

BACS: Background Aware Continual Semantic Segmentation

Mostafa ElAraby
DIRO, Mila - Quebec AI Institute
Université de Montréal
Montreal, Canada
 {elarabim}@mila.quebec

Ali Harakeh
DIRO, Mila - Quebec AI Institute
Université de Montréal
Montreal, Canada

Liam Paull
CIFAR AI Chair
DIRO, Mila - Quebec AI Institute
Université de Montréal
Montreal, Canada

Abstract—Semantic segmentation plays a crucial role in enabling comprehensive scene understanding for robotic systems. However, generating annotations is challenging, requiring labels for every pixel in an image. In scenarios like autonomous driving, there’s a need to progressively incorporate new classes as the operating environment of the deployed agent becomes more complex. For enhanced annotation efficiency, ideally, only pixels belonging to new classes would be annotated. This approach is known as Continual Semantic Segmentation (CSS). Besides the common problem of classical catastrophic forgetting in the continual learning setting, CSS suffers from the inherent ambiguity of the background, a phenomenon we refer to as the “background shift”, since pixels labeled as background could correspond to future classes (forward background shift) or previous classes (backward background shift). As a result, continual learning approaches tend to fail. This paper proposes a Backward Background Shift Detector (BACS) to detect previously observed classes based on their distance in the latent space from the foreground centroids of previous steps. Moreover, we propose a modified version of the cross-entropy loss function, incorporating the BACS detector to down-weight background pixels associated with formerly observed classes. To combat catastrophic forgetting, we employ masked feature distillation alongside dark experience replay. Additionally, our approach includes a transformer decoder capable of adjusting to new classes without necessitating an additional classification head. We validate BACS’s superior performance over existing state-of-the-art methods on standard CSS benchmarks.

Keywords—Continual Learning, Semantic Segmentation, Catastrophic Forgetting, Background Shift, Incremental Learning.

I. INTRODUCTION

A typical assumption in training deep neural network models is the availability of the entire dataset at training time. However, in many applications, incrementally learning a new stream of classes without forgetting previously-learned knowledge is required for achieving human-level intelligence. When we fine-tune a model on a new set of classes, it may forget previously acquired knowledge (catastrophic forgetting) [1]–[3]. In recent years, CNN have shown tremendous progress on many computer vision tasks, including semantic segmentation, with all classes learned jointly in a single shot [4]–[6]. However, during deployment on a robot, for example, novel classes may emerge which

necessitate updating the model. For that purpose, Continual Semantic Segmentation (CSS) [7]–[10] learn new classes incrementally without a need for retraining on data from previous classes.

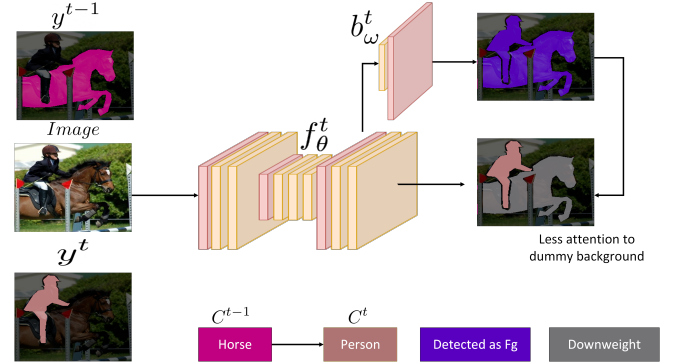


Figure 1: Background Aware Continual Semantic Segmentation (BACS) framework overview. A backward background detector b_{ω}^t down weights background pixels detected that have appeared as classes in previous steps to avoid the collapse of old classes into the background.

One of the main challenges of class-incremental learning is the catastrophic forgetting of previously observed data points distributions while learning distributions with data from new classes. Existing class incremental learning literature creates a trade-off between learning new knowledge (plasticity) and keeping previously acquired knowledge (stability) [11]. CSS poses more challenges exacerbating the catastrophic forgetting phenomenon. In CSS, annotations are present only for new classes. By default, everything else is labeled as background to make it easier and more realistic during the data collection and annotation. The collapse of old and future classes into the background label results in a shift in the background distribution during training which exacerbate further the forgetting of old classes. We distinguish between two types of background shifts, forward and backward. In the forward background shift, the distribution of the background class is shifting towards the current step’s new classes. On the other hand, in the backward background shift, the distribution of old classes is moving to the current

step’s background class. The collapse of old and future classes into the background class causes a misalignment of features collected from previous steps and exacerbates catastrophic forgetting.

Recent works attempt to address both background shift and catastrophic forgetting using a custom Knowledge Distillation (KD) [12] technique used extensively in CSS. CSS literature [9], [10], [13], [14] focuses on teacher-student KD in all classes while taking into account the background shift. However, the more we constrain our student network, the better the stability at the expense of the network’s plasticity. Moreover, the network’s performance on new classes will rely heavily on the order and number of classes used in the initial step, making it harder to start the learning process from a few classes. For the background shift problem, current literature relies on the pseudo-labeling of existing data to detect previous classes, causing potential overfitting of errors coming from old classes known as the problem of confirmation bias in pseudo-labeling [15].

