CAM Back Again: Large Kernel CNNs from a Weakly Supervised Object Localization Perspective

Shunsuke Yasuki, Masato Taki
Graduate School of Artificial Intelligence and Science, Rikkyo University, Japan
Project Page: https://github.com/snskysk/CAM-Back-Again

Abstract

Recently, convolutional neural networks (CNNs) with large size kernels have attracted much attention in the computer vision field, following the success of the Vision Transformers. Large kernel CNNs have been reported to perform well in downstream vision tasks as well as in classification performance. The reason for the high performance of large kernel CNNs in downstream tasks has been attributed to the large effective receptive field (ERF) produced by large size kernels, but this view has not been fully tested. We therefore revisit the performance of large kernel CNNs in downstream task, focusing on the weakly supervised object localization (WSOL) task. WSOL, a difficult downstream task that is not fully supervised, provides a new angle to explore the capabilities of the large kernel CNNs. Our study compares the modern large kernel CNNs ConvNeXt, RepLKNet, and SLaK to test the validity of the naive expectation that ERF size is important for improving downstream task performance. Our analysis of the factors contributing to high performance provides a different perspective, in which the main factor is feature map improvement. Furthermore, we find that modern CNNs are robust to the CAM problems of local regions of objects being activated, which has long been discussed in WSOL. CAM is the most classic WSOL method, but because of the above-mentioned problems, it is often used as a baseline method for comparison. However, experiments on the CUB-200-2011 dataset show that simply combining a large kernel CNN, CAM, and simple data augmentation methods can achieve performance (90.99% MaxBoxAcc) comparable to the latest WSOL method, which is CNN-based and requires special training or complex post-processing.

1. Introduction

Convolutional Neural Networks (CNNs) dominated the computer vision in the 2010s [20, 21]. In recent years, however, CNN's position has been threatened by the superior performance of Vision Transformers (ViTs) [14, 24], which has inspired research efforts to narrow the performance gap between CNNs and ViTs [13, 23]. The study of large kernel CNNs is one such example. Multi-head self-attention (MHSA), which is considered one of the key mechanism of Transformers, collects information from a large area on the image. It is expected that introducing a larger kernel to CNNs will approximate MHSA's ability to model long-range dependencies. Large kernel CNNs such as RepLKNet [13] have recently attracted a lot of attention because they actually show good performance not only in image classification but also in downstream tasks such as object detection and segmentation. The success of large kernel CNNs is said to be due to their large size kernel and the large effective receptive field (ERF) [27] they produce [13, 23]. A large ERF can be interpreted as a reflection that the output is able to collect information from a larger area on the input, and is considered a sign that the CNN has acquired the ability to model long-range dependencies similar to MHSA [35]. The larger ERFs are also expected to improve performance in downstream vision tasks [6, 7, 48]. The view that the high performance of large kernel CNNs is due to large ERFs is a compelling argument given the interpretation of ERF size.

However, there is insufficient validation to support this view, and the following perspectives have not been fully investigated. First, experimental results on the impact of ERFs on downstream tasks are insufficient; it is not certain whether larger ERFs improve downstream task performance. Second, there is a lack of research on factors other than ERFs. For example, ViTs with large ERFs do not always perform well in downstream tasks due to the breakdown of spatial coupling of images caused by patch embedding and the lack of class identifiability in the attention map [40]. Given these facts, it remains difficult to position the large ERF as a main factor based solely on the results of existing studies, so it is worth further investigating from a new angle what factors are responsible for the performance of large kernel CNNs.

To explore the role of large kernel CNNs from a new perspective, we focus on the weakly supervised object lo-

calization (WSOL) task as a downstream task. WSOL is a task that aims to perform localization using only image-level class labels and is of interest because of its potential to reduce annotation costs. This task is not a fully supervised task; moreover, it requires discriminative localization of object classes, which requires the generation of high-quality feature maps. In addition, CNN-based WSOLs often suffer from partial localization problems [40], making feature maps a challenge to improve. We consider that testing large kernel CNNs with WSOL as the downstream task, which is subject to these difficulties, will allow for investigations that are sensitive to the nature of the feature maps.

Based on these ideas, this paper examines the validity of the existing view that ERF size is important for improving the performance of downstream tasks by comparing latest CNNs. In addition, we analyze other high performance factors behind it. Our main contributions are as follows:

- We present a new perspective by testing the common but little tested conjecture that ERFs are the main factor of the high performance of large kernel CNNs. Specifically, we show that it is difficult to position the size of ERFs as the main factor of performance improvement, at least for WSOL, and that the high performance of large kernel CNNs is due to improved feature maps.
- We find that the feature map improvements described above are effective in avoiding the CAM problem of activating local areas of objects, which has long been discussed in WSOL.
- WSOL experiments on the CUB-200-2011 dataset show that simply combining a latest large kernel CNN with the most classic WSOL method (CAM) and simple data augmentation techniques can achieve performance comparable to state-of-the-art CNN-based WSOL methods. Furthermore, the new simple proposed method based on our findings outperforms the CNN-based state-of-the-art score by 0.43%.

