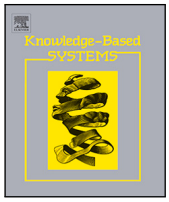
Knowledge-Based Systems 296 (2024) 111911

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Maximizing discrimination capability of knowledge distillation with energy function

Seonghak Kim[1], Gyeongdo Ham[1], Suin Lee, Donggon Jang, Daeshik Kim *

School of Electrical Engineering, Korea Advanced Institute of Science and Technology, 291, Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea

ARTICLE INFO

Keywords:
Knowledge distillation
Energy function
Temperature adjustment
Data augmentation
Resource-limited device

ABSTRACT

To apply the latest computer vision techniques that require a large computational cost in real industrial applications, knowledge distillation methods (KDs) are essential. Existing logit-based KDs apply the constant temperature scaling to all samples in dataset, limiting the utilization of knowledge inherent in each sample individually. In our approach, we classify the dataset into two categories (i.e., low energy and high energy samples) based on their energy score. Through experiments, we have confirmed that low energy samples exhibit high confidence scores, indicating certain predictions, while high energy samples yield low confidence scores, meaning uncertain predictions. To distill optimal knowledge by adjusting non-target class predictions, we apply a higher temperature to low energy samples to create smoother distributions and a lower temperature to high energy samples to achieve sharper distributions. When compared to previous logit-based and feature-based methods, our energy-based KD (Energy KD) achieves better performance on various datasets. Especially, Energy KD shows significant improvements on CIFAR-100-LT and ImageNet datasets, which contain many challenging samples. Furthermore, we propose high energy-based data augmentation (HE-DA) for further improving the performance. We demonstrate that higher performance improvement could be achieved by augmenting only a portion of the dataset rather than the entire dataset, suggesting that it can be employed on resource-limited devices. To the best of our knowledge, this paper represents the first attempt to make use of energy function in knowledge distillation and data augmentation, and we believe it will greatly contribute to future research.

1. Introduction

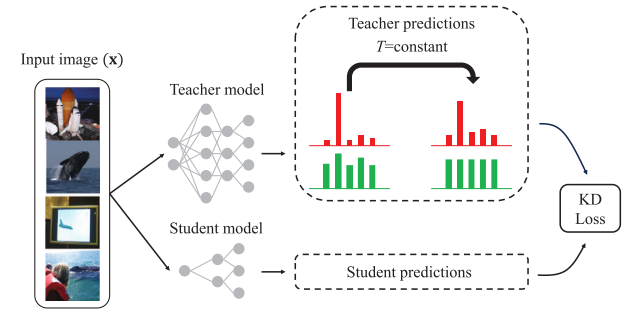
In recent years, computer vision has witnessed significant advancements, notably in areas like image classification [1,2], object detection [3,4], and image segmentation [5,6], primarily driven by the emergence of deep learning. However, the high-performance requirements of these deep learning models have led to their substantial size, resulting in significant computational costs. This poses challenges for practical deployment in real-world industries. To address these limitations, model compression methods, including model pruning [7], quantization [8], and knowledge distillation (KD) [9], have been proposed. Among these, KD stands out for its superior performance and ease of implementation, making it widely adopted in various computer vision applications. KD involves training a lightweight student model by distilling meaningful information from a more complex teacher model, enabling the student model to achieve performance similar to that of the teacher model.

Since its introduction by Hinton [10], KD has evolved into two main approaches: logit-based [11] and feature-based [12] distillation. Logit-based methods use final predictions for training the student, while feature-based methods leverage information from intermediate layers. Although feature-based methods are generally known to outperform logit-based ones, they may be challenging to use in real-world applications due to potential privacy and safety concerns associated with accessing intermediate layers of the teacher model. Hence, this paper focuses on logit-based distillation, offering practical advantages for real-world applications by not requiring access to intermediate layers.

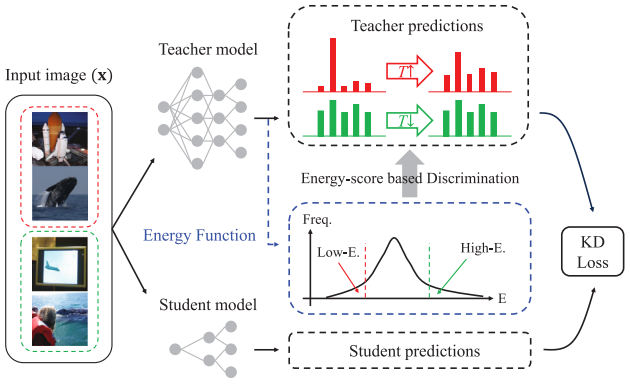
We propose a novel logit-based distillation method designed for easy integration into existing logit-based KDs. This method significantly enhances performance by maximizing the utilization of teacher knowledge by the students. As illustrated in Fig. 1, applying an energy function to each image categorizes the entire dataset into low-energy and high-energy samples. We then apply different temperature scaling to the separated samples, employing high temperature for low-energy and low temperature for high-energy samples. This approach results in smoother distributions from low-energy samples and sharper distributions from high-energy samples, effectively adjusting non-target

* Corresponding author.
E-mail addresses: hakk35@kaist.ac.kr (S. Kim), rudeh6185@kaist.ac.kr (G. Ham), suinlee@kaist.ac.kr (S. Lee), jdg900@kaist.ac.kr (D. Jang), daeshik@kaist.ac.kr (D. Kim).
[1] Authors contributed equally.

https://doi.org/10.1016/j.knosys.2024.111911
Received 14 February 2024; Received in revised form 25 April 2024; Accepted 4 May 2024
Available online 8 May 2024



(a) Knowledge distillation with constant temperature scaling



(b) Knowledge distillation with different temperature scaling based on the energy scores

Fig. 1. Schematic diagram of conventional knowledge distillation and our method: (a) constant temperature scaling, (b) different temperature scaling. Our method receives the energy score of each sample from the blue dashed line.

class predictions for optimal knowledge distillation. Consequently, our method significantly improves the performance of the student model.

In addition, we propose High Energy-based Data Augmentation (HE-DA) to further enhance performance. Unlike previous augmentation-based KD methods that apply augmentation to the entire dataset, HE-DA achieves similar or even better performance by utilizing only 20% to 50% of the dataset, offering practical advantages in terms of storage and computational cost.

Through extensive experimentation on commonly used classification datasets, such as CIFAR-100 [13], TinyImageNet, and ImageNet [14], we have verified that our proposed methods outperform existing state-of-the-art approaches, particularly demonstrating strengths in the case of challenging datasets, such as CIFAR-100-LT and ImageNet.

2. Related works

2.1. Knowledge distillation

Knowledge distillation (KD) is a technique used to enhance the performance of lightweight student networks by leveraging the dark knowledge embedded in large teacher networks. Over the years, KD methods have evolved to narrow the performance gap between student and teacher models by utilizing both final predictions, known as logits-based distillation [10,11,15–17], and intermediate features, known as features-based distillation [12,18–28].

Previous works on logits-based distillations include the following: DML [15] proposed a mutual learning strategy for collaboratively teaching and learning between student and teacher models; TAKD [16] introduced a multi-step method with an intermediate-size network (i.e., assistant network) to bridge the gap between teachers and students; DKD [11] decomposed the soft-label distillation loss into two components: target class knowledge distillation (TCKD) and non-target class knowledge distillation (NCKD), enabling each part to independently harness its effectiveness; Multi KD [17] proposed multi-level prediction alignment, containing instance, batch, and class levels, and prediction augmentation. While these approaches emphasize effective knowledge transfer, they do not consider dividing entire datasets or provide mechanisms to distinguish and transfer knowledge from specific samples.

FitNet [18] was groundbreaking as it leveraged not only the final outputs but also intermediate representations. Since the introduction of FitNet, various feature-based KD methods have emerged as follows: AT [22] prompted the student to mimic the attention map of the teacher network; PKT [19] employed various kernels to estimate the probability distributions, employing different divergence metrics for distillation; RKD [20] focused on transferring the mutual relations of data samples; CRD [21] framed the objective as contrastive learning for distillation; VID [23] took a different approach by maximizing mutual information; OFD [24] introduced a novel loss function incorporating teacher transform and a new distance function; Review KD [12] introduced a review mechanism that leverages past features for guiding current ones through residual learning. Additionally, they incorporated attention-based fusion (ABF) and hierarchical context loss (HCL) to further enhance performance.

More recently, several high-performing feature-based distillation methods have emerged [25–28]. Unlike conventional feature-based methods, which primarily rely on measuring model similarity using isotropic L2 distance, FCFD [25] focused on optimizing the functional similarity between teacher and student features. By considering the anisotropic nature of neural network operations, FCFD ensured that the student learns more effectively from the teacher. Z. Guo introduced CAT-KD [26], which achieves both high interpretability and competitive performance by transferring class activation maps. This approach is grounded in the understanding that the CNN's ability to distinguish category regions is vital for its performance. J. Li proposed LSH-TL [27], which utilizes two algorithms by applying teacher labels and designing feature temperature parameters. This method addresses the challenge of training data not aligning with the ground-truth distribution of classes and features. Lastly, SAKD [28] was presented, employing the sparse attention mechanism by treating student features as queries and teacher features as key values. It utilizes sparse attention values through random deactivation to adjust the feature distances.

While feature-based methods have demonstrated superior performance compared to logits-based ones, attributed to their ability to leverage more information from intermediate layers rather than relying solely on final predictions, they often present challenges in real-world applications. This is due to privacy and safety concerns associated with accessing the intermediate layers of the teacher model. Therefore, in this paper, we shift our focus to logits-based distillation, which offers practical advantages for real-world deployment. To achieve comparable or even superior results to feature-based KDs, we introduce a novel perspective to knowledge distillation: regulating knowledge transfer based on the energy scores of samples.

Our method distinguishes itself from previous KD methods by its compatibility with existing state-of-the-art methods. Unlike previous approaches, which are often limited to standalone use and cannot be easily combined with emerging methods, our approach seamlessly integrates with various methods, providing potential for further performance enhancement. Moreover, our method exhibits significant performance improvements when confronted with challenging samples, rendering it particularly suitable for real-world scenarios. The Energy-based KD (Energy KD) proposed herein represents a significant advancement in the development of more effective and efficient knowledge distillation techniques.

