

# Squeeze-and-Remember Block

Rinor Cakaj

*Signal Processing*

*Robert Bosch GmbH & University of Stuttgart*

71229 Leonberg, Germany

Rinor.Cakaj@de.bosch.com

Jens Mehnert

*Signal Processing*

*Robert Bosch GmbH*

71229 Leonberg, Germany

JensEricMarkus.Mehnert@de.bosch.com

Bin Yang

*ISS*

*University of Stuttgart*

70550 Stuttgart, Germany

bin.yang@iss.uni-stuttgart.de

**Abstract**—Convolutional Neural Networks (CNNs) are important for many machine learning tasks. They are built with different types of layers: convolutional layers that detect features, dropout layers that help to avoid over-reliance on any single neuron, and residual layers that allow the reuse of features. However, CNNs lack a dynamic feature retention mechanism similar to the human brain’s memory, limiting their ability to use learned information in new contexts. To bridge this gap, we introduce the “Squeeze-and-Remember” (SR) block, a novel architectural unit that gives CNNs dynamic memory-like functionalities. The SR block selectively memorizes important features during training, and then adaptively re-applies these features during inference. This improves the network’s ability to make contextually informed predictions. Empirical results on ImageNet and Cityscapes datasets demonstrate the SR block’s efficacy: integration into ResNet50 improved top-1 validation accuracy on ImageNet by 0.52% over dropout2d alone, and its application in DeepLab v3 increased mean Intersection over Union in Cityscapes by 0.20%. These improvements are achieved with minimal computational overhead. This shows the SR block’s potential to enhance the capabilities of CNNs in image processing tasks.

**Index Terms**—Convolutional Neural Network (CNN), Dynamic Memory, Feature Retention, Image Processing

## I. INTRODUCTION

Convolutional Neural Networks (CNNs) have revolutionized the field of machine learning, especially in image processing. Central to their functionality are convolutional layers, which extract a hierarchy of features from images, and fully connected layers (FC), which interpret these features for classification [1], [2]. In addition, attention mechanisms, such as the Squeeze-and-Excitation (SE) block [3], which also inspired the naming of our proposed block, refine the feature processing to further improve model performance.

Despite their advanced capabilities, CNNs still lack one crucial aspect: a memory mechanism similar to that of the human brain. Specifically, the ability to encode, store and retrieve information as seen in the hippocampus and related structures [4]. This gap limits their ability to dynamically use learned information.

To bridge this gap, we introduce the “Squeeze-and-Remember” (SR) block, a novel architectural unit for CNNs. This unit is designed to store high-level features during the training phase and then retrieve these features during inference. Such features may include the texture of a cliff or the architectural elements of a church. During inference, the SR

block adaptively integrates previously undetected or relevant high-level features into feature maps, providing dynamic and context-aware addition to new inputs.

The SR block, as shown in Figure 1, begins its operation with a  $1 \times 1$  convolution on the input feature map  $X$ . This convolution has two purposes: to reduce the depth of the feature map for computational efficiency, and to extract essential information into a compacted representation  $\bar{X}$ .

The compacted feature map  $\bar{X}$  is then processed by a two-layer Fully Connected Network (FCN). The FCN’s task is to convert this compressed feature map into a set of weights corresponding to  $P$  memory blocks in the SR block’s memory  $M$ . These memory blocks  $M_i$  are designed to span a wide range of features, to ensure a comprehensive representation across the feature space. The FCN adaptively weights these blocks based on the input, ensuring that the weighted sum  $\bar{M}$  contains high-level features.

The final output feature map  $\hat{X}$  is obtained by adding  $\bar{M}$  to the original input  $X$ . This fusion results in a final output that is composed of both existing and integrated features.

In our experiments, we found that the optimal placement of the SR block is within the deeper layers of CNNs. These layers, which typically process more abstract and class-relevant features, benefit from the SR block’s ability to augment these features with specific, class-relevant information from its memory.

During training, the key components of the SR block: the  $1 \times 1$  convolution, FCN, and memory blocks are optimized by backpropagation. This tuning process allows the block to identify and encode essential high-level features within the memory blocks. During inference, the SR block uses this training to selectively weight memory blocks based on incoming feature maps, allowing effective use of previously learned features in new contexts. The details and impact of this feature retrieval mechanism are discussed further in Section IV-D.

Unlike Recurrent Neural Networks (RNNs) [5], [6], which are highly effective in sequential data processing through feature reuse, the SR block is specifically designed for CNNs that handle static inputs such as images. It allows these networks to memorize features during training and utilize these during inference.

While weights and biases in layers such as convolutions [1] and batch normalization [7] may look like they store

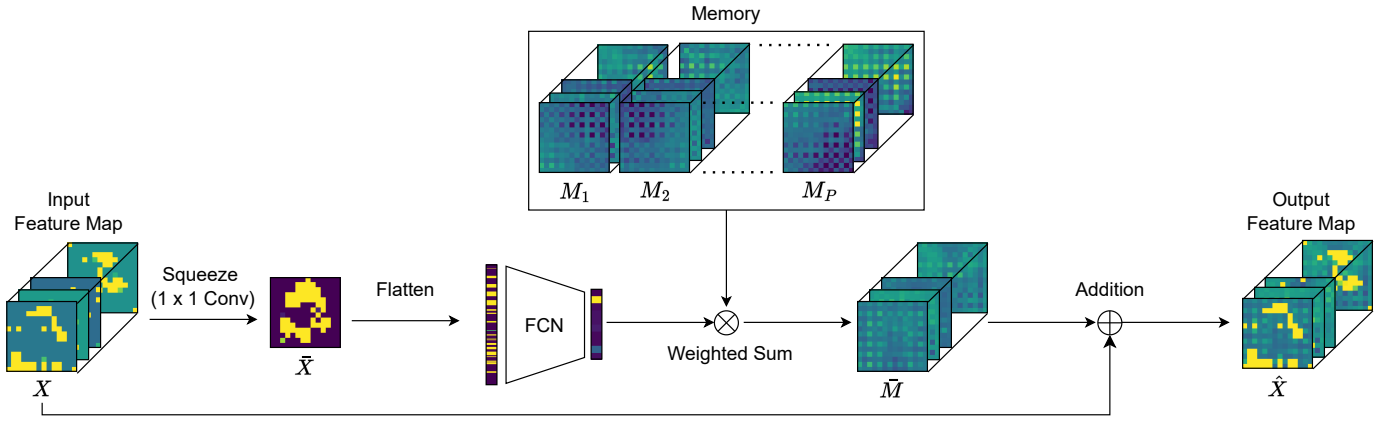


Fig. 1. The Squeeze-and-Remember (SR) Block Architecture, operating in three stages: (1) “Squeeze”: Applies a  $1 \times 1$  convolution to input feature map  $X$ , yielding a reduced feature map  $\tilde{X}$  that retains essential characteristics of  $X$ . (2) “Remember”: In this stage,  $\tilde{X}$  is flattened and passed through a Fully Connected Network (FCN) for weight calculation, followed by the computation of a weighted sum from memory blocks  $M_1$  to  $M_P$ . These memory blocks represent a feature-spanning set, designed to capture a comprehensive range of features. The computed weighted sum,  $\tilde{M}$ , contains high-level, possibly undetected or enhanced features such as cliff textures or architectural details. (3) “Add”: This final step adds  $\tilde{M}$  to  $X$ , producing the output feature map  $\hat{X}$ .

information, they lack adaptability. Each input is processed with the same set of “memories”, unlike the SR block, which selectively applies memorized features based on the context of the input.

Despite its advanced functionality, the SR block is computationally efficient. It easily integrates into existing CNN architectures without significant modifications and adds minimal computational overhead to the model (see Section III-E).