2. Related Work

2.1. Deep Convolutional Neural Networks

Research on large convolutional kernels [22, 29] has left behind problems such as poor classification performance and saturation of benefits due to increased kernel size. In recent years, however, inspired by the success of Vision Transformers, there has been a growing movement to re-examine CNN architectures. ConvNeXt [25] introduced 7x7 convolution kernel that is larger than the previous CNN, influenced by Swin Transformer's 7-12 size window strategy [24]. This allowed the CNN to outperform the Swin Transformer. RepLKNet [13] achieved large kernel models up to

31×31 by training large and small kernels in parallel, and SLaK [23] achieved large kernel models over 51×51 by using a training method with vertical and horizontal rectangular kernels. These can be seen as studies that improve CNNs into models with long-range dependencies, such as Vision Transformers, in terms of kernel size expansion. Hornet [30] is an attempt to give CNNs long-range dependencies through gated convolution and recursive design. We study the essential factors that enable large kernel CNNs, which have been increasingly successful in recent years, to perform well in downstream tasks. The view that ERF is the main factor in these high performances is commonly held, but the validity of this view has not been adequately verified. For example, [12] investigate the relationship between ERFs and downstream tasks (object detection) and argue that larger ERFs do not necessarily improve the performance of downstream tasks and that the benefits of larger ERFs may saturate as the model scales up. While these views are encouraging and support our argument, the comparison aspect of the Spatial Token Mixer (STM) is strong and differs from the direction of our research.

2.2. Weakly Supervised Object Localization (WSOL)

Weakly supervised object localization (WSOL) is a task that aims to localize objects from only image-level labels, and localization is usually performed using a classifier. A representative study of WSOL is CAM [50], which generates heat maps for localization from a model with global average pooling. However, CAMs generated from CNNs trained for classification tend to localize and activate small discriminative portions of objects, and various improvements to this problem have been investigated. Improvements to solve some problems in CAM include data augmentation strategies [10, 32, 38, 42, 45], devising training methods [8, 15, 40, 46, 47], specialized post-processing methods using only trained classifiers [5, 18, 31], and higher resolution heatmaps [4, 33, 41], Classification and localization separation [28, 36, 43], and other approaches. Following this stream of improvements, we re-examine CAM, a classic and problematic method, in the light of modern CNNs.

3. Performance Testing of Large Kernel Models in WSOL Tasks

By focusing on WSOL as a downstream task, this study aims to give a new perspective on the essential reasons for the performance of modern CNNs. To the best of our knowledge, there are no research reports on WSOL tasks using large kernel CNNs, so testing the performance of large kernel CNNs in WSOL tasks is also a novel effort.

Fig. 1 shows the WSOL scores for the traditional CNN (ResNet50 [16]) and the latest CNNs on the CUB-200-2011 dataset. As a WSOL method, we use CAM, which
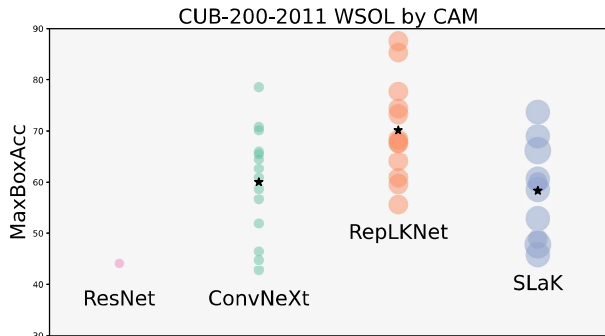
Figure 1. WSOL scores (MaxBoxAcc) for CNNs on the CUB-200-2011 dataset. Marker size indicates kernel size. All models pretrained on ImageNet1K are fine-tuned for classifying the CUB-200-2011 dataset. For ConvNeXt, RepLKNet, and SLaK, scores are measured for various models with different pre-training data, pixel size for fine tuning, and number of model parameters.

is often adopted as a baseline, and the evaluation method is MaxBoxAcc [9]. All of these models, trained 400 epochs, are not WSOL-optimized (classification-optimized) but show significant overall improvements simply by using the latest CNN as the backbone model. See Secs. 8 and 9 for details on fine-tuning and WSOL evaluation.

The RepLKNet model with the highest score in Fig. 1 is the 31B1K384 in [13], fine-tuned using the same data augmentation method as in the pre-training. The fine-tuning was done with 100 epochs instead of the 400 epochs normally used in RepLKNet and recorded a MaxBoxAcc of 90.99. Tab. 1 shows our best score and scores of other CNN-based WSOL methods. Our CAM with a RepLKNet backbone scores significantly better than WSOL-optimized CAMs with a ResNet backbone, beating all pre-2021 methods. Our model also scores competitively with the newer WSOL methods. Because of various problems with CAM, such as the tendency to activate only local regions of objects, various improvements have been attempted, such as those shown in Tab. 1. Those improvements range from data augmentation strategies to more complex training methods and post-processing. However, we have shown that simply combining latest large kernel CNNs with the most classic WSOL methods (CAM) and simple data augmentation methods can achieve performance comparable to state-of-the-art CNN-based WSOL methods. In the following sections, we examine the reasons for the performance improvement from several perspectives.