Furthermore, adding new classes in each step, we need to update our architecture to accommodate new classes. CSS baselines use a separate classification head where we initialize a new head per new upcoming set of classes without knowing the step number during inference (class-incremental setup). Random initialization of new heads could provoke a misalignment among background class features learned by the previous model [9], affecting the stability of the network if not initialized using the background class weights.

Our paper first identifies the drawback of using pseudo-labeling and output KD during the training and its effect on the plasticity of the network. First, we introduce a backward background detector, BACS, network connected to our latent space representation to detect if the pixel is a “true” background or corresponds to an old class from any previously observed step, as shown in fig. 1. Next, we incorporate the output of BACS into our loss function to mitigate the effect of the background shift. Moreover, we propose a Masked Knowledge Distillation (MKD) on the penultimate layer’s features that only focus on background pixels detected as foreground. Lastly, we use a transformer decoder as the output classification layer to avoid the initialization trick of new heads and to make the model use fewer parameters than the multi-classifier setup.

In summary, our contributions are the following:

- We propose a backward background shift detector trained to detect the background and foreground of each step based on the distance from saved foreground prototypes.
- We introduce a variation of the cross-entropy loss function that uses the output of BACS during training to mitigate catastrophic forgetting.
- We propose a feature-based MKD that uses the information from the down-weighted background pixels.

- Finally, we suggest using a transformer-like decoder to avoid both initializing new heads and copying the weights of the background class [9].

We demonstrate the effectiveness of BACS in addressing the background shift problem and handling long steps while improving by a large margin both the plasticity and stability of the network compared to existing baselines [10], [14], [16], [17]. Moreover, our method adds new classes by simply appending a single parameter without perturbing the background class’s feature space, resulting in better stability. Finally, we perform a set of ablation studies to show the effect of each proposed component on our results.

II. RELATED WORK

This section will summarize the most important related work in semantic segmentation, continual learning, and continual semantic segmentation CSS.

A. Semantic Segmentation

Early methods [18]–[20] for semantic segmentation relied on classifying patches of input images, then refining the predictions based on the context. Later, fully convolutional networks (FCN) [21], [22] enable the substitution of fully connected layers with convolutional ones to produce spatial maps. Compared with traditional methods, the FCN model improved various semantic segmentation tasks achieving end-to-end semantic segmentation. The U-Net [23] improved over the FCN by learning context and spatial information using convolutional upsampling. Current methods [21], [24]–[30] focus on learning multi-scale feature aggregation from existing pre-trained convolutional networks. Recent methods [31]–[34] relied on the attention score to learn the connections between image contexts. Another set of methods attempted to fuse various receptive fields of view using atrous convolutions [4], [35] and spatial pyramids with dilated convolutions [36], [37] in an encoder-decoder setup. Following the success of transformers in natural language processing [38], an adapted transformer architecture has become prevalent for computer vision tasks [39]–[42], showing an improvement over existing architectures. Existing transformer architectures in semantic segmentation either focus on multi-scale feature fusion [43]–[46] or contextual feature aggregation [47], [48]. In our work, we extend DeepLabV3 [4] to have a dynamically extendable decoder similar to the decoder of Segformer [45]. Our proposed decoder enables us to extend our model with novel classes with low memory and compute requirements.

B. Continual Learning

Continual learning focuses on learning from a sequential data stream to gradually extend acquired knowledge without forgetting old knowledge. Data can stem from different domains (covariate shift) or different tasks. In some literature, continual learning is also called lifelong learning [49]–[52],

sequential learning [53]–[55] or incremental learning [56]–[59]. The major challenge in continual learning is learning new tasks without suffering from catastrophic forgetting. This problem originates from the plasticity-stability dilemma [60] in deep learning systems. Plasticity refers to the ability to integrate new knowledge and stability in retaining previous knowledge while learning new ones. In CSS, we focus on the task incremental learning setting since incremental tasks share the same background class. In incremental learning, we can distinguish three methods: replay-based, regularization-based, and Parameter isolation methods.

Replay methods rely on storing samples from old tasks or generating pseudo-samples with a generative model, then replaying them while learning new tasks to avoid forgetting. The end-to-end incremental learning [61] relies on keeping a balanced memory of previously seen classes and merges them with the new task data in an end-to-end learning fashion. Dark Experience Replay (DER) [62] relies on replaying previous tasks’ data saved along with their logits from earlier tasks, thus matching the network’s output with its past through the optimization trajectory. Gradient Episodic Memory (GEM) [63] uses exemplars to solve a constrained optimization problem that chooses gradient updates in the direction of learning new tasks while retaining knowledge of previously seen classes. We can consider replay data as some low-resource training with few data points. In our proposed framework, we mitigate the catastrophic forgetting using Dark Experience Replay (DER) [62] as it provides a solid baseline empirically shown to converge to flatter minima compared to vanilla experience replay.