Incorporating the SR block into a ResNet50 model, using dropout2d, enhanced the top-1 validation accuracy on the ImageNet dataset by 0.52% compared to using dropout2d alone. Additionally, for semantic segmentation tasks on the Cityscapes dataset, adding the SR block to the DeepLab v3 model, using with a ResNet50 backbone, resulted in a 0.20% increase in mean Intersection over Union compared to the baseline model. These results highlight the effectiveness of the SR block, while maintaining a low computational footprint (see Section IV).

## II. RELATED WORK

This section provides an overview of the basic building blocks of Deep Neural Networks (DNNs). In particular, we will focus on CNNs, and introduce the concept of memory-augmented networks.

### A. Memory-Augmented Networks

Memory-augmented networks, such as Neural Turing Machines (NTMs) [8] and Differentiable Neural Computers (DNCs) [9] are particularly effective in tasks involving sequential data, such as text and time-series analysis. These architectures give networks the ability to do complex data manipulation tasks, including copying and sorting. They achieve this by integrating neural networks with external memory components [10], [11].

In contrast to these architectures, which focus on sequential tasks, our SR block offers a unique solution for non-sequential

tasks, such as image classification and semantic image segmentation. The SR block diverges from traditional memory-augmented networks by learning to encode features during training, which are then adaptively recalled during inference.

### B. Feature Extraction

The convolutional layer is the cornerstone of CNNs and are used as feature extraction mechanism. In the early layers, these convolutional units detect simple patterns such as edges, corners, and curves [1]. As the network deepens, convolutional layers progressively learn to recognize more complex and abstract features, including object classes and semantic structures [2]. In convolutional layers, the receptive field initially targets local regions and progressively broadens in deeper layers to cover larger image areas.

### C. Feature Evaluation or Classification

Following feature extraction, CNNs use FC layers for feature evaluation or classification. These layers interpret the flattened feature maps from convolutional layers, classifying inputs into distinct categories [2]. The broad receptive field of FC layers plays an important role in evaluating extracted features to a classification output.

### D. Feature Weighting or Attention

Attention mechanisms in CNNs mirror the human cognitive process by prioritizing the most informative parts of an input. Techniques like the SE block [3] and CBAM [12] perform channel-wise and spatial feature recalibration to highlight more important features. The evolution and increasing importance of attention mechanisms [13] highlight their role in enhancing CNN performance.

### E. Feature Reuse

Residual and recurrent connections in DNNs allows feature reuse, giving the network the possibility to use extracted

features in subsequent layers [5], [14]. This is particularly important in CNNs where convolutional layers act as both feature extractors and filters. The introduction of residual connections [14] shows the effectiveness of this strategy in deep network architectures.

#### F. Feature Normalization

Feature normalization is an important aspect of DNN training [15]–[18]. It stabilizes and accelerates the training process. Batch normalization (BN) [7] standardizes feature distributions within mini-batches, reducing internal covariate shifts and thus allowing smoother training dynamics.

#### G. Feature Pooling

Pooling layers reduce the spatial dimensions of feature maps through operations like max pooling [19] and average pooling [20]. This dimensionality reduction not only contributes to translation invariance but also decreases computational load, making the network more efficient.

#### H. Feature Hiding

Regularization techniques such as dropout [21] and drop-connect [22] prevent overfitting by randomly dropping neurons or weights during training. While this is effective in FC layers, the efficacy of dropout in convolutional layers is limited due to the spatial correlation of their activation units. This has led to the development of structured dropout methods like dropblock [23] and dropout2d [24], which target semantic information and encourage the network to learn distributed representations.

In summary, while each of these components has individually advanced CNN capabilities, our SR block represents an innovation by introducing a memory-like mechanism for feature retention and adaptive application in non-sequential tasks like image classification.

### III. SQUEEZE-AND-REMEMBER BLOCK

The *Squeeze-and-Remember* (SR) block introduces a novel approach to improve CNNs. It is designed to memorize high-level features during training and adaptively recall them during inference. The SR block contains three core components:

- 1) **Squeeze:** Compresses the feature map channels to extract essential features.
- 2) **Remember:** Compute the importance of memory blocks based on the compressed feature map.
- 3) **Add:** Add memorized high-level features back into the original feature map.

#### A. Squeeze

The initial step in the SR block compresses the input feature map  $X \in \mathbb{R}^{C \times H \times W}$  to both reduce computational complexity and extract essential features. This compression is done through a  $1 \times 1$  convolution. It is chosen for its efficiency in channel-wise feature compression while minimizing additional computational load.

During training, the  $1 \times 1$  convolution learns to extract key features from the feature map that best represent the

input as a whole. This learning process is facilitated by initializing the weights of this convolution uniformly within the range  $(-\sqrt{k}, \sqrt{k})$ , where  $k = 1/C$ , ensuring a balanced and effective start for the training process.

The compressed feature map  $\bar{X} \in \mathbb{R}^{1 \times H \times W}$ , representing these extracted features, is then flattened for further processing in the “Remember” stage of the SR block.

#### B. Remember

In this step, the flattened feature map  $\bar{X}$  is processed by a two-layer FCN. The first layer linearly transforms the feature map. The subsequent layer, assigns weights to the memory blocks using a softmax activation function. The weights of the FCN are initialized using the same uniform distribution as in the “Squeeze” step.

In the training phase, the FCN in the “Remember” step plays an important role in identifying the optimal combination of memory blocks. It learns to adaptively weight and merge these blocks. This step constructs and integrates high-level relevant features into the network.

#### C. Add

The memory component of the SR block, represented as  $M$ , includes  $P$  memory blocks ( $M_i \in \mathbb{R}^{C \times H \times W}$ ). These are initially zero-filled. The blocks are trained to become a feature-spanning set, which captures a wide range of class-dependent features. The weights generated by the FCN output are used to create a weighted sum of the memory blocks, resulting in the feature map  $\hat{M}$ . This map is then added to the original feature map  $X$ . This process effectively allows the network to “remember” learned features in a dynamical way.

The memory blocks operate as a unique feature-spanning set, with each block  $M_i$  contributing to a composite feature map  $\hat{M}$  based on the input-dependent weights computed by the FCN. This adaptability of the SR block to different inputs is a key innovation that moves beyond static feature extraction.

In summary, the Squeeze-and-Remember block introduces a novel way of manipulating features in CNNs, allowing networks to adaptively reuse learned features and apply them in new contexts.

#### D. Hyperparameters

The design of new CNN architectures involves important decisions about hyperparameters and layer configurations. We focused on three key aspects for the SR block: (1) the optimal integration point in the network architecture, (2) the optimal number of neurons in the hidden layer of the FCN, and (3) the optimal number of memory blocks.

ResNets [14] and VGGs [25] are structured into several modules. Our experiments suggest that the SR block is most effective when placed after the third or fourth module. More generally, this means that the SR block should be inserted in deeper layers. This positioning allows the SR block to use more abstract, class-relevant features, and add class-specific memory content to the feature maps.

For CIFAR-100, we set the FCN’s hidden layer in ResNets to contain 8 neurons, while for ImageNet, we experimented

with 16 and 32 neurons. The number of memory blocks was varied between 2 to 12 for CIFAR and ImageNet. We observed that the performance of the network does not necessarily correlate with a higher number of memory blocks (see Section IV-B).

#### E. Model and Computational Complexity

The SR block introduces additional parameters, mainly in the memory component  $M$ . The total number of parameters added is calculated as:

$$\underbrace{C}_{1 \times 1\text{-Conv}} + \underbrace{H \cdot W \cdot U + U \cdot P}_{\text{FCN}} + \underbrace{P \cdot C \cdot H \cdot W}_{\text{Memory}}, \quad (1)$$

where  $C$  represents the channel count,  $H$  and  $W$  the height and width of the feature map, and  $U$  the number of neurons in the FCN’s hidden layer. For ResNets and VGGs on CIFAR-100, the relative increase in model parameters is typically less than 1%. For ImageNet, due to higher resolution feature maps, the increase ranges from about 1.59% to 7.87%. As detailed in Section IV, the performance improvement of the SR block is due to its feature memory capability, not just the increase in parameters.

