# Conditional Variational Autoencoder Data Augmentation for Robustness to Instance-Dependent Noise

Eugene Kim
eykim@ucsd.edu
University of California, San Diego
La Jolla, California, USA

## Abstract

The effectiveness of deep learning heavily relies on high-quality labeled data. In practice, obtaining accurate and noise-free labels is a challenging and time-consuming task, and often, artificially generated label noise is unrepresentative to real-world label noise. In this paper, the use of instance-dependent noise (IDN) will be discussed as a benchmark for modeling real-world noise and a new deep generative approach for this problem will be proposed. Specifically, the conditional variational autoencoder (CVAE) is introduced for data augmentation on desired class labels. The proposed method is evaluated on the MNIST dataset and the experimental results demonstrate that it outperforms several state-of-the-art techniques under less noisy conditions. The CVAE has the ability to encode generalized class-conditional features which helps mitigate the influence of noisy labels. The results show that the CVAE-based data augmentation is robust to small percentages of noise, highlighting its potential for improving the performance of image classification in real-world scenarios.

***CCS Concepts:*** • **Computing Methodologies** → *Machine Learning*; • **Applied Computing** → *Computer Vision*; *Data Augmentation*;

***Keywords:*** data augmentation, generative models, instance-dependent noise, image classification

**ACM Reference Format:**
Eugene Kim. 2023. Conditional Variational Autoencoder Data Augmentation for Robustness to Instance-Dependent Noise. In *Proceedings of* . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn



, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

Deep Neural Networks (DNNs) have proven to be highly effective in many vision tasks due to its ability to learn various complex patterns. However, the performance significantly drops when noise is introduced to the training label as the network begins to learn independent of the label accuracy. Thus, DNNs rely heavily on high-quality labeled data, which is highly infeasible to obtain due to time limitations, human error, image corruption [3], and web-based noise [4]. Two main approaches have been taken in the past to evaluate noisy labels. The first approach involves experimenting in a controlled environment by producing synthetic noise [13], and the second approach involves using data with untraceable real-world noise such as Clothing-1M [12] and WebVision[7]. The controlled experiment provides a proper benchmark to evaluate a model's performance but is often unrepresentative of real-world noise. On the other hand, the real-world noise datasets are not suitable for evaluation on different noise levels. Manual labeling must be done to adjust the noise level, but real-world noise datasets have unknown and fixed size noise [4]. To address the problems in both synthetic and real-world noise, this paper makes use of the controlled instance-dependent noise (IDN), a form of synthetic noise that is dependent on each image rather than the respective classes [2].

Symmetric noise, class-conditional noise (CCN), and IDN are the three common methods of producing synthetic noise. As the name suggests, symmetric noise changes the example's label independently and uniformly to a random label. However, this violates the trend for real-world noise according to the research done by Xiao et al. on Clothing-1M [12]. The authors concluded that clothes (labels) such as hoodies are more prone for misclassication compared to jackets and windbreakers. Similarly, it is more likely for numbers 2 and 5 to be misclassified compared to number 0 based on Figure 1 (See page 2). Therefore, the corruption probability for the class label 0 ($p_0$) should be certainly smaller than that of class label 2 ($p_2$) or class label 5 ($p_5$). This problem serves as motivation for utilizing CCN [8]. CCN assigns different corruption probabilities for each class based on its respective feature map. However, real-world noise goes beyond the class-dependent assumption and more closely follows

noise distribution based on each instance [4]. Thus, this paper leverages controlled experiments on IDN to evaluate a model's robustness to noise [2].