2.2. Energy-based learning

Energy-based machine learning models have a long history, beginning with the Boltzmann machine [29,30], a network of units with associated energy for the entire network. Energy-based learning [31–33] offers a unified framework for various probabilistic and non-probabilistic learning approaches. Recent research [34] demonstrated the use of energy functions to train generative adversarial networks (GANs), where the discriminator utilizes energy values to differentiate between real and generated images. Xie [35–37] also established the connection between discriminative neural networks and generative random field models. Subsequent studies have explored the application of energy-based models in video generation and 3D geometric patterns. Liu [38] demonstrated that non-probabilistic energy scores can be directly used in a score function for estimating out-of-distribution (OOD) uncertainty. They show that these optimization goals fit more naturally into an energy-based model than a generative model and enable the exploitation of modern discriminative neural architectures.

Building upon these prior works, our proposed framework extends the use of non-probabilistic energy values to knowledge distillation and data augmentation. Notably, our framework provides different knowledge for low energy and high energy samples, representing a novel contribution.

3. Methods

3.1. Background

Our approach revolves around categorizing each sample in the dataset into two groups: low-energy and high-energy groups. These groups are determined by the energy function $E(\cdot)$, which maps the input $\mathbf{x}_i$, having dimension d, to a single, non-probabilistic scalar value. (i.e., $E(\mathbf{x}_i) : \mathbb{R}^d \rightarrow \mathbb{R}$) [31]. The representation of the energy function with the neural network f is as follows:

$$E(\mathbf{x}_i; f) = -T^E \cdot \log \sum_{j=1}^{K} e^{z_j^f(\mathbf{x}_i)/T^E}, \tag{1}$$

where $z_j^f(\mathbf{x}_i)$ indicates the logit corresponding to the j class label of input image sample i using the neural network f, T^E is the temperature parameter for the energy score, and K denotes the total number of classes. Appendix C contains a list of all the symbols mentioned in this paper.

The motivation behind segregating categories according to energy scores is that we can regard input data with low likelihood as high-energy samples [39]. This can be achieved by utilizing the data's density function $p(\mathbf{x}_i)$ expressed by the energy-based model [31,40].

$$p(\mathbf{x}_i) = \frac{e^{-E(\mathbf{x}_i;f)/T^E}}{\int_{\mathbf{x}} e^{-E(\mathbf{x}_i;f)/T^E}}, \tag{2}$$

where $\mathbf{x}$ denotes the entire dataset and the denominator can be disregarded since it remains constant independently (i.e., $\int_{\mathbf{x}} e^{-E(\mathbf{x}_i;f)/T^E} = C$). Therefore, it can be expressed by

$$\log p(\mathbf{x}_i) = -\frac{E(\mathbf{x}_i; f)}{T^E} - \log C. \tag{3}$$

This equation shows that the energy function is proportional to the log likelihood function. In other words, samples with lower energy have a higher probability of occurrence, indicating a *certain image*, while samples with higher energy have a lower probability of occurrence, referring a *uncertain image*. This distinguishable nature of the energy function can be effectively utilized to categorizes samples, thereby facilitating optimal knowledge distillation. To visually explore the relationship between energy scores and image, refer to Fig. 2, which shows images associated with low and high energy, respectively. Consequently, the energy score, being a valuable tool for dataset division, can be employed in both knowledge distillation (KD) and data augmentation (DA) separately. Further elaboration on each method will be provided in the subsequent sections.

3.1.1. Low and high energy samples

To validate the insights gained from the energy scores, it is valuable to visualize the images belonging to both the low-energy and high-energy.

Fig. 2 illustrates the categorization of ImageNet based on the energy score of each image, dividing them into categories of low-energy and high-energy samples. The red boxes depicts images with low-energy scores, effectively representing their respective classes. We have denoted this category as *certain images*. On the other hand, the green boxes displays images with high-energy scores, indicating either a confused label or a mixture of different objects. These images have been designated as *uncertain images*.

Fig. 3 demonstrates the average predictions for *certain* and *uncertain* images. Certain images exhibit high confidence scores and possess insufficient knowledge about non-target predictions, while uncertain images showcase low confidence scores and a relatively uniform distribution. It is worth noting that the predictions presented in Fig. 3 support the classification of each image as either certain or uncertain. These findings align with prior research [39] that higher energy levels are associated with out-of-distribution (OOD) data. An important difference is that we categorize low-energy and high-energy data within the same dataset based on our criteria. Additional images are available in Appendix B.

As a result, it is reasonable to utilize higher temperature scaling for low-energy samples to create smoother predictions and lower temperature scaling for high-energy samples to achieve sharper predictions during the distillation process. This ensures that the teacher model optimally transfers its knowledge to the student model.

3.2. EnergyKD: Energy-based knowledge distillation

Utilizing the mentioned energy score, we propose an Energy-based Knowledge Distillation (Energy KD), where the differences between low and high energy allow effective transfer of knowledge. Specifically, we obtain the energy score for each image sample through the logits of pre-trained teacher models using Eq. (1). After classifying the images into low-energy and high-energy groups based on their energy scores, we apply distinct softmax temperature scaling to each group, thereby enhancing the student model's ability to learn a more diverse range of information.

First of all, we consider teacher network $f_{\mathcal{T}}$ and student network f_S, which maps input image $\mathbf{x}_i$ with dimension d to number of classes K (i.e., $f_{\mathcal{T},S} : \mathbb{R}^d \rightarrow \mathbb{R}^K$) as follows:

$$f_{\mathcal{T}} = f_{\mathcal{T}}(\mathbf{x}_i; \theta_{\mathcal{T}}) \tag{4}$$

$$f_S = f_S(\mathbf{x}_i; \theta_S) \tag{5}$$

Here, $\theta_{\mathcal{T}}$ and θ_S are the parameters of each model. The energy score $\mathcal{E}$ for each sample $\mathbf{x}_i$ can be calculated using the teacher network $f_{\mathcal{T}}$ as follows:

$$\mathcal{E}_{\mathcal{T}}^{(i)} = E(\mathbf{x}_i; f_{\mathcal{T}}). \tag{6}$$

We can obtain the energy score for all images in the training dataset and arrange them in ascending order. Subsequently, images with lower energy values are categorized into the certain group, while those with higher energy values are assigned to the uncertain group.

In contrast to the conventional KD loss $\mathcal{L}_{\text{KD}}$, which employs the same temperature scaling for all images, as indicated by

$$\mathcal{L}_{\text{KD}}(\mathbf{x}_i; f_S, f_{\mathcal{T}}, T) = \mathcal{D}_{\text{KL}}\left(\sigma\left(\frac{z^{f_{\mathcal{T}}}}{T}\right) \middle\| \sigma\left(\frac{z^{f_S}}{T}\right)\right), \tag{7}$$

where $\mathcal{D}_{\text{KL}}$ denotes Kullback–Leibler divergence, $\sigma(\cdot)$ is the softmax function, T is the temperature scaling factor, and $z^{f_{\mathcal{T}}}, z^{f_S}$ indicate the logit using the teacher network $f_{\mathcal{T}}$, student network f_S, respectively.

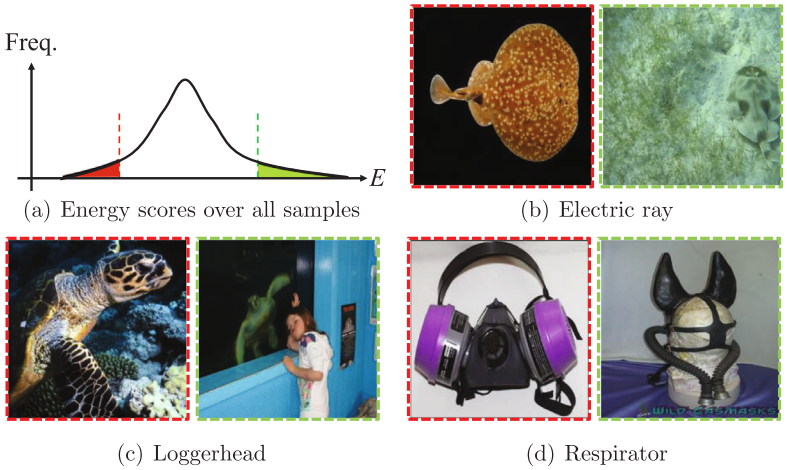


Fig. 2. ImageNet samples categorized according to their energy scores obtained from ResNet32x4. The red boxes belong to the certain images and have low energy scores, accurately representing their assigned labels. The green boxes are relative to the uncertain images and have high energy scores, not clearly reflecting their assigned labels.

Our method adjusts the confidence of predictions based on the energy score, enabling the student to acquire a broader range of knowledge. This adjustment can be utilized by simply changing the temperature scaling factor ($T \rightarrow T_{\text{ours}}$) as follows:

$$\mathcal{L}_{\text{ours}}(\mathbf{x}_i; f_S, f_\mathcal{T}, T_{\text{ours}}) = \mathcal{D}_{\text{KL}}\left(\sigma\left(\frac{z^{f_\mathcal{T}}}{T_{\text{ours}}}\right)\middle\|\sigma\left(\frac{z^{f_S}}{T_{\text{ours}}}\right)\right) \tag{8}$$

$$T_{\text{ours}} = \begin{cases} T + T_{(-)}, & \mathcal{E}_\mathcal{T}^{(i)} \geq \mathcal{E}_\mathcal{T}^{\text{high}} = \mathcal{E}_\mathcal{T}[-N \cdot r] \\ T + T_{(+)}, & \mathcal{E}_\mathcal{T}^{(i)} \leq \mathcal{E}_\mathcal{T}^{\text{low}} = \mathcal{E}_\mathcal{T}[N \cdot r] \\ T, & \text{else,} \end{cases} \tag{9}$$

where $\mathcal{E}_\mathcal{T}^{\text{high}}$ and $\mathcal{E}_\mathcal{T}^{\text{low}}$ are constant values that define the range of high-energy and low-energy classifications. $T_{(-)}$ is a negative integer used to decrease the temperature, facilitating the transfer of more target predictions for uncertain samples. Conversely, $T_{(+)}$ is a positive integer used to increase the temperature, enabling certain samples to incorporate more non-target knowledge. N represents the total number of training samples, and to establish the range based on energy, we employed a percentage of the total samples denoted by $r = \{0.2, 0.3, 0.4, 0.5\}$. The parentheses $[\cdot]$ indicate the index of the array, as shown in Fig. 4 for a more intuitive understanding.

