Knowledge Transferred Fine-Tuning: Convolutional Neural Network Is Born Again With Anti-Aliasing Even in Data-Limited Situations

SATOSHI SUZUKI¹, SHOICHIRO TAKEDA², NAOKI MAKISHIMA¹, ATSUSHI ANDO¹,², RYO MASUMURA¹, (Member, IEEE), AND HAYARU SHOUNO³, (Member, IEEE)
¹NTT Computer and Data Science Laboratories, NTT Corporation, Yokosuka 239-0847, Japan
²NTT Human Informatics Laboratories, NTT Corporation, Yokosuka 239-0847, Japan
³Department of Informatics, The University of Electro-Communications, Chofu 182-8585, Japan
Corresponding author: Satoshi Suzuki (satoshi.suzuki.xv@hco.ntt.co.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas under Grant JP19H04982, and in part by the Grant-in-Aid for Scientific Research (A) under Grant JP18H04106.

ABSTRACT Anti-aliased convolutional neural networks (CNNs) are models that introduce blur filters to intermediate representations of CNNs to achieve high accuracy in image recognition tasks. A promising way to prepare a new anti-aliased CNN is to introduce blur filters to the intermediate representations of pre-trained (non anti-aliased) CNNs, since many researchers have released them online. Although this scheme can build the new anti-aliased CNN easily, the blur filters drastically degrade the pre-trained representations. Therefore, to take full advantage of the benefits of introducing blur filters, fine-tuning using massive amounts of training data is often required. This can be problematic because the training data is often limited. In such a “data-limited” situation, the fine-tuning does not bring about a high performance because it induces overfitting to the limited training data. To tackle this problem, we propose “knowledge transferred fine-tuning.” Knowledge transfer is a technique that utilizes the representations of a pre-trained model to help ensure generalization in data-limited situations. Inspired by this concept, we transfer knowledge from intermediate representations in a pre-trained CNN to an anti-aliased CNN while fine-tuning. The key idea of our method is to transfer only the essential knowledge for image recognition in the pre-trained CNN using two types of loss: pixel-level loss and global-level loss. The former loss transfers the detailed knowledge from the pre-trained CNN, but this knowledge may contain “aliased” non-essential knowledge. The latter loss, on the other hand, is designed to increase when the pixel-level loss transfers non-essential knowledge while ignoring the essential knowledge, i.e., it penalizes the pixel-level loss. Experimental results demonstrate that the proposed method using just 25 training images per class on ImageNet 2012 can achieve higher accuracy than a conventional pre-trained CNN.

INDEX TERMS Anti-aliasing, convolutional neural network, data-limited situation, knowledge transfer.

I. INTRODUCTION
Introducing blur filters to image recognition models often plays a crucial role in generalizing because they can relax the difference in the image objects’ scale or position [1], [2]. For example, Neocognitron [1] achieves robustness against image position shifts by introducing blur filters to an ancestor model, Cognitron [3]. Recent studies have demonstrated that modern convolutional neural networks (CNNs) [4]–[6] can also improve their accuracies by introducing blur filters to
intermediate representations.1 For example, Zhang [7] proposed an “anti-aliased CNN” that introduces blur filters to conventional down-sampling operations such as pooling and stride convolution. These down-sampling operations usually ignore the sampling theorem [8], and thus high-frequency signals alias into low-frequencies during sampling. In the anti-aliased CNN, the (low-pass) blur filter suppresses such aliasing effects caused by the down-sampling. As a result, the anti-aliased CNNs can achieve higher accuracy in image recognition tasks than conventional CNNs without blur filters. Building on this success, various studies [9]–[12] have extended the anti-aliased CNN and demonstrated that blur filters work well for many image recognition tasks. To sum up, introducing blur filters is a promising way to improve the accuracy of CNNs.

While the studies above implicitly assumed the training of anti-aliased CNNs from scratch, many researchers, universities, and companies have released pre-trained (non anti-aliased) CNNs online. Therefore, it might be possible to prepare a new anti-aliased CNN model by introducing blur filters to a pre-trained one without training from scratch. Although this approach is promising, the blur filters drastically degrade the pre-trained representations. Thus, to take full advantage of the benefits of introducing blur filters, fine-tuning [13]–[15] is required to build new anti-aliased representations from degraded pre-trained representations with massive training data. This can be difficult because the training data in image recognition is often limited. In such a “data-limited” situation, the fine-tuning cannot bring about a high performance because it induces overfitting to the limited training data.

In this paper, we propose a novel fine-tuning method for anti-aliased CNNs in the data-limited situation,2 called “knowledge transferred fine-tuning,” which combines fine-tuning with knowledge transferring by means of teacher-student collaboration. Knowledge transfer [17] shares the representations and constrains the decision boundary of a student model from that of a teacher model. As a result, it avoids overfitting to the limited training data [17], [18] and effectively builds the representations in the student model [19], [20]. These advantages of the knowledge transfer are a good fit for dealing with the fine-tuning problems of the anti-aliased CNN. Therefore, with the concept of knowledge transfer as a basis, our method fine-tunes an anti-aliased CNN while transferring knowledge from a pre-trained (non anti-aliased) CNN, which has not been overfitted to the limited training data, as a teacher model.

Our method transfers the knowledge from the intermediate representations in the pre-trained CNN for affecting the degraded pre-trained representations directly. Our aim is to transfer only essential knowledge from the pre-trained CNN, as these representations may contain non-essential “aliased” knowledge. To achieve this aim, our method transfers the knowledge using two types of loss: pixel-level loss and global-level loss. The pixel-level loss transfers knowledge by directly calculating the distance between the representations in the teacher and student. This loss is almost the same as the one used in an existing method called FitNets [20], which transfers the detailed knowledge from the teacher model. However, since the detailed knowledge may contain aliased knowledge, the pixel-level loss (i.e., FitNets) risks transferring non-essential knowledge rather than essential knowledge. To mitigate this risk, we introduce a novel loss called “global-level loss.” In the global-level loss, two functions, $f_t$ and $f_s$, extract coarse knowledge that contributes to recognition from the intermediate representations of the teacher and student, respectively. When the pixel-level loss transfers non-essential knowledge while ignoring essential knowledge, the global-level loss increases because $f_t$ has difficulty extracting the knowledge that contributes to recognition from the intermediate representations of the student. In other words, this loss can penalize the pixel-level loss. As a result, the complementary behavior of these losses makes it possible to transfer the essential knowledge of the teacher model maximally. Our method prepares an anti-aliased CNN anew by introducing blur filters to a pre-trained CNN and then fine-tunes it with the recognition loss (the standard loss used in image recognition), the pixel-level loss, and the global-level loss. We evaluated the proposed method on two natural image recognition datasets, ImageNet 2012 [21] and Caltech-256 [22], and found that it is more accurate than the simple fine-tuning using only the recognition loss.

The main contributions of this paper are as follows.