The primary contribution of this paper is using a data augmentative approach to introduce new information to the dataset at lower noise levels. This paper uses a deep generative model called Conditional Variational Autoencoder (CVAE) for extended data augmentation in the training phase. CVAE consist of two main components, encoder and decoder. The encoder aims to find a lower-dimensional latent space representation of the input data conditioned by a class label [5]. The decoder maps the latent vector back to the original space by minimizing the reconstruction and KL divergence loss. This paper aims to use the decoder of CVAE to generate new samples while maintaining a balanced dataset. In comparison to other state-of-the-art methods, the model trained on the extended dataset made by CVAE achieved the highest accuracy for 10% percent noise. The experimental results conclude that CVAE's encoder is robust to small percentage ($\approx$ 10%) of noise during training. These results aim to promote further research in data augmentation using deep generative models for noisy labeled data.

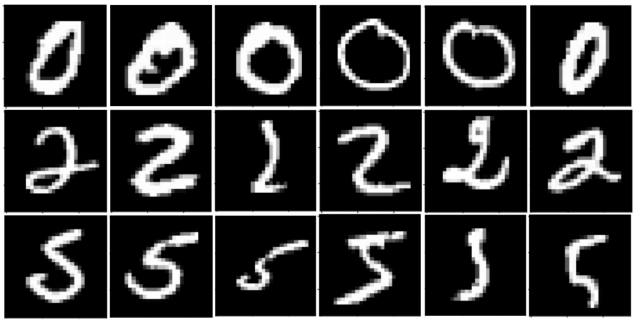

**Figure 1.** Examples of *0* (first row), *2* (second row), and *8* (third row). Assigning a uniform probability to each class does not provide an accurate representation of the corruption probability.

## 2 Dataset

### 2.1 Definitions

These definitions are adapted from Section 2.1 of Chen et al.'s paper [2].

For a classification task with $c$ classes, let $D_{X,Y} = \{(\mathbf{x_i}, y_i)\}_{i=1}^{n}$ with the feature $\mathbf{x}_i \in D_{X,Y}$ and its corresponding label $y_i \in \{1, ..., c\}$. Assuming that $\mathbf{Y}$ is not observable due to noise, $\bar{D}_{X,\bar{Y}} = \{(\mathbf{x_i}, \bar{y}_i)\}$ is i.i.d. drawn from the observation distribution $\bar{Y} \sim \bar{D}_{X,\bar{Y}}$. Let $f : \mathbb{P}^c$ be the defined as a DNN that outputs a probability distribution over all classes $c$ where $\mathbb{P}^c = \{s \in \mathbb{R}_+^c : \|s\|_1 = 1\}$. The probability is assumed to be obtained by a softmax function at output of $f$.

### 2.2 Symmetric Noise and CCN Limitations

Many controlled environments in the past have been done on symmetric noise or class-conditional noise (CCN) due to its simplicity in nature [8, 10]. However, both types of synthetic noise do not accurately represent the corruption probability distribution of real-world data due to their assumptions. With regards to CCN, it makes two major assumptions that do not the follow the real-world noise distribution based on prior research and observation [4, 12].

First, CCN assumes that every instance has an equal corruption probability within each class. In Figure 1 on this page, the third example of the number 2 is more likely to be mislabeled as the number 1 compared to the second example. The corruption probability for the third example should be greater than that of the second example. By assuming all instances of each class have equal corruption probability, CCN does not closely follow the real-world noise distribution [12].

Second, CCN assumes that every instance of a respective class share the same features as the other instances. For CCN, the feature space $X$ conditioned on the true label $Y$ has no influence on the label noise $\bar{Y}$ [2]. However, $\bar{Y}$ should be dependent on both $X$ and $Y$ to represent real-world noise. In Figure 1, the first example of the number 2 has features similar to the number 6 while the third example of the number 2 closely resembles the number 1. Thus, the probability for the first example to be mislabeled as the number 6 should not be the same as the second third example.

Symmetric noise adds one more critical assumption to the two already made by CCN. It strictly assumes that the corruption probability $p_c$ is uniformly distributed among all classes. According to prior research done by Xiao et al. on Clothing-1M [12], certain articles of clothing such as jackets and windbreakers are more likely to be mislabeled compared to hoodies. Furthermore, in Figure 2 (See page 3), the t-Distributed Stochastic Neighbor Embedding (t-SNE) plots reveal more overlapping clusters for the numbers 4 and 9 compared to the numbers 0 and 1.