4. Analysis of Feature Maps and Weights

In this section, we analyze what is responsible for the large improvement in WSOL scores in CAMs with modern CNN backbones, in terms of feature maps and weights. To do so, we first review the problems that CAM faces in the WSOL task. Next, we show our observations on how the state of

Model	Backbone	MaxBoxAccV1	V2
CAM(cvpr,2016) [50]	ResNet	73.2	63
HaS(iccv,2017) [32]	ResNet	78.1	64.7
ACoL(cvpr,2018) [45]	ResNet	72.7	66.4
DDT(cvpr,2021) [37]	VGG16	84.55*	–
F-CAM(+CAM) (wacv,2022) [4]	ResNet	90.3	79.4
F-CAM(+LayerCAM) (wacv,2022)	ResNet	92.4	82.7
F-CAM(+LayerCAM) (wacv,2022)	VGG	91	84.3
CREAM(cvpr,2022) [39]	InceptionV3	90.43*	64.9*
BridgeGAP(cvpr,2022) [19]	VGG	93.17*	80.1*
Ours	RepLKNet	90.99	76.72

Table 1. Summarizes the best score of our CAM with RepLKNet backbone and the scores of other CNN-based WSOL methods for WSOL on the CUB-200-2011 benchmark dataset. * indicates that the threshold selection for bounding box generation has not been calculated using the different dataset (fullsup dataset [9]) than the training data or the final prediction data.

the feature maps and weights of the latest CNN model addresses the existing problems of CAM.

4.1. Preliminary: Problems Faced by CAMs in WSOL Tasks

Class Activation Map (CAM) [50] is a method for localizing object regions in an image that belong to a specific category using CNN with Global Average Pooling (GAP) layer—See Sec. 11 for CAM calculations. It is a typical method in WSOL task and is often used as a baseline method. A well-known problem with CAM methods is their tendency to generate locally activated maps. The last CAM in Fig. 2 looks like a local activation map for a bird's head, which does not allow for correct localization. [3] show how these problems in CAMs occur based on the components of the CAM. Considering their arguments in light of recent WSOL studies, there are two essential problems with CAM. First, because the product of activated regions and weights contributes to the logit due to the GAP layer, CNN training tends to increase the weights corresponding to feature maps that activate local regions important for discrimination. This effect enhances the contribution of local feature maps to the CAM. Second, if feature maps corresponding to negative weights are activated in object areas—especially areas that are not important for classification—, localization is further narrowed down to discriminable areas of the object. Fig. 2 illustrates these problems using actual feature maps.

4.2. Modern CNNs Can Address Existing Problems with CAM

While there are two problems that CAM faces in the WSOL task as shown in Fig. 2, WSOL scores by CAMs using the latest CNN model have improved dramatically (Fig. 1). Furthermore, checking the generated CAMs, we can see that ConvNeXt and RepLKNet, in particular, generate CAMs in which the entire object is activated (Fig. 3). These facts sug-
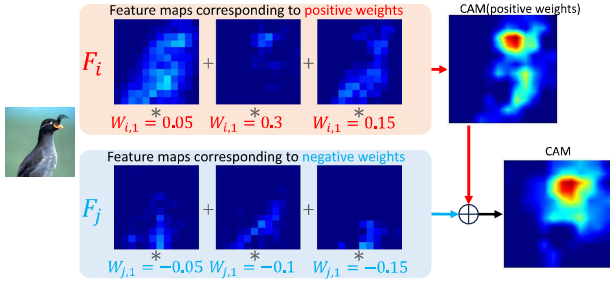
Figure 2. Illustration of the problem that CAMs generated from CNN classifiers tend to locally activate discriminative parts of objects. The figure for F_i represents the problem between the activation area size and the weight size: feature maps with smaller activation regions tend to have larger weights and larger contributions to the CAM. The figure for F_j, on the other hand, represents the local activation region problem. This occurs because feature maps corresponding to negative weights activate non-discriminative regions within the object region. See Sec. 12 for more information.
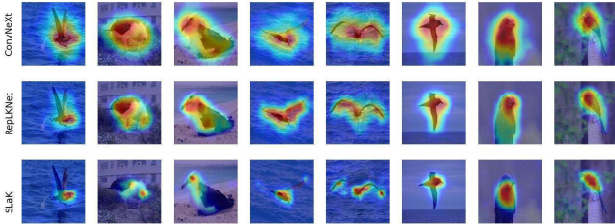

Figure 3. Examples of CAM generation results for the CUB-200-2011 dataset by ConvNeXt, RepLKNet, and SLaK. Among the latest CNNs, ConvNeXt and RepLKNet tend to globally activate the entire object.

gest that modern CNNs can address the existing problems in CAMs.

4.3. Solving the Activation Site Problem for Feature Maps Corresponding to Negative Weights

We show through qualitative and quantitative observations how the latest CNNs address the problem of negative weights shown in F_j in Fig. 2. Fig. 4 show feature maps with large positive and negative weights for three different latest CNNs. In the following, we refer to the feature maps corresponding to the positive and negative weights as F_{pos} and F_{neg}. ConvNeXt and SLaK have many of F_{neg} activated, while RepLKNet has fewer maps activated. F_{neg}, which has the effect of suppressing the activation of the final logit for the correct category, should not be activated in the classification task. Thus, the RepLKNet feature maps are in a desirable state from the classification task perspective. Such a deactivated state is also desirable for WSOL tasks to avoid this localization problem. RepLKNet seems to avoid this problem by its tendency to automatically deactivate F_{neg}.

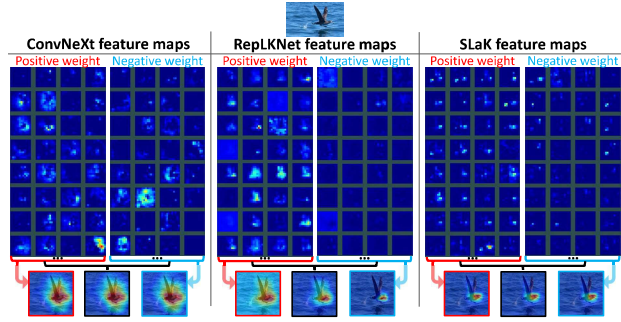On the other hand, F_{neg} can still have a negative im-

Figure 4. Each image group represents F_{pos} and F_{neg} obtained from one input image for each of ConvNeXt, RepLKNet and SLaK. The red-framed heatmap represents the CAM generated from F_{pos} only, the blue-framed heatmap represents the CAM generated from F_{neg} only, and the center heatmap represents the normal CAM.