1) We present a knowledge transferred fine-tuning. This method transfers the knowledge from a pre-trained CNN to its anti-aliased version by using the pixel-level and global-level losses while performing fine-tuning.

2) We show that ResNet-18 and 34 trained by our method achieve higher accuracy than the pre-trained CNN in a 25 images/class situation evaluated with ImageNet 2012 [21] dataset. The simple fine-tuning could not achieve this even in a 500 images/class situation.

3) We demonstrate that, interestingly, the global-level loss works particularly well for achieving high accuracy when the datasets in the pre-training and the fine-tuning are different, i.e., in a transfer learning scenario. On the basis of this finding, we introduce a new hypothesis for aliased signals in the pre-trained representations.

II. RELATED WORK

We first introduce anti-aliased CNN and its variants in Subsection II-A. Then, we review existing methods of knowledge transfer in Subsection II-B.

---

1 “Intermediate representations” here refers to the intermediate layer outputs of a CNN.

2 A preliminary version of this paper was presented at the IEEE International Conference on Image Processing (ICIP 2021) [16]. In the current paper, we have included new experimental settings (see Subsection IV-B) and analyzed the effects of our method from viewpoints other than accuracy (Subsection IV-A5). Further, we have compared our method with other well-known knowledge transfer methods and found they are less effective in our problem setting (Subsection IV-A4). These new experiments, analyses, and comparisons may be suggestive for future research directions.
A. CONVOLUTIONAL NEURAL NETWORK AND ANTI-ALIASING

Local connectivity and weight sharing are central ideas of the neural networks used for image recognition. These ideas are usually embodied as convolutional neural networks (CNNs) [1], [23]. In recent studies, CNNs have been scaled up and shown impressive performance [4]–[6].

Most modern CNNs are built using the same principles: alternating convolution and down-sampling operations. Here, down-sampling operations, e.g., pooling and stride convolution, progressively reduce the resolution of intermediate representations and give CNNs efficient computation and a larger receptive field [24]. However, interestingly, they usually ignore the classic sampling theorem [8]. This theorem states that the sampling rate must be at least twice the highest frequency of the signal. If it is not, high-frequency signals alias into low-frequencies during sampling. Therefore, when down-sampling a signal, the textbook solution is to anti-alias by low-pass filtering the signal. Even though this is a well-known fact, down-sampling operations in modern CNNs typically are without anti-aliasing [4]–[6].

Zhang [7] pointed out this issue and proposed an “anti-aliased” CNN that introduces Gaussian blur filters to conventional down-sampling operations. In the anti-aliased CNN, the (low-pass) blur filters suppress the aliasing effects caused by the down-sampling. As a result, it can obtain better intermediate representations than conventional CNNs without blur filters. One major benefit of the anti-aliased CNNs is their higher accuracy, but they are also known to obtain shift-invariant representations [7], [25], which makes them harder to affect by small spatial image shift or translation.

Various studies [9]–[12] have built on the initial success of the anti-aliased CNN through a number of extensions. Li et al. [9] proposed wavelet integrated CNNs (WaveC-Nets) in which down-sampling operations are replaced with a discrete wavelet transform decomposition. WaveC-Nets have a stronger theoretical justification compared to anti-aliased CNNs because they can directly determine high-frequency signals that are removed by wavelet transform decomposition to fit the sampling theorem. Sinha et al. [10] determined the strength of blur filters in a curriculum manner [26] and demonstrated that they work well for many image recognition tasks. Zou et al. [11] proposed content-aware anti-aliasing, in which a low-pass filter is adaptively applied for different spatial locations instead of a single Gaussian blur filter. Moving beyond the usual image recognition, Xie et al. [12] showed that denoising of intermediate representations is effective in the context of adversarial training [27], [28], which aims to improve robustness to adversarial examples [27], [29]. Here, note that denoising has almost the same anti-aliasing effect as blur filters. The above studies have shown that introducing blur filters is a promising way to improve CNN models. However, they have not considered training in a data-limited situation, which is the focus of this paper.

B. KNOWLEDGE TRANSFER

Knowledge transfer was originally devised as a technique for training a lightweight model, i.e., a shallow and/or narrow neural network, by means of teacher-student collaboration [30], [31]. Specifically, it trains a lightweight student model by transferring knowledge from a heavyweight teacher model. In a pioneering work, Hinton et al. [17] proposed a knowledge transfer method in which the output of a student mimics that of a teacher for the same input image. It shares representations and constrains the decision boundary of the student model from that of the teacher model and thereby brings certain advantages to the student one. Both Hinton et al. [17] and Kimura et al. [18] have demonstrated knowledge transfer methods that avoid overfitting to the limited training data. In another study, Furlanello et al. [19] used the same CNN architecture for both the teacher and student models and iteratively trained the CNN from scratch using the knowledge transfer. Interestingly, their trained CNN achieved a higher accuracy than the originally trained one, thus demonstrating that the knowledge transfer can effectively build the representations in the student model. On the basis of these studies, we combine fine-tuning for anti-aliased CNNs in the data-limited situation with the knowledge transfer.

As described in Section I, our method transfers the knowledge from the intermediate representations in the pre-trained CNN. Such knowledge transfer is a somewhat different methodology from the method proposed by Hinton et al. [17], though its benefits are almost the same. In a well-known pioneering work, Romero et al. [20] proposed FitNets, which transfers the knowledge from intermediate representations for effectively building a student model that has a very deep architecture. It transfers the knowledge by calculating the distance between the intermediate representations of the teacher and student and bringing them closer. This knowledge transfer method is almost the same as one of our proposed losses (the pixel-level loss). After FitNets, several variants of the knowledge transfer method from intermediate representations have been proposed [32]–[35]. For example, Zagoruyko and Komodakis [32] proposed transferring the knowledge in the form of spatial attention, where spatial attention maps are computed by summing the intermediate representations along the channel dimension. Huang and Wang [33] transformed the intermediate representations to the form of the Gram matrix [36] and then transferred the knowledge. Heo et al. [34] transferred the knowledge by using sign values of the intermediate representations rather than their exact values. Interestingly, these methods have shown that transforming the intermediate representations and mitigating the constraining effect improves the performance of the knowledge transfer. However, they only performed evaluations with massive training data and did not consider the data-limited situation, which is the main target of our

3We describe detailed formulations of the existing methods [17], [32]–[34] in the Appendix.
work. We argue that these previous methods are insufficient for our presumed use case because, in data-limited situations, the student model needs more knowledge from the teacher for helping with generalization compared to the situation with massive training data.

In contrast to these existing knowledge transfer methods, our proposed method transfers knowledge by means of the pixel-level and global-level losses. The complementary behavior of these losses transfers the essential knowledge of the teacher model maximally.