This paper addresses the assumptions made by CCN and symmetric noise by going beyond class-level assumptions and using instance-dependent noise (IDN).

### 2.3 IDN Generator

Real-world noise is semantically and visually more consistent with the true label [4]. To capture this distribution, methods like controlled web-based noise has been introduced by Jiang et al. [4]. However, this method requires several human annotators to verify the labels. Due to limited resources, this paper will use a model-based IDN method from Section 3.1 of Chen et al.'s paper to generate noise [2]. For each instance, the IDN generator uses the output of a DNN from $T$ epochs to determine which ones are difficult to classify. The generator will then take the highest probability label (different from the correct label) and use it as the noisy label for that

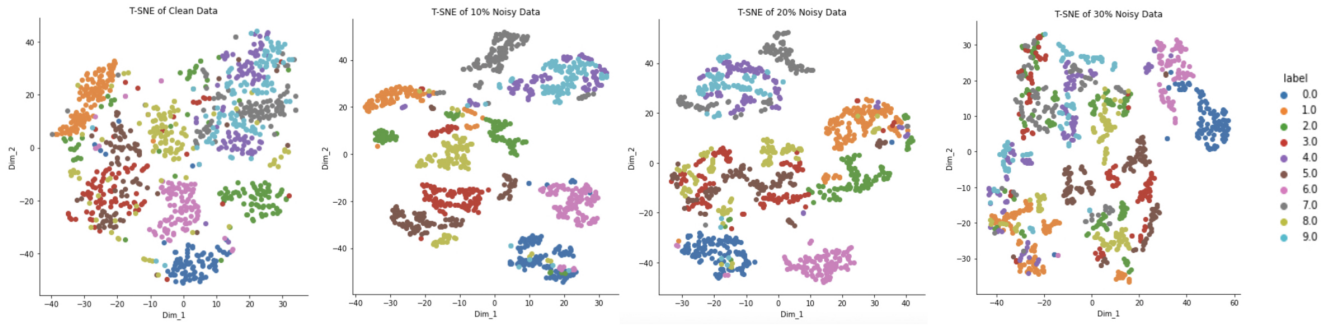


**Figure 2.** t-SNE plots of MNIST dataset in varying noise levels

instance. This process allows $\bar{Y}$ to be dependent on both $X$ and $Y$ in a controlled environment.

In this paper, a vanilla convolutional neural network (CNN) is used as the DNN for the IDN generator. The CNN filters are capable of learning important features on the instance-level [9]. The lower levels of CNN tend to learn basic class-level features such as general shapes, edges, and corners. As the network gets deeper, the filters begin to learn more complex patterns on the instance-level. The predictions made by the DNN is dependent on both $X$ and $Y$. Therefore, it is reasonable to use DNNs to produce IDN.

Jiang et al. defines the score of mislabeling as $N(x)$ and the potential noisy label $\tilde{y}(x)$ from the DNN as follows [4]:

$$S = \sum_{t=1}^{T} S^t/T \in \mathbb{R}^{n\times c},$$

$$N(x_i) = \max_{k\neq y_i} S_{i,k}, \tilde{y}(x_i) = \operatorname{argmax}_{k\neq y_i} S_{i,k},$$

where $S^t = [f^t(x_i)]_{i=1}^n$ refers to the DNN's output $t$-th epoch. Given a corruption probability $p$, we can take the $p$ highest probability outputs different from the true label and flip the labels. This ensures that the noise level can be controlled for proper benchmark comparisons.

In Figure 2, the t-SNE plots of varying noise levels are displayed. The data points from 10% and 20% noise is similar if not better clustered than the clean data. This is expected since there are instances where certain numbers resemble another class better than the true class. The IDN generator flips this label allowing for closely-related instances to be clustered. This closely resembles real-world noise since noisy labels are visually and semantically consistent with the true labels [4].