The computational overhead of the SR block is low. In comparative tests, a standard ResNet50 and a ResNet50 with the SR block on CIFAR-100 (mini-batch of 128 images) showed minimal difference in processing times (377ms vs. 379ms, respectively). For ImageNet (mini-batch of 5 images), the times were 61.74ms for the standard ResNet50 and 62.89ms for the variant with the SR block, tested on a NVIDIA Quadro RTX 3000 GPU.

This trend holds for CPU inference times as well, which are particularly relevant for embedded device applications. For a  $32 \times 32$  pixel input image, the inference time increased marginally from 41.45ms for the standard ResNet-50 to 43.01ms with the SR block.

### IV. EXPERIMENTS

In this section, we evaluate the effectiveness of our Squeeze-and-Remember block in enhancing CNNs for supervised image recognition and semantic segmentation tasks. Our empirical studies primarily focus on two key areas: image recognition using CIFAR-100 and ImageNet datasets, and semantic segmentation on the Cityscapes dataset. We compare the performance of various CNN architectures using the SR block against their standard configurations. Additionally, we evaluate the synergistic effect of integrating the SR block with regularization techniques like dropout [21], dropout2D [24], and dropblock [23].

For clarity and consistency, our model descriptions follow a specific naming convention. For example, “ResNet50 + D” denotes a ResNet50 model trained with dropout, “D2D” for dropout2D, and “DB” for dropblock. Models that include the SR block are indicated by adding “SR” to their name, such as “ResNet50 + SR” or “ResNet50 + D2D + SR”.

#### A. Experiments on CIFAR

Our evaluation of the SR block was performed on the CIFAR-10/100 datasets, containing 50,000 training and 10,000 test color images of  $32 \times 32$  pixels. We utilized a range of CNN architectures for our experiments, including ResNet18/34/50 [14] and VGG16/19-BN [25]. Detailed results for CIFAR-10 are provided in Appendix VI.

To ensure robustness and reproducibility, each model was trained and evaluated five times with different random seeds. This affects the network initialization, data ordering, and augmentation processes. We present the mean test accuracy and its standard deviation. The data split was 90% for training and 10% for validation. The best performing model on the validation set was selected for the final evaluation.

Baseline networks were trained using data augmentation, weight decay, early stopping, and varying dropout techniques (dropout [21], dropout2d [24], dropblock [23]). We also investigated the performance improvement across ResNet architectures when the Squeeze-and-Remember block was used alongside SE [3] and CBAM blocks [12]. The VGG models (VGG16-BN and VGG19-BN) followed a similar training regimen, with dropout applied only in their fully connected (FC) layers. Comprehensive implementation details are available in Appendix VII-A.

Table I presents our findings on the CIFAR-100 dataset. In this table, **M** indicates the SR block’s integration point, **P** the number of memory blocks, and **OH** the additional parameter overhead. Notably, all ResNet and VGG models incorporating the SR block showed consistent improvements in accuracy. For example, ResNet50 using the SR block showed an increase of 1.05% in accuracy.

A key observation from our study is the improved performance of the SR block when used together with regularization techniques like dropout, dropout2D, and dropblock. This synergistic effect suggests that the benefits of each of these methods complement each other. We think that the regularization methods may influence the SR block’s memory blocks to focus on class-dependent instead of instance-specific features. This can contribute to the accuracy gains.

Importantly, the benefits of the SR block go beyond simply adding more parameters to the network. This is evident when we look at how the SR block improves performance compared to simply making the network larger, such as the more than 10% increase in parameters when moving from ResNet34 to ResNet50. These comparisons show that the effectiveness of the SR block comes from its ability to remember and use features in a unique way, not just from adding more parameters.

Table II present the effect of adding the Squeeze-and-Remember block to SE and CBAM on CIFAR-100. It significantly improves performance across different ResNet models. For example, using the SR block with CBAM in ResNet34 boosts accuracy by 0.53%. When the SR block is combined with the SE block in ResNet50, there’s a 0.43% increase in accuracy.

TABLE I

MEAN TEST ACCURACY (TEST ACC.) AND STANDARD DEVIATION FOR CIFAR-100 EXPERIMENTS. “M” AND “P” INDICATE THE SR BLOCK’S INTEGRATION POINT AND NUMBER OF MEMORY BLOCKS, WHILE “OH” REFERS TO THE PARAMETER INCREASE PERCENTAGE DUE TO THE SR BLOCK.

Model	M	P	Test Acc.	OH
R18			76.47% $\pm$ 0.20%	
R18 + D			77.09% $\pm$ 0.26%	
R18 + D + SR	3	10	<b>77.26%</b> $\pm$ <b>0.11%</b>	1.47%
R18 + D2D			77.59% $\pm$ 0.24%	
R18 + D2D + SR	4	12	<b>77.93%</b> $\pm$ <b>0.25%</b>	0.88%
R18 + DB			78.01% $\pm$ 0.20%	
R18 + DB + SR	3	2	<b>78.12%</b> $\pm$ <b>0.15%</b>	0.30%
R18 + SR	4	2	76.74% $\pm$ 0.33%	0.15%
R34			77.07% $\pm$ 0.45%	
R34 + D			77.60% $\pm$ 0.49%	
R34 + D + SR	4	4	<b>77.96%</b> $\pm$ <b>0.38%</b>	0.16%
R34 + D2D			78.71% $\pm$ 0.28%	
R34 + D2D + SR	3	6	<b>78.81%</b> $\pm$ <b>0.28%</b>	0.46%
R34 + DB			78.68% $\pm$ 0.22%	
R34 + DB + SR	3	2	<b>78.84%</b> $\pm$ <b>0.22%</b>	0.16%
R34 + SR	3	6	77.22% $\pm$ 0.30%	0.46%
R50			76.17% $\pm$ 0.69%	
R50 + D			77.47% $\pm$ 0.40%	
R50 + D + SR	4	6	<b>77.72%</b> $\pm$ <b>0.29%</b>	0.21%
R50 + D2D			78.58% $\pm$ 0.58%	
R50 + D2D + SR	3	8	<b>78.68%</b> $\pm$ <b>0.55%</b>	0.55%
R50 + DB			78.34% $\pm$ 0.56%	
R50 + DB + SR	4	4	<b>78.43%</b> $\pm$ <b>0.29%</b>	0.14%
R50 + SR	3	2	77.22% $\pm$ 0.74%	0.14%
VGG16			72.48% $\pm$ 0.35%	
VGG16 + SR	2	8	<b>72.87%</b> $\pm$ <b>0.27%</b>	0.42%
VGG19			71.34% $\pm$ 0.14%	
VGG19 + SR	4	10	<b>71.42%</b> $\pm$ <b>0.27%</b>	0.10%

The improvements from the SR block show its ability to act as a supplementary component that memorizes important features during training, which are then adaptively recalled during inference. This improves the ability of the network to make predictions by reintegrating high-level features, extending the scope of feature enhancement beyond recalibration.

### B. ImageNet

We evaluated the SR block on the ImageNet 2012 classification dataset [26], which contains 1.2 million training, 50,000 validation, and 150,000 test images across 1,000 categories. Notably, regularization methods like dropout2d require longer training on ImageNet to converge effectively [23]. We run our experiments four times with varied random seeds for 450 epochs using ResNet50 and report top-1 classification accuracy on the validation set. Implementation details are provided in Appendix VII-B.