Regularization-based approaches for continual learning avoid saving samples from previous tasks, prioritizing privacy, and reducing the memory used to store previous samples. Instead, these approaches add a regularization term to consolidate previous knowledge while learning new classes. The initial idea was to save the output of an earlier task model given a new input image to alleviate catastrophic forgetting [64]. That same approach has been re-introduced by Learning Without Forgetting (LwF) [65] using a Knowledge Distillation (KD) loss while training on new tasks to preserve the decision boundary of the neural network. However, these methods are vulnerable to domain shift between tasks [50], as they tend to keep the feature space near its counterpart trained on the previous task. For that reason, shallow autoencoders are used to constrain task features in their corresponding learned low dimensional space [66], thus reducing the negative effect of domain shift between tasks.

In our proposed framework, we use a soft teacher-student knowledge distillation on the penultimate layer that only focuses on features belonging to old classes without affecting the plasticity of the network.

C. Continual Semantic Segmentation

In continual semantic segmentation, the background class might include pixels associated with previously observed classes from earlier steps, as in fig. 1, exacerbating catastrophic forgetting. Most of the existing work [9], [10], [14], [17], [67]–[69] keeps the old network from the previous step and uses KD to keep old step information. Using pseudo-labeling and a teacher-student approach makes starting with a few classes harder, as the optimization procedure tends to keep the network close to its weights in the initial step. Pod distillation loss [14] regularizes the current step’s network output with the previous one at each output layer and discards uncertain pseudo-labels. Later Pod distillation was extended using an object replay buffer to support long sequences [13]. Matching class representation prototypes and repulsing different classes’ representation [17] improved the representation space and reduced the effect of catastrophic forgetting compared with the classic knowledge distillation loss. [70] proposed a custom convolutional layer that fuses weights from the previous step layer with the current one to reduce the effect of catastrophic forgetting. Another set of methods [16], [71], [72] has developed several replay-based methods for CSS. Finally, a separate feature extractor is created per step and frozen at the end of the step, thus reducing the catastrophic forgetting to its minimum while using binary cross-entropy loss to avoid the background shift [73].

Our proposed approach keeps a small buffer of examples and their corresponding logits from previous steps. We use BACS to differentiate between the background and old classes instead of pseudo-labeling that would accumulate errors. In the subsequent section, we explain how we detect and disentangle the background class from previously observed foreground classes in continual semantic segmentation.

III. PROPOSED METHOD

A. Notation and Problem Setting

In Continual Semantic Segmentation (CSS), we observe a set of incremental classes $t = 1 \dots T$, each having a dataset D^t . In each incremental step t , the model has to learn a set of classes C^t using dataset D^t consisting of a set of 2D images x^t of size $N = H \times W$ and 2D ground truth per pixel labels $y_i^t \in C^t \forall i = 1 \dots N$. The label space of each step t consists of new classes C^t and a dummy background class c_{bg} , which is associated with every pixel that was not labeled. We assume there is no intersection between label spaces C of various steps, resulting in a dummy background class c_{bg} possibly containing future and old classes. We denote the collapse of old classes into the background class by *backward background shift* and the collapse of new classes by *forward background shift*. We follow the same set of scenarios considered in [9], [14] grouped into three modes

sequential, disjoint, and overlap. The sequential mode keeps the ground truth of all labels $C^{1:t} \triangleq \bigcup_{t'=1}^t C^{t'}$ observed so far. On the other hand, the disjoint mode keeps only the ground truth of the current step C^t while having the backward background shift. The third and final mode is overlap that uses all images having at least a class from C^t with everything else as c_{bg} and thus suffers from both backward and forward background shift. In our work, we consider the overlap mode, which is the most realistic and challenging.

We define our semantic segmentation model f_θ^t to be a mapping from input image x^t to per-pixel predictions, as shown in fig. 2. We denote the free logits before the softmax layer for input x^t by $z^t \in \mathbb{R}^{N \times |C^{1:t}|}$ ($z^t[i, c]$ denotes the logit value for pixel i that corresponds to class c), and the hidden penultimate layer per pixel features for input x^t by $j^t \in \mathbb{R}^{N \times |d|}$ where d is the size of the decoder’s hidden representation. Our model f_θ^t consists of a convolutional feature extractor h^t and a decoder g^{t1} . Existing baselines’ architecture uses a decoder g^t consisting of another feature extractor and a separate classifier per step t ; thus, we do not need to know the step id during inference. However, due to the forward background shift, random initialization of new classifiers would perturb the feature space from the background to the new classes. For that reason, recent methods [9], [13], [14], [16], [17], [73] used *unbiased initialization trick* that uses the background class weights to initialize new heads in a way that incentivize the model to predict new classes as the background class. Furthermore, we denote \mathcal{M} as the exemplar balanced memory consisting of a small set of samples from previous steps $D^{1:t-1}$ along with their corresponding logits at its step t , and use it as a replay for the current step.

CSS faces the same inherent challenges of continual learning; the model forgets old classes while learning new ones. Without a forward or backward background shift, we can adapt continual learning methods to achieve acceptable results in the sequential mode. However, CSS comes with two additional challenges. The first one is the background shift which we split into forward and backward for simplicity. The second is the initialization of new heads in a forward background shift where the network might observe future classes as a background class. Existing baselines [9], [10], [14], [17] solves the first challenge by either pseudo-labeling using the previous step network f_θ^{t-1} to detect backward background shift or to train separate feature extractors and classifiers using binary cross-entropy while freezing old steps’ heads [73]. For the forward background shift, the parameters of new classifiers are initialized precisely as the background class to reduce the feature space perturbation,

¹Convention f_θ^t means model trained at the end of step t . Still, we use a single network trained at the end of all steps during inference. On the other hand, output heads are separate per step in the conventional multi-convolutional setup [10]

thus decreasing its counter-effect on the background class.