Feature map weights	ImageNet1K MacBoxAcc		CUB-200-2011 MacBoxAcc	
	positive	negative	positive	negative
ConvNeXt	61	61.2	67.3	66.6
RepLKNet	60.1	7.4	79	30.9
SLaK	55.5	54.5	66.6	65.7

Table 2. Measured WSOL scores for CAMs generated from only F_{pos} (positive) and from only F_{neg} (negative). Experiments on two benchmark datasets, ImageNet-1K and CUB-200-2011.

pact on CAMs in ConvNeXt and SLaK. We therefore check the CAMs generated from F_{pos} and F_{neg} only. These are illustrated as CAMs in the red and blue boxes at the bottom of Fig. 4. In RepLKNet, we can see that there is a big difference between the F_{pos} and F_{neg} cases. On the other hand, for ConvNeXt and SLaK, there is no such difference. This phenomenon is also evident in the WSOL scores (Tab. 2). These results suggest that in ConvNeXt and SLaK, the activation patterns of F_{neg} are strongly aligned with those of F_{pos}. Thus, contrary to what one would expect from Fig. 2, the problems that F_{pos} and F_{neg} have, respectively, no longer occur independently. And if, as in ConvNeXt, the CAM generated from F_{pos} tends to globally activate the entire object, problems on the F_{neg} side are less likely to occur, thus avoiding the CAM problem as a result.

4.4. Resistance to Activation Area Size and Weight Size Problems

In this section, we present an analysis of how modern CNNs avoid the problem for F_i in Fig. 2. First, we check the relationship between the activation areas and weights of the feature maps (Fig. 5). Assume that the larger (smaller) the size of the activation region, the smaller (larger) the weights are, regardless of whether they are positive or negative. In this case, Fig. 5 is expected to draw a mountain-like shape. The experimental results Fig. 5 show that the ConvNeXt and SLaK plots have the expected mountainous shape. ConvNeXt is more peaked than SLaK, so it tends to generate
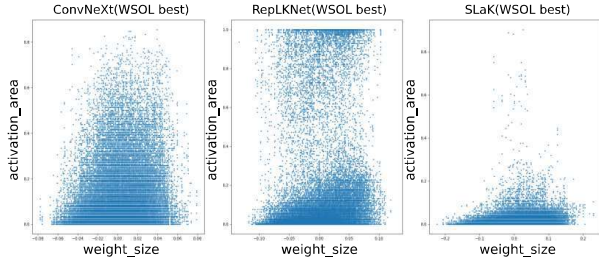
Figure 5. Relationship between activation areas and weights of feature maps. The activation area is calculated by binarizing the feature map with a threshold of 10 and calculating the percentage of pixels that exceed the threshold. Experiments on ConvNeXt, RepLKNet, and SLaK WSOL top-score models fine-tuned on the CUB-200-2011 dataset. See Sec. 13 for results in the non-best-scoring models.
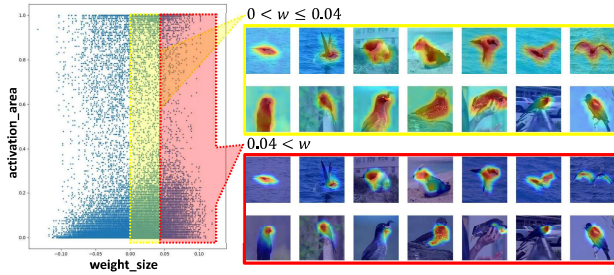

Figure 6. The illustration compares CAMs generated from feature maps corresponding to small weights—$0 < w \leqq 0.04$ (yellow box)—and those generated from feature maps corresponding to large weights—$0.04 < w$ (red box)—for RepLKNet.

more global CAMs by increasing the activation area of the feature maps. On the other hand, RepLKNet shows a pillar-like shape rather than the undesirable mountainous shape.

However, as shown in Fig. 6, even in RepLKNet, feature maps corresponding to small weights tend to activate the CAM globally, whereas feature maps corresponding to large weights tend to activate the CAM locally. Thus, the problem of small activation areas giving large weights is not completely solved in RepLKNet.

Despite these unfavorable conditions, CAMs generated by ConvNeXt and RepLKNet tend to be globally activated. This is likely due to the significantly improved global nature of the feature maps themselves, which reduces the impact from the problems described so far. To confirm this, we perform a principal component analysis of the resulting feature maps. Fig. 7 visualizes the features of the first principal component (hereafter referred to as PC1) obtained by principal component analysis of the feature maps generated from the bird images.

From Fig. 7, we observe that PC1 in ConvNeXt tends to activate global regions that surround the object somewhat extra widely. RepLKNet's PC1, on the other hand, surrounds the object without waste, giving it an ideal shape for

a localization map. SLaK's PC1 tends to activate locally, and there is also a noticeable tendency to activate regions that are unrelated to the object.

Furthermore, from the distribution of contribution rate of PC1 Fig. 8, we can see that PC1 can explain more than half of the feature map in ConvNeXt and RepLKNet. Thus, by producing globally activated feature maps, ConvNeXt and RepLKNet avoid the local activation problem due to the size of activation regions and weights.