Table III shows the top-1 validation accuracy on ImageNet. Our experiments on ResNet50 models show that incorporating the SR block with dropout2d regularization improves the network’s performance, as seen by a 0.52% increase in accuracy over the baseline with dropout2d alone. Removing the SR block from a trained ResNet50, regularized with dropout2d and using 10 memory blocks, led to a decrease of 0.4%

TABLE II

MEAN TEST ACCURACY (TEST ACC.) AND STANDARD DEVIATION FOR CIFAR-100 EXPERIMENTS. COLUMNS “M” AND “P” DENOTE THE INTEGRATION POINT OF THE SR BLOCK AND THE NUMBER OF MEMORY BLOCKS. THE “OH” COLUMN INDICATES THE PERCENTAGE INCREASE IN PARAMETERS ATTRIBUTABLE TO THE USE OF FEATURE-MODIFYING BLOCKS SUCH AS SE, CBAM, AND SR.

Model	M	P	Test Acc.	OH
ResNet18 + CBAM	-	-	76.71% $\pm$ 0.28%	0.78%
<b>ResNet18 + CBAM + SR</b>	3	6	76.96% $\pm$ 0.24%	1.66%
ResNet18 + SE	-	-	76.96% $\pm$ 0.24%	0.79%
ResNet18 + SE + SR	3	2	76.86% $\pm$ 0.21%	1.09%
ResNet34 + CBAM	-	-	77.20% $\pm$ 0.33%	0.74%
<b>ResNet34 + CBAM + SR</b>	3	6	<b>77.73%</b> $\pm$ <b>0.21%</b>	1.05%
ResNet34 + SE	-	-	77.73% $\pm$ 0.22%	0.76%
<b>ResNet34 + SE + SR</b>	4	2	<b>77.93%</b> $\pm$ <b>0.23%</b>	0.84%
ResNet50 + CBAM	-	-	78.73% $\pm$ 0.54%	10.61%
<b>ResNet50 + CBAM + SR</b>	3	6	<b>79.06%</b> $\pm$ <b>0.29%</b>	12.28%
ResNet50 + SE	-	-	78.30% $\pm$ 0.30%	10.68%
<b>ResNet50 + SE + SR</b>	4	4	<b>78.73%</b> $\pm$ <b>0.30%</b>	11.24%

TABLE III

TOP-1 VALIDATION ACCURACY ON IMAGENET. “M” AND “P” INDICATE THE SR BLOCK’S INTEGRATION POINT AND NUMBER OF MEMORY BLOCKS, WHILE “OH” REFERS TO THE PARAMETER INCREASE PERCENTAGE DUE TO THE SR BLOCK.

Model	M	P	Validation Acc.	OH
R50			76.87% $\pm$ 0.15%	
R50 + D			77.53% $\pm$ 0.10%	
R50 + D + SR	3	2	<b>77.64%</b> $\pm$ <b>0.10%</b>	1.59%
R50 + D2D			77.27% $\pm$ 0.12%	
R50 + D2D + SR	3	2	<b>77.70%</b> $\pm$ <b>0.06%</b>	1.59%
R50 + D2D + SR	3	10	<b>77.79%</b> $\pm$ <b>0.02%</b>	7.87%
R50 + D2D + SR	3	20	77.76% $\pm$ 0.08%	15.73%
R50 + DB			77.40% $\pm$ 0.11%	
R50 + DB + SR	3	8	<b>77.47%</b> $\pm$ <b>0.11%</b>	6.31%
R50 + SR	3	10	76.90% $\pm$ 0.07%	7.87%

in accuracy. This result shows that the SR block positively impacts performance.

It is important to note that improvements are achieved even with a minimal configuration of 2 memory blocks, achieving a 0.43% increase in accuracy. However, scaling up to 20 memory blocks does not necessarily yield proportional benefits, potentially leading to inefficiencies due to channel redundancies. Despite increasing the number of parameters, the SR block maintains a low computational overhead, preserving a small impact on inference time (see Section III-E).

The SR block shows its most significant performance gains when combined with dropout2d. We think that this improvement is due to the channel dropping mechanism of dropout2d, which likely encourages the memory blocks within the SR block to capture more representative, class-specific features.

### C. Cityscapes

For semantic segmentation, we utilized the Cityscapes dataset [27], which contains high-quality, finely annotated images from 50 European cities, categorized into 19 semantic classes (2,975 training, 500 validation, and 1,525 test images).

TABLE IV

COMPARATIVE mIoU AND mAcc METRICS ON CITYSCAPES. THIS TABLE SHOWS THE IMPROVED MEAN INTERSECTION OVER UNION (mIoU) AND MEAN ACCURACY (mAcc) ACHIEVED BY INCORPORATING THE SR BLOCK INTO THE F-CONV AND DLv3 MODELS.

Model	mIoU	mAcc
F-CONV	77.33% $\pm$ 0.16%	85.02% $\pm$ 0.26%
F-CONV + SR	<b>77.53% <math>\pm</math> 0.23%</b>	<b>85.15% <math>\pm</math> 0.29%</b>
DLv3	79.76% $\pm$ 0.32%	86.73% $\pm$ 0.32%
DLv3 + SR	<b>79.96% <math>\pm</math> 0.19%</b>	<b>86.90% <math>\pm</math> 0.23%</b>

We run the experiments five times with varied random seeds using the MMSegmentation Framework [28], with implementation notes available in Appendix VII-C.

Table IV shows the SR block’s effect on semantic segmentation. In the Fully Convolutional Network (F-CONV) [29] (M=2, P=3) and DeepLab v3 (DLv3) [30] (M=2, P=2), we observed improvements in Mean Intersection over Union (mIoU) and Mean Accuracy (mAcc). The SR block increased F-CONV’s mIoU and mAcc by 0.2% and 0.13%, with 12.85% more parameter and 1.90% computational increase. For DLv3, mIoU and mAcc improved by 0.2% and 0.17%, with 24.83% more parameters and 3.81% more computational cost.

#### D. Interpretation of the SR mechanism

In this section, we investigate the functional dynamics of the SR block integrated into a ResNet50 architecture. Our analysis focuses on three areas: the softmax activation in the FCN, the influence of the SR block on the feature maps, and the characteristics of the memory blocks. The ResNet50 using a SR block with 10 memory blocks, was trained on the ImageNet dataset with dropout2d regularization.

**Softmax Activations:** We analyzed fifty samples each from the “cliff”, “pug”, and “church” classes, focusing on the mean and standard deviation of their softmax activations (see Figure 2). The results highlight different class-specific patterns in weight distribution by the FCN, showing a preference for using multiple memory blocks in the recall process. Notably, the “cliff” class shows a varied but consistent use of several blocks. This pattern differs from “pug” or “church” classes. This approach of combining memory blocks, allows effective high-level feature construction.

**Feature Map Modifications:** Figure 3 shows the SR block’s ability to modify feature maps. Pre- and Post-SR channel-wise mean activations, along with their Mean Activation Delta, reveal the class-specific feature additions by the block. The Mean Activation Delta shift by 2% to 10%, indicating an adaptive addition of high-level features. These can be more pronounced textures in “cliff” landscapes or architectural features in “church” images.

**Memory Block Analysis:** Figure 4 presents the mean channel activity for each of the SR block’s ten memory blocks. By computing the channel means within each memory block, we observe a variety of activation patterns. These patterns suggest that the memory blocks represent a comprehensive

feature-spanning set. The SR block combines these blocks to recall high-level feature for each class.

#### E. Comparative Analysis with SE/CBAM Blocks

This section compares the Squeeze-and-Remember block against the Squeeze-and-Excitation [3] and Convolutional Block Attention Module [12]. It highlights the unique contribution of the SR block to dynamic feature addition in CNNs.

1) *Initial Feature Processing:* The SE block uses global average pooling to distill spatial information into a global channel descriptor. As an extension, CBAM introduces spatial attention, which allows networks to focus on important features across both channel and spatial dimensions. In contrast, the SR block utilizes a  $1 \times 1$  convolution to extract essential features for the subsequent “Remember” phase. This step focuses on feature extraction over aggregation.

