been proposed to date. Typically, models that perform well on one dataset or a specific noise transition method do not perform well on other datasets or different noise scenarios. In other words, existing robust loss functions tend to overfit to the dataset or specific types of noise.

Among these methods, classic works include Mean Absolute Error (MAE) (Ghosh et al., 2017). However, when used alone, MAE often suffers from underfitting, meaning that while it does not overfit to noisy samples, it also does not effectively learn from correctly labeled samples. Similar works include Generalized Cross Entropy (GCE) (Zhang & Sabuncu, 2018) and Symmetric Cross Entropy (SCE) (Zhang & Sabuncu, 2018), but they still face underfitting issues.

A more innovative work is SCE proposed by (Wang et al., 2019) , which combines Reverse Cross Entropy (RCE) and Cross Entropy (CE) to create SCE, making it both robust to noise and capable of fitting correctly labeled samples. Inspired by the above, Ma et al. (2020) proposed Active Passive Loss (APL), which posits that any loss function, once normalized, can become a robust loss function for noise. It divides the loss function into two types: *Active Loss*: Aims to maximize the probability of the sample being classified into a specific label; *Passive Loss*: Aims to minimize the probability of the sample being classified into other labels.

Combining Active Loss and Passive Loss can address the underfitting problem of robust loss functions. However, this method still only has significant effects on certain types of noise and does not perform well for noise in real-world scenarios.

2.2 Research on Vision Transformers

The Transformer was initially applied to the field of Natural Language Processing (NLP) (Dosovitskiy et al., 2020). Due to its self-attention mechanism, which effectively models the relationships between semantic elements, the Transformer swiftly dominated the NLP domain upon its introduction, surpassing and eventually completely replacing traditional Recurrent Neural Network (RNN) models such as LSTM (Hochreiter, 1997). Building upon the Transformer architecture, models like BERT have enabled the use of techniques such as pre-training and fine-tuning in NLP, significantly enhancing the accuracy of various NLP tasks (Devlin, 2018).

Transformers and self-attention mechanisms have also been applied to the field of computer vision, outperforming traditional Convolutional Neural Networks (CNNs) in certain tasks. For instance, in the image classification task, the Vision Transformer (ViT) model divides an image into distinct patches, then flattens each patch from two dimensions to one dimension following a specific rule, and subsequently inputs them into the Transformer network (Dosovitskiy et al., 2020). This model has achieved accuracy comparable to or even surpassing advanced CNN models on datasets such as ImageNet (Russakovsky et al., 2015) and other large-scale proprietary datasets. However, the training time required to achieve such high precision is notably long and daunting.

The Visual Transformer (VT) (B. Wu et al., 2020) posits that convolutional operations are suitable for extracting low-level concepts from images, while Transformers excel at handling high-level semantics. This model also adopts a Transformer backbone, utilizing point-wise convolutions and self-attention mechanisms to transform images into high-level semantic representations suitable for Transformer processing, thereby enabling the learned representations to be applied to downstream tasks such as image classification and semantic segmentation. Subsequently, models like DeiT (Touvron et al., 2021), TNT (Han et al., 2021), HAT (Wang et al., 2020), Lite-Transformer (Z. Wu et al., 2020), and CVT (Wu et al., 2021) have progressively secured positions in tasks such as object detection, image recognition, and semantic segmentation, showing strong potential to replace traditional convolutional networks.

However, a significant challenge in applying Transformers to the field of computer vision is that Transformers require their input to be a sequence of vectors, whereas images are inherently two-dimensional or three-dimensional (i.e., grayscale or color images). Consequently, how to convert two-dimensional information into a one-dimensional format without loss is a key issue that experts in the field have been collaboratively addressing. Analysis indicates that Transformers can capture global information, whereas convolutions are limited to modeling local information based on the size of the convolutional kernel. Therefore, Transformers can be considered as an extension of convolutions, and the functionality achieved by the self-attention mechanism is, to some extent, less than that of convolutions, leading to the view of Transformers as a specialized form of convolutions. Nevertheless, the ability of Transformers to process and model high-level semantics surpasses that of other networks. This paper aims to leverage this capability to address the task of label noise.