Fig. 4 illustrate how our method divides the entire dataset into low-energy and high-energy samples. To facilitate understanding, we employ a dataset comprising 10 samples and set the percentage parameter r to 0.4. Eq. (6) calculates the energy score for each data sample, and then we arrange in ascending order (i.e., $\mathcal{E}_\mathcal{T}^{(1)} < \mathcal{E}_\mathcal{T}^{(2)} < \cdots < \mathcal{E}_\mathcal{T}^{(10)}$). To aid comprehension, we introduce two different indexes: negative and positive indexes. As shown in Eq. (9), $\mathcal{E}_\mathcal{T}^{\text{high}}$ is set to $\mathcal{E}_\mathcal{T}[-4]$ and $\mathcal{E}_\mathcal{T}^{\text{low}}$ is set to $\mathcal{E}_\mathcal{T}[4]$. Consequently, samples with energy score $\mathcal{E}_\mathcal{T}[-4], \mathcal{E}_\mathcal{T}[-3], \mathcal{E}_\mathcal{T}[-2]$, and $\mathcal{E}_\mathcal{T}[-1]$ (equivalent to $\mathcal{E}_\mathcal{T}[7], \mathcal{E}_\mathcal{T}[8], \mathcal{E}_\mathcal{T}[9]$, and $\mathcal{E}_\mathcal{T}[10]$), equal to or greater than $\mathcal{E}_\mathcal{T}[-4]$, belong to high-energy samples, Conversely, samples with energy score $\mathcal{E}_\mathcal{T}[1], \mathcal{E}_\mathcal{T}[2], \mathcal{E}_\mathcal{T}[3]$, and $\mathcal{E}_\mathcal{T}[4]$, equal to or less than $\mathcal{E}_\mathcal{T}[4]$, belong to low-energy sample.

Hinton's paper [10] introducing the concept of KD softened the probabilities by setting the temperature used for softmax higher than normal ($T = 1$). These relative probabilities of incorrect answers provide valuable insights into the generalization tendencies of the complex model [10]. Since Hinton's work, this information in soft targets has been termed *dark knowledge* [9,41,42]. Following this line of research, we also adopted the term *dark knowledge* to grasp the useful probability information. Previous research indicated that, in order to enhance the performance of the KD, *dark knowledge* must be appropriately distributed [43]. Our approach can increase important *dark knowledge* about non-target classes in the low energy sample while increasing predictions of the target class in the high-energy sample.

3.3. HE-DA: High energy-based data augmentation

We propose an additional technique, High Energy-based Data Augmentation (HE-DA), where data augmentation is selectively applied only to image samples belonging to high-energy groups, which have already been classified for Energy KD. In conventional knowledge distillation (KD), data augmentation (DA) is frequently employed across the entire dataset to enhance the generalization and performance of the student model. However, the straightforward application of DA may result in a significant increase in computational costs due to the doubling of the dataset.

To efficiently apply DA to KD, we present an augmentation method that focuses on specific samples (i.e., uncertain samples) instead of augmenting the entire dataset. This approach is rooted in the concept that certain samples already contain sufficient information, whereas uncertain samples require additional information to elucidate ambiguous content. Eq. (3) and Fig. 2 demonstrates that high-energy samples correspond to uncertain samples. Consequently, our focus is directed towards augmenting the samples within the high-energy group, aiming to provide students with more information to enhance their performance.

In our Energy KD approach, we sorted the energy scores obtained from the teacher model in ascending order. The samples with lower values are categorized as part of the low-energy dataset $\mathbf{x}_{\text{low}}$ within the entire dataset $\mathbf{x}$, while the samples with higher values are classified as belonging to the high-energy dataset $\mathbf{x}_{\text{high}}$ as follows:

$$\mathbf{x} = \{\mathbf{x}_{\text{low}}, \mathbf{x}_{\text{else}}, \mathbf{x}_{\text{high}}\} \tag{10}$$

$$\mathbf{x}_i = \begin{cases} \mathbf{x}_{\text{high}}, & \mathcal{E}_\mathcal{T}^{(i)} \geq \mathcal{E}_\mathcal{T}^{\text{high}} = \mathcal{E}_\mathcal{T}[-N \cdot r] \\ \mathbf{x}_{\text{low}}, & \mathcal{E}_\mathcal{T}^{(i)} \leq \mathcal{E}_\mathcal{T}^{\text{low}} = \mathcal{E}_\mathcal{T}[N \cdot r] \\ \mathbf{x}_{\text{else}}, & \text{else,} \end{cases} \tag{11}$$

where $\mathbf{x}_{\text{else}}$ represents datasets that do not belong to either low-energy or high-energy data and the parentheses $[\cdot]$ denote the index of the array, as mentioned earlier. For a clearer understanding, please refer back to Fig. 4. We exclusively apply augmentation to samples that were classified as part of the high energy $\mathbf{x}_{\text{high}}^{\text{aug}}$ as follows:

$$\mathbf{x}_{\text{high}}^{\text{aug}} = G_{\text{aug}}(\mathbf{x}_{\text{high}}), \tag{12}$$

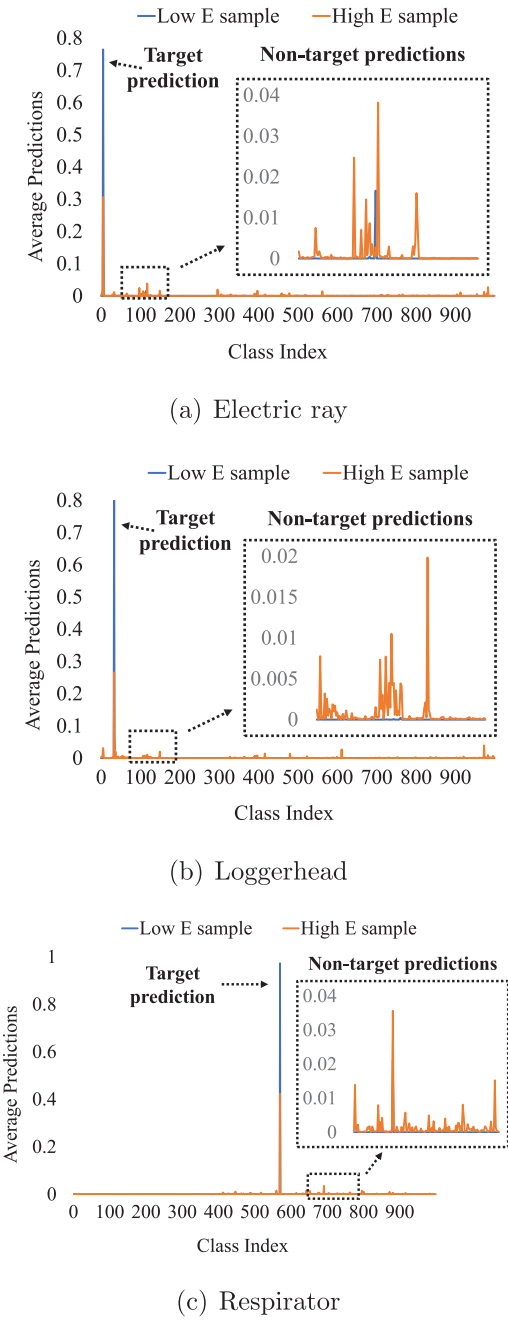
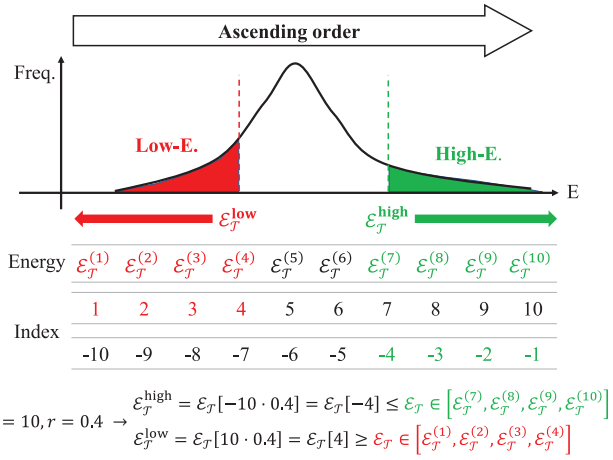


Fig. 3. Average predictions for particular classes with low energy (blue line) and high energy (red line) samples. Low energy samples exhibit high confidence scores and lack substantial dark knowledge, whereas high energy samples display low confidence scores and have inordinate knowledge.

where G_{aug} refers the data augmentation function, here, we applied CutMiX [44] and MixUp [45].

Despite utilizing only an additional 20% to 50% of the training data, our method outperforms existing approaches, which use the entire dataset, yielding superior results while simultaneously reducing computational costs. It is noteworthy that our approach demonstrates higher performance than applying data augmentation only to low-energy samples or applying data augmentation to both low and high-energy samples. Please refer to Section 4.5 for more details. The results with MixUp are included in Appendix B.



$$N = 10, r = 0.4 \rightarrow \begin{aligned} \mathcal{E}_{\mathcal{T}}^{\text{high}} &= \mathcal{E}_{\mathcal{T}}[-10 \cdot 0.4] = \mathcal{E}_{\mathcal{T}}[-4] \leq \mathcal{E}_{\mathcal{T}} \in \left[\mathcal{E}_{\mathcal{T}}^{(7)}, \mathcal{E}_{\mathcal{T}}^{(8)}, \mathcal{E}_{\mathcal{T}}^{(9)}, \mathcal{E}_{\mathcal{T}}^{(10)}\right] \\ \mathcal{E}_{\mathcal{T}}^{\text{low}} &= \mathcal{E}_{\mathcal{T}}[10 \cdot 0.4] = \mathcal{E}_{\mathcal{T}}[4] \geq \mathcal{E}_{\mathcal{T}} \in \left[\mathcal{E}_{\mathcal{T}}^{(1)}, \mathcal{E}_{\mathcal{T}}^{(2)}, \mathcal{E}_{\mathcal{T}}^{(3)}, \mathcal{E}_{\mathcal{T}}^{(4)}\right] \end{aligned}$$

Fig. 4. Energy distribution across the entire datasets. This illustrated example assumes that there are 10 image samples and sets the percentage of the total samples to 40%.