III. PROPOSED METHOD

Figure 1 shows an overview of the knowledge transferred fine-tuning. We prepare an anti-aliased CNN anew by introducing blur filters to a pre-trained CNN. Therefore, the pre-trained and anti-aliased CNNs have almost the same architecture, but the anti-aliased CNN has blur filters. Our method aims to transfer only the essential knowledge from the “aliased” intermediate representations in the pre-trained CNN (teacher model) to the anti-aliased CNN (student model) using a “transfer unit.” In this paper, we define “essential” to mean without aliasing signals, and “non-essential” to mean with many aliasing signals. In our method, as in the usual knowledge transfer methods [17], [20], [32]–[34], the teacher and student models receive the same image as input, and then we transfer the knowledge with the transfer unit. The configuration of the transfer unit is shown in Figure 2.

For transferring only the essential knowledge, the transfer unit calculates two types of loss: pixel-level loss and global-level loss. The pixel-level loss tries to transfer the detailed knowledge, which leads the student model to inherit the recognition ability of the teacher model [20]. However, this knowledge, which leads the student model to inherit the recognition ability of the teacher model [20]. However, this risks transferring non-essential knowledge from the teacher model, since the detailed knowledge may contain aliased knowledge. To mitigate this risk, the global-level loss penalizes the pixel-level loss if it transfers non-essential knowledge while ignoring essential knowledge.

These two losses are formulated as follows. First, we assume \( t, s \in \mathbb{R}^{C \times H \times W} \) are the tensors. Since the teacher and student models use the convolution operation for calculating the intermediate representations, \( t \) and \( s \) can be regarded as the intermediate representations in the teacher model and student model, respectively. \( C, H, \) and \( W \) represent the number, height, and width of the feature maps. The pixel-level loss calculates the distance between the representations as

\[
L_{\text{pixel}} = d(t, s),
\]

where \( d \) is a function that calculates the distance between two representations (e.g., mean absolute error (MAE)). The global-level loss transfers the coarse knowledge that contributes to image recognition. If the pixel-level loss transfers non-essential knowledge while ignoring essential knowledge, the global-level loss increases. We define the global-level loss as

\[
L_{\text{global}} = d(f_t(t), f_s(s)).
\]

In this formulation, we use the same function \( d \) as Equation (1) for calculating the distance, but other functions could be used as well. \( f_t \) and \( f_s \) are trainable functions that extract and output coarse information contributing to recognition. In our implementation, we construct them as small CNNs with \( 1 \times 1 \) filter size convolution, global average pooling [37], and fully connected layers, as shown in Figure 2. We set the number of input and output channels of the \( 1 \times 1 \) filter size convolution to be the same. The global average pooling summarizes all pixel-level information of the feature maps by averaging, and thus, the representation after this pooling is a tensor consisting of \( 1 \times 1 \) resolution feature maps. Therefore, the global-level loss can transfer coarse (independent of the pixel-level information) knowledge extracted from the summarized representation. If any guidance is not for \( f_t \), it cannot extract recognition information and the global-level loss may be trapped in a trivial solution, e.g., \( f_t(t) = f_s(s) = 0 \). To avoid this, we guide \( f_t \) by “auxiliary” recognition loss \( L_{\text{aux-recog}} \), as shown in Figure 2. \( L_{\text{aux-recog}} \) is a recognition loss calculated with the output of \( f_t \) and an image label. Therefore, the dimension of the \( f_t \) and \( f_s \) outputs becomes the same as the number of classes. The auxiliary recognition loss forces \( f_t \) to maximally extract the coarse recognition information from the representation \( t \). If the pixel-level loss transfers non-essential knowledge while ignoring essential knowledge, \( f_s \) has difficulty extracting the recognition information because the recognition information in \( s \) shrinks compared to that in \( t \). Therefore, \( L_{\text{global}} \) can penalize the pixel-level loss for transferring too much non-essential knowledge. In the experiments, we show that the complementary behavior of these two losses works well for fine-tuning the anti-aliased CNN.

Finally, we fine-tune the anti-aliased CNN with both the pixel-level and global-level losses and the recognition loss \( L_{\text{recog}} \), which is calculated with the output of the student and an image label. Softmax cross-entropy [38] is usually utilized as \( L_{\text{recog}} \) (and \( L_{\text{aux-recog}} \)). For example, the recognition loss \( L_{\text{recog}} \) calculated with softmax cross-entropy is defined as

\[
L_{\text{recog}} = - \sum_k l_k \cdot \log y_k.
\]

Here, \( o = \text{student}(x) \), and \( y_i = \frac{\exp(o_i)}{\sum_j \exp(o_j)} \).

In Equation (4), \( \text{student}(x) \) and \( l \) indicate the output of the student model and the one-hot vector corresponding to the label of input image \( x \), respectively. As shown in Figure 1, the knowledge transferred fine-tuning uses several transfer units, so we average all losses calculated in the transfer units, as

\[
L_{\text{kuf}} = \frac{1}{N} \sum_n (L^n_{\text{pixel}} + L^n_{\text{global}}).
\]
where \( N \) is the number of transfer units and \( L^{\text{pixel}}_n \) and \( L^{\text{global}}_n \) indicate the losses calculated on the \( n \)th transfer unit. The final loss is defined as

\[
L = L^{\text{recog}} + \lambda \cdot L^{\text{ktf}},
\]

where \( \lambda \) is a hyper-parameter that determines the effect of the pixel-level and global-level losses. \( f_t \) and \( f_s \) are trained by \( L^{\text{aux-recog}} \) and \( L^{\text{global}} \), respectively. Their optimizations are performed at the same time as the optimization of the student model using Equation (6). Note that the auxiliary recognition loss has nothing to do with the anti-aliased CNN, since this loss is calculated with the output of \( f_t \) and an image label. As such, it is not included in Equation (6).

The above method is only used in the training phase. During the inference phase, we can remove the teacher model and transfer units, and use only the student model as an anti-aliased CNN. Therefore, our method does not affect inference cost, e.g., inference time and memory usage.

IV. EXPERIMENTS

To evaluate the proposed knowledge transferred fine-tuning, we conduct experiments on image recognition tasks in two scenarios: one in which the pre-training and fine-tuning are done with the same dataset and the other in which they are done with different datasets. In both scenarios, we use the CNNs trained with ImageNet 2012 [21] as the pre-trained ones. Therefore, in the first scenario, we continue to use ImageNet 2012 for fine-tuning, and in the second scenario, we use Caltech-256 [22] for fine-tuning.

A. EXPERIMENTS ON ImageNet

In this subsection, we describe the experiment on the first scenario, i.e., experiments on ImageNet 2012. We evaluated the quantitative effects of our method and the simple fine-tuning, which only uses the recognition loss, in the data-limited situation. We further compared the proposed method with other fine-tuning methods including fine-tuning with frozen parameters, other knowledge transfer methods, and variations of the proposed method.