3. Model Construction

As previously mentioned, Instance-Dependent Noise (IDN) is a form of noise that is closer to real-world scenarios compared to Class-Conditional Noise (CCN). IDN presents greater task difficulty and has more promising application prospects. This study aims to address IDN noise, with both the assumed conditions and experimental settings referencing (P. Chen et al., 2021).

P. Chen et al. (2021) proposed that IDN noise is highly susceptible to being fitted and is characterized by poor memorization effects and instance-level memorization. This renders traditional CCN noise handling methods (i.e., pre-training followed by correction) ineffective, as the poor memorization effects adversely affect the model's pre-training performance.

To overcome this challenge, inspired by the VT model's approach of transforming images into high-level semantic information, this study employs point-wise convolution to abstract semantic concepts from images. It then utilizes the self-attention mechanism within the Transformer (Dosovitskiy et al., 2020) backbone to model the semantic information within image groups.

We anticipate that, under a low ratio of IDN noise (e.g., a 40% noise ratio), the majority of the semantic information extracted from image groups via point-wise convolution will be valid. Specifically, when the noise ratio is below 50%, the correct semantic concepts within image groups will outweigh the noisy semantic labels. Subsequently, the Transformer's self-attention mechanism is used to model the correlations among the semantic information within image groups. Images with strong correlations are considered to have correct labels, whereas those with weak correlations are deemed to have incorrect labels. This approach allows for the identification and differentiation of noisy samples within the training set, thereby achieving noise reduction.

3.1 Model Framework

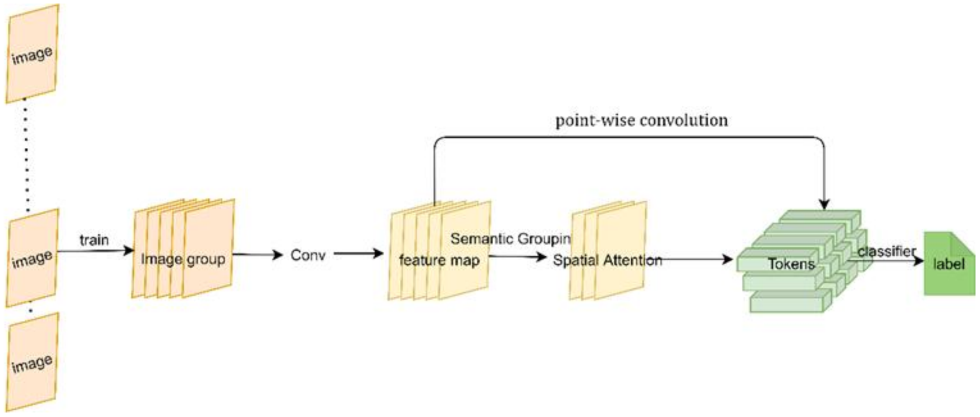

Figure 3.1: Model Training Process

The training process of the model is illustrated in Figure 3.1. Initially, m samples with labels $\overline{Y} \in \overline{\mathcal{Y}}$ are selected to form an image group, and a group label $\overline{Y}$. is assigned to this image group. Subsequently, convolutional and Multi-Layer Perceptron (MLP) layers are employed to extract features from each image within the image group, generating feature maps. Utilizing point-wise convolution, n tokens are generated for each image. As a result, a single image group processed in this manner can produce $m \times n$ tokens. These tokens represent high-level semantic concepts suitable for processing by the Transformer.

The tokens generated are then input into the Transformer network. Leveraging the self-attention mechanism within the Transformer, a self-attention coefficient matrix is produced for the tokens. By summing these coefficients along the key dimensions, the weight of each token is determined. A higher token weight indicates a stronger correlation between the token and the group label $\overline{Y}$. Subsequently, the tokens are passed through a fully connected network and a Softmax layer to obtain the predicted label probabilities. Finally, the cross-entropy loss function and backpropagation algorithm are utilized to update the model parameters.

The testing process of the model is depicted in the above figure. To ensure consistency in the input distribution for the Transformer, the concept of image groups is maintained during testing. First, the test image is subjected to various data augmentation methods (such as downsampling, random pixel disappearance, contrast adjustment, random cropping, etc.) to generate m different images, thereby forming an image group (Cubuk et al., 2019; Nishi et al., 2021). This image group is then input into the previously trained network to obtain the final prediction result.