2) *From Channel Reweighting to Adaptive Feature Recall:* The SE block recalibrates feature maps by computing weights through a bottleneck structure with sigmoid activation, based on the global context of input features. CBAM extends this by incorporating sequential channel and spatial attention mechanisms. In contrast, the “Remember” stage of the SR block utilizes an FCN not for feature recalibration, but for adaptively weighting and merging memory blocks to construct and reintegrate high-level features.

3) *Feature Recalibration vs Feature Integration:* SE and CBAM scale the original feature map with learned importance weights to highlight significant features. Conversely, the SR block adds memorized features through its “Add” operation. This goes beyond the scope of feature recalibration.

## V. CONCLUSION AND FUTURE WORK

In this work, we introduced the Squeeze-and-Remember block, an innovative architectural unit designed to improve the feature-recall capabilities of CNNs. The SR block allows a network to adaptively add memorized features into feature maps.

Incorporating the Squeeze-and-Remember block into CNNs improves performance on benchmarks like CIFAR-100 and ImageNet. A ResNet50 model with the SR block outperforms the same model with just dropout2d by 0.52% in top-1 validation accuracy on ImageNet. Additionally, our findings highlight the SR block’s effectiveness in boosting accuracy by up to 0.54% in ResNet models when used with feature-modifying mechanisms like SE and CBAM. This shows its value as a component that improves feature processing beyond recalibration. In semantic segmentation, where the SR block is added to a ResNet50-backed DeepLab v3 model, increases the mean Intersection over Union by 0.20% on Cityscapes. Notable, this is achieved with low extra computational demand.

Future directions include the development of training strategies that more efficiently train the memory blocks with important features. Addressing current limitations, such as the zero-initialization of memorization blocks, could further increase the impact of the SR block.

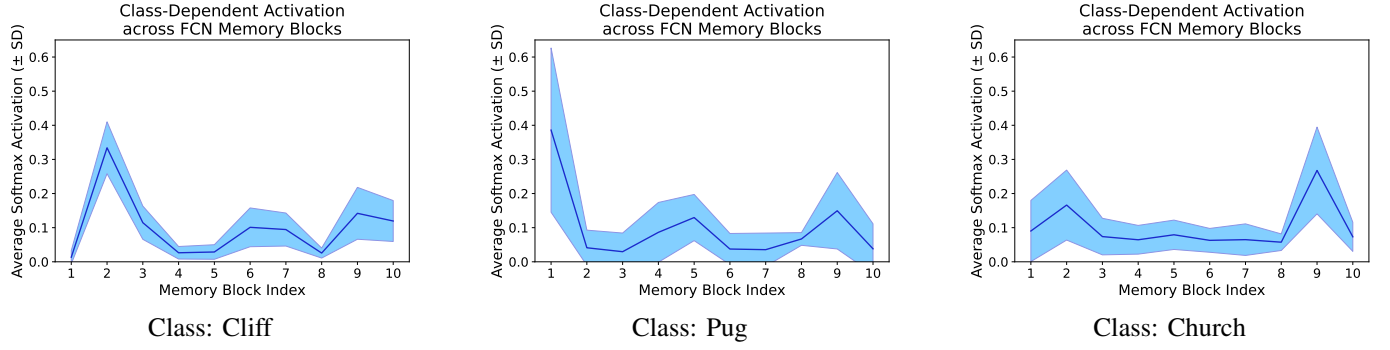


Fig. 2. Class-Conditional FCN Activation Patterns: Softmax activation means and standard deviations across ten memory blocks for “cliff”, “pug”, and “church” illustrate different strategies of memory usage. The “cliff” class shows diverse but consistent activation in contrast to “pug” and “church”.

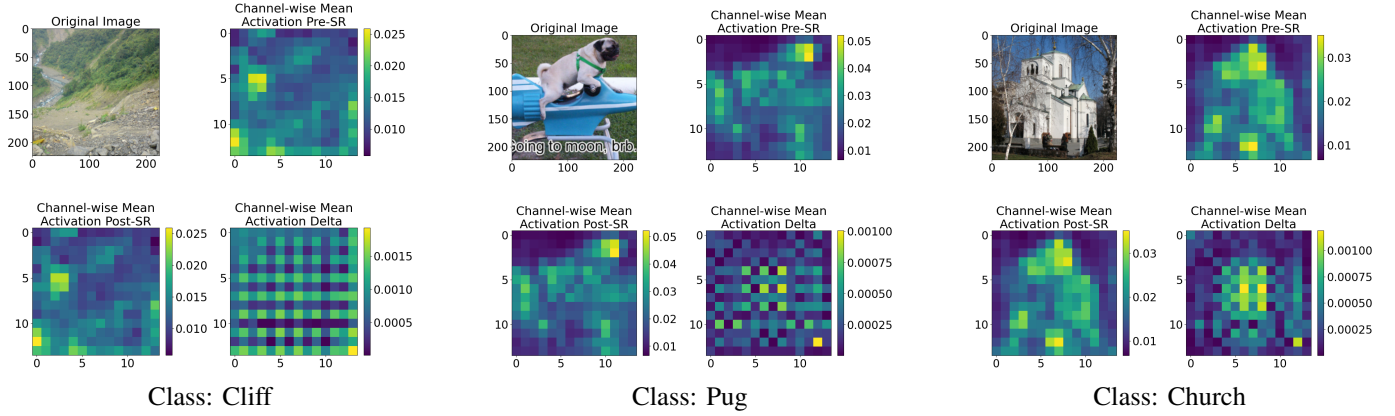


Fig. 3. Different Impact of SR Block on Feature Maps: The figure illustrates class-dependent feature map transformations for “Cliff”, “Pug”, and “Church” classes in the SR block. It contrasts channel-wise activations before and after SR processing, with Mean Absolute Differences.

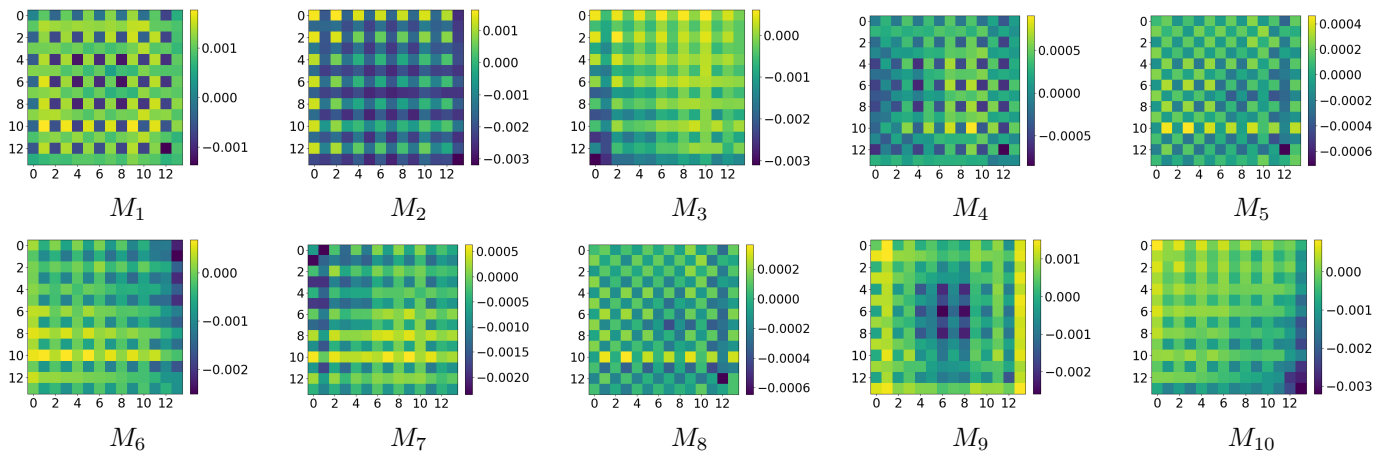


Fig. 4. Feature Encoding in SR-Enhanced ResNet50: The figures  $M_1$  through  $M_{10}$  show the average channel activity in the ten memory blocks in the SR block. The variation in activation patterns demonstrates the memory blocks complex encoding capabilities. Each block selectively capturing different aspects of the feature spectrum.