B. Backward Background Shift Detector

This section proposes an effective method to tackle background shift in CSS. First, we introduce our backward background shift detector b_ω that follows a Siamese network architecture [74] comparing input pixel representation with each foreground representation saved as a running mean per step t (prototype) P^t . The backward background shift detector maps the encoder’s feature space $h^t(x^t)$ to multi-label classification detecting whether a pixel belongs to the foreground of any of the classes previously observed using a binary probability $Fg^t \in \mathbb{R}^{N \times |t|}$. The higher the probability Fg^t at step t , the higher the probability of being a foreground at step t . The proposed backward background detector b_ω consists of a convolutional feature projector K , trained only in the initial step, followed by a set of 1×1 convolutional filters ϕ_Φ^t ². We maintain prototype P^t as a representative of the foreground classes in step t by computing a running mean of the projection’s network output $K(h^t(x^t))$ during the training phase.

Following the analysis from [75], deeper layers contribute the most to forgetting, and thus we need to only train the projection network K on the first step. For subsequent steps, we only train step t ’s output convolutional filter ϕ_Φ^t to disentangle Fg^t from anything else since we only have the ground truth C^t , and everything else is collapsed to c_{bg} . For previous step heads, we freeze their corresponding output heads $\phi_\Phi^{1:t-1}$ and rely on the prototypes P^t to retain previous steps’ knowledge in detecting their corresponding foreground. We use the binary focal loss [76] $\ell_{bacs}^t(x^t, y^t)$ to train each detection head with the projection network K only trained on the first step. Hence, our detector can identify backward background shift based on the maximum logit value from each trained head. Next, we introduce our loss function that uses the backward background shift detector.

C. Backward Background Shift Aware Loss

With the background shift, a conventional cross-entropy loss would exacerbate catastrophic forgetting as the model learns to associate old classes with the dummy background. Hence, we introduce a backward background shift aware loss function that uses the output of our background detector heads to guide the training. Our new loss, inspired by the focal loss function [76], focuses on the actual background class by using a focal weight based on the maximum output from our backward background shift detector and hence ignores the dummy background detected that has high foreground probability. A focal weight hyper-parameter γ smoothly adjusts the rate at which we down-weight the dummy background. Moreover, gradients do not propagate through the output of the backward background detector

²here ϕ_Φ^t denotes a separate convolutional filter per each step

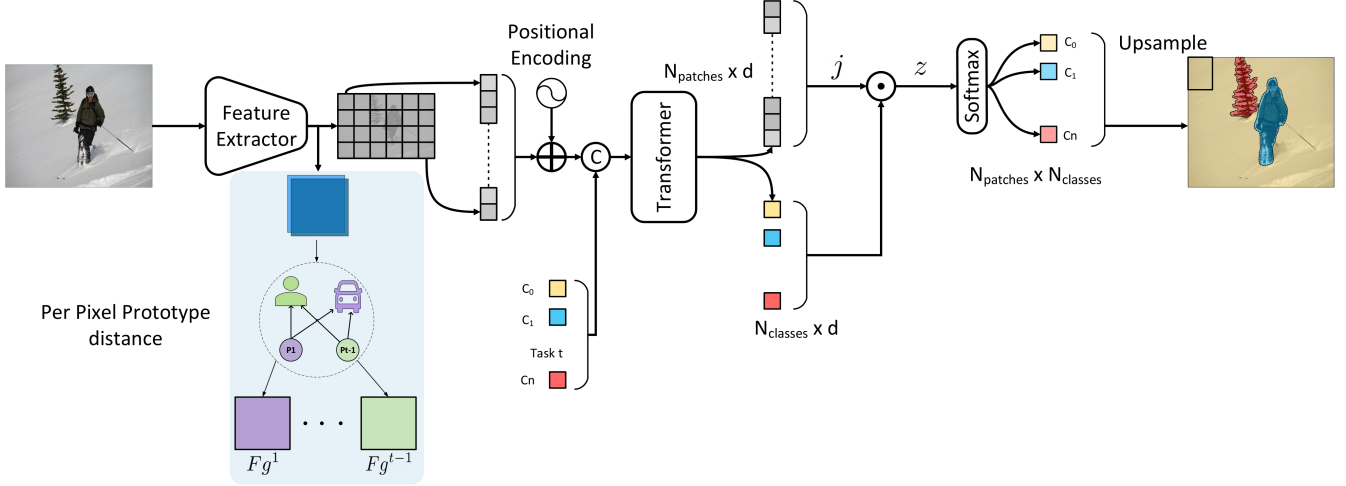


Figure 2: Our continual learning framework BACS consists of the backward background detector, shown in blue, and a transformer decoder. The backward background detector compares the latent space of each pixel with a per-step centroid to detect the foreground. The maximum output probability of all heads, $\max Fg^{1:t-1}$, is used to reduce the emphasis on pixels that belong to old classes collapsing to the background class in step t ground truth. Next, the transformer decoder allows the addition of new classes by initializing new class tokens.