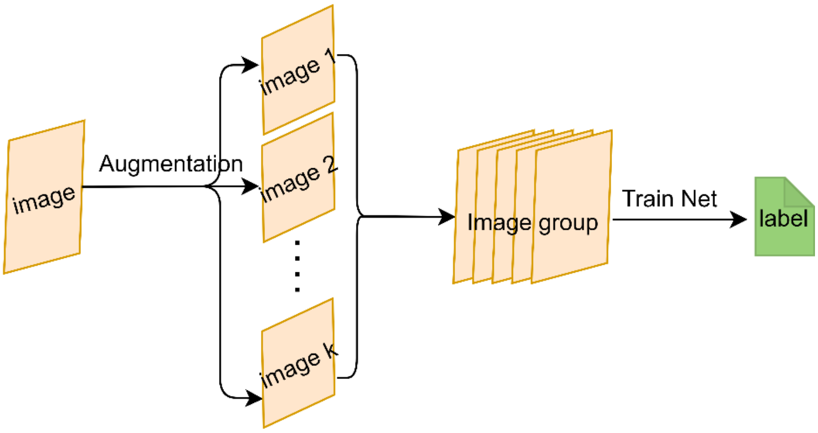

Figure 3.2: Model Training Process

3.2 Detailed Model Description

3.2.1 Image Group

An image group is defined as a collection of k images with labels $\overline{Y} \in \overline{\mathcal{Y}}$, sharing a common group label $\overline{Y}$. The dataset is represented as $\mathcal{D}_{\mathcal{X},\overline{\mathcal{Y}}} = \left(x_i, \overline{y}_i\right)_{i=1}^{n}$, where $\overline{\mathcal{Y}}$ defines the sample label space XXX and $\mathcal{Y}$ defines the true sample space XXX. Consequently, an image group can be defined as $\mathrm{I_c} = I_{c,1}, I_{c,2}, \dots I_{c,m}$, where $\mathrm{I_{c,m}} = \left(\mathrm{x_{(c,m)}}, \overline{y}_{\mathrm{(c,m)}}\right)$.

3.2.2 Tokenizer

The Tokenizer is a mechanism that transforms feature maps into tokens. The principle is as follows: define the feature map as X. Each pixel $\mathrm{X_p} \in \mathcal{R}^{C}$ is mapped to one of L semantic groups using point-wise convolutions. Within each semantic group, spatial pooling is applied to obtain a token, denoted as T. The formula is presented below:

$$T = \overbrace{\mathrm{SOFTMAX}_{\mathrm{HW}}(XW_A)^T}^{\mathrm{A} \in R^{HW\dot{L}}} X$$

Here, $W_A \in \mathcal{R}^{C \times \mathcal{L}}$ is multiplied with X to generate semantic groups. A softmax function is then applied to produce A, which represents spatial attention. This attention map A is subsequently multiplied with X to generate L tokens.

3.3.3 Transformer

After generating the tokens, the Transformer is utilized to model the relationships between them. Leveraging the self-attention mechanism within the Transformer, a self-attention coefficient matrix is produced for the tokens. By summing these coefficients along the key dimensions, the weight of each token is determined. A higher token weight signifies a stronger correlation between the token and the group label $\overline{Y}$. Subsequently, the tokens are input into a fully connected network and a Softmax layer to obtain the predicted label probability. Finally, the cross-entropy loss function and backpropagation algorithm are employed to update the model parameters.

4. Experiments and Results

4.1 Dataset Description

The MNIST dataset is a handwritten digit recognition dataset comprising 60,000 images of handwritten digits. Out of these, 50,000 images are designated as the training set and 10,000 images as the test set. Each image is a 28×28-pixel grayscale image representing handwritten digits from 0 to 9. The noisy labels are generated strictly following the method outlined by P. Chen et al. (2021).

CIFAR-10 consists of 60,000 color images, each being a 32×32-pixel three-channel color photograph. Of these, 50,000 images are allocated as the training set, further divided into five training batches of 10,000 images each, and the remaining 10,000 images form a separate test batch. The dataset encompasses ten classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The noisy labels in this dataset are also generated according to the methodology proposed by P. Chen et al. (2021).

Clothing1M is a real-world noisy dataset collected from several online shopping websites. Each image in this dataset is automatically labeled based on the surrounding environment during the scraping process, resulting in inherently noisy annotations. The dataset comprises one million images across 14 categories. However, only 72,409 images are manually selected by the authors as clean data, which are subsequently divided into validation and test sets.

4.2 Evaluation Metrics

This study employs the commonly used performance evaluation metric in the Learning with Noisy Labels (LNL) domain: accuracy. The accuracy is calculated using the following formula [60]:

$$Accuracy = \left(\frac{Number\ of\ correctly\ classified\ images}{Total\ number\ of\ classified\ images}\right) * 100\%$$

4.3 Experimental Results

The proposed model was evaluated on three benchmark datasets for handling Instance-Dependent Noise (IDN): MNIST, CIFAR-10, and Clothing1M. The experiments were conducted using Python 3.8, PyTorch 1.4, torchvision 0.7, toolkit 10.0, and Tesla V100 GPUs. The experimental results are presented as follows.

Method	10%	20%	30%	40%
CE	94.07	85.62	75.75	65.83
Forward	93.93	85.39	76.29	68.30
Co-teaching	95.77	91.07	86.20	79.30
GCE	94.56	86.71	78.32	69.78
DAC	94.13	85.63	75.83	65.59
SEAL	94.75	93.63	88.52	80.73
Ours	**94.80**	**93.74**	**89.30**	**82.12**

Table 4.1: MNIST Experimental Results

Method	10%	20%	30%	40%
CE	91.25	86.34	80.87	75.68
Forward	91.06	86.35	78.87	71.12
Co-teaching	91.22	87.28	84.33	78.72
GCE	90.57	86.44	81.54	76.71
DAC	90.94	86.16	80.88	74.80
SEAL	**91.32**	**87.79**	**85.30**	82.98
Ours	91.24	86.98	83.30	**83.12**

Table 4.2: CIFAR-10 Experimental Results

Method	Accuracy
CE*	68.94
Forward*	69.84
Co-teaching*	70.15
GCE*	69.09
CE	69.07
SEAL	**73.40**
Ours	**72.90**

Table 4.3: Clothing1M Experimental Results

From the tables above, it is evident that the proposed model surpasses the baseline across all three datasets. Additionally, it significantly outperforms classical robust loss function methods and the traditional Co-teaching network. However, it slightly underperforms compared to the SEAL method, which is based on pseudo-label training. Specifically, on the MNIST dataset, our model achieves state-of-the-art (SOTA) results. On CIFAR-10, the model attains SOTA performance at a high noise label ratio (40%) but performs slightly below SEAL at lower noise label ratios. On the Clothing1M dataset, our model outperforms CCN-based models but still lags behind the IDN-based SEAL model.

These results demonstrate the effectiveness of our model in resisting the impact of IDN-type noise. Furthermore, the results suggest that our model is more suitable for datasets with lower image complexity (e.g., MNIST) and those with a large number of fine-grained images (e.g., Clothing1M). Conversely, the model exhibits slightly poorer performance on the more ambiguous CIFAR-10 dataset.

4.4 Discussion on Model Effectiveness

This study posits that the key to resisting the impact of IDN noise lies in learning representations suitable for classification before the classifier layer. Figure 4-1 illustrates the dimensionality reduction visualization using ResNet34 trained for three epochs on the CIFAR-10 dataset with 40% IDN noise:

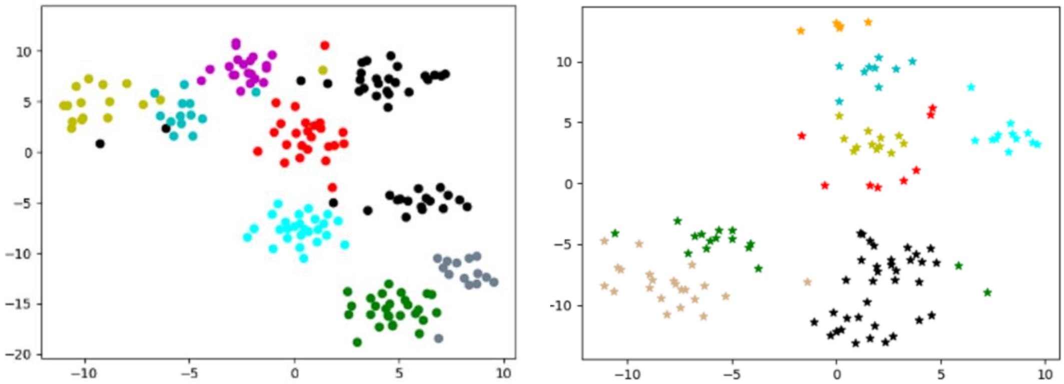

Figure 4.1: Dimensionality Reduction Visualization

From this observation, it becomes evident that the IDN noise causes ResNet34 to prematurely overfit to the noise, resulting in compromised decision boundaries. The model fails to learn discriminative representations suitable for classification tasks. In contrast, Figure 4-2 demonstrates the outcome after processing with our proposed model.

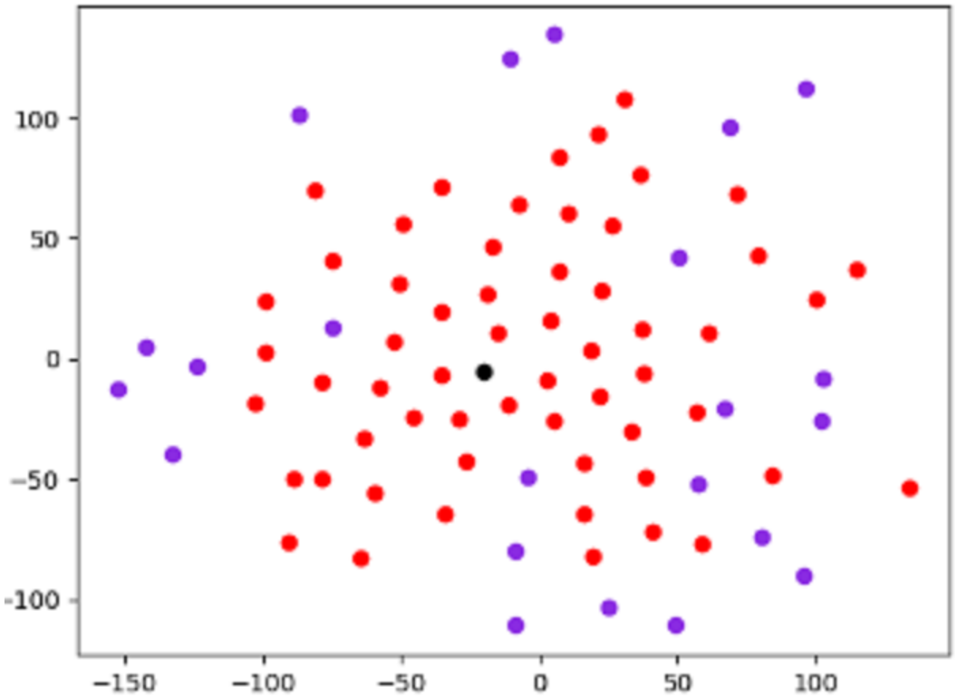

Figure 4.2: T-SNE Dimensionality Reduction Visualization

In this visualization, red data points represent correctly labeled samples, while purple data points denote incorrectly labeled samples. It is evident that after processing with our model, the red samples are closer to the centroid (represented by black points). By defining sample weights based on the distance of sample points to the centroid, the likelihood of the model overfitting to noise is reduced.

The Image Group concept in our model is akin to Mixup (Liang et al., 2018; Vaswani, 2017; Zhang et al., 2018), which involves directly overlaying multiple images to generate augmented data samples between classes. Mixup can be understood as a regularization technique that forces the model to learn simpler decision boundaries by generating inter-class samples. Additionally, Mixup focuses on modeling local decision boundaries through neighborhood risk minimization, thereby avoiding the pitfalls of global empirical risk minimization that can lead to local optima, resulting in superior model performance (Zhang et al., 2020). In our model, the tokens from multiple images within a group embody this idea, enabling the model to leverage features from the same class across different images for modeling. This approach utilizes neighborhood risk minimization to learn representations that are more consistent with the characteristics of each class.

5. Discussion and Future Work

In the field of computer vision, addressing the data hunger problem has long been a focal point of scholarly attention and research. Learning with Noisy Labels (LNL) has emerged as a rapidly developing method to mitigate data scarcity issues. Since the Transformer architecture achieved groundbreaking success in computer vision, its potential across various tasks remains a fertile ground for further exploration and research.

The proposed model aims to provide an effective approach and research experience for employing Transformers in the context of noisy label learning. By leveraging the transferability of Vision Transformers (ViT), the Tokenizer's ability to transform semantic information, and the augmentation-like effects of Image Groups, the model effectively mitigates the impact of IDN noise and facilitates robust learning. Although the model has achieved SOTA results on certain datasets, there remain areas for further refinement. Future work should include more comprehensive ablation studies to investigate the actual roles of the Tokenizer and Image Group within the model. Additionally, employing visualization techniques to elucidate the mechanisms behind these components would be beneficial.

These research outcomes can also be extended to the emerging framework of Masked Autoencoders (MAE) [65]. Inspired by the reconstruction of corrupted images, MAE evolves into a process of masking certain pixels and subsequently reconstructing them. The relationship between masked and unmasked pixels aligns with the IDN noise assumption—both rely on the pixels themselves or the image itself. However, MAE operates at the pixel level, whereas IDN operates at the instance level. Nevertheless, a group of pixels within an image can be viewed as a subset of the image, effectively constituting a new small image. This provides a novel perspective for integrating IDN methodologies into MAE research, offering new avenues to address the data hunger problem more effectively. Moving forward, we intend to explore this angle to identify more efficient solutions to the data scarcity issue.

References:

Berk, R., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society Series a-Statistics in Society*, *172*(1), 191-211. https://doi.org/10.1111/j.1467-985X.2008.00556.x

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, *32*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, *33*, 9912-9924.

Chen, P., Ye, J., Chen, G., Zhao, J., & Heng, P.-A. (2021). Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. Proceedings of the AAAI Conference on Artificial Intelligence,

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. Proceedings of the IEEE/CVF international conference on computer vision, Chen, Y., Shen, X., Hu, S. X., & Suykens, J. A. (2021). Boosting co-teaching with compression regularization for label noise. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations,

Gao, Q., Li, D., Wang, Y., Zhao, C., Li, M., Xiao, J., Kang, Y., Lin, H., & Wang, N. (2024). Analysis of intestinal flora and cognitive function in maintenance hemodialysis patients using combined 16S ribosome DNA and shotgun metagenome sequencing. *Aging Clin Exp Res*, *36*(1), 28. https://doi.org/10.1007/s40520-023-02645-y

Ghosh, A., Kumar, H., & Sastry, P. S. (2017). Robust loss functions under label noise for deep neural networks. Proceedings of the AAAI conference on artificial intelligence,

Goldberger, J., & Ben-Reuven, E. (2017). Training deep neural-networks using a noise adaptation layer. International conference on learning representations,

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., & Gheshlaghi Azar, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, *33*, 21271-21284.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, *31*.

Han, J., Luo, P., & Wang, X. (2019). Deep self-learning from noisy labels. Proceedings of the IEEE/CVF international conference on computer vision,

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *Advances in neural information processing systems*, *34*, 15908-15919.

Herm, L. V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, *69*, 102538. https://doi.org/ARTN10.1016/j.ijinfomgt.2022.102538

Hochreiter, S. (1997). Long Short-term Memory. *Neural Computation MIT-Press*.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., & Fei-Fei, L. (2018). Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. International conference on machine learning,

Kumar, A., & Ithapu, V. K. (2020). Secost:: Sequential co-supervision for large scale weakly labeled audio event detection. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Kundu, J. N., Venkat, N., & Babu, R. V. (2020). Universal source-free domain adaptation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Lee, K.-H., He, X., Zhang, L., & Yang, L. (2018). Cleannet: Transfer learning for scalable image classifier training with label noise. Proceedings of the IEEE conference on computer vision and pattern recognition,

Li, J., Socher, R., & Hoi, S. C. (2020). DivideMix: Learning with Noisy Labels as Semi-supervised Learning. International Conference on Learning Representations,

Li, R., Jiao, Q., Cao, W., Wong, H.-S., & Wu, S. (2020). Model adaptation: Unsupervised domain adaptation without source data. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., & Li, L.-J. (2017). Learning from noisy labels with distillation. Proceedings of the IEEE international conference on computer vision,

Liang, D., Yang, F., Zhang, T., & Yang, P. (2018). Understanding mixup training methods. *IEEE Access*, *6*, 58774-58783.

Liang, J., Hu, D., & Feng, J. (2020). Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. International conference on machine learning,

Lin, J., Yang, A., Bai, J., Zhou, C., Jiang, L., Jia, X., Wang, A., Zhang, J., Li, Y., & Lin, W. (2021). M6-10t: A sharing-delinking paradigm for efficient multi-trillion parameter pretraining. *arXiv preprint arXiv:2110.03888*.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., & Bailey, J. (2020). Normalized loss functions for deep learning with noisy labels. International conference on machine learning,

Malach, E., & Shalev-Shwartz, S. (2017). Decoupling" when to update" from" how to update". *Advances in neural information processing systems*, *30*.

Menon, A. K., Rawat, A. S., Reddi, S. J., & Kumar, S. (2020). Can gradient clipping mitigate label noise? International Conference on Learning Representations,

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., & Tewari, A. (2013). Learning with noisy labels. *Advances in neural information processing systems*, *26*.

Nishi, K., Ding, Y., Rich, A., & Hollerer, T. (2021). Augmentation strategies for learning with noisy labels. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. Proceedings of the IEEE conference on computer vision and pattern recognition,

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. (2017). Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Reddy, M. D. M., Basha, M. S. M., Hari, M. M. C., & Penchalaiah, M. N. (2021). Dall-e: Creating images from text. *UGC Care Group I Journal*, *8*(14), 71-75.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., & Rabinovich, A. (2014). Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211-252.

Saito, K., Kim, D., Sclaroff, S., Darrell, T., & Saenko, K. (2019). Semi-supervised domain adaptation via minimax entropy. Proceedings of the IEEE/CVF international conference on computer vision,

Scott, C., Blanchard, G., & Handy, G. (2013). Classification with asymmetric label noise: Consistency and maximal denoising. Conference on learning theory,

Shu, Y., Cao, Z., Long, M., & Wang, J. (2019). Transferable curriculum for weakly-supervised domain adaptation. Proceedings of the AAAI conference on artificial intelligence,

Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., & Fergus, R. (2014). Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. International conference on machine learning,

Vahdat, A. (2017). Toward robustness against label noise in training deep discriminative neural networks. *Advances in neural information processing systems*, *30*.

Vaswani, A. (2017). Attention is all you need. *Advances in neural information processing systems*.

Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., & Belongie, S. (2017). Learning from noisy large-scale datasets with minimal supervision. Proceedings of the IEEE conference on computer vision and pattern recognition,

Wang, H., Wu, Z., Liu, Z., Cai, H., Zhu, L., Gan, C., & Han, S. (2020). Hat: Hardware-aware transformers for efficient natural language processing. *arXiv preprint arXiv:2005.14187*.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., & Bailey, J. (2019). Symmetric cross entropy for robust learning with noisy labels. Proceedings of the IEEE/CVF international conference on computer vision,

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. Proceedings of the IEEE/CVF international conference on computer vision,

Wu, Z., Liu, Z., Lin, J., Lin, Y., & Han, S. (2020). Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*.

Xiao, T., Xia, T., Yang, Y., Huang, C., & Wang, X. (2015). Learning from massive noisy labeled data for image classification. Proceedings of the IEEE conference on computer vision and pattern recognition,

Yang, S., Wang, Y., Herranz, L., Jui, S., & van de Weijer, J. (2023). Casting a bait for offline and online source-free domain adaptation. *Computer Vision and Image Understanding*, *234*, 103747.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., & Sugiyama, M. (2019). How does disagreement help generalization against label corruption? International conference on machine learning,

Zhang, H., Cisse, M., Dauphin, Y., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk management. 6th Int. Conf. Learning Representations (ICLR),

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., & Zou, J. (2020). How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*.

Zhang, Z., & Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, *31*.

Zheltonozhskii, E., Baskin, C., Mendelson, A., Bronstein, A. M., & Litany, O. (2022). Contrast to divide: Self-supervised pre-training for learning with noisy labels. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,