4.5. Difference in CAM Quality between RepLKNet and ConvNeXt

The discussion thus far has shown that the long-discussed problem of CAMs tending to be activated locally is solved in RepLKNet and ConvNeXt. However, RepLKNet tends to generate CAMs that are optimal for localization, whereas ConvNeXt's CAMs tend to be activated somewhat more broadly for the target. The reason for this difference can be understood from observations of the feature maps so far in this paper.

First, as can be seen from the feature map plot Fig. 4, RepLKNet's feature maps look cleaner than ConvNeXt's. Checking F_{pos}, we can see that many feature maps activate object regions in RepLKNet, while ConvNeXt's feature maps have more diverse and noisy activation patterns. For quantitative material on this observation, see Sec. 14. A similar trend can also be seen in the distribution of PC1 contribution rates (Fig. 8). ConvNeXt's PC1 contribution rate tends to be much lower than RepLKNet. This is indicative of the diversity of ConvNeXt's feature map, and such noisy activation is likely to blur the CAM contours, leading to overly broad localization.

Similar results were also obtained in an experiment comparing WSOL scores using the binarized PC1 as a localization map. Interestingly, the original WSOL method using PC1 outperformed the state-of-the-art performance of CNN-based WSOL (Tab. 1) by 0.43% (Tab. 3). See Sec. 16 for details.

Model	MaxBoxAcc	MaxBoxAccV2	IoU threshold
ConvNeXt+PC1	83.3	76.5	82
RepLKNet+PC1	**93.6***	80.4	34
SLaK+PC1	30.5	38.2	17

Table 3. CUB-200-2011 WSOL scores based on PC1.

5. Analysis of the Relationship between Kernel Size, ERF, WSOL score, etc.

With respect to why latest CNN models perform well on downstream tasks, it is an interesting study to investigate in detail the previously simple expectation that ERF size is important for performance improvement. By analyzing the relationship between ERF size and downstream task scores and other factors, this chapter examines the validity of the
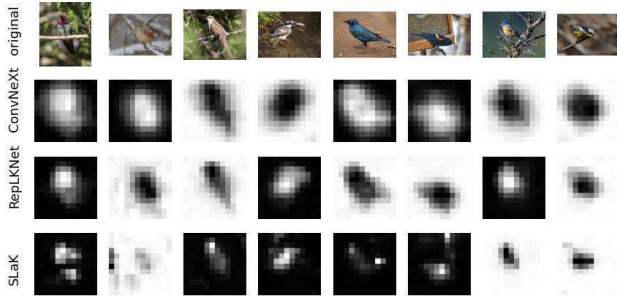
Figure 7. PC1 features obtained by principal component analysis of feature maps are visualized. For a subset of the CUB-200-2011 dataset, Feature maps obtained from the final convolution layer of the model with the highest WSOL score are used in the analysis. Pixel size for all models is 12×12.
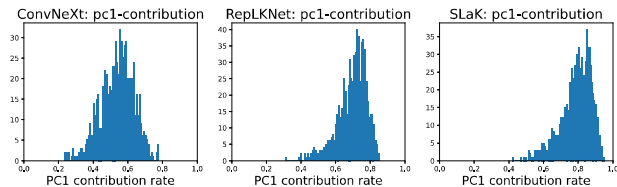

Figure 8. The distribution of PC1 contribution rates obtained by principal component analysis of feature maps are visualized. For 1000 fullsup data [9] of CUB-200-2011, feature maps obtained from the final convolution layer of the model with the highest WSOL score are used in the analysis.
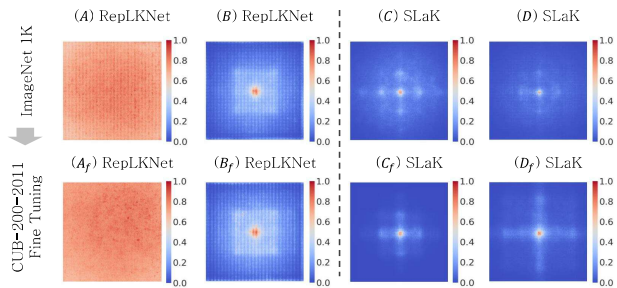

Figure 9. A, A_f, B, B_f, visualize the effective receptive fields of RepLKNet with a kernel size of 31, C, C_f, SlaK with a kernel size of 51, D, D_f, SlaK with a kernel size of 61—A:31B1K224, B:31B22K384, C:SLaKtiny, D:SLaKtiny. 100 CUB-200-2011 images are sampled and resized to 1024×1024 to measure the contribution of pixels on the input image to the center point of the feature map generated in the final convolution layer.

commonly held view that large ERF sizes are important for high performance.