using stop gradients to avoid backpropagating the dummy background to our projector. Our proposed loss consists of two parts; the first one learns to distinguish between the actual background class and the foreground while avoiding the dummy background class. The second is used to disentangle new classes from everything else.

$$\ell_{bg_fg}^{\theta^t}(x^t, y^t) = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (1 - \max_{\tilde{t} \in \{1:t-1\}} Fg^{\tilde{t}})^{\gamma} \log \hat{z}^t[i, y_i^t] \quad (1)$$

where \hat{z}^t contains the probability of either being a background or a foreground

$$\hat{z}^t[i, y_i^t] = \begin{cases} \sum_{k \in C^t} z^t[i, k] & \text{if } y_i^t \neq c_{bg} \\ z^t[i, y_i^t] & \text{if } y_i^t = c_{bg}. \end{cases} \quad (2)$$

Our intuition from eq. (1) is that the focal loss from our backward background detector will guide the training to distinguish between the actual background and everything else without affecting old classes. Furthermore, to simplify the learning, we compare it against the probability of being any of the other classes as shown in eq. (2).

The second part focuses on distinguishing new classes from everything else. We use the same unbiased cross-entropy loss [10] to simultaneously compare the new classes' probabilities to the sum of the probabilities of background and old classes. Moreover, the first part of the loss learns the actual background, which avoids the confusion between old classes and the actual background when we use the unbiased cross-entropy loss [10] alone.

$$\ell_{new}^{\theta^t}(x^t, y^t) = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \log \hat{z}^t[i, y_i^t], \quad (3)$$

where \hat{z}^t represents the probability of being any of the old classes or the background $\{c_{bg} \cup C^{1:t-1}\}$ versus new classes.

$$\hat{z}^t[i, y_i^t] = \begin{cases} \sum_{k \in C^{1:t-1}} z^t[i, k] & \text{if } y_i^t \in \{c_{bg} \cup C^{1:t-1}\} \\ z^t[i, y_i^t] & \text{if } y_i^t \in \{C^t\}. \end{cases} \quad (4)$$

Our loss function in subsequent steps used to learn new classes without pseudo-labeling becomes:

$$\ell_{BACS}^{\theta^t} = \ell_{new}^{\theta^t}(x^t, y^t) + \ell_{bg_fg}^{\theta^t}. \quad (5)$$

Our intuition is that we can update the model to predict the new classes and simultaneously manage to distinguish between dummy backgrounds belonging to different steps and the actual background. It is worth noting that direct pseudo-labeling underperforms due to the erroneous high-confidence predictions [15] that will keep propagating throughout the training affecting the network's plasticity. For that reason [14] uses a threshold and an entropy loss to decide which pseudo-label to keep. On the other hand, our proposed strategy uses the backward background shift detector to distinguish foreground and background while ignoring the dummy background in the loss function.

D. Decoder Multi-Classifer Initialization

We now described how we add new classes to our decoder without having to initialize new classifiers as background. Existing CSS methods initialize new classifiers as a

background class incentivizing the network to predict new classes as background. This initialization trick hinders the convergence of the network to new classes and stabilizes the background’s feature space at the expense of new classes. Furthermore, the multi-classifier setup tends to be specialized in the last set of classes, creating a bias towards new classes [77]. To address these two drawbacks of the initialization trick, we use a transformer decoder inspired by the Segmenter architecture [48], where a set of tokens represent our classes. First, we process class tokens, shown in colour in fig. 2, jointly with patch embeddings, shown in grey in fig. 2, through a transformer decoder layer. The output patches are passed through a dot-product operation to retrieve the final classification output. To add new classes, we append new class tokens to the input of the transformer decoder initialized as the average of all previous class tokens. Using class tokens makes adding new classes easier and reduces the bias toward new classes. Also, it improves the plasticity of our model with fewer parameters, as shown in fig. 2.

E. Catastrophic Forgetting

In addition to the background shift and the initialization of classifiers, we tackle catastrophic forgetting by using a masked knowledge distillation and Dark Experience Replay (DER) [62]. An effective method to tackle catastrophic forgetting is to set constraints using the previous step’s model weights f_{θ}^{t-1} . These constraints enforce f_{θ}^t to produce similar behavior on previously observed classes. A common constraint is to add soft knowledge distillation on the output layer [12], [65]. However, due to the forward background shift where the teacher f_{θ}^{t-1} considers new classes as a background, a representation space drift occurs, affecting the stability of the network. For that purpose, unbiased knowledge distillation [10] compares the output of the teacher for the background class with either being a background or a new class in f_{θ}^t . Moreover, regularizing output-level knowledge distillation would reduce the plasticity of the network and force it to rely on a large number of initial classes. We propose a feature-based masked knowledge distillation to constrain the feature space of old classes without limiting the plasticity of the network. In that proposed knowledge distillation, we distill the knowledge of the penultimate layer j^{t-1} for only a dummy background selected by our backward background detector. To summarize, our masked distillation loss with δ a tuned threshold becomes:

$$\ell_{kd}^{\theta^t}(x^t, y^t) = (\max_{\tilde{t} \in \{1:t-1\}} Fg^{\tilde{t}} > \delta) \cdot \|(j^{t-1})^2 - (j^t)^2\|. \quad (6)$$

Furthermore, we adapt dark experience replay [62] to mitigate catastrophic forgetting while ignoring the background class as it might contain old or future classes. Following [78], we use a balanced loss-aware reservoir sampling strategy. In brief, we make room for new examples by discarding

less-critical samples in the buffer. The importance score is computed based on the number of samples belonging to that same class in the buffer and the loss value of the trained network denoting its difficulty.