## 3 Method

### 3.1 Conditional Variational Autoencoder

This paper uses a deep generative model called Conditional Variational Autoencoder (CVAE) for class-conditioned data augmentation. CVAE consists of two main components, the encoder and decoder. Similar to the traditional Variational Autoencoder (VAE), the encoder attempts to find a lower-dimensional latent space representation of the input image while the decoder attempts to reconstruct new samples using this latent vector [6]. CVAEs share similar traits with Bayesians model such as learning the underlying data distribution. To be more specific, the model tries to find the latent space of the input parameterized by $\mu$ and $\sigma$. This latent vector is then used in the decoding phase to generate new samples by minimizing the reconstruction and KL divergence error. The reconstruction loss is measured by taking the distance between the input to the generated output while the KL divergence measures the difference between the distribution of the latent space and a prior distribution. Formally, the objective of the VAE is

$$E[\log P(X|z)] - D_{KL}[Q(z|X)\|P(z)] \tag{1}$$

where $Q(z|X)$ is the encoder and $P(X|z)$ is the decoder. Under some encoding error, we want to maximize the log likelihood of $P(X)$. The CVAE has an additional conditioning variable $c$ for an objective function of

$$E[\log P(X, c|z)] - D_{KL}[Q(z|X, c)\|P(z|c)] \tag{2}$$

where $Q(z|X, c)$ is the new encoder and $P(X, c|z)$ is the new decoder. The full architecture is displayed in Figure 3. Unlike traditional VAEs, CVAEs generate instances conditioned on a class label enabling the model to generate new samples of a specified class [5]. This is essential for maintaining a balanced dataset during data augmentation. Under low noise conditions, CVAE is able to increase the size of the original dataset with new meaningful instances. There are other generative models such as Conditional Generative Adversarial Networks (cGAN) that can also be used for data augmentation. According to studies done by Thekumparampil et al. on cGAN, the authors determined that traditional cGANs will not only generate misclassified samples but also poor quality ones when trained on noisy labels [11]. However,

CVAE is known to be more robust to noise created by adversarial attacks [1]. Thus, this paper uses CVAEs for data augmentation on noisy labels.