5.1. Does the Larger the Kernel Size, the Larger the ERF?

According to theoretical analysis of ERF [27], ERF grows linearly with kernel size. However, there are many factors that determine the actual size (kernel size, dilated convolu-
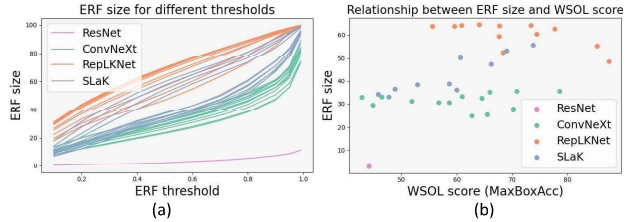

(a)   (b)

Figure 10. (a) shows the ERFs of the various fine-tuning models for the CUB-200-2011 dataset. ERF size is measured at 11 different threshold values of [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99]. (b) shows the relationship between WSOL score (MaxBoxAcc) and ERF size. Here, ERF size is the Area Under the Curve (AUC) of (a).

tion, subsampling, layer depth, skip connections, etc. [27]), and a larger kernel size does not necessarily mean a larger ERF. Thus, when comparing RepLKNet with a kernel size of 31 and SLaK with a kernel size of 51, we cannot necessarily say that the latter ERF is larger. In this regard, [23] shows that the ERF of the latter is larger.

However, as shown in Fig. 9, we found several cases in which the ERF of RepLKNet with a smaller kernel size is larger than that of SLaK. Detailed quantitative verification (Fig. 10 (a)) also confirms that the ERF size of SLaK is smaller than that of RepLKNet in many cases. The ERF size for the SLaK model group, shown in blue in the graph, is smaller than the ERF size for the RepLKNet model group, shown in orange. Furthermore, the ERF sizes of some ConvNeXt models (green) exceed those of some SLaK models with smaller ERF sizes. One reason for the relatively small ERF despite the large kernel size of the SLaK model may be due to its architectural design. Reflecting SLaK's use of decomposition into vertical and horizontal rectangular kernels, Fig. 9 shows that SLaK's ERF shape is more cross-shaped than RepLKNet's ERF.

In summary, the ERFs of the latest CNN models are overall much larger than those of classic CNN models such as ResNet, but we observed that the kernel sizes of these reported models are not necessarily the dominant determinant of ERF size.

5.2. Does a Larger ERF Result in a Higher WSOL Score?: Analysis of Conditions that Result in a Higher WSOL Score

Fig. 10 (b) shows the relationship between ERF size and WSOL score (MaxBoxAcc) for various models. While the overall trend appears to be for larger ERFs to have higher WSOL scores, observations for each group of architectures indicate that the relationship between ERFs and scores is unstable. For example, RepLKNet is the architecture group with the highest WSOL scores, but within this group, models with higher scores tend to have smaller ERF sizes. Also,

some models in ConvNeXt and SLaK have WSOL scores comparable to ResNet, which has a significantly smaller ERF size. These results suggest that the relationship between ERF size and WSOL score is not so simple. See Sec. 9 for MaxBoxAccV2.

Fig. 11 also provides insight into the relationship between WSOL scores and training data rather than ERF size. To summarize what can be seen from the direction of the arrows in each graph in Fig. 11, WSOL scores tend to be lower for pre-trained models and for models pre-trained by larger dataset. The WSOL scores for larger training pixel size settings are higher for the CUB-200-2011 dataset and lower for the ImageNet1K dataset.

Thus, the relationship between training data and WSOL scores was seen in some cases with different trends across datasets and in others with consistent trends regardless of the dataset. Other influencing factors, such as changes in training data and training settings, also make the importance of ERFs in performance improvement uncertain. While the overall ERF size increases for ConvNeXt, SLaK, and RepLKNet in that order, ConvNeXt and RepLKNet CAMs are more likely to be activated globally Fig. 3. Given these observations, it becomes increasingly difficult to justify the relationship between ERF and performance improvement. Thus, the above results do not support the existing expectation that ERF size is important for the performance of downstream tasks in latest CNN models, and it is difficult to position ERF size as the main driver of performance improvement. See Sec. 10 for additional analysis on ERF.

6. Discussion

6.1. Why ConvNeXt and RepLKNet are More Likely to Generate Globally Activated CAMs

CAM tends to be global for ConvNeXt and RepLKNet and local for SLaK, but the reason for this is not clear. Therefore, in this section we discuss the origin of this trend based on experimental results.

First, what makes the interpretation of CAM trends difficult is the fact that CAM activation is local for SLaK, which has a larger kernel size than RepLKNet, whereas it is global for ConvNeXt, which does not have as large a kernel size. For these reasons, it is difficult to expect that the large kernel is the main cause of global CAM. Also, despite the fact that ERF sizes are larger for ConvNeXt, SLaK, and RepLKNet, in that order, ConvNeXt and RepLKNet are the two that achieve large CAM.

We therefore focus on the inherent properties of each CNN architecture. To see the impact of architectural design, we prepare models with randomly initialized weights and compare and analyze the feature maps obtained from them. The first line of Fig. 12 shows the relationship between the GAP values of the feature maps and the weight sizes. This
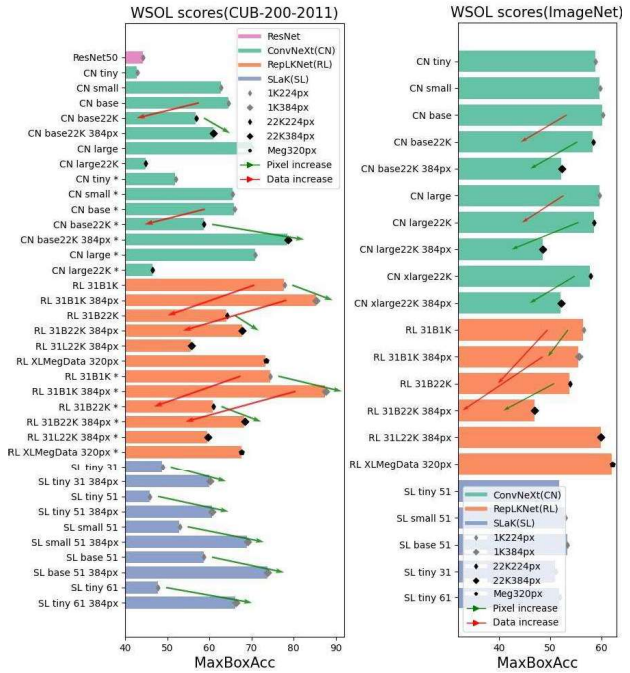