To summarize, our loss function becomes the sum of our background aware cross-entropy loss $\ell_{BACS}^{\theta^t}$, experience replay $\ell_{replay}^{\theta^t}$, dark knowledge replay $\ell_{der}^{\theta^t}$, penultimate layer knowledge distillation $\ell_{kd}^{\theta^t}$ and backward background detector training loss function $\ell_B^{\omega^t}$.

$$\ell^{\theta^t} = \underbrace{\ell_{BACS}^{\theta^t}}_{\text{classification}} + \underbrace{\alpha \ell_{der}^{\theta^t} + \beta \ell_{der++}^{\theta^t} + \kappa \ell_{kd}^{\theta^t}}_{\text{forgetting}} + \ell_B^{\omega^t} \quad (7)$$

In eq. (7), we have a set of hyper-parameters α , β , and κ tuned to guide the training to determine whether the focus is on stability or plasticity.

IV. EXPERIMENTS AND RESULTS

A. Setup and Datasets

For a fair comparison, we use the same setup, hyper-parameters, and datasets as [9]. However, we propose testing on more challenging setups where we start from a few classes (e.g., 5, 2).

We evaluate our proposed approach on the most challenging overlap mode with forward and backward background shifts. For Pascal-VOC, we evaluate several dataset setups, e.g., (15 – 1), which means we start our first task with 15 classes, then increment one class (for a total of 6 tasks). Furthermore, we evaluate more challenging setups, including (10 – 1), (5 – 3) and (2 – 1). Similarly, Cityscapes (14 – 1) means we start initially with 14 classes followed by increments of 1.

We evaluate our proposed method on two datasets: Pascal-VOC [79] and Cityscapes [80]. VOC contains an explicit background class and 20 classes, 10,582 training images, and 1,449 testing images. Cityscapes contain 19 classes and a set of unlabelled classes that we consider background taken from 21 cities. For all datasets, we use random resize, crop augmentation, and random horizontal flip during training following [9], [14]. The final image size for Pascal-VOC and Cityscapes is 512×512 . Hyper-parameters were tuned on a validation set created as a subset of the training set made of 20% of the images.

1) *Evaluation Metrics:* We use the mean Intersection-Over-Union (mIoU) as our evaluation metric. The IoU is defined as $\text{IoU} = \frac{\text{true positive}}{\text{true positive} + \text{false negative} + \text{false positive}}$, which quantifies the accuracy of our method. To evaluate our method’s ability to preserve previously learned information, we calculate the mIoU for the classes of the initial task. Additionally, We measure the IoU of newly added classes to denote the plasticity of the framework following [9] metrics.

Table I: Experimental results on challenging setups of Pascal VOC 2012. BACS outperforms existing baselines with a large margin on challenging setups where we start with a small number of classes and small increments.

	VOC 10-1 (11 tasks)			VOC 15-1 (6 tasks)			VOC 5-3 (6 tasks)			VOC 5-1 (16 tasks)			VOC 2-1 (19 tasks)		
Method	0-10	11-20	all	0-15	16-20	all	0-5	6-20	all	0-5	6-20	all	0-2	3-20	all
LwF-MC	4.65	5.90	4.95	6.40	8.40	6.90	20.91	36.67	24.66	N/A	N/A	N/A	N/A	N/A	N/A
ILT	7.15	3.67	5.50	8.75	7.99	8.56	22.51	31.66	29.04	N/A	N/A	N/A	N/A	N/A	N/A
MiB	12.25	13.09	12.65	34.22	13.50	29.29	52.2	42.1	45.01	11.47	9.45	10.03	21.57	7.93	9.88
PLOP	44.03	15.51	30.45	65.12	21.11	54.64	17.48	19.16	18.68	0.12	9.0	6.46	0.01	5.22	4.47
PLOP + UCD	42.3	28.3	35.3	66.3	21.5	55.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
PLOPLong	61.06	18.56	40.83	72.06	26.66	61.2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
ReCALL	59.5	46.7	54.8	65.7	47.8	62.7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
RCIL	55.4	15.1	34.3	70.6	23.7	59.4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
BACS	64.1	36.9	51.13	72	44.32	65.4	46.61	44.8	45.3	35.2	30.4	31.8	41.05	29.7	31.35

2) *Implementation Details*: In order to facilitate replication of our experimental setup, we have made available our implementation based on PyTorch Lightning, which can be accessed via a public GitHub repository³.