4. Experiments

The performance of our method is evaluated by comparing it to previous knowledge distillations such as KD [10], AT [22], OFD [24], CRD [21], FitNet [18], PKT [19], RKD [20], VID [23], DML [15], TAKD [16], DKD [11], ReviewKD [12], Multi KD [17], FCFD [25], CAT-KD [26], LSH-TL [27], and SAKD [28] considering various architectural configurations including ResNet [46], WideResNet [47], VGG [48], MobileNet [49], and ShuffleNet [50,51]. Details of the implementation can be found in Appendix A. *All experiments were conducted three times, and the reported results represent the average values.*

4.1. Datasets

CIFAR-100 [13] is a widely used dataset for image classification, consisting of 100 classes. The samples have a resolution of 32 × 32 pixels, and the dataset includes 50,000 training images and 10,000 testing images.

ImageNet [14] is a comprehensive dataset extensively employed for image classification. It comprises 1,000 classes, and the samples are of size 224 × 224 pixels. The training set is notably large, containing 1.28 million images, while the test set consists of 5,000 images.

TinyImageNet is a scaled-down version of ImageNet, featuring 200 classes with images sized 64 × 64 pixels. The dataset includes 500 training images, 50 validation images, and 50 testing images for each class.

4.2. Effect of EnergyKD

Table 1 displays the results obtained using the same architecture for both teacher and student models on the CIFAR-100 dataset, while Table 2 showcases the results obtained with different architectures. Previous methods can be categorized into two types: feature-based methods and logit-based methods, and the results from previous papers on each method are recorded.

The tables consistently demonstrate that the application of our method to previous logit-based KD results in higher performance compared to not applying it, regardless of structural differences between the student and teacher models. In the case of vanilla KD, our method (Energy KD) yields a performance gain of up to 1.6.

Additionally, when integrating our method into recently developed logit-based methods like DKD and Multi KD (Energy DKD and Energy Multi), we observe performance gains of up to 0.5 and 0.6, respectively. Notably, our method, despite being logit-based, outperforms the state-of-the-art feature-based KD for all same architectures. Additionally,

Table 1
Top-1 accuracy (%) on the CIFAR-100 test sets when using teacher and student models with the same architectures. Our results, highlighted in **bold**, demonstrate exceptional performance compared to the results obtained without employing our method.

Teacher	WRN-40–2	WRN-40–2	ResNet56	ResNet110	ResNet32 × 4	VGG13
Acc.	75.61	75.61	72.34	74.31	79.42	74.64
Student	WRN-16–2	WRN-40–1	ResNet20	ResNet32	ResNet8 × 4	VGG8
Acc.	73.26	71.98	69.06	71.14	72.50	70.36
FitNet	73.58	72.24	69.21	71.06	73.50	71.02
PKT	74.54	73.54	70.34	72.61	73.64	72.88
RKD	73.35	72.22	69.61	71.82	71.90	71.48
CRD	75.48	74.14	71.16	73.48	75.51	73.94
AT	74.08	72.77	70.55	72.31	73.44	71.43
VID	74.11	73.30	70.38	72.61	73.09	71.23
OFD	75.24	74.33	70.98	73.23	74.95	73.95
ReviewKD	76.12	75.09	71.89	73.89	75.63	74.84
FCFD	76.34	75.43	71.68	–	76.80	74.86
CAT-KD	75.60	74.82	71.62	73.62	76.91	74.65
LSH-TL	76.42	74.50	71.50	74.10	76.73	74.11
SAKD	75.86	75.00	71.93	73.92	76.16	74.66
DML	73.58	72.68	69.52	72.03	72.12	71.79
TAKD	75.12	73.78	70.83	73.37	73.81	73.23
KD	74.92	73.54	70.66	73.08	73.33	72.98
Energy KD	**75.45**	**74.28**	**71.30**	**73.68**	**74.60**	**73.73**
DKD	76.24	74.81	71.97	74.11	76.32	74.68
Energy DKD	**76.66**	**74.97**	**72.10**	**74.11**	**76.78**	**74.90**
Multi KD	76.63	75.35	72.19	74.11	77.08	75.18
Energy Multi	**77.19**	**75.70**	**72.76**	**74.60**	**77.31**	**75.56**

Table 2
Top-1 accuracy (%) on the CIFAR-100 test sets when using teacher and student models with the different architectures. Our results, highlighted in **bold**, demonstrate exceptional performance compared to the results obtained without employing our method.

Teacher	WRN-40–2	ResNet50	ResNet32 × 4	ResNet32 × 4	VGG13
Acc.	75.61	79.34	79.42	79.42	74.64
Student	ShuffleNetV1	MobileNetV2	ShuffleNetV1	ShuffleNetV2	MobileNetV2
Acc.	70.50	64.60	70.50	71.82	64.60
FitNet	73.73	63.16	73.59	73.54	64.14
PKT	73.89	66.52	74.10	74.69	67.13
RKD	72.21	64.43	72.28	73.21	64.52
CRD	76.05	69.11	75.11	75.65	69.73
AT	73.32	58.58	71.73	72.73	59.40
VID	73.61	67.57	73.38	73.40	65.56
OFD	75.85	69.04	75.98	69.82	68.48
ReviewKD	77.14	69.89	77.45	77.78	70.37
FCFD	77.81	71.07	78.12	78.20	70.67
CAT-KD	77.35	71.36	78.26	78.41	69.13
LSH-TL	76.62	68.02	75.80	76.62	68.17
SAKD	76.77	–	76.32	77.21	–
DML	72.76	65.71	72.89	73.45	65.63
TAKD	75.34	68.02	74.53	74.82	65.63
KD	74.83	67.35	74.07	74.45	67.37
Energy KD	**75.90**	**68.97**	**75.20**	**75.87**	**68.65**
DKD	76.70	70.35	76.45	77.07	69.71
Energy DKD	**77.06**	**70.77**	**76.89**	**77.55**	**70.19**
Multi KD	77.44	71.04	77.18	78.44	70.57
Energy Multi	**77.76**	**71.32**	**77.82**	**78.64**	**70.89**

for different architectures, we observe that our method yields results almost comparable to those of the state-of-the-art feature-based methods.

These results suggest that our method holds the potential for seamless integration into future logit-based methods, providing a pathway to further enhance performance. This underscores the superiority of our method for real-world applications, particularly in scenarios where utilizing an intermediate layer is challenging.

Table 3 presents the performance of our methods on ImageNet. Notably, even on ImageNet, considered a more challenging dataset than CIFAR-100, our method demonstrates significant improvements over other distillation methods. This improvement is attributed to the optimization of knowledge distillation for this challenging dataset, achieved by applying different temperatures to high- and low-energy samples based on the energy score of the images. On a Top-1 basis, our method achieved a performance improvement of up to 0.6% over ReviewKD and up to 0.93% over DKD. For detailed hyperparameter settings, please refer to Appendix A.

4.3. Temperature ablations

Earlier, we applied different temperatures to low-energy and high-energy samples. To assess the feasibility of employing distinct temperature scaling for both energy samples, we conducted temperature ablation experiments on each sample, as presented in Table 4. Adjusting the temperature for both energy types yielded superior results compared to modifying the temperature for only one energy type, whether low-energy or high-energy.

Table 3
Top-1 and Top-5 accuracy (%) on the ImageNet validation. In the row above, ResNet50 is the teacher and MobileNetV1 is the student. In the row below, ResNet34 is the teacher and ResNet18 is the student. The best results are highlighted in **bold** and the second best underlined.

Distillation			Features				Logits		
R50-MV1	Teacher	Student	AT	OFD	CRD	ReviewKD	KD	DKD	Energy DKD
Top-1	76.16	68.87	69.56	71.25	71.37	72.56	68.58	72.05	**72.98**
Top-5	92.86	88.76	89.33	90.34	90.41	91.00	88.98	91.05	**91.31**
R34-R18	Teacher	Student	AT	OFD	CRD	ReviewKD	KD	DKD	Energy DKD
Top-1	73.31	69.75	70.69	70.81	71.17	71.61	70.66	71.70	**72.21**
Top-5	91.42	89.07	90.01	89.98	90.13	90.51	89.88	90.41	**90.81**

Table 4
Left: Performance evaluated based on the sample type. 'Low-' indicates the application of temperature scaling only to low energy samples (i.e., high T), while 'High-' signifies the utilization of temperature scaling solely for high energy samples (i.e., low T)., Right: Comparing the effectiveness of temperature gradation with the temperature utilized in the performance analysis of our approach.

Teacher Student	ResNet32 × 4 ResNet8 × 4	ResNet32 × 4 ShuffleNetV2	Teacher Student	WRN-40–2 WRN-16–2	ResNet32 × 4 ResNet8 × 4	VGG13 MobileNetV2	ResNet32 × 4 ShuffleNetV2
Low-	73.27	75.38	KD	75.06	73.33	67.37	74.45
High-	73.88	75.34	Gradation	75.49	74.32	68.57	75.74
Ours	**74.60**	**75.87**	Ours	**75.45**	**74.60**	**68.65**	**75.87**

Table 5
Sensitivity analysis on the values of the percentage (r).