1) DATASET

We used the ImageNet 2012 [21] dataset for the natural image recognition task. ImageNet 2012 has a total of 1,000 classes.
TABLE 1. Accuracy of proposed and simple fine-tuning methods in data-limited situations. Bold values indicate higher accuracy than the pre-trained CNN (indicated under the cell of the model name).

| Images / class | 10  | 25  | 50  | 100 | 200 | 500 | 750 |
|---------------|-----|-----|-----|-----|-----|-----|-----|
| ResNet-18     |     |     |     |     |     |     |     |
| Simple fine-tuning | 67.13 % | 67.41 % | 67.88 % | 68.35 % | 68.72 % | 69.56 % | 70.08 % |
| Proposed method | 69.71 % | 70.03 % | 70.22 % | 70.30 % | 70.32 % | 70.40 % | 70.44 % |
| ResNet-34     |     |     |     |     |     |     |     |
| Simple fine-tuning | 70.79 % | 70.99 % | 71.37 % | 71.70 % | 71.82 % | 72.16 % | 73.06 % |
| Proposed method | 73.15 % | 73.56 % | 73.64 % | 73.75 % | 73.84 % | 73.88 % | 73.93 % |
| VGG-16        |     |     |     |     |     |     |     |
| Simple fine-tuning | 70.58 % | 70.79 % | 70.83 % | 71.04 % | 71.08 % | 71.44 % | 71.65 % |
| Proposed method | 71.22 % | 71.44 % | 71.52 % | 71.65 % | 71.74 % | 71.84 % | 71.88 % |

3) COMPARISON WITH SIMPLE FINE-TUNING

We first evaluated the quantitative effects of the proposed and simple fine-tuning methods in the data-limited situation. Note that the simple fine-tuning is distinct from ours in that \( \lambda = 0 \) in Equation (6). The hyper-parameter \( \lambda \) was tuned in ResNet-18 with the 50 images/class situation. The relationship between \( \lambda \) and the recognition accuracy is shown in Figure 3. As we can see in the figure, \( \lambda = 70 \) was the optimal. To check the feasibility of our method, we used this parameter across all experiments.

Table 1 lists the recognition accuracies of the two methods. As we can see in the table, the proposed method consistently obtained a higher accuracy than the simple fine-tuning. Further, our method could obtain a higher accuracy than the pre-trained CNN in many cases (indicated in bold). Remarkably, ResNet-18 and 34 achieved higher accuracy than the pre-trained CNN even in the 25 images/class situation. The simple fine-tuning could not achieve this even in the 500 images/class situation. These results demonstrate that our method works well for fine-tuning the anti-aliased CNN in the data-limited situation.

The effect of our method on VGG-16 seemed to be somewhat small. We found that this was partly due to the sub-optimal hyper-parameter \( \lambda = 70 \) for VGG-16, as this \( \lambda \) was tuned in ResNet-18 with the 50 images/class situation. We searched for the optimal \( \lambda \) from among \( \lambda \in \{50, 60, 70, 80, 90, 100\} \) again and found that it was 400 in the 50 images/class situation, where it could achieve 71.61 %, which is a higher accuracy than the pre-trained CNN. Therefore, even in VGG-16, the proposed method may have a similar effect as in ResNet-18 and 34. It is worth pointing out that even with a sub-optimal \( \lambda \), the proposed method showed a gain compared to the simple fine-tuning.

4) OPTIMIZATION DETAILS

We chose to use ResNet-18, ResNet-34 [6], and VGG-16 [5] as pre-trained CNN architectures. Since these models are simple, but have almost all the components used in modern CNNs, they are good testbeds. For the experiments, we used a pre-trained CNN on ImageNet 2012 as a teacher and the anti-aliased CNN that introduces \( 3 \times 3 \) size blur filters to the pre-trained CNN [7] as a student. We placed four transfer units (\( N = 4 \)) in the convolution layers just before each down-sampling layer. For example, we introduced them to the 5th, 9th, 13th, and 17th convolution layers in ResNet-18. In the experiments, we set the same optimization setting as in the pre-training, except for the learning rate and the training epoch. Therefore, the batch size and the momentum parameter were set to 256 and 0.9. The learning rate was set to 0.001 for ResNet and 0.0001 for VGG. These rates are \( 10^{-2} \) times the initial learning rates. We set the learning rate of the transfer units (\( f_k \) and \( f_s \)) to 0.01, since they train from initialized weights. The training epoch was set to 30. The hyper-parameter \( \lambda \) in Equation (6) was chosen from \( \lambda \in \{50, 60, 70, 80, 90, 100\} \). These candidates of \( \lambda \) are based on previous studies [32, 33].

We used the MAE as the function \( d \) in Equations (1) and (2). We also used the softmax cross-entropy as the recognition loss \( L_{\text{recog}} \) and the auxiliary recognition loss \( L_{\text{aux\_recog}} \).

and consists of about 1.28 million training images and 50,000 validation images.

In the experiments, we randomly selected 10, 25, 50, 100, 200, 500, and 750 training images per class from the full training images for preparing data-limited situations. Unfortunately, the test server for ImageNet 2012 could not be browsed at the time of writing, so we report recognition accuracies on the validation images, following previous studies [6, 7].

4 To ensure reproducibility, we used exactly the same images as those in Kayhan et al.’s work [39] for fine-tuning in the 50 images/class situation. The list of these images can be found online at https://github.com/oskyhn/CNNs-Without-Borders.
TABLE 2. Accuracy of proposed and various fine-tuning methods on ResNet-18 in the 50 images/class situation. The best accuracy is highlighted in bold.

| Method                                      | Accuracy   |
|---------------------------------------------|------------|
| Pre-trained CNN (ResNet-18)                 | 69.76 %    |
| Simple fine-tuning                          | 67.88 %    |
| Fine-tuning (1st layer frozen)              | 67.80 %    |
| Fine-tuning (5 shallow layers frozen)       | 67.73 %    |
| Fine-tuning (9 shallow layers frozen)       | 67.42 %    |
| Knowledge transfer from output [17]         | 68.91 %    |
| Attention transfer [32]                     | 69.14 %    |
| Gram matrix [33]                            | 68.07 %    |
| Activation boundary transfer [34]           | 68.03 %    |
| Proposed method (only pixel-level loss)     | 70.07 %    |
| Proposed method (only global-level loss)    | 70.01 %    |
| Proposed method                             | 70.22 %    |

TABLE 3. Consistency values of proposed and other methods on ResNet-18. The proposed method and the simple fine-tuning are fine-tuned with the 50 images/class situation. The best value is highlighted in bold.

| Method                                      | Consistency |
|---------------------------------------------|-------------|
| Pre-trained CNN (ResNet-18)                 | 85.30 ± 0.035 |
| Simple fine-tuning                          | 87.41 ± 0.119 |
| Proposed method                             | **88.81 ± 0.055** |
| Anti-aliased ResNet-18 [7]                  | 87.52 ± 0.085 |

hyper-parameters to maximize accuracy will be the focus of future work.