Figure 11. WSOL scores (MaxBoxAcc) for the CUB-200-2011 dataset on the left and the ImageNet1K dataset on the right. Green arrows connect models that differ only in training pixel size settings, and red arrows connect models that differ only in pre-training settings. * denotes models that introduced the same data augmentation strategy as the pre-training in fine tuning.

figure is evaluated against feature maps generated from 100 random bird images (CUB-200-2011 dataset) for each of ConvNeXt, RepLKNet, and SLaK with randomly initialized weights. Each model is selected for achieving the highest WSOL score when fine-tuning on the CUB-200-2011 dataset.

The GAP values reflect the area of the activated region through global pooling. For SLaK, the relationship between GAP values and weights is distributed as a 2D Gaussian centered at the origin. ConvNeXt and RepLKNet, on the other hand, have a widely distributed bias in the direction of the GAP values. The Gaussian shape of SLaK strongly resembles the general trend of CNNs, where the smaller (larger) the activated regions in the feature map, the larger (smaller) the weights tend to be. Thus, for the same reasons explained in Sec. 4.4, SLaK inherently has a tendency to produce locally activated CAMs from the time of initialization. On the other hand, in the plots for ConvNeXt and RepLKNet, large weights do not necessarily imply small GAP values, thus avoiding the general CNN tendency mentioned above. Thus, already at the time of initialization, it is suggested that ConvNeXt and RepLKNet have a strong tendency to generate globally activated feature maps.

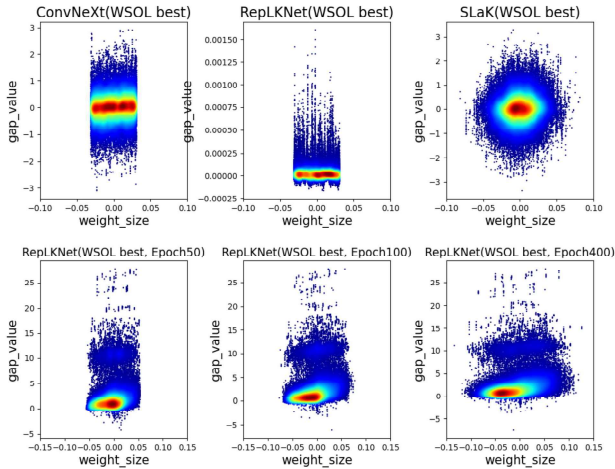Furthermore, compared to ConvNeXt, RepLKNet's dis-

Figure 12. This figure shows the relationship between the GAP values and weights that the feature maps have. The first row shows the results for the three models with the highest WSOL scores in the CUB-200-2011 dataset. The weights are randomly initialized and plotted against 100 feature maps obtained from 100 random images selected from the same dataset. The second row represents the 50, 100, and 400 epoch states in RepLKNet of the first row. See Sec. 13 for results in the non-best-scoring models.

tribution is skewed toward positive GAP values. ConvNeXt's GAP values are centered at 0 and spread evenly in both positive and negative directions. In contrast, RepLKNet's GAP values are most concentrated in positive GAP values and the entire distribution is long in the positive direction. This characteristic may have resulted in feature maps with even larger GAP values and activation area sizes through training. Note that the distribution of RepLKNet changes during the training process, as shown in the second row of Fig. 12—50, 100, and 400 epochs from left to right. From these plots, we can see that the separation of the RepLKNet distribution into two peaks, upper and lower, observed in Fig. 5, occurs during the training process.

6.2. Impact of Data Augmentation Strategies

The latest CNN innovations are not limited to those related to architecture. For example, the training of RepLKNet utilizes training techniques such as AdamW optimizer [26], Stochastic Depth [17], and data augmentation techniques such as RandAugment [11], mixup [44], CutMix [42], Rand Erasing [49]. Of these, CutMix is known to be effective in improving the performance of downstream tasks including WSOL. CutMix forces the classifier to focus on a wider range of cues, including non-discriminative parts of the object, through a cut-and-paste operation on the patches. As a result, it is expected to learn a spatially distributed representation [42]. Therefore, the impact of CutMix on WSOL performance of latest CNNs needs to be checked.

A comparison of the differences in WSOL scores after fine-tuning with and without data augmentation processing, including CutMix, showed that the range of variation in WSOL scores with the additional processing ranged from -5.8% to +4.0%, with no positive or negative bias (Tab. 4). In addition, the pre-trained model RepLKNet XL MegData 320px is the only ImageNet-1K classification model for which no data augmentation was performed during the pre-training process. If this model is used as the pre-training model and no data augmentation is applied in fine-tuning, the WSOL score is 73.4%. This is higher than 7 of the other 11 models that received data augmentation during the pre-training or fine-tuning (or both) process.

Based on these results, we conclude that the architecture itself, more than the data augmentation strategy, contributes to the improved performance of the downstream tasks of the latest CNNs.