Teacher Student		WRN-40–2 WRN-16–2	ResNet32 × 4 ResNet8 × 4	ResNet32 × 4 ShuffleNetV2	VGG13 MobileNetV2
KD		74.92	73.33	74.45	67.37
Energy KD	$r = 0.1$	75.24	**74.62**	75.32	**68.85**
	$r = 0.2$	**75.45**	74.34	**75.87**	68.65
	$r = 0.3$	75.29	74.60	75.76	68.75
	$r = 0.4$	75.37	74.16	75.38	68.81
	$r = 0.5$	75.06	74.28	75.65	68.70

We further explored the application of more varied temperatures across the entire dataset. To achieve this, we divided the entire CIFAR-100 dataset into 10 segments (i.e., $\mathbf{x} \rightarrow [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \dots, \mathbf{x}_{10}]$) based on their energy scores and applied different temperatures (i.e., $T_1, T_2, \dots, T_n, \dots, T_{10}$) to each segment. Specifically,

$$\mathbf{x} = \begin{cases} \mathbf{x}_1 \rightarrow T_1 = T_{\min} \\ \quad \mathbf{x}_2 \rightarrow T_2 \\ \quad\quad \vdots \\ \quad \mathbf{x}_n \rightarrow T_n \\ \quad\quad \vdots \\ \mathbf{x}_{10} \rightarrow T_{10} = T_{\max} \end{cases} . \tag{13}$$

where $T_{\min}$ and $T_{\max}$ represent the minimum and maximum temperatures within the temperature range, respectively. We refer to this as *Temperature Gradation* because we sequentially and gradually increase the temperature. (i.e., $T_1 < T_2 < \cdots < T_n < \cdots < T_{10}$)

Table 4 demonstrates that the method with two different temperatures applied to both energy types achieves performance comparable to the *Temperature Gradation*, which employs a broader range of temperature scaling. From these experiments, we can conclude that addressing certain and uncertain images is more meaningful than dealing with images in between. Additional details about each experiment can be found in Appendix A.

4.4. Sensitivity analysis

Table 5 demonstrates that Energy KD consistently outperforms KD across all values of the ratio (r). Additionally, it is noteworthy that most models achieve optimal results at relatively low values of r. This suggests that prioritizing the processing of images with extreme energy values is more important than processing the majority of images.

4.5. Contribution of HE-DA

Table 6 showcases the outstanding performance of the High-Energy-based Data Augmentation (HE-DA) method on the CIFAR-100 dataset. Performance is evaluated by applying HE-DA to vanilla KD (refer to the upper table) and DKD (refer to the lower table), a state-of-the-art logit-based method. Results for augmenting the entire dataset (i.e., 100%) are obtained from previous papers [52,53].

In the case of vanilla KD, we achieve comparable performance to applying data augmentation to the entire dataset (i.e., 100%), despite applying HE-DA to only 20% of the data (i.e., $r = 0.2$) for most models. The optimal performance of our method is reached when HE-DA is applied to 40%–50% of the data, resulting in a performance improvement of up to 2.87 over vanilla KD and up to 0.71 over that of data augmentation on the full dataset. Concerning DKD, our method attains a performance improvement of up to 1.86 over the baseline DKD and a performance improvement of up to 0.68 over that of data augmentation on the full dataset. Notably, when applying basic data augmentation methods (i.e., augmentation on the entire dataset) to DKD, some models perform worse than without augmentation. In contrast, our method consistently achieves performance improvements across all models.

We extended our experiments to a more challenging dataset, Tiny-Imagenet, to evaluate the performance of HE-DA, and the results are presented in Table 7. These results for TinyImagenet closely mirror those obtained for the CIFAR-100 dataset, showcasing excellent performance.

For vanilla KD (refer to the upper table), our method outperforms 100% data augmentation despite applying only 20%, with our best performance demonstrating an improvement of up to 1.57 over vanilla KD and up to 0.75 over 100% data augmentation. Moving to DKD (refer to the lower table), our method achieves a performance improvement of up to 1.33 over basic DKD and up to 1.27 over 100% data augmentation. Our method consistently delivers excellent performance across all models.

Fig. 5 illustrates the performance variations based on sample types (i.e., low-energy, high-energy, and mixed-energy samples) for two different teacher-student pairs (i.e., VGG13-MobileNetV2 and ResNet32x4-ShuffleNetV2). These experiments clearly demonstrate that exclusively utilizing high-energy samples (marked as the gray line in the graph) results in higher performance compared to using low-energy samples (marked as the blue line) or a mix of samples, including half low-energy and half high-energy samples (marked as the orange line), for all augmentation rates ($r = 0.2 \sim 0.5$), which indicates the amount of additionally augmented samples for each type.

Table 6
Performance evaluated when applying High Energy-based Data Augmentation (HE-DA) to the CIFAR-100 test sets. The best results are highlighted in **bold** and the second best underlined. Δ^* denotes the performance difference between the best result among various rates ($r\%$) of our method and the result without augmentation, while Δ^{**} denotes the performance difference between the best result and full data augmentation (100%).

Teacher	WRN-40–2	ResNet56	ResNet32 × 4	VGG13	VGG13	ResNet32 × 4
Acc.	75.61	72.34	79.42	74.64	74.64	79.42
Student	WRN-16–2	ResNet20	ResNet8 × 4	VGG8	MobileNetV2	ShuffleNetV2
Acc.	73.26	69.06	72.50	70.36	64.60	71.82
KD*	74.92	70.66	73.33	72.98	67.37	74.45
w/ CutMix** (100%)	75.34	70.77	74.91	74.16	68.79	76.61
w/ HE-DA (20%)	75.27	70.90	74.84	74.04	68.06	76.57
w/ HE-DA (30%)	75.54	71.15	74.85	74.17	68.62	76.87
w/ HE-DA (40%)	75.72	**71.43**	75.13	74.42	68.69	77.16
w/ HE-DA (50%)	**75.95**	71.26	**75.22**	**74.54**	**69.13**	**77.32**
Δ^*	+1.03	+0.77	+1.89	+1.56	+1.76	+2.87
Δ^{**}	+0.61	+0.66	+0.31	+0.38	+0.34	+0.71
DKD*	76.24	71.97	76.32	74.68	69.71	77.07
w/ CutMix** (100%)	75.72	71.59	76.86	75.14	70.81	78.81
w/ HE-DA ($r\%$)	**76.40**	**72.21**	**77.23**	**75.55**	**71.28**	**78.93**
Δ^*	+0.16	+0.24	+0.91	+0.87	+1.57	+1.86
Δ^{**}	+0.68	+0.62	+0.37	+0.41	+0.47	+0.12

Table 7
Performance evaluated when applying High Energy-based Data Augmentation (HE-DA) to the TinyImageNet test sets. The best results are highlighted in **bold** and the second best underlined. Δ^* denotes the performance difference between the best result among various rates ($r\%$) of our method and the result without augmentation, while Δ^{**} denotes the performance difference between the best result and full data augmentation (100%).

Teacher	WRN-40–2	ResNet56	ResNet32 × 4	VGG13	VGG13	ResNet32 × 4
Acc.	61.28	58.37	64.41	62.59	62.59	64.41
Student	WRN-16–2	ResNet20	ResNet8 × 4	VGG8	MobileNetV2	ShuffleNetV2
Acc.	58.23	52.53	55.41	56.67	58.20	62.07
KD*	58.65	53.58	55.67	61.48	59.28	66.34
w/ CutMix** (100%)	59.06	53.77	56.41	62.17	60.48	67.01
w/ HE-DA (20%)	59.16	54.09	56.70	61.87	60.16	67.27
w/ HE-DA (30%)	59.36	53.59	56.57	62.36	60.63	67.45
w/ HE-DA (40%)	59.54	**54.52**	57.02	62.24	60.59	67.25
w/ HE-DA (50%)	**59.69**	53.99	**57.13**	**62.51**	**60.85**	**67.64**
Δ^*	+1.04	+0.94	+1.46	+1.03	+1.57	+1.30
Δ^{**}	+0.63	+0.75	+0.72	+0.34	+0.37	+0.63
DKD*	59.66	54.39	58.57	63.12	61.70	67.37
w/ CutMix** (100%)	59.92	54.01	59.23	63.12	62.73	67.97
w/ HE-DA ($r\%$)	**60.43**	**55.28**	**59.58**	**63.78**	**63.03**	**68.25**
Δ^*	+0.77	+0.89	+1.01	+0.66	+1.33	+0.88
Δ^{**}	+0.51	+1.27	+0.35	+0.66	+0.30	+0.28

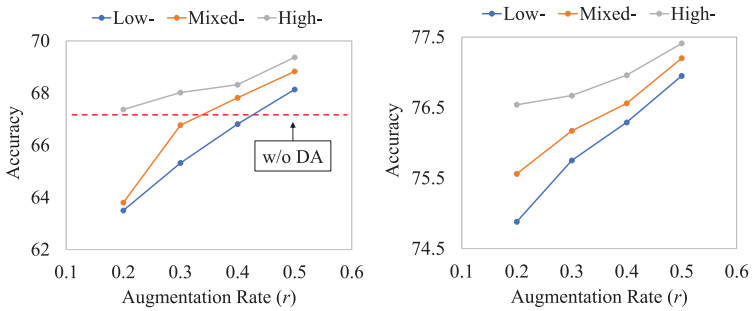


Fig. 5. Performance variations according to the sample types: low, high, and mixed energy. (a): VGG13/MobileNetV2, (b): ResNet32x4/ShuffleNetV2.

It is worth noting that when mixed samples of low-energy and high-energy data are employed (i.e., 50:50), the performance falls between the gray line (i.e., accuracy of high-energy samples) and the blue line (i.e., accuracy of low-energy samples). Using only high-energy data yields superior results, while using only low-energy data leads to lower performance, suggesting that reducing the additional low-energy data by augmentation positively affects performance. The reason behind this could be the learning model's proficiency in understanding low-energy samples. Additional augmentation for low-energy samples, which already includes certain information for correct classification, might lead to confusion, hindering the learning process of the student model. Thus, for optimal performance, it is reasonable to decrease the quantity of augmented data for low-energy samples and rely solely on augmented data for high-energy samples.