Following previous work [9], [14], we use a Resnet-101 backbone [5] pre-trained on ImageNet [81], and instead of using DeepLab v3 [6] decoding head we use segmenter [48] decoder with dynamic class tokens. We compared the number of parameters between our proposed decoder having 55M parameters with DeepLab v3 [6] decoder having 58M parameters on Pascal VOC 15 – 1 final step showing the effectiveness of our decoder with fewer parameters.

We optimize the network for all baselines using SGD with an initial learning rate of 10^{-2} reduced to 10^{-3} in subsequent tasks with momentum 0.9. We use the same learning rate schedule, data augmentation, and class order as [9], [14]. For the memory size, we use $|\mathcal{M}| = 300$ for both VOC and cityscapes. We train the network for 30 epochs with batch size 24 distributed on two GPUs for each task.

We select EWC [82], LwF-MC [65], and ILT [9] to compare general results on VOC with results exacerbated from SSUL [73]. We compare specific CSS experiments against our re-implementation for both MiB [10], PLOP [14], and joint training as an upper bound.

B. Quantitative and Qualitative Evaluation

1) *Pascal VOC 2012*: Quantitative results in table I compare our method with existing baselines on different scenarios, including the challenging 5 – 3, 10 – 1, 15 – 1 and the two tasks scenarios 15 – 5, 19 – 1 evaluated in the overlap mode having both backward and forward background shift. We show that our proposed framework outperforms both PLOP [14] and MiB [9] in terms of the plasticity on all tasks while retaining the knowledge acquired on old classes, especially in tasks starting with a small number

of classes. In both PLOP [14] and MiB [9], the pseudo-labeling and constrained teacher-student strategy limit the network’s plasticity resulting in degraded performance on new classes. In BACS, we mitigate this problem by discarding the pseudo-labeling strategy and simply relying on our backward background detector to solve the background shift problem.

2) *Cityscapes*: In table II, we also evaluate our method with the challenging Cityscapes setup having 19 classes with the unlabelled classes folded into the background class. We evaluate our method using a 14 – 1 setup. Here, BACS performs better than MiB and PLOP, specifically on new classes.

Table II: Continual Semantic Segmentation results on Cityscapes 14-1 in Mean IoU (%).

Method	14-1 (6 tasks)		
	1-14	15-19	all
MiB	56	15.7	46.5
PLOP	55.6	10	44.3
BACS	55.7	19.47	46.6

C. Impact of Class Ordering

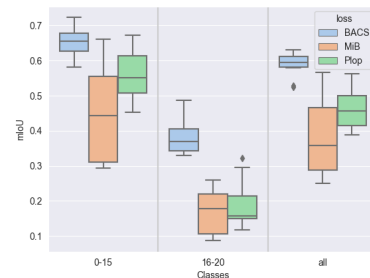


Figure 3: mIoU Evaluation of 10 different class orderings between BACS, MiB, and PLOP.

³<https://github.com/mostafaelaraby/BACS-Continual-Semantic-Segmentation>

Here, we analyze the effect of class ordering on the performance of various CSS frameworks. We perform experiments on ten different class orders on the challenging VOC 15 – 1 setup to analyze our robustness to the class order. In fig. 3, we display the mean and standard deviation of different class orders for various methods [10], [14]. Experimental results show the robustness of our method to class order in the challenging overlap mode.

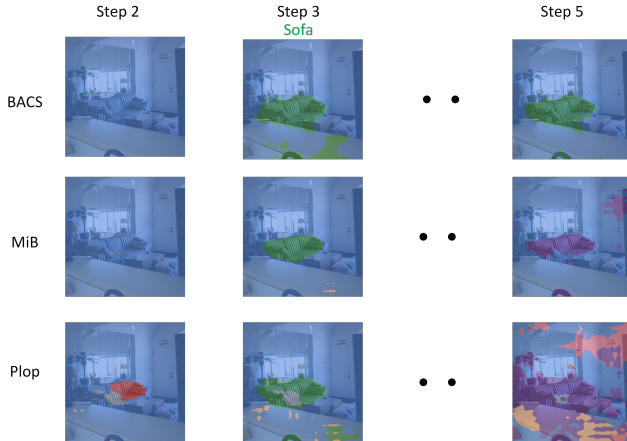


Figure 4: Qualitative comparison between BACS, MiB and PLOP on 15 – 1 VOC setup. **Left column:** Predictions after learning two tasks, not including the upcoming sofa class. **Middle column:** Predictions after incrementing the sofa class. **Right column:** Predictions at the end of the training.

D. Ablations

Table III: Ablation study for BACS on VOC 15-1.

Configurations				15-1 (6 tasks)		
BACS	DER	MKD	Dec	0-15	16-20	all
✓	✓	✓	✓	72	44.3	65.4
✓	✓	✗	✓	68.8	29.7	59.5
✓	✓	✗	✗	58.82	40.82	54.54
✓	✗	✓	✓	55	7	43.8

Here, we analyze the effect of each proposed component of BACS on VOC 15 – 1 in the overlap mode. table III compares the results of each ablation case where we disable one or two of the components. The first row shows the result of BACS with all the components, including dark experience replay (DER), masked knowledge distillation(MKD), and our transformer decoder (Transformer Decoder (Dec)). When we disable MKD in the second row, the performance on both old and new classes deteriorates, showing the benefits of soft knowledge distillation for both plasticity and stability. On the other hand, using the multi-classifier setup instead of Dec improves the plasticity due to the

increased number of parameters and fails to reduce the effect of the forward background shift due to the initialization trick used. In Dec, we add a single parameter per new class, whereas the multi-classifier setup initializes a new 1×1 convolutional head. Finally, in the last row, we show the effect of using DER on stabilizing the training regarding stability and plasticity.