## REFERENCES

- [1] 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 Computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [2] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, pp. 611 – 629, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:49389393>
- [3] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 7132–7141. [Online]. Available: [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Hu\\_Squeeze-and-Excitation\\_Networks\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html)
- [4] S. Ackerman, "Discovering the Brain." *Washington (DC): National Academies Press (US)*, 1992, 8, Learning, Recalling, and Thinking. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK234153/>
- [5] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167278919305974>
- [6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [7] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456. [Online]. Available: <http://proceedings.mlr.press/v37/ioffe15.html>
- [8] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *CoRR*, vol. abs/1410.5401, 2014. [Online]. Available: <http://arxiv.org/abs/1410.5401>
- [9] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwinska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. P. Agapiou, A. P. Badia, K. M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis, "Hybrid computing using a neural network with dynamic external memory," *Nat.*, vol. 538, no. 7626, pp. 471–476, 2016. [Online]. Available: <https://doi.org/10.1038/nature20101>
- [10] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1410.3916>
- [11] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 1378–1387. [Online]. Available: <http://proceedings.mlr.press/v48/kumar16.html>
- [12] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11211. Springer, 2018, pp. 3–19. [Online]. Available: [https://doi.org/10.1007/978-3-030-01234-2\\_1](https://doi.org/10.1007/978-3-030-01234-2_1)
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.90>
- [15] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *CoRR*, vol. abs/1607.06450, 2016. [Online]. Available: <http://arxiv.org/abs/1607.06450>
- [16] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," *CoRR*, vol. abs/1607.08022, 2016. [Online]. Available: <http://arxiv.org/abs/1607.08022>
- [17] Y. Wu and K. He, "Group Normalization," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11217. Springer, 2018, pp. 3–19. [Online]. Available: [https://doi.org/10.1007/978-3-030-01261-8\\_1](https://doi.org/10.1007/978-3-030-01261-8_1)
- [18] R. Cakaj, J. Mehnert, and B. Yang, "Spectral Batch Normalization: Normalization in the Frequency Domain," in *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–10.
- [19] D. Scherer, A. C. Müller, and S. Behnke, "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition," in *Artificial Neural Networks - ICANN 2010 - 20th International Conference, Thessaloniki, Greece, September 15-18, 2010, Proceedings, Part III*, ser. Lecture Notes in Computer Science, K. I. Diamantaras, W. Duch, and L. S. Iliadis, Eds., vol. 6354. Springer, 2010, pp. 92–101. [Online]. Available: [https://doi.org/10.1007/978-3-642-15825-4\\_10](https://doi.org/10.1007/978-3-642-15825-4_10)
- [20] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten Digit Recognition with a Back-Propagation Network," in *Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989]*, D. S. Touretzky, Ed. Morgan Kaufmann, 1989, pp. 396–404. [Online]. Available: <http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network>
- [21] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2670313>
- [22] L. Wan, M. D. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, "Regularization of Neural Networks using DropConnect," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conference Proceedings, vol. 28. JMLR.org, 2013, pp. 1058–1066. [Online]. Available: <http://proceedings.mlr.press/v28/wan13.html>
- [23] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," in *Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018*. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/file/7edcfb2d8f6a659ef4cd1e6c9b6d7079-Paper.pdf>
- [24] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using Convolutional Networks," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 648–656. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7298664>
- [25] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015. [Online]. Available: <https://doi.org/10.1007/s11263-015-0816-y>
- [27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [28] M. Contributors, "MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and benchmark," <https://github.com/open-mmlab/mms Segmentation>, 2020.
- [29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12,*



2015. IEEE Computer Society, 2015, pp. 3431–3440. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7298965>

- [30] L. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” *CoRR*, vol. abs/1706.05587, 2017. [Online]. Available: <http://arxiv.org/abs/1706.05587>
- [31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/bd8ca288fee7f92f2bfa9f7012727740-Abstract.html>
- [32] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J. Big Data*, vol. 6, p. 60, 2019. [Online]. Available: <https://doi.org/10.1186/s40537-019-0197-0>
- [33] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 1026–1034. [Online]. Available: <https://doi.org/10.1109/ICCV.2015.123>

## VI. EVALUATING SR BLOCK ON CIFAR-10

In complement to our main manuscript, we present extended experiments on the CIFAR-10 dataset, consisting of 50,000 training images and 10,000 test images of 32x32 pixel resolution. We utilized a range of CNN architectures for our experiments, including ResNet18/34/50 [14] and VGG16/19-BN [25].

To ensure robustness and reproducibility, each model was trained and evaluated five times using different random seeds, impacting network initialization, data ordering, and augmentation processes. We present the mean test accuracy and its standard deviation for these trials. The data split comprised 90% for training and 10% for validation, with the best-performing model on the validation set chosen for the final evaluation.

Baseline networks, including various ResNet configurations, were subjected to data augmentation, weight decay, early stopping, and varying dropout techniques (dropout, dropout2d, dropblock). The VGG models (VGG16-BN and VGG19-BN) followed a similar training regimen, with dropout applied exclusively in their fully connected (FC) layers. Detailed implementation details are available in Appendix VII-A.

Our CIFAR-10 experiments show consistent accuracy improvements across various models, similar to our CIFAR100 findings. The performance enhancements from adding the SR block are detailed in Table V, including where it’s integrated (**M**), the number of memory blocks used (**P**), and the additional parameters it introduces (**OH**). Importantly, combining the SR block with regularization techniques boosts performance further than using these methods on their own.

## VII. DETAILED METHODOLOGICAL FRAMEWORK

### A. Methodology for CIFAR Datasets

Our CIFAR experiments used PyTorch 1.10.1 [31] on an Nvidia GeForce 1080Ti GPU. We used the same training

TABLE V  
MEAN TEST ACCURACY (TEST ACC.) AND STANDARD DEVIATION FOR CIFAR-10 EXPERIMENTS. “M” AND “P” INDICATE THE SR BLOCK’S INTEGRATION POINT AND NUMBER OF MEMORY BLOCKS, RESPECTIVELY, WHILE “OH” REFERS TO THE PARAMETER INCREASE PERCENTAGE DUE TO THE SR BLOCK.

Model	M	P	Test Acc.	OH
R18			94.04% ± 0.08%	
R18 + D			94.29% ± 0.13%	
R18 + D + SR	3	6	<b>94.36% ± 0.17%</b>	0.89%
R18 + D2D			94.40% ± 0.03%	
R18 + D2D + SR	3	10	<b>94.56% ± 0.13%</b>	1.48%
R18 + DB			94.43% ± 0.14%	
R18 + DB + SR	4	6	<b>94.52% ± 0.17%</b>	0.45%
R18 + SR	4	4	93.93% ± 0.37%	0.34%
R34			93.69% ± 0.30%	
R34 + D			94.55% ± 0.31%	
R34 + D + SR	3	10	<b>94.62% ± 0.29%</b>	0.77%
R34 + D2D			94.30% ± 0.14%	
R34 + D2D + SR	3	8	<b>94.50% ± 0.12%</b>	0.23%
R34 + DB			94.38% ± 0.29%	
R34 + DB + SR	4	2	<b>94.60% ± 0.21%</b>	0.08%
R34 + SR	3	8	94.18% ± 0.73%	0.62%
R50			93.31% ± 0.36%	
R50 + D			94.09% ± 0.16%	
R50 + D + SR	3	2	<b>94.16% ± 0.23%</b>	0.15%
R50 + D2D			93.99% ± 0.29%	
R50 + D2D + SR	3	6	<b>94.01% ± 0.29%</b>	0.42%
R50 + DB			93.97% ± 0.31%	
R50 + DB + SR	3	2	<b>94.21% ± 0.19%</b>	0.15%
R50 + SR	3	6	93.79% ± 0.41%	0.42%
VGG16			93.27% ± 0.12%	
VGG16 + SR	4	8	<b>93.53% ± 0.24%</b>	0.10%
VGG19			93.21% ± 0.08%	
VGG19 + SR	2	12	<b>93.39% ± 0.15%</b>	0.48%