Furthermore, it is worth noting that the accuracy of high-energy results remains relatively stable, regardless of the variation in augmentation rate, which is hyperparameters that determine the amount of augmented data. In other words, when dealing with high-energy data, results from augmenting 10% to 20% of the dataset show no significant difference compared to those from augmenting 40% to 50%. However, for low-energy data, augmenting 10% to 20% may result in lower performance compared to even results with no augmentation at all (marked as the red dash line), suggesting that the accuracy from using high-energy data is not significantly affected by changes in the

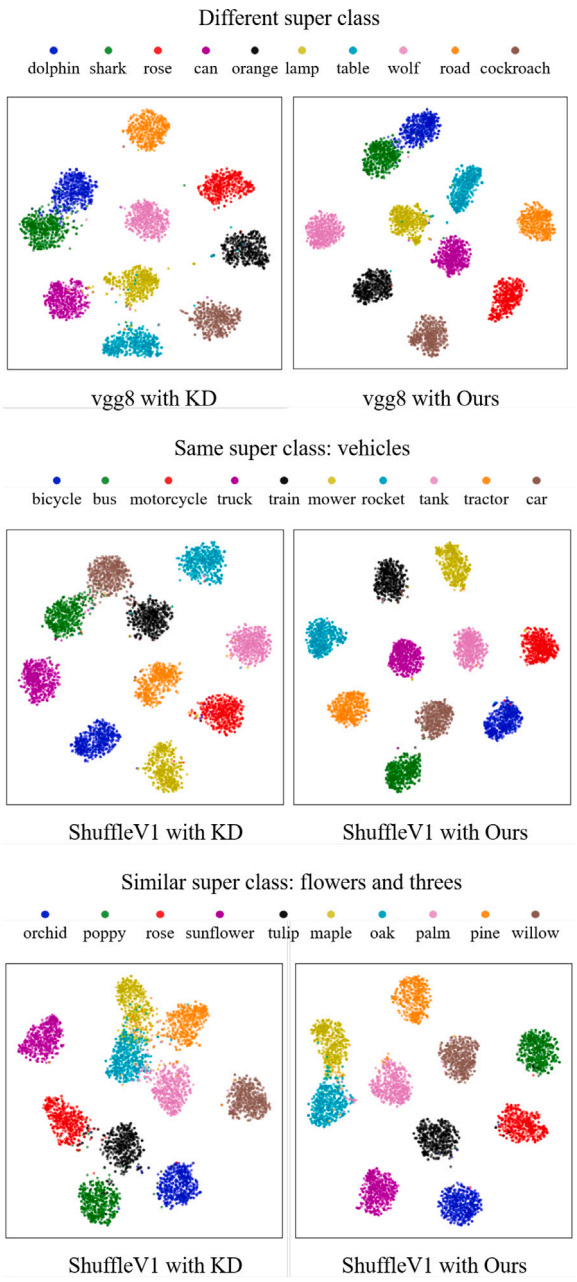


Fig. 6. Feature representations from the penultimate layer of student networks on the some classes of CIFAR-100 dataset. Upper: Different super class, Middle: Same super class (vehicles), Lower: Similar super class (flowers and threes).

augmentation rate. This characteristic can be particularly valuable in scenarios where computational resources are severely limited because using only a small amount of additional high-energy data is enough to achieve better results than previous KD methods.

4.6. tSNE and correlation

Fig. 6 displays tSNE results for various classes of CIFAR-100. The upper figure illustrates clustering for different super classes, showing little similarity between classes. The middle figure showcases clustering for the same super class (vehicles) containing similar classes, and the lower figure displays clustering for similar super classes (flowers and trees). In all figures, we observe that the representations produced by our method are closer for the same class and exhibit less overlap from other classes.

Table 8
Computational costs are measured according to the rate at which data augmentation is applied. The percentage rise is computed based on the value of $r = 0.1$.

r	0.1	0.2	0.3	0.4	0.5	1.0
% ↑	0.0%	3.94%	5.52%	8.78%	14.17%	33.26%

Table 9
Top-1 accuracy (%) on the CIFAR-100-LT datasets, employing both identical and distinct architectures for teacher and student models.

Teacher Student	ResNet32 × 4 ResNet8 × 4	VGG13 VGG8	VGG13 MobileNetV2	ResNet32 × 4 ShuffleNetV2
KD	72.51	71.59	65.19	72.93
DKD	73.60	73.25	66.73	74.09
ReviewKD	73.42	73.02	66.36	74.11
Energy KD	**73.97**	**73.73**	**67.08**	**74.61**

Therefore, our method demonstrates better clustering ability compared to KD, enhancing the discriminability of deep features.

The essence of knowledge distillation lies in how closely the predictions of the student model align with those of the teacher model, given the information provided by the teacher. Fig. 7 visually illustrates this concept by comparing the correlation matrices of the student and teacher logits. Darker colors represent larger differences between the matrices, while lighter colors indicate smaller differences. In other words, the lighter the color, the better the student model mimics the teacher model and produces similar results, demonstrating its capability to yield superior outcomes. In contrast to previous KD, the application of our Energy KD induces the student to generate logits that are more similar to the teacher, thereby ensuring outstanding student performance.

4.7. Computational costs

Table 8 presents the computational cost in relation to the percentage of augmentation applied, specifically referring to the learning time per epoch on the CIFAR-100 datasets. The table shows that applying augmentation to the entire dataset results in a 33.26% increase in computational cost. When applying augmentation to 40%–50% of the dataset (which produces the peak performance of our method), we notice a more modest increase in computational expenses, ranging from 8.78% to 14.17%. These results demonstrate that our approach excels not only in terms of performance but also in terms of efficiency. Details regarding the computing infrastructure used for this experiment are introduced in Appendix A.

4.8. Long-tailed dataset

Table 9 presents the experimental results using the CIFAR-100 long-tail (LT) dataset. CIFAR-100-LT and CIFAR-100 are used for training and testing, respectively. The experiments involved four different architecture pairs. In the case of the long-tail type, exponential decay is applied, and an imbalance factor of 0.5 is used. More detailed experimental parameters are provided in Appendix A. The table illustrates that our method outperforms state-of-the-art DKD and ReviewKD methods. These results highlight the effectiveness of our approach, even when dealing with challenging datasets, as observed in experiments on ImageNet datasets with significant differences in the number of samples among classes.

5. Conclusions

In this paper, we introduce a novel perspective by incorporating the energy score of a sample, a factor traditionally overlooked.

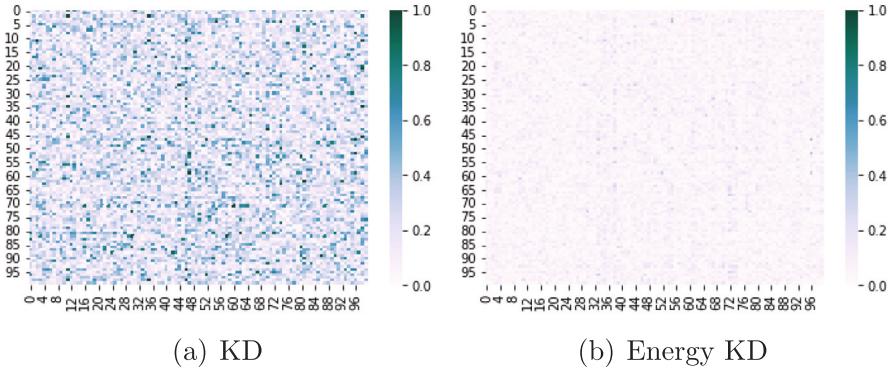


Fig. 7. Correlation disparities between the logits of the student and teacher. Energy KD shows smaller disparities than KD.

Our approach classifies datasets into low-energy and high-energy samples based on their energy scores, applying higher temperatures to low-energy samples and lower temperatures to high-energy samples. In comparison to both logit-based and feature-based methods, our EnergyKD consistently outperforms on various datasets. Notably, on challenging datasets such as CIFAR-100-LT and ImageNet, EnergyKD demonstrates significant performance gains, establishing its effectiveness in real-world scenarios. Furthermore, when coupled with High Energy-based Data Augmentation (HE-DA), it not only enhances performance but also maintains computational efficiency. We anticipate that our framework, offering a new perspective by considering the energy score of samples in both knowledge distillation and data augmentation, will pave the way for prosperous future research in model compression. However, this paper focuses solely on image classification. Extending our approach to various computer vision tasks, such as object detection and semantic segmentation, is part of our future work. For semantic segmentation, which involves pixel-by-pixel classification, devising an energy score function for each pixel will enable us to develop a segmentation-specific method.

CRediT authorship contribution statement

Seonghak Kim: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gyeongdo Ham:** Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Suin Lee:** Visualization, Investigation, Formal analysis, Data curation. **Donggon Jang:** Conceptualization. **Daeshik Kim:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Implementation details

A.1. Hyperparameters

CIFAR-100: A batch size of 64 and 360 epochs were used. The learning rate began at 0.05 and reduced by 0.1 at every 150, 180, 210, 240, 270, and 300 epochs. In addition, Tables A.10 and A.11 outline the hyperparameters utilized for Energy KD and Energy DKD and Table A.12 demonstrates the hyperparameters applied in Energy KD for the CIFAR-100-LT.

ImageNet: A batch size of 256 and 150 epochs were used. The learning rate began at 0.1 and decreased by 0.1 at every 30, 60, 90, and 120 epochs. Moreover, Table A.13 provides details about the hyperparameters applied for Energy DKD.

A.2. Rationale for temperature range

Temperature scaling in KD has been employed to soften the probability distribution from the teacher model. Since its introduction by Hinton [10], this temperature scaling has been considered a hyperparameter that can be adjusted according to the training datasets. Typically, this temperature is set to 4 for CIFAR100 in many KD research studies [11,17].

Therefore, we set the middle temperature to 4 and $T_{(\pm)}$ as half of the middle temperature (i.e., $T_{(\pm)} = \pm 2$) because we have found that adjusting the temperature by only one unit (i.e., $T_{(\pm)} = \pm 1$) is insufficient to effectively modify the probability distribution, particularly in the case of the CIFAR100 dataset. In other words, using $T_{(\pm)} = \pm 1$ weakens the advantage of Energy KD. As shown in Table A.14, this observation is validated through experiments using different values for $T_{(\pm)}$, which demonstrate lower performance when using ± 1 compared to using ± 2.