4) COMPARATIVE ANALYSIS

In this subsection, we present a comparison of our method and existing methods. We examined fine-tuning with frozen parameters of certain shallow layers [14], [15], several knowledge transfer methods other than ours, and variations of the proposed method. The fine-tuning with frozen parameters contributes to avoiding overfitting in the data-limited situation. As the other knowledge transfer methods, we used transferring knowledge from the output distribution [17]–[19], spatial attention maps [32], Gram matrix of the intermediate representations [33], and sign values of the representations [34]. The formulation of these methods and the hyper-parameter settings are provided in the Appendix. One variation of the proposed method uses only pixel-level loss (λ = 60 was the optimal) and the other uses only global-level loss (λ = 70 was the optimal).

Table 2 lists the comparison results on ResNet-18 in the 50 images/class situation. The fine-tuning methods with frozen parameters (second to fifth rows) could not obtain higher accuracy than the simple fine-tuning (67.88 %). The blur filters also degrade the representations of shallow layers, so freezing the parameters makes it difficult to build new anti-aliased representations from the degraded representations. We can see in the table that by increasing the number of freezing parameters, the accuracy of the anti-aliased CNN decreased. Therefore, although freezing the parameters is a promising way to avoid overfitting in the training of conventional CNNs, it is difficult to derive the same benefit in the fine-tuning of anti-aliased CNNs.

Next, we evaluate the other knowledge transfer methods (sixth to ninth rows), some of which are improved versions of the method proposed by Romero it et al. [20] (i.e., pixel-level loss in our method). Interestingly, all of them obtained higher accuracy than the simple fine-tuning but not the pre-trained CNN (69.76 %) or the fine-tuning with pixel-level loss (70.07 %). In other words, while the other knowledge transfer methods could not perfectly transfer the knowledge from the representations in the pre-trained CNN, the pixel-level loss was able to achieve this. This may be because these four methods mitigate the constraining effect of the knowledge transfer compared with the pixel-level loss. As discussed in Subsection II-B, such methods may be insufficient in data-limited situations. For example, the knowledge transferring from output distribution [17] does not directly affect the degraded intermediate representations, and thus mitigates the constraining effect compared with the pixel-level loss. Therefore, it seems inadequate for building the new anti-aliased intermediate representations from the degraded intermediate ones. This suggests the correctness of our method that transfers the knowledge from the intermediate representations. Furthermore, the other three knowledge transfer methods [32]–[34] do not directly transfer the pixel-level knowledge, but rather transform it into spatial attention maps, a Gram matrix, or sign values. The results in Table 2 show that mitigating the constraining effect of the knowledge transfer by these transforming techniques is not effective in the data-limited situation.

We also compared the proposed method with its variations. As we can see in Table 2, all were more accurate than the pre-trained CNN, unlike the other methods. This suggests that the proposed method can effectively build the new anti-aliased representations from the degraded ones while avoiding overfitting. We also found that using only pixel-level loss or only global-level loss achieved good results. This suggests that each loss transfers almost all the essential knowledge from the pre-trained CNN, even though they only focus on detailed or coarse knowledge. However, the best accuracy was observed when our method used both losses. This suggests that each loss by itself is not sufficient to transfer full essential knowledge, and the complementary behavior of the pixel-level loss and the global-level loss works well for achieving high image recognition accuracy.

5) MEASURING SHIFT-INVARIANCE BY CONSISTENCY

In previous studies, the anti-aliased CNNs achieved a high shift-invariant property [7], [25], i.e., they were less affected by small spatial shift or translation compared to conventional ones. A natural question about our method is whether an anti-aliased CNN fine-tuned with it can retain this...
shift-invariant property along with the recognition accuracy. We evaluate this point by calculating the consistency, which measures the shift-invariance of the CNN.

The consistency was used as proposed in Zhang [7]. Intuitively, consistency measures how often a CNN outputs the same recognition result, given two different shifts on the same image. It is formulated as follows:

$$C = \frac{1}{|D|} \sum_{x \in D} G_{h_1, w_1, h_2, w_2}(x),$$

(7)

$$G_{h_1, w_1, h_2, w_2}(x) = \mathbb{1} \{\text{argmax}(o_{h_1, w_1}) = \text{argmax}(o_{h_2, w_2})\},$$

(8)

where $x$ is the image extracted from the whole dataset $D$, and $h_1, w_1$ (height/width) and $h_2, w_2$ parameterize the shifts. $o_{h, w}$ is the output of the student model, i.e., student$(x_{h, w})$. In this formulation, we denote the input image that is shifted by $h, w$ as $x_{h, w}$. In the experiment, we selected the shift parameters $h_1, w_1, h_2, w_2$ at random. $\mathbb{1}$ denotes the indicator function.

Table 3 shows the consistency values of several CNN models based on ResNet-18. We selected $h_1, w_1, h_2, w_2$ at random in this experiment, so we calculated the consistency five times and averaged them. The consistency values of the proposed and simple fine-tuning methods are from the 50 images/class situation. We used the anti-aliased ResNet-18 model provided by Zhang [7] that was trained from scratch using all training data in ImageNet 2012. It achieved a higher recognition accuracy (71.69%) than the anti-aliased CNN fine-tuned with our method (70.22%), but since it uses much more training data than what is available in the data-limited situation, it is not considered a direct competitor.

As we can see in Table 3, our proposed method obtained higher consistency values than the pre-trained CNN. This suggests that our method can retain the shift-invariant property of anti-aliasing. More interestingly, it also obtained a higher consistency than the anti-aliased ResNet-18. Since the simple fine-tuning and anti-aliased ResNet-18 have almost the same consistency values, this shift-invariant property of our method is probably not from fine-tuning. These findings indicate that our proposed method may have an important component for making a shift-invariant CNN model.

### B. Experiments on Caltech-256

In this subsection, we evaluate our method with the second scenario, where we fine-tuned the ImageNet 2012 pre-trained CNN with Caltech-256 [22]. This scenario is often referred to as “transfer learning.”

1) **Dataset**

We used the Caltech-256 [22] dataset for the natural image recognition task. Caltech-256 consists of a total of 30,607 images split into 256 classes.

In the experiments, we randomly selected 5, 10, 15, 20, 30, and 45 training images per class from the full training images for preparing a data-limited situation. We used the remaining images not sampled for training as the test images.

2) **Optimization Details**

The optimization in this experiment is quite similar to that in Subsection IV-A, so here we explain only the optimization details that differ. We used ResNet-18 [6] as a pre-trained CNN architecture. The learning rate was set to 0.001 in all but the final fully connected layer (FC in Figure 1), where it was set to 0.01. This is because the numbers of classes between ImageNet 2012 and Caltech-256 are different, so we needed to initialize the fully connected layer to fit them. The training epoch was set to 60 to ensure sufficient convergence. We used $\lambda = 70$ as the hyper-parameter of the proposed method across all experiments in this subsection. Note that this hyper-parameter was tuned with the 50 images/class situation on ImageNet 2012.