E. Initialization of New Tokens

Table IV: Initialization of new tokens corresponding to newly added classes. Random where we randomly initialize a new token per class, background means initializing new tokens as the background class and Mean, which averages all the old tokens to generate new ones.

Token Initialization	VOC 15-1 (6 tasks)		
	0-15	16-20	all
Random	67.1	30	58.51
Background	68.3	33.83	60
Mean	72	44.3	65.4

In table IV, we show that Transformer Decoder (Dec) is less sensitive to the way we initialize our tokens compared to the multi-classifier setup sensitivity discussed in [10]. In the first row, we show the effect of random initialization of new tokens, which slightly perturb the feature space due to the forward background shift. Next, we show the effect of initializing as a background similar to the initialization trick [10] but this time applied on a single parameter token. Finally, we compare our proposed initialization of tokens that reduces the feature space perturbation by simply taking the mean of all previous class tokens. We empirically show the effect of different token initialization on the stability of the network throughout the training.

V. CONCLUSION

In this work, We proposed a new framework Background Aware Continual Semantic Segmentation (BACS) to address three principal challenges in Continual Semantic Segmentation (CSS): catastrophic forgetting, background shift, and the initialization of new class heads. We proposed four main contributions, including a backward background shift detector, a variation of cross-entropy loss, a masked feature-based Masked Knowledge Distillation (MKD) and finally, a transformer decoder. These main contributions demonstrate significant improvements in handling a large number of tasks, particularly when starting from a small set of classes. Our experiments confirm that each component contributes substantially to the overall effectiveness of the model, outperforming existing methods across standard benchmarks.

REFERENCES

- [1] A. Robins, “Catastrophic forgetting, rehearsal and pseudorehearsal,” *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

- [2] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.
- [3] S. Thrun, "Lifelong Learning Algorithms," in *Learning to Learn*, S. Thrun and L. Y. Pratt, Eds. Springer, 1998, pp. 181–209. [Online]. Available: https://doi.org/10.1007/978-1-4615-5529-2_8
- [4] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation," *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: <http://arxiv.org/abs/1802.02611>
- [5] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. J. Smola, "ResNeSt: Split-attention Networks," *CoRR*, vol. abs/2004.08955, 2020. [Online]. Available: <https://arxiv.org/abs/2004.08955>
- [6] S. C. Yurtkulu, Y. H. Sahin, and G. B. Ünal, "Semantic Segmentation with Extended DeepLabv3 Architecture," in *27th Signal Processing and Communications Applications Conference, SIU 2019, Sivas, Turkey, April 24-26, 2019*. IEEE, 2019, pp. 1–4. [Online]. Available: <https://doi.org/10.1109/SIU.2019.8806244>
- [7] F. Özdemir, P. Furstahl, and O. Göksel, "Learn the new, keep the old: Extending pretrained models with new anatomy and images," *CoRR*, vol. abs/1806.00265, 2018. [Online]. Available: <http://arxiv.org/abs/1806.00265>
- [8] F. Özdemir and O. Göksel, "Extending Pretrained Segmentation Networks with Additional Anatomical Structures," *CoRR*, vol. abs/1811.04634, 2018. [Online]. Available: <http://arxiv.org/abs/1811.04634>
- [9] U. Michieli and P. Zanuttigh, "Incremental Learning Techniques for Semantic Segmentation," *CoRR*, vol. abs/1907.13372, 2019. [Online]. Available: <http://arxiv.org/abs/1907.13372>
- [10] F. Cermelli, M. Mancini, S. R. Bulò, E. Ricci, and B. Caputo, "Modeling the Background for Incremental Learning in Semantic Segmentation," *CoRR*, vol. abs/2002.00718, 2020. [Online]. Available: <https://arxiv.org/abs/2002.00718>
- [11] J. G. Zilly, A. Achille, A. Censi, and E. Frazzoli, "On Plasticity, Invariance, and Mutually Frozen Weights in Sequential Task Learning," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 12 386–12 399. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/hash/6738fc33dd0b3906cd3626397cd247a7-Abstract.html>
- [12] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: <http://arxiv.org/abs/1503.02531>
- [13] A. Douillard, Y. Chen, A. Dapogny, and M. Cord, "Tackling Catastrophic Forgetting and Background Shift in Continual Semantic Segmentation," *CoRR*, vol. abs/2106.15287, 2021. [Online]. Available: <https://arxiv.org/abs/2106.15287>
- [14] —, "PLOP: Learning without Forgetting for Continual Semantic Segmentation," *CoRR*, vol. abs/2011.11390, 2020. [Online]. Available: <https://arxiv.org/abs/2011.11390>
- [15] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and Confirmation Bias in Deep Semi-supervised Learning," *CoRR*, vol. abs/1908.