Moreover, Table A.15 indicates that results with $T_{(\pm)} \pm 3$ become similar to those obtained with $T_{(\pm)} \pm 2$ when controlling the percentage parameter (r), suggesting that comparable performance can be achieved with a greater temperature difference than 2 by adjusting the percentage. Among these options, we choose $T_{(\pm)} \pm 2$ because it shows sufficiently better results compared to previous KD methods.

A.3. Pseudo code

Algorithm 1 provides the pseudo-code of Energy KD in a PyTorch-like [54] style.

A.4. Information for temperature ablations

The left part of Table 4 in main paper indicates that our approach, involving the categorization of samples into low and high energy groups based on their energy values, and subsequently applying distinct temperatures to each group, outperformed strategies where different temperatures were exclusively applied to either the low energy or high energy samples.

Furthermore, the right part of Table 4 in main paper presents a performance evaluation conducted by dividing the entire dataset into multiple partitions instead of just two, accompanied by a wider range of temperatures. The outcomes indicate that the differences between our

Table A.10
The hyperparameters used in the experiments for CIFAR-100 using Energy KD.

Student	WRN-16–2	ResNet20	ResNet8 × 4	VGG8	MobileNetV2	ShuffleNetV2
$T_{(+)}$	2.0	2.0	2.0	2.0	2.0	2.0
$T_{(-)}$	2.0	2.0	2.0	2.0	2.0	2.0
r	0.2	0.3	0.3	0.2	0.2	0.2
Teacher	WRN-40–2	ResNet56	ResNet32 × 4	VGG13	VGG13	ResNet32 × 4

Table A.11
The hyperparameters used in the experiments for CIFAR-100 using Energy DKD.

Student	VGG8	MobileNetV2	ShuffleNetV2
$T_{(+)}$	2.0	1.0	4.0
$T_{(-)}$	2.0	1.0	1.0
r	0.1	0.1	0.3
Teacher	VGG13	VGG13	ResNet32 × 4

Table A.12
The hyperparameters used in the experiments for CIFAR-100-LT using Energy DKD.

Student	ResNet8 × 4	VGG8	MobileNetV2	ShuffleNetV2
$T_{(+)}$	2.0	1.0	1.0	2.0
$T_{(-)}$	2.0	1.0	1.0	2.0
r	0.2	0.2	0.2	0.2
Teacher	ResNet32 × 4	VGG13	VGG13	ResNet32 × 4

Table A.13
The hyperparameters used in the experiments for ImageNet.

Student	ResNet18	MobileNetV1
$T_{(+)}$	1.0	1.0
$T_{(-)}$	1.0	1.0
r	0.2	0.2
Teacher	ResNet34	ResNet50

Table A.14
Sensitivity on the values of temperature difference $T_{(\pm)}$.

Teacher Student		ResNet32 × 4 ResNet8 × 4
KD		73.33
Energy KD	$T_{(\pm)} = \pm1$	74.04
	$T_{(\pm)} = \pm2$	**74.60**
	$T_{(\pm)} = \pm3$	74.44

Table A.15
Sensitivity on the values of the percentage (r) in case of $T_{(\pm)} = \pm3$.

Teacher Student		ResNet32 × 4 ResNet8 × 4
Energy KD ($T_{(\pm)} = \pm3$)	$r = 0.1$	74.30
	$r = 0.2$	**74.44**
	$r = 0.3$	74.28
	$r = 0.4$	74.29
	$r = 0.5$	74.17

approach, applying distinct temperatures solely to the two extremes, and the temperature gradient across the entire dataset are minor. This suggests that our attention should be directed towards high and low-energy samples, making the broader division unnecessary.

In this section, we offer comprehensive details employed to conduct these experiments. We employed the following temperature scaling for all experiments conducted on CIFAR-100.

$$T_{\text{ours}} = \begin{cases} 2, & \text{high energy samples} \\ 6, & \text{low energy samples} \\ 4, & \text{else.} \end{cases} \tag{A.1}$$

Algorithm 1 Energy KD

```
# x: input images
# model_s, model_t: student and teacher model
# T : temperature scaling parameter
# T_ours : our temperature scaling parameter
# E_high : high energy threshold
# E_low : low energy threshold
# E() : energy function

o_s = model_s(x) # logits from student model
o_t = model_t(x) # logits from teacher model

# original KD
# p_s = F.softmax(o_s/T)
# p_t = F.softmax(o_t/T)

E_t= E(o_t) # energy values from teacher's logits

for i, E_i in enumerate(E_t):
   if E_i < E_low: # low energy samples
      T_ours[i] = T[i] + T_(+)
   elif E_i > E_high: # high energy samples
      T_ours[i] = T[i] + T_(-)
   else:
      T_ours[i] = T[i]

p_s = softmax(o_s/T_ours) # predictions from student
p_t = softmax(o_t/T_ours) # predictions from teacher

L_ours(x; T_ours) = nn.KLDivLoss((p_s)||(p_t))
```

In case of applying low energy samples only, the temperature is as follows:

$$T_{\text{ours}} = \begin{cases} 4, & \text{high energy samples} \\ 6, & \text{low energy samples} \\ 4, & \text{else.} \end{cases} \tag{A.2}$$

When applying high energy samples only, the temperature is as follows:

$$T_{\text{ours}} = \begin{cases} 2, & \text{high energy samples} \\ 4, & \text{low energy samples} \\ 4, & \text{else.} \end{cases} \tag{A.3}$$

For temperature gradation, the applied temperature is given by

$$T_{\text{ours}} = \begin{cases} T_{\text{min}} = 2.0 \\ T_2 = 2.5 \\ T_3 = 3.0 \\ T_4 = 3.5 \\ T_5 = 4.0 \\ T_6 = 4.0 \\ T_7 = 4.5 \\ T_8 = 5.0 \\ T_9 = 5.5 \\ T_{\text{max}} = 6.0. \end{cases} \tag{A.4}$$

A.5. Computing source

The experimental server used in the study had the following hardware specifications: (CPU) AMD EPYC 7742; (GPU) NVIDIA RTX 3090;

and (RAM) 1000 GB. The experiments were conducted using CUDA version 11.4 and PyTorch version 1.4.1.

Appendix B. Additional experiments

B.1. Low and high energy samples

Fig. B.8 illustrates images from the ImageNet Dataset divided into low-energy and high-energy groups. Within the red boxes belonging to the low-energy group, the samples clearly exhibit their labels. In contrast, the samples in the green boxes do not accurately represent their labels and have the potential to cause confusion.

Based on these observations, dividing the dataset into two groups based on energy scores and utilizing them for knowledge distillation (KD) and data augmentation (DA) holds significant value.

B.2. High Energy-based MixUp: HE-MixUp

For energy-based data augmentation, we partition the entire dataset into two categories: samples with high energy and those with low energy. Our primary focus lies in utilizing solely the high-energy samples for augmentation, a method known as HE-DA (High Energy-based Data Augmentation). To evaluate the impact of our HE-DA technique, we employed the widely recognized CutMix approach, renowned for its

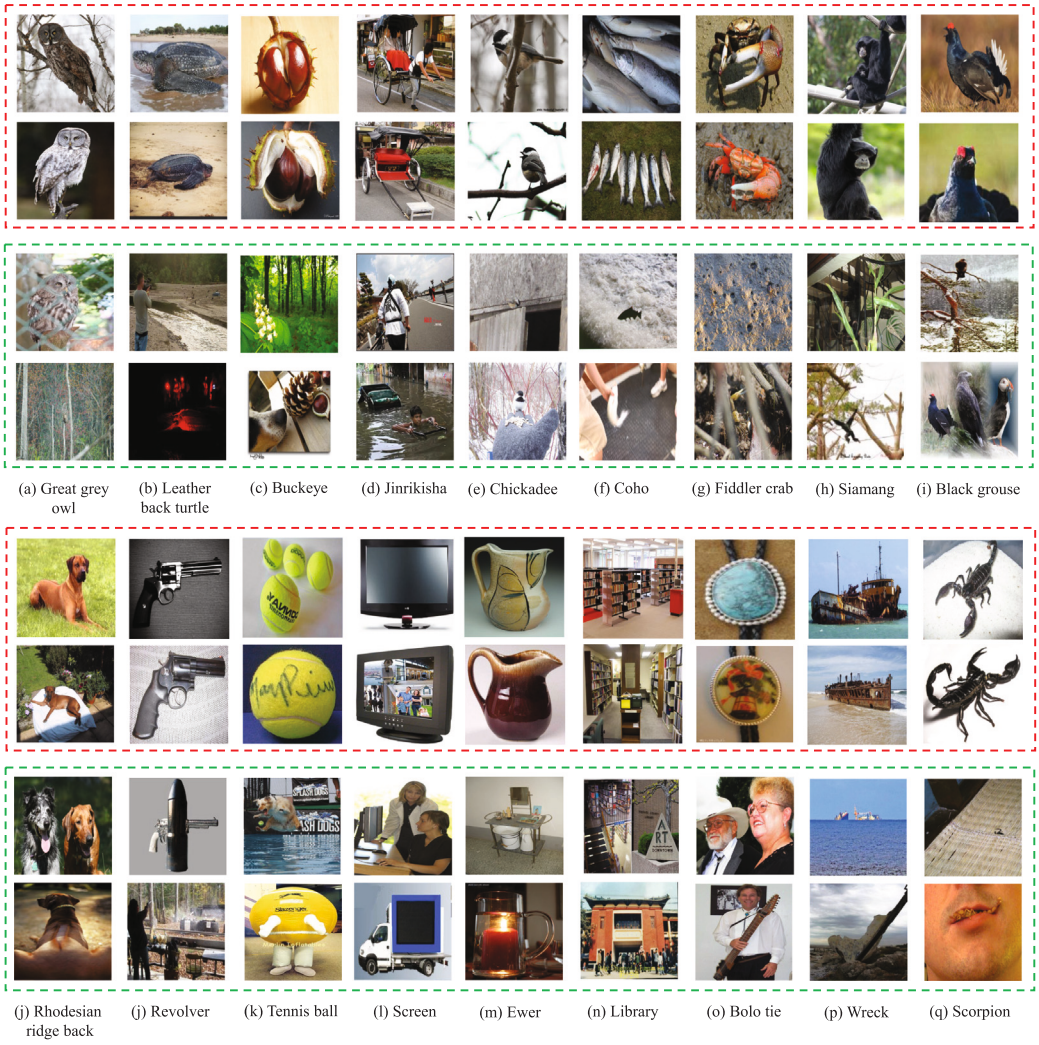


Fig. B.8. ImageNet samples belonging to 18 categories ((a) ~ (q)). Each sample is divided into low-energy (Red boxes) and high-energy groups (Green boxes) based on the energy scores computed from ResNet32x4. We can observe that the low-energy group is clearly distinguished by their labels, whereas the high-energy group lacks these clear distinctions. Our method applies different temperature scaling to each group to convey appropriate knowledge from the teacher to the student model.