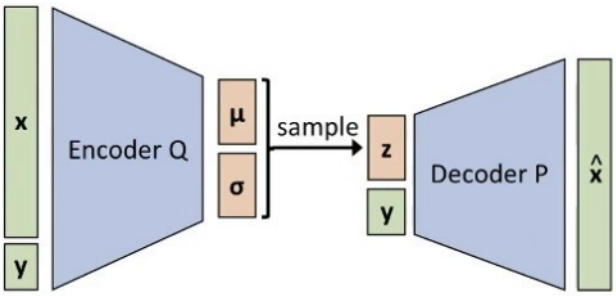


**Figure 3.** Architecture of CVAE consisting of encoder and decoder conditioned on class $y$

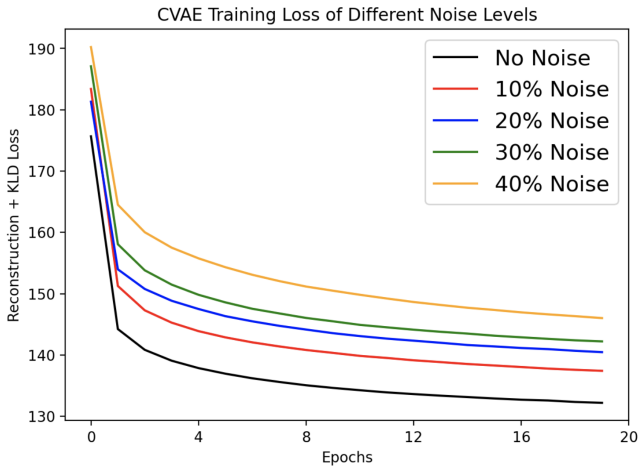


**Figure 4.** CVAE Training Loss throughout training on IDN with varying noise levels

### 3.2 Implementation Details

This paper uses a balanced MNIST dataset with 42000 instances, excluding artificially generated samples. Four CVAEs are trained with noise levels ranging from 10% to 40%, increasing in increments of 10%. For each noise level, we generate an additional 42000 instances using the trained CVAEs. The training phase requires minimal model refinement beyond the typical hyperparameter tuning of learning rate and batch size. To generate the outcomes in Figure 4, a learning rate of 0.0004 and a batch size of 16 are employed on the all models. The consistently decreasing trend in loss indicates that CVAE training remains stable, regardless of the level of noise. The shape remains consistent as the noise increases, but the total loss linearly increases. The increase in total loss is most likely influenced by reconstruction loss and not the KL Divergence loss. While, the KL Divergence loss is independent of the actually input, the reconstruction loss directly deals with finding the distance between the input and the generated output. Thus, as noise increases, the generated image from the latent space becomes distorted leading to a high reconstruction loss. This serves as motivation to potentially experiment with different reconstruction loss functions. Aside from loss functions to improve robustness, there have been attempts in the past to use other forms of generative models. One common model is the cGANS which have been mentioned before. However, cGANs are notorious for mode collapse and training instability even with small amounts of noise [11]. Therefore, CVAE's are superior to cGANs when training on noisy labels. As Figure 4 shows, the CVAE's steady loss curve enables the models to quickly reach a local minimum within 20 epochs.

After training the CVAEs, Figure 5 displays generated number samples (3 and 9) at different noise levels. In 10% and 20% noisy label conditions, the trained CVAE's provide new and meaningful data to the original dataset. The model has learned the underlying structure and patterns of the number 3 and 9. As the noise level increases, the underlying structure of the number becomes harder to discern. A solution to this problem might be to use traditional VAEs to generate images. The objective function of traditional VAEs, as shown in Equation 1 on Page 3, is independent of the label. Thus, the overall loss is guaranteed to be the same in all noisy label conditions. However, the generated images will require manually labeling and does not guarantee a balanced dataset. Since the output is not conditioned on a class, there may be significantly more outputs of the number 0 compared to the rest. Therefore, CVAE is implemented as the deep generative model for data augmentation.

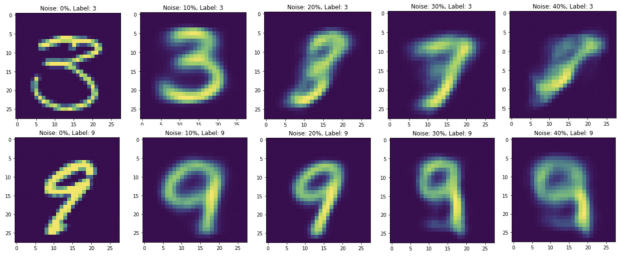

**Figure 5.** CVAE generated numbers (3 and 9) from MNIST on varying noise levels (clean, 10%, 20%, 30%, and 40% noise)

## 4 Experimental Results

In this paper, ResNet18 is employed as the evaluation model to assess the performance of each augmented dataset at different levels of noise. Figure 6 displays the training/validation accuracy and loss over 10 epochs. The highest validation accuracy was typically observed prior to the fifth epoch,