setup for all models, with a batch size of 128 and the SGD optimizer with a momentum of 0.9. The initial learning rate was set to 0.1, with adjustments based on the dataset: for CIFAR-10, it was reduced by a factor of 0.1 at epochs 90 and 136, along with a weight decay of  $1e - 4$  [14]; for CIFAR-100, it was reduced by 0.2 at epochs 60, 120, and 160, using a weight decay of  $5e - 4$ . These parameter settings match those recommended in [14], without further hyperparameter optimization. Data augmentation techniques included random cropping (size 32 with padding 4), random horizontal flipping, and normalization, as outlined in [32], with normalization being the only preprocessing step for the test set. Network initialization followed the Kaiming uniform approach [33], with VGG’s linear layers similarly initialized and its convolutional layers set to a Gaussian distribution ( $\mu = 0$ ,  $\sigma = \sqrt{2/n}$ ).

### B. Methodology for ImageNet Datasets

In our experiments, we used PyTorch 1.10.1 [31] with four Nvidia GeForce 1080Ti GPUs. All models were trained with the SGD optimizer, set with a momentum of 0.9 and an initial learning rate of 0.1, which was reduced by a factor of 0.1 at epochs 75, 150, 225, 300, and 375. A weight decay factor of  $1e - 4$  was applied, and the batch size was standardized at 128.

For data augmentation, the training dataset was randomly cropped to  $224 \times 224$ , randomly flipped horizontally, and normalized. The validation dataset was resized to  $256 \times 256$ , center cropped to  $224 \times 224$ , and normalized. Our setup matches the configuration specified in the PyTorch ImageNet training example<sup>1</sup>.

### C. Methodology for Cityscapes Datasets

In the Cityscapes experiments, PyTorch 1.10.1 [31] and four Nvidia GeForce 1080Ti GPUs were used. Using the MM-Segmentation Framework [28], we utilized FCON [29] and DeepLab v3 (DLv3) [30] on the dataset [27], which consists of 2,975 training, 500 validation, and 1,525 testing images across 19 semantic classes. Training involved resizing, random cropping, flipping, photometric distortion, normalization, and padding. Testing employed multi-scale flip augmentation and normalization.

The FCON model, with a ResNet50 backbone, used an Encoder-Decoder architecture with FCONHead as the decode and auxiliary heads. DLv3, also on a ResNet50 backbone, incorporated an Atrous Spatial Pyramid Pooling (ASPP) head with dilations of 1, 12, 24, and 36, and an FCONHead for the auxiliary head. Both models used SyncBN and a dropout ratio of 0.1, with the auxiliary head contributing 40% to the total loss.

Optimization was via SGD (learning rate 0.01, momentum 0.9, weight decay 0.0005). A polynomial decay learning rate policy was applied (power 0.9, minimum learning rate  $1e-4$ ), over 80,000 iterations with checkpoints and evaluations (focusing on mIoU) every 8,000 iterations.

## VIII. EXTENDED ANALYSIS OF FCN OUTPUT WEIGHTING ON MEMORY BLOCKS IN RESNET50

This section builds on the results of the main paper by exploring the weighting mechanism of the FCN output on memory blocks within ResNet50 models using dropblock and dropout2d regularization. We performed a detailed analysis of the softmax activation patterns at the FCN output to further understand this aspect.

We selected five classes from the ImageNet dataset to represent a wide range of semantic and visual diversity: “goldfish”, “pug”, “plane”, and “cliff” as used also in the SE networks [3], with the addition of “church” to increase the diversity of the dataset. For each class, fifty samples from the validation set were analyzed to compute the mean and standard deviation of softmax activations at the FCN final layer. Figures 5, 6, and 7 show these activations for ResNet50 models augmented with SR blocks of different memory sizes and trained with dropout2d and dropblock. They demonstrate the adaptability of the SR block in processing different class features.

We contrast them with the average activations across all 1000 classes using the images from the validation. An important observation is that the SR block does not equal the number of memory blocks to the number of classes; rather, it

uses an ensemble of memory blocks. In particular, each class influences the weighting of the memory blocks in a different way, thus controlling the feature enrichment process in a class-specific way.

<sup>1</sup><https://github.com/pytorch/examples/blob/main/imagenet>

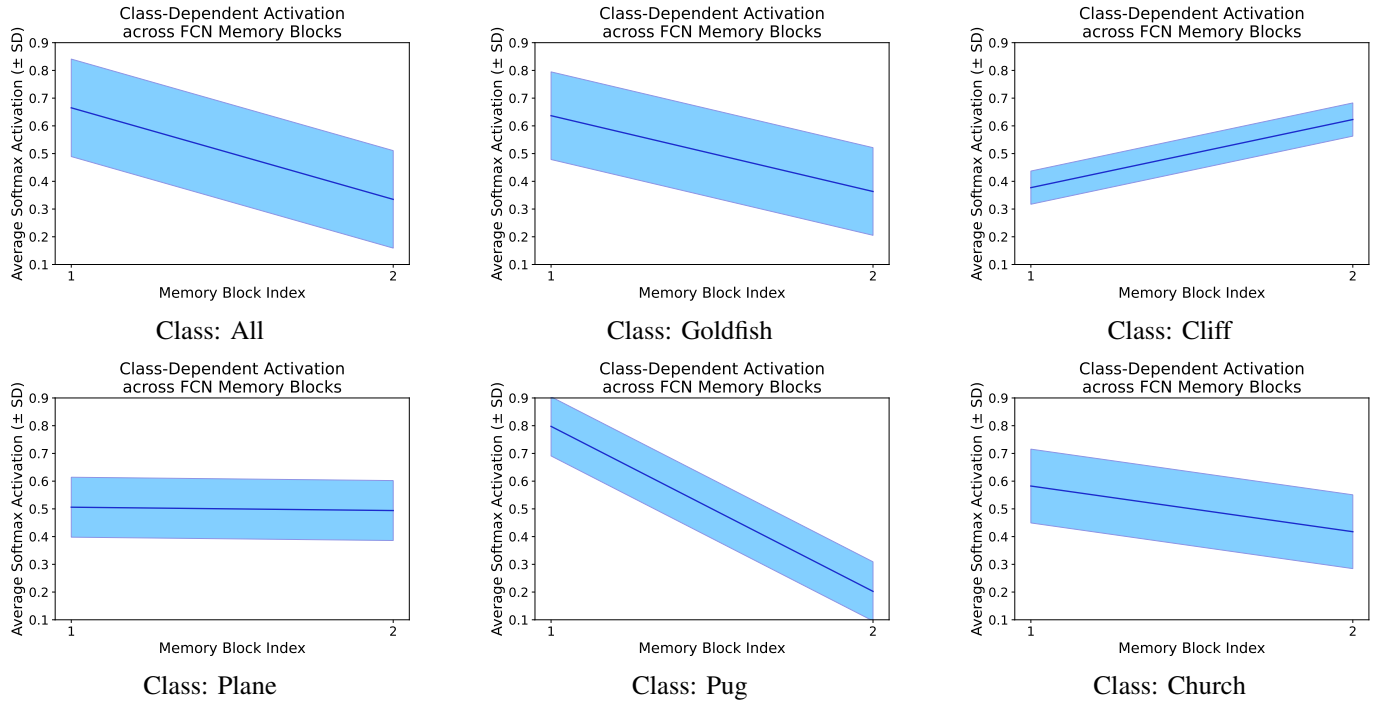


Fig. 5. This figure shows the means and standard deviations of softmax activations for “all”, reflecting the activation across the validation set, and for specific classes, including “goldfish”, “cliff”, “plane”, “pug”, and “church”. It illustrates the unique memory usage within a ResNet50 using a SR block with two memory blocks and trained with dropout2d. Notably, “cliff” shows a wide range of activation behaviors that differ from the more uniform patterns of “pug” and “church”, highlighting the ability of the FCN to adapt feature processing in a class-dependent manner.