Table B.16
Performance evaluated when applying High Energy-based MixUp (HE-MixUp) to the CIFAR-100 test sets, using teacher and student models with the same and different architectures. The best results are highlighted in **bold** and the second best underlined.

Teacher Student	WRN-40–2 WRN-16–2	ResNet56 ResNet20	ResNet32 × 4 ResNet8 × 4	VGG13 VGG8	VGG13 MobileNetV2	ResNet32 × 4 ShuffleNetV2
KD	74.92	70.66	73.33	72.98	67.37	74.45
KD+MixUp (100%)	75.37	70.95	74.46	73.68	68.51	76.13
HE-MixUp (20%)	74.97	70.44	74.42	73.58	67.19	75.76
HE-MixUp (30%)	75.21	71.17	74.60	74.01	68.25	76.11
HE-MixUp (40%)	75.37	**71.16**	74.65	74.02	68.74	**76.24**
HE-MixUp (50%)	**75.65**	71.07	**74.92**	**74.05**	**69.00**	76.40

superior effectiveness compared to MixUp. In this section, we expand our experimentation by applying the MixUp technique to provide additional validation. Our goal is to showcase the adaptability of our approach across various augmentation techniques.

Table B.16 illustrates the performance of our approach when integrated with MixUp. Similar to the results observed with CutMix, the result presents evidence that applying augmentation to only a subset of the data (ranging from $r = 0.2$ to $r = 0.5$) leads to comparable or even greater improvements compared to augmenting the entire dataset.

Appendix C. Nomenclature

Symbols	Descriptions
$\mathbf{x}$	Entire input data
$\mathbf{x}_{\text{low}}$	Data with low energy score
$\mathbf{x}_{\text{high}}$	Data with high energy score
$\mathbf{x}_{\text{else}}$	Data not with low or high energy
$\mathbf{x}^{\text{aug}}$	Data by augmentation
T^{E}	Temperature for energy function
T, T_{ours}	Temperature scaling factor
$T_{(+)}$	Positive integer for T increase
$T_{(-)}$	Negative integer for T decrease
z^f_j	jth class label's logit with network f
$z^{f_\mathcal{T}}$	Logits with teacher network
$z^{f_\mathcal{S}}$	Logits with student network
$\mathcal{E}_\mathcal{T}$	Energy score using teacher network
$\mathcal{E}^{\text{high}}_\mathcal{T}$	Threshold for high energy range
$\mathcal{E}^{\text{low}}_\mathcal{T}$	Threshold for low energy range
d	Dimension
K	Number of classes
N	Number of samples
C	Constant value
$E(\cdot)$	Energy function
$p(\cdot)$	Density function
$G_{\text{aug}}(\cdot)$	Augmentation function
$\sigma(\cdot)$	Softmax function
$[\cdot]$	Array index
f	Neural network
$f_\mathcal{T}$	Teacher network
$f_\mathcal{S}$	Student network
$\theta_\mathcal{T}$	Parameters for teacher
$\theta_\mathcal{S}$	Parameters for student
r	Percentage

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[2] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 116–131.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).

[4] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[6] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[7] J. Liu, B. Zhuang, Z. Zhuang, Y. Guo, J. Huang, J. Zhu, M. Tan, Discrimination-aware network pruning for deep model compression, IEEE Trans. Pattern Anal. Mach. Intell. 44 (8) (2021) 4035–4051.

[8] Y. Zhou, S.-M. Moosavi-Dezfooli, N.-M. Cheung, P. Frossard, Adaptive quantization for deep neural network, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.

[9] J. Gou, B. Yu, S.J. Maybank, D. Tao, Knowledge distillation: A survey, Int. J. Comput. Vis. 129 (2021) 1789–1819.

[10] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.

[11] B. Zhao, Q. Cui, R. Song, Y. Qiu, J. Liang, Decoupled knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11953–11962.

[12] P. Chen, S. Liu, H. Zhao, J. Jia, Distilling knowledge via knowledge review, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5008–5017.

[13] A. Krizhevsky, G. Hinton, et al., Toronto, ON, CanadaToronto, ON, Canada, 2009.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.

[15] Y. Zhang, T. Xiang, T.M. Hospedales, H. Lu, Deep mutual learning, in: CVPR, 2018.

[16] S.I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, H. Ghasemzadeh, Improved knowledge distillation via teacher assistant, in: AAAI, 2020.

[17] Y. Jin, J. Wang, D. Lin, Multi-level logit distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24276–24285.

[18] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, 2014, arXiv preprint arXiv:1412.6550.

[19] N. Passalis, M. Tzelepi, A. Tefas, Probabilistic knowledge transfer for lightweight deep representation learning, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2020) 2030–2039.

[20] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3967–3976.

[21] Y. Tian, D. Krishnan, P. Isola, Contrastive representation distillation, 2019, arXiv preprint arXiv:1910.10699.

[22] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2016, arXiv preprint arXiv:1612.03928.

[23] S. Ahn, S.X. Hu, A. Damianou, N.D. Lawrence, Z. Dai, Variational information distillation for knowledge transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9163–9171.

[24] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, J.Y. Choi, A comprehensive overhaul of feature distillation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1921–1930.

[25] D. Liu, M. Kan, S. Shan, X. CHEN, Function-consistent feature distillation, in: The Eleventh International Conference on Learning Representations, ICLR, 2023, URL https://openreview.net/forum?id=pgHNOcxEdRI.

[26] Z. Guo, H. Yan, H. Li, X. Lin, Class attention transfer based knowledge distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11868–11877.

[27] J. Li, Z. Tang, K. Chen, Z. Cui, Knowledge distillation based on fitting ground-truth distribution of images, Appl. Sci. 14 (8) (2024) 3284.

[28] Z. Guo, P. Zhang, P. Liang, SAKD: Sparse attention knowledge distillation, Image Vis. Comput. 146 (2024) 105020.

[29] D.H. Ackley, G.E. Hinton, T.J. Sejnowski, A learning algorithm for Boltzmann machines, Cogn. Sci. 9 (1) (1985) 147–169.

[30] R. Salakhutdinov, H. Larochelle, Efficient learning of deep Boltzmann machines, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 693–700.

[31] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang, A tutorial on energy-based learning, Predict. Struct. Data 1 (2006).

[32] M. Ranzato, C. Poultney, S. Chopra, Y. Cun, Efficient learning of sparse representations with an energy-based model, Adv. Neural Inf. Process. Syst. 19 (2006).

[33] M. Ranzato, Y.-L. Boureau, S. Chopra, Y. LeCun, A unified energy-based framework for unsupervised learning, in: Artificial Intelligence and Statistics, PMLR, 2007, pp. 371–379.

[34] J. Zhao, M. Mathieu, Y. LeCun, Energy-based generative adversarial network, 2016, arXiv preprint arXiv:1609.03126.

[35] J. Xie, Y. Lu, R. Gao, S.-C. Zhu, Y.N. Wu, Cooperative training of descriptor and generator networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (1) (2018) 27–45.

[36] J. Xie, Z. Zheng, R. Gao, W. Wang, S.-C. Zhu, Y.N. Wu, Learning descriptor networks for 3d shape synthesis and analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8629–8638.

[37] J. Xie, S.-C. Zhu, Y.N. Wu, Learning energy-based spatial-temporal generative convnets for dynamic patterns, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2019) 516–531.

[38] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, Adv. Neural Inf. Process. Syst. 33 (2020) 21464–21475.

[39] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, Adv. Neural Inf. Process. Syst. 33 (2020) 21464–21475.
[40] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, K. Swersky, Your classifier is secretly an energy based model and you should treat it like one, 2019, arXiv preprint arXiv:1912.03263.
[41] D.Y. Park, M.-H. Cha, D. Kim, B. Han, et al., Learning student-friendly teacher networks for knowledge distillation, Adv. Neural Inf. Process. Syst. 34 (2021) 13292–13303.
[42] C. Yuan, R. Pan, Obtain dark knowledge via extended knowledge distillation, in: 2019 International Conference on Artificial Intelligence and Advanced Manufacturing, AIAM, IEEE, 2019, pp. 502–508.
[43] X.-C. Li, W.-S. Fan, S. Song, Y. Li, S. Yunfeng, D.-C. Zhan, et al., Asymmetric temperature scaling makes larger networks teach well again, Adv. Neural Inf. Process. Syst. 35 (2022) 3830–3842.
[44] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6023–6032.
[45] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, 2017, arXiv preprint arXiv:1710.09412.
[46] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016.
[47] S. Zagoruyko, N. Komodakis, Wide residual networks, in: BMVC, 2016.
[48] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: ICLR, 2015.
[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobilenetV2: Inverted residuals and linear bottlenecks, in: CVPR, 2018.
[50] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.
[51] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 116–131.
[52] H. Wang, S. Lohit, M.N. Jones, Y. Fu, What makes a" good" data augmentation in knowledge distillation-a statistical perspective, Adv. Neural Inf. Process. Syst. 35 (2022) 13456–13469.
[53] S. Kim, G. Ham, Y. Cho, D. Kim, Robustness-reinforced knowledge distillation with correlation distance and network pruning, 2023, arXiv preprint arXiv:2311.13934.
[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).