3) **Results**

We evaluated the quantitative effects of the proposed and comparison methods in the data-limited situation. Table 4 lists the experimental results. The first row shows the accuracies of fine-tuned conventional ResNet-18 (i.e., without blur filters), and the remaining rows show the accuracies of the anti-aliased CNN. First, our method achieved the best accuracy in all experimented situations, which demonstrates that it is also effective in scenarios such as transfer learning. It is worth noting that this result of our method was obtained without hyper-parameter tuning in this experimental setting. The tuned hyper-parameter $\lambda = 70$ is the optimal value in the 50 images/class situation on “ImageNet 2012,” not Caltech-256. This implies that our method is robust for tuning hyper-parameters. Further, the hyper-parameter tuning with Caltech-256 may bring more accurate image recognition for our method. The simple fine-tuning did not achieve higher accuracy than the conventional (no blur) model in the 5, 10, and 15 images/class situations. This result is probably due...

### Table 4. Accuracy of proposed and other comparison methods in data-limited situation. The CNN models are pre-trained on ImageNet 2012 and fine-tuned on Caltech-256. The best accuracy is highlighted in bold.

| Images / class | 5   | 10  | 15  | 20  | 30  | 45  |
|---------------|-----|-----|-----|-----|-----|-----|
| ResNet-18 (No blur) | 51.68% | 68.01% | 73.12% | 74.74% | 76.65% | 78.03% |
| Simple fine-tuning | 50.82% | 67.65% | 72.85% | 75.02% | 77.58% | 79.58% |
| Only pixel-level loss | 52.63% | 69.15% | 73.60% | 75.96% | 77.60% | 79.48% |
| Proposed method | **53.44%** | **69.80%** | **74.21%** | **76.18%** | **77.84%** | **79.63%** |
to overfitting to the limited training data, and is consistent with the results in Subsection IV-A. In fact, by increasing the training data, the performance of the simple fine-tuning became close to that of our method and achieved higher accuracy than the conventional model.

We also evaluated the proposed method only using the pixel-level loss and found that it achieved the second-best accuracy following the one using both losses in the data-limited situation. This result is consistent with those in Table 2. However, we found that the gap between the two methods was larger than in Table 2, which implies that the complementary behavior of the pixel-level and global-level losses is more important in transfer learning. One reasonable hypothesis is that the aliased signals in the pre-trained CNN somewhat contribute to recognizing ImageNet 2012 but not Caltech-256. In the pre-training with ImageNet 2012, the CNN is trained only with the objective of maximizing recognition accuracy. Consequently, the CNNs tend to use any available intermediate representations to maximize accuracy, even those that are not interpretable to humans, such as aliased signals. Several previous studies have shown this tendency of CNN [40], [41]. Therefore, the pre-trained CNN transforms the aliased signals through the processing layer-by-layer for contributing to ImageNet 2012 recognition. However, the intermediate representations obtained by this transformation are useless for Caltech-256 since they are tuned only for ImageNet 2012 and not for general image recognition. As a result, the global-level loss, which penalizes the pixel-level loss for transferring too much non-essential knowledge, became more significant. Furthermore, by increasing the training data, the gap between our method with and without the global-level loss became small. This result is reasonable in that, by increasing training data, the anti-aliased CNN becomes able to build the new anti-aliased representations without the complementary behavior of the proposed losses.

4) EVALUATION OF ANTI-ALIASING EFFECT OF OUR METHOD

Finally, we qualitatively evaluate the anti-aliasing effect of our knowledge transferred fine-tuning. Figure 4 shows the input image and the corresponding intermediate representations of three CNNs, which are the models shown in the first (“ResNet-18”), third (“Only pixel-level loss”), and fourth (“Proposed method”) rows in Table 4. All CNNs were fine-tuned in the five images/class situation. We show some feature maps from the intermediate representation of the 9th convolution layer. Each row shows the feature map of the same channel. For example, the feature maps shown at the top are on the 44th channel of the intermediate representation. Graphs under the intermediate representations indicate line profiles of the pixel value.

We can see in Figure 4 that, in conventional ResNet-18, many aliased signals were in the intermediate representations. In contrast, the other two CNNs (Proposed method, Only pixel-level loss) clearly suppressed the aliasing effects, since they have the (low-pass) blur filters. This result is consistent with Zhang’s [7], and such anti-aliasing effects lead the anti-aliased CNN to achieve high recognition accuracy. The difference between the proposed method and the method with only pixel-level loss is rather slight, but our method (i.e., using the pixel-level and global-level losses) shows more anti-aliasing effects than when using only pixel-level loss. For clarity, we show line profiles of the pixel value under the feature maps. These results demonstrate the effectiveness of our method in that the global-level loss penalizes the pixel-level loss if it transfers the non-essential “aliased” knowledge too much.

Overall, these results in Subsections IV-A and IV-B demonstrate that the knowledge transferred fine-tuning is
simple yet effective for fine-tuning anti-aliased CNN in data-limited situations. This effect occurs because our method succeeds in building the new anti-aliased intermediate representations from the degraded ones. The simple fine-tuning, which only uses the recognition loss, induces overfitting to the limited training data and does not outperform our method. We therefore conclude that our method is essential for preparing the anti-aliased CNN in data-limited situations.

V. CONCLUSION

We have developed a knowledge transferred fine-tuning for building anti-aliased convolutional neural networks (CNNs) in data-limited situations. In such situations, the simple fine-tuning cannot achieve high performance because it induces overfitting to the limited training data. In our method, we prepare an anti-aliased CNN anew by introducing blur filters to a readily available pre-trained CNN. On the basis of the idea of knowledge transfer, our method transfers knowledge from the pre-trained CNN that has not been overfitted to the limited training data while fine-tuning the anti-aliased CNN. We aim to transfer only essential knowledge from the pre-trained CNN because the intermediate representations may include non-essential “aliased” knowledge. To achieve this aim, our method transfers the knowledge using two types of loss: pixel-level loss and global-level loss. The pixel-level loss transfers detailed knowledge, while the global-level loss transfers coarse recognition knowledge and penalizes the pixel-level loss if it transfers non-essential knowledge while ignoring essential knowledge.

Our evaluations on the ImageNet 2012 dataset showed that the knowledge transferred fine-tuning achieves higher accuracy than the simple fine-tuning and other comparison methods. The results on ResNet-18 and 34 indicate that our method can achieve higher accuracy than the pre-trained CNN even in the 25 images/class situation. The simple fine-tuning cannot achieve this even in the 500 images/class situation. This suggests that our method may greatly enhance the feasibility of anti-aliased CNNs because it can build a new anti-aliased CNN even in data-limited situations. Furthermore, in the fine-tuning with Caltech-256, the knowledge transferred fine-tuning achieves higher accuracy than the comparison methods, which demonstrates the effectiveness of the knowledge transferred fine-tuning in scenarios such as transfer learning.