Pre-trained RepLKNet model	without DA	with DA	Difference
31B1K 224px	77.68	74.39	-3.30
31B1K 384px	85.30	87.47	+2.17
31B22K 224px	64.07	60.86	-3.21
31B22K 384px	67.73	68.31	+0.59
31L22K 384px	55.59	59.58	+3.99
XLMegData 320px	73.35	67.59	-5.76

Table 4. Results summarizing the differences in WSOL scores for six RepLKNet models—ImageNet1K classification models—with different training setups, with and without data augmentation (DA) strategies when fine-tuning on the CUB-200-2011 dataset.

7. Conclusion

In this paper, we found that when combined with latest large kernel CNNs, simple CAM method in WSOL tasks performs well competitively with state-of-the-art CNN-based WSOL methods. In addition, we revisited the main factors behind the high performance exhibited by latest large kernel CNNs from the perspective of the WSOL task. As a result, we present some material that precludes us from placing the size of the ERF as the main factor behind the high performance. We then provide a new view that high performance in WSOL tasks is due to the inherent capability of the architecture (e.g., structures that facilitate the generation of feature maps with large GAP values or activation regions) and the resulting improvement in feature maps. We also confirmed that the CAM problem of local regions of objects being activated, which has long been discussed in WSOL, is effectively solved by the improved feature map. Finally, we hope that our research, which evaluates both architecture-specific inherent characteristics and tendencies acquired through training from a WSOL perspective, will lead to a deeper understanding of the properties required for better vision models and facilitate systematic architecture development.

References

[1] bethgelab. Toolbox of model-vs-human. https://github.com/bethgelab/model-vs-human. 2

[2] GitHub - Evaluating Weakly Supervised Object Localization Methods Right. https://github.com/clovaai/wsolevaluation/issues/42. 2

[3] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, page 618–634, Berlin, Heidelberg, 2020. Springer-Verlag. 3, 2

[4] Soufiane Belharbi, Aydin Sarraf, Marco Pedersoli, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. F-cam: Full resolution class activation maps via guided parametric upscaling. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3727–3736, 2022. 2, 3

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 1

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*, page 833–851, Berlin, Heidelberg, 2018. Springer-Verlag. 1

[8] Zhiwei Chen, Changan Wang, Yabiao Wang, Guannan Jiang, Yunhang Shen, Ying Tai, Chengjie Wang, Wei Zhang, and Liujuan Cao. Lctr: On awakening the local continuity of transformer for weakly supervised object localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):410–418, 2022. 2

[9] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. to appear. 3, 6, 2, 5

[10] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4256–4271, 2021. 2

[11] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2020. 8

[12] Jifeng Dai, Min Shi, Weiyun Wang, Sitong Wu, Linjie Xing, Wenhai Wang, Xizhou Zhu, Lewei Lu, Jie Zhou, Xiaogang Wang, Yu Qiao, and Xiaowei Hu. Demystify transformers convolutions in modern image deep networks, 2022. 2

[13] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11953–11965, 2022. 1, 2, 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[15] Bai Haotian, Zhang Ruimao, Wang Jiong, and Wan Xiang. Weakly supervised object localization via transformer with implicit spatial calibration. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[17] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *Computer Vision – ECCV 2016*, pages 646–661, Cham, 2016. Springer International Publishing. 8

[18] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 2

[19] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14258–14267, 2022. 3

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 − 90, 2012. 1

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[22] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12316–12325, 2021. 2

[23] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Mykola Pechenizkiy, Decebal Constantin Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview, 2023. 11th International

Conference on Learning Representations, ICLR 2023, ICLR 2023 ; Conference date: 01-05-2023 Through 05-05-2023. 1, 2, 6

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 1, 2

[25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022. 2

[26] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 8

[27] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 4905–4913, Red Hook, NY, USA, 2016. Curran Associates Inc. 1, 6

[28] Shakeeb Murtaza, Soufiane Belharbi, Marco Pedersoli, Aydin Sarraf, and Eric Granger. Discriminative sampling of proposals in self-supervised transformers for weakly supervised object localization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1–11, 2023. 2

[29] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters — improve semantic segmentation by global convolutional network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1743–1751, 2017. 2

[30] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser-Lam Lim, and Jiwen Lu. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[31] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2

[32] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *International Conference on Computer Vision (ICCV)*, 2017. 2, 3

[33] Thanos Tagaris, Maria Sdraka, and Andreas Stafylopatis. High-resolution class activation mapping. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4514–4518, 2019. 2

[34] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision? *ArXiv*, abs/2105.07197, 2021. 2

[35] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao. Internimage: Exploring large-scale vision foundation models with

deformable convolutions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14408–14419, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1

[36] Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S. Kevin Zhou, and Shuguang Cui. Shallow feature matters for weakly supervised object localization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5989–5997, 2021. 2

[37] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 88:113–126, 2019. 3

[38] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6488–6496, 2017. 2

[39] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. Cream: Weakly supervised object localization via class re-activation mapping. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9427–9436, 2022. 3

[40] Yuan Yao, Fang Wan, Wei Gao, Xingjia Pan, Zhiliang Peng, Qi Tian, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13, 2022. 1, 2

[41] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[42] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. 2, 8

[43] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13457–13466, 2020. 2

[44] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. 8

[45] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 2, 3

[46] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *European Conference on Computer Vision*. Springer, 2018. 2

[47] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *European Conference on Computer Vision*. Springer, 2020. 2

[48] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6877–6886, 2021. 1

[49] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07): 13001–13008, 2020. 8

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 2, 3