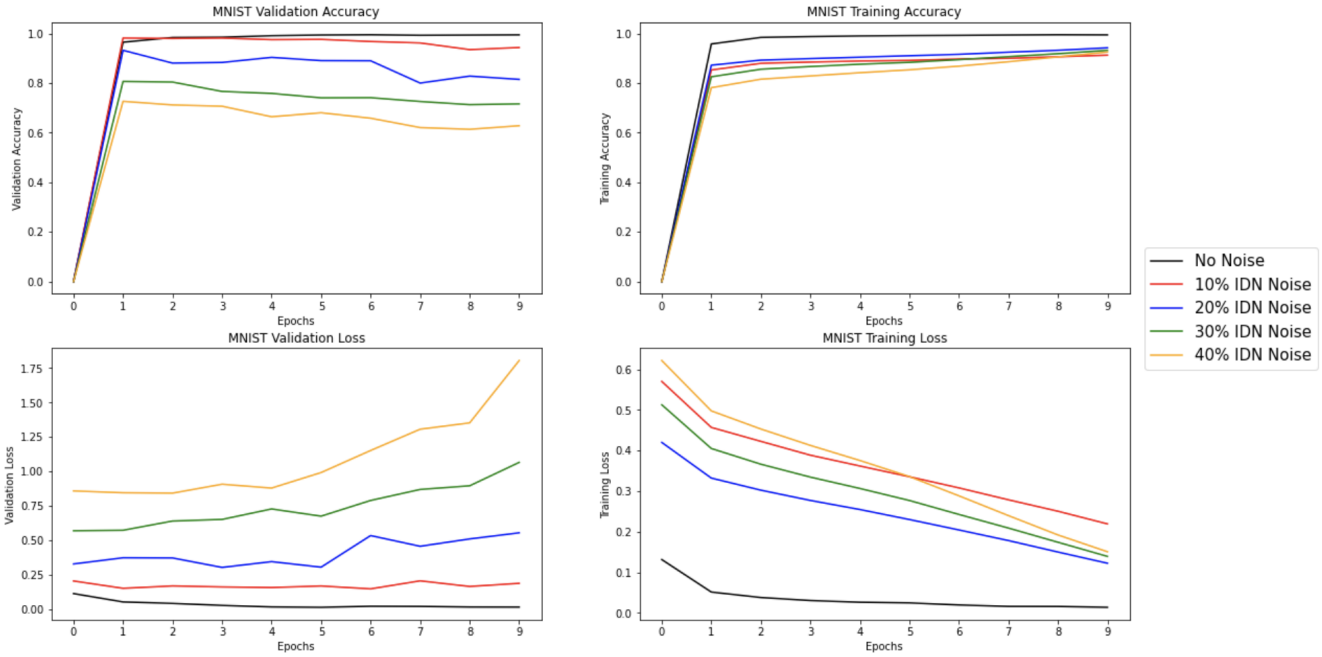


**Figure 6.** Training/Validation Accuracy throughout training on IDN with varying noise levels.

**Table 1.** Testing accuracy (%) on MNIST dataset under different levels of IDN. The * marks published results from Chen et al. [2].

| Method | 10% | 20% | 30% | 40% |
|---|---|---|---|---|
| CE* | 94.07 | 85.62 | 75.75 | 65.83 |
| | ±0.29 | ±0.56 | ±0.09 | ±0.56 |
| Forward* | 93.93 | 85.39 | 76.29 | 68.30 |
| | ±0.14 | ±0.92 | ±0.81 | ±0.42 |
| Co-teaching* | 95.77 | 91.07 | 86.20 | 79.30 |
| | ±0.03 | ±0.19 | ±0.35 | ±0.84 |
| GCE* | 94.56 | 86.71 | 78.32 | 69.78 |
| | ±0.31 | ±0.47 | ±0.43 | ±0.58 |
| DAC* | 94.13 | 85.63 | 75.82 | 65.69 |
| | ±0.02 | ±0.56 | ±0.58 | ±0.78 |
| DMI* | 94.21 | 87.02 | 76.19 | 67.65 |
| | ±0.12 | ±0.42 | ±0.64 | ±0.73 |
| SEAL* | 96.75 | 93.63 | 88.52 | 80.73 |
| | ±0.08 | ±0.33 | ±0.15 | ±0.41 |
| **CVAE** | **97.02** | 88.53 | 75.82 | 60.46 |

after which it decreased as a result of overfitting. The best-performing model on the validation set was selected for evaluation on the test set.