In future work, we plan to improve the proposed method so that it can achieve higher accuracy without hyper-parameter tuning. While the hyper-parameter tuned with other models also showed good accuracy, it was found to be suboptimal. We plan to investigate a method that can determine hyper-parameters from the behavior of losses during training. There are various approaches to determining the hyper-parameter [42]–[44] in multi-task learning [45]. However, these approaches aim to determine the balance of losses in multi-task learning, which makes it difficult to change the hyper-parameter significantly. As discussed in Subsection IV-A3, the optimal hyper-parameters are quite different between ResNet and VGG-16. Therefore, we consider that developing a new methodology is needed for determining hyper-parameter for the knowledge transferred fine-tuning. Finally, we also plan to investigate the generalization ability of our method to other application scenarios than those used in this paper.

APPENDIX

FORMULATION OF OTHER KNOWLEDGE TRANSFER METHODS

In this appendix, we formulate the other knowledge transfer methods used in Subsection IV-A4 and describe the tuned hyper-parameters. The following four methods were used in our comparisons:

1) output distribution [17], [18],
2) spatial attention maps [32],
3) Gram matrix of the representations [33], and
4) sign values of the representations [34].

Hereafter, we formulate them in order.

First, in the knowledge transfer from the output distribution [17]–[19], the loss for fine-tuning is formulated as

\[ L = L_{\text{recog}} + \lambda \cdot L_{\text{out}}. \]

\( L_{\text{out}} \) is the softmax cross-entropy loss with a temperature between the outputs of the teacher and student, and is defined as

\[ L_{\text{out}} = -\sum_k y_{k,T} \log \hat{y}_{k,T}. \]

Here, \( y_{i,T} = \frac{\exp(o_{i,T})}{\sum_j \exp(o_{j,T})}, \hat{y}_{i,T} = \frac{\exp(o'_{i,T})}{\sum_j \exp(o'_{j,T})}, \)

where \( o = \text{student}(x) \), and \( o' = \text{teacher}(x) \).

Second, in the knowledge transfer from the spatial attention maps [32], the loss for fine-tuning is formulated as

\[ L = L_{\text{recog}} + \lambda \cdot L_{\text{at}}. \]

Strictly speaking, we need to calculate the loss \( L_{\text{at}} \) on each transfer unit. We omit this for the sake of clarity. \( L_{\text{at}} \) transfers the knowledge from the spatial attention maps as

\[ L_{\text{at}} = d(\text{AT}(t), \text{AT}(s)). \]

Here, \( \text{AT}(A) = \frac{F(A)}{||F(A)||_2}, F(A) = \sum_i |A_i|. \)

\( A_i \in \mathbb{R}^{H \times W} \) indicates the \( i \)th channel of the intermediate representation. Therefore, in Equation (15), \( F(A) \)
means simply summing up the intermediate representation along the channel direction, and AT(A) normalizes it. As the distance function, the mean squared error (MSE) was used in Zagoruyko and Komodakis [32], but for a fair comparison, we used the same function as the proposed method, i.e., the MAE. We tuned the hyper-parameter $\lambda \in \{50, 60, 70, 80, 90, 100\}$, and found $\lambda = 70$ to be optimal.

Third, in the knowledge transfer from the Gram matrix of the intermediate representation [33], the loss for fine-tuning is formulated as

$$L = L_{\text{recog}} + \lambda \cdot L_{\text{gram}}.$$  \hspace{1cm} (16)

We omit the calculation on each transfer unit for the sake of clarity, the same as Equation (13). $L_{\text{gram}}$ transfers the knowledge from the Gram matrix as

$$L_{\text{gram}} = d(\text{Gram}(t), \text{Gram}(s)), \hspace{1cm} (17)$$

where the function Gram($t$) transfers the intermediate representations to the Gram matrix. For example, Gram($t$) is $t_{\text{flat}} \cdot t_{\text{flat}}$. Here, $t_{\text{flat}} \in \mathbb{R}^{C \times (H \times W)}$ is a 2-D tensor that is flattened from its 3-D version, $t \in \mathbb{R}^{C \times H \times W}$. This Gram matrix is known to represent the texture information of the input image [36]. As the distance function, the mean squared error (MSE) was used in Huang and Wang [33], but for a fair comparison, we used the same function as the proposed method, i.e., the MAE. We tuned the hyper-parameter $\lambda \in \{50, 60, 70, 80, 90, 100\}$, and found $\lambda = 60$ to be optimal.

Finally, in the knowledge transfer from the sign values of the representations, the loss is defined as

$$L = L_{\text{recog}} + \lambda \cdot L_{\text{abt}}.$$  \hspace{1cm} (18)

We omit the calculation on each transfer unit for the sake of clarity, the same as Equation (13). $L_{\text{abt}}$ transfers the knowledge from sign values of the representations as

$$L_{\text{abt}} = \frac{1}{M} \sum_{i,j,k} m_{i,j,k}, \hspace{1cm} (19)$$

where

$$m_{i,j,k} = \begin{cases} d(\mu, s_{i,j,k}) & \text{if } t_{i,j,k} > 0 \text{ and } |s_{i,j,k}| < \mu, \\ d(-\mu, s_{i,j,k}) & \text{if } t_{i,j,k} \leq 0 \text{ and } |s_{i,j,k}| < \mu, \\ 0 & \text{otherwise}, \end{cases} \hspace{1cm} (20)$$

where $\mu$ is the margin parameter and $M = C \times H \times W$. $t_{i,j,k}$ and $s_{i,j,k}$ are elements of representations existing in row $j$ and column $k$ on the $i$th channel. Equation (20) can constrain so that the sign values of the representations of the teacher and student models match each other by properly setting $\mu$. In accordance with Heo it et al. [34], we set $\mu = 4.0$. Heo it et al. [34] used the MSE as the distance function, but for a fair comparison, we used the same function as the proposed method, i.e., the MAE. In this method, we used the intermediate representations before the Rectified Linear Unit (ReLU) non-linearity [46], since this method needs a negative representation value in Equation (20). Please note that the representations used in our proposed method occur after ReLU. Finally, we tuned the hyper-parameter $\lambda \in \{50, 60, 70, 80, 90, 100\}$, and found $\lambda = 80$ to be optimal.

**ACKNOWLEDGMENT**

The authors thank Prof. Hideaki Kimata of Kagokai University, Ryuichi Tanida of NTT Human Informatics Laboratories, and Takahiro Kawashima of SOKENDAI for fruitful discussion about the manuscript.

**REFERENCES**

[1] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” Biol. Cybern., vol. 36, no. 4, pp. 193–202, Apr. 1980.

[2] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

[3] K. Fukushima, “Cognitron: A self-organizing multilayered neural network,” Biol. Cybern., vol. 20, pp. 121–136, Sep. 1975.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097–1105.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015, pp. 1–14.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[7] R. Zhang, “Making convolutional networks shift-invariant again,” in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 7324–7334.

[8] H. Nyquist, “Certain topics in telegraph transmission theory,” Trans. Amer. Inst. Elect. Eng., vol. 47, no. 2, pp. 617–644, Apr. 1928.