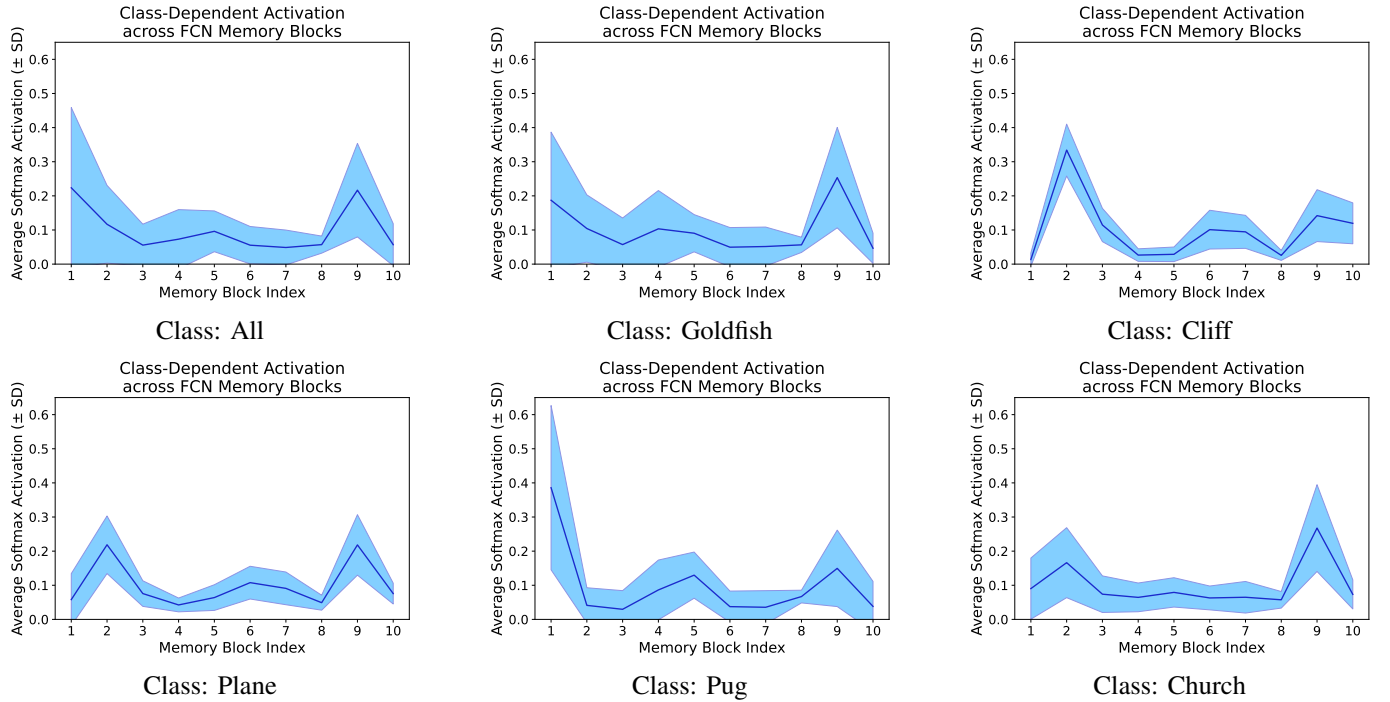


Fig. 6. This figure shows the means and standard deviations of softmax activations for “all”, reflecting the activation across the validation set, and for specific classes, including “goldfish”, “cliff”, “plane”, “pug”, and “church”. It illustrates unique memory usage strategies within a ResNet50 using a ten-memory SR block and trained with dropout2d. Notably, “cliff” exhibits a wide range of activation behaviors that differ from the more uniform patterns of “pug” and “church”, highlighting the ability of the FCN to adapt feature processing in a class-dependent manner.

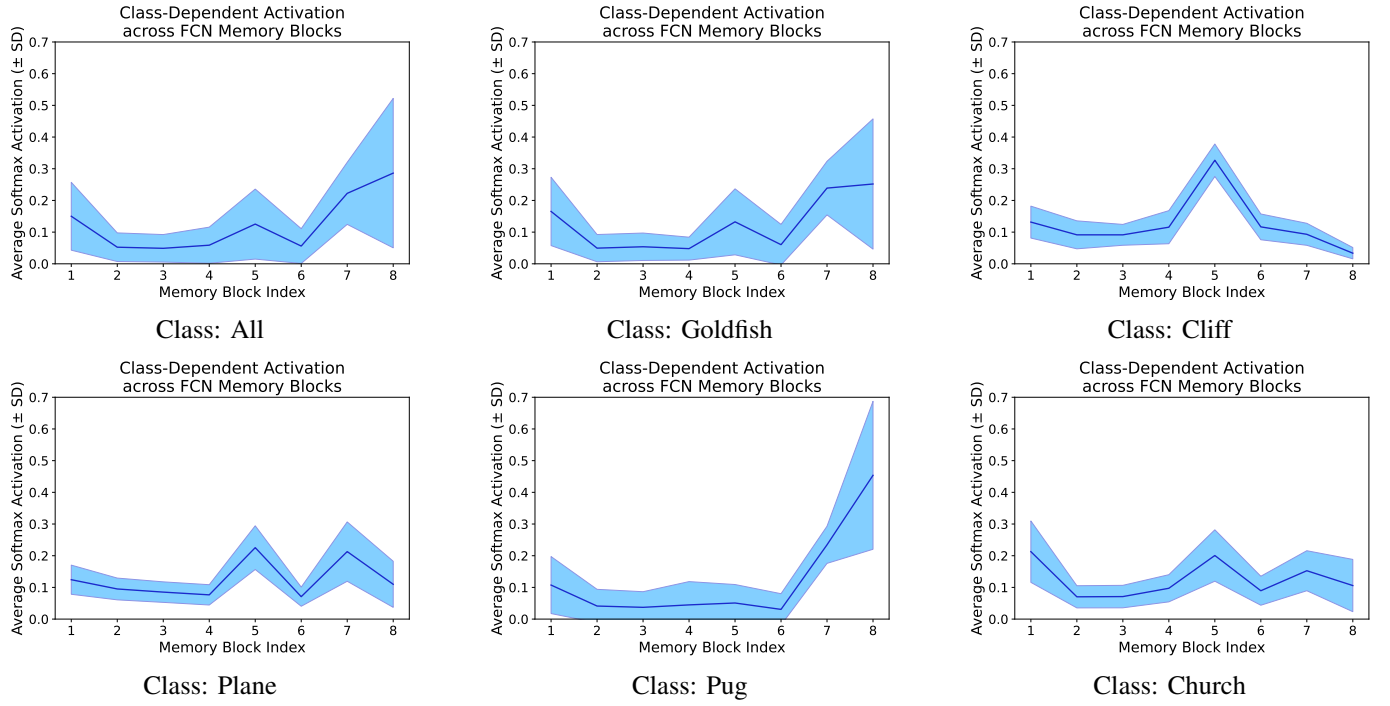


Fig. 7. This figure shows the means and standard deviations of softmax activations for “all”, reflecting the activation across the validation set, and for specific classes, including “goldfish”, “cliff”, “plane”, “pug”, and “church”. It illustrates unique memory usage strategies within ResNet50 using an eight-memory SR block and trained with dropblock. Notably, “cliff” exhibits a wide range of activation behaviors that differ from the more uniform patterns of “pug” and “church”, highlighting the ability of the FCN to adapt feature processing in a class-dependent manner.