To ensure fair comparisons, the benchmark used to evaluate all results in Table 1 is the same, and the IDN is generated consistently across all methods. Asides from the CVAE, all reported findings are sourced from Chen et al.'s study [2].

The application of CVAE for data augmentation produces the most precise outcomes for 10% noise. CVAE's results remain competitive with other techniques up to 20% noise; nevertheless, the performance rapidly deteriorates as the noise level surpasses 20%.

For the 10% noise condition, the significant improvement in the model's accuracy after data augmentation reveals that the generated data are unseen so that they are informative for the training. One possible reason for this superior performance is due to the mechanism of the IDN generator. Since the IDN generator swaps the true class label with the next most similar one, both class labels may provide useful information.

As an example, take the third instance of the number 2 in Figure 1, Page 2. The generator switches the label of the number 2 with that of number 1 because of their shared features. We suppose that this swap is not harmful to the training set since this instance can provide new information to features associated with the label 1. This case occurs only when the shared features exhibit a high degree of similarity.

In low noise scenarios, the generator maintains discriminative features of each class as discussed above. Figure 2 depicts this visually, where the t-SNE plot illustrates well-defined clusters in the 10% noise condition. Consequently, the CVAE can learn the underlying patterns within each class so that it can generate unseen data.

## 5 Conclusion

This paper aims to investigate the importance of utilizing controlled experiments with IDN as opposed to CCN or symmetric noise. To achieve this objective, we put forth a deep generative model, namely CVAE, for data augmentation. Our findings suggest that CVAEs performs well under low IDN conditions due to the underlying mechanisms of IDN. We hope to facilitate further research around learning in situations involving IDN.

## References


[1] Alexander Camuto, Matthew Willetts, Stephen Roberts, Chris Holmes, and Tom Rainforth. 2021. Towards a Theoretical Understanding of the Robustness of Variational Autoencoders. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*, Arindam Banerjee and Kenji Fukumizu (Eds.). PMLR, 3565–3573. https://proceedings.mlr.press/v130/camuto21a.html

[2] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. 2021. Beyond Class-Conditional Assumption: A Primary Attempt to Combat Instance-Dependent Label Noise. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (May 2021), 11442–11450. https://doi.org/10.1609/aaai.v35i13.17363

[3] Dan Hendrycks and Thomas G. Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *CoRR* abs/1903.12261 (2019). arXiv:1903.12261 http://arxiv.org/abs/1903.12261

[4] Lu Jiang, Di Huang, and Weilong Yang. 2019. Synthetic vs Real: Deep Learning on Controlled Noise. *CoRR* abs/1911.09781 (2019), 1–2. arXiv:1911.09781 http://arxiv.org/abs/1911.09781

[5] Diederik P. Kingma, Danilo Jimenez Rezende, Shakir Mohamed, and Max Welling. 2014. Semi-Supervised Learning with Deep Generative Models. *CoRR* abs/1406.5298 (2014). arXiv:1406.5298 http://arxiv.org/abs/1406.5298

[6] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]

[7] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. 2017. WebVision Database: Visual Learning and Understanding from Web Data. arXiv:1708.02862 [cs.CV]

[8] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. 2017. Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach. arXiv:1609.03683 [stat.ML]

[9] Shaeke Salman and Xiuwen Liu. 2019. Overfitting Mechanism and Avoidance in Deep Neural Networks. *CoRR* abs/1901.06566 (2019). arXiv:1901.06566 http://arxiv.org/abs/1901.06566

[10] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Push the Student to Learn Right: Progressive Gradient Correcting by Meta-learner on Corrupted Labels. *CoRR* abs/1902.07379 (2019). arXiv:1902.07379 http://arxiv.org/abs/1902.07379

[11] Kiran Koshy Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. 2018. Robustness of Conditional GANs to Noisy Labels. arXiv:1811.03205 [stat.ML]

[12] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning From Massive Noisy Labeled Data for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *CoRR* abs/1805.07836 (2018). arXiv:1805.07836 http://arxiv.org/abs/1805.07836