[9] Q. Li, L. Shen, S. Guo, and Z. Lai, “Wavelet integrated CNNs for noise-robust image classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 501–509.

[10] S. Sinha, A. Garg, and H. Larochelle, “Curriculum by smoothing,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 21653–21664.

[11] X. Zou, F. Xiao, Z. Yu, and Y. J. Lee, “Delving deeper into anti-aliasing in ConvNets,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2020, pp. 1–13.

[12] C. Xie, Y. Wu, L. V. D. Maaten, A. L. Yuille, and K. He, “Feature denoising for improving adversarial robustness,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 501–509.

[13] S. Kornblith, J. Shlens, and Q. V. Le, “Do better ImageNet models transfer better?” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2661–2671.

[14] H. Azizpour, A. S. Razavian, J. Sullivan, A. Makki, and S. Carlsson, “Factors of transferability for a generic ConvNet representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 9, pp. 1790–1802, Sep. 2016.

[15] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, “SpotTune: Transfer learning through adaptive fine-tuning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4805–4814.

[16] S. Suzuki, S. Takeda, R. Tanida, H. Kimata, and H. Shounou, “Knowledge transferred fine-tuning for anti-aliased convolutional neural network in data-limited situation,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2021, pp. 864–868.

[17] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in Proc. NIPS Deep Learn. Represent. Learn. Workshop, 2015, pp. 1–15.

[18] A. Kimura, Z. Graharnani, K. Takeuchi, T. Iwata, and N. Ueda, “Few-shot learning of neural networks from scratch by pseudo example optimization,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2018, p. 1–12.

[19] T. Furlanello, Z. Lipton, M. Tschannen, L. Jet, and A. Anandkumar, “Born again neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 1–10.

[20] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, and Y. Bengio, “FitNets: Hints for thin deep nets,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015, pp. 1–15.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.
[22] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” California Inst. Technol., Pasadena, CA, USA, Tech. Rep. 7694, 2007.

[23] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Neural Comput., vol. 1, no. 4, pp. 541–551, 1989.

[24] F. Yu, V. Koltun, and T. Funkhouser, “Dilated residual networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 472–480.

[25] H. Higuchi, S. Suzuki, and H. Shouno, “Measuring shift-invariance of convolutional neural network with a probability-incorporated metric,” in Proc. Int. Conf. Neural Inf. Process. (ICONIP), 2021, pp. 719–728.

[26] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in Proc. Int. Conf. Mach. Learn. (ICML), 2009, pp. 1–8.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015, pp. 1–11.

[28] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2018, pp. 1–28.

[29] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2014, pp. 1–10.

[30] C. Buscail, R. Caruana, and A. Niculescu-Mizil, “Model compression,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2006, pp. 1–7.

[31] J. Ba and R. Caruana, “Do deep nets really need to be deep?,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2014, pp. 1–9.

[32] S. Zagoruyko and N. Komodakis, “Paying more attention to attentiveness transfer,” 2017, pp. 1–13.

[33] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A selectivity transfer,” 2017, pp. 1–10.

[34] O. Sener and V. Koltun, “Multi-task learning as multi-objective optimization,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2018, pp. 1–12.

[35] Z. Chen, V. Badrinarayanan, C. Lee, and A. Rabinovich, “Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 1–10.

[36] R. Cipolla, Y. Gal, and A. Kendall, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 14274–14285.

[37] A. Byas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 1–12.

[38] H. Wang, X. Wu, Z. Huang, and E. P. Xing, “High-frequency component of activation boundaries formed by hidden neurons,” in Proc. AAAI Conf. Artif. Intell. (AAAI), 2021, pp. 3779–3787.

[39] O. Semih Kayhan and J. C. van Gemert, “On translation invariance in deep learning models resistant to adversarial attacks,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015, pp. 1–28.

[40] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” in Proc. Int. Conf. Mach. Learn. (ICML) Deep Learn. Workshop, 2015, pp. 1–5.

[41] Z. Huang and N. Wang, “Like what you like: Knowledge distill via neuron selectivity transfer,” 2017, arXiv:1707.01219.

[42] B. Heo, M. Lee, S. Yun, and J. Y. Choi, “Knowledge transfer via distillation of activation boundaries formed by hidden neurons,” in Proc. AAAI Conf. Artif. Intell. (AAAI), 2019, pp. 3779–3787.

[43] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” Int. J. Comput. Vis., vol. 129, pp. 1789–1819, Mar. 2021.

[44] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2414–2423.

[45] M. Lin, Q. Chen, and S. Yan, “Network in network,” 2013, arXiv:1312.4400.

[46] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[47] O. Semih Kayhan and J. C. van Gemert, “On translation invariance in CNNs: Convolutional layers can exploit absolute spatial location,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 14274–14285.

[48] A. Byas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 1–12.

[49] H. Wang, X. Wu, Z. Huang, and E. P. Xing, “High-frequency component helps explain the generalization of convolutional neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8684–8694.

[50] Z. Chen, V. Badrinarayanan, C. Lee, and A. Rabinovich, “Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 1–10.

[51] R. Cipolla, Y. Gal, and A. Kendall, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7482–7491.

[52] O. Sener and V. Koltun, “Multi-task learning as multi-objective optimization,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2018, pp. 1–12.

[53] R. Caruana, “Multitask learning,” Mach. Learn., vol. 28, pp. 41–75, Dec. 1997.

[54] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” in Proc. Int. Conf. Mach. Learn. (ICML) Deep Learn. Workshop, 2015, pp. 1–5.
RYO MASUMURA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in engineering from Tohoku University, Sendai, Japan, in 2009, 2011, and 2016, respectively. Since joining Nippon Telegraph and Telephone Corporation (NTT), in 2011, he has been engaged in research on speech recognition, spoken language processing, and natural language processing. He is a member of the Acoustic Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ), the Natural Language Processing (NLP), and the International Speech Communication Association (ISCA). He received the Student Award and the Awaya Kiyoshi Science Promotion Award from the ASJ in 2011 and 2013, respectively, the Sendai Section Student Awards The Best Paper Prize from the Institute of Electrical and Electronics Engineers (IEEE) in 2011, the Yamashita SIG Research Award and the SIG-NL Excellent Paper Award from the IPSJ in 2014 and 2018, the Young Researcher Award and the Paper Award from the Association for the NLP in 2015 and 2020, the ISS Young Researcher’s Award in Speech Field, and the ISS Excellent Paper Award from the Institute of Electronic, Information and Communication Engineers (IEICE) in 2015 and 2018.

HAYARU SHOUNO (Member, IEEE) received the Ph.D. degree in engineering from Osaka University, Osaka, in 1999. He is currently a Professor at the Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo. His research interests include the field of computer vision and machine learning, including neural networks. He is an Action Editor of Neural Networks and an Elected Governor of the Asia Pacific Neural Network Society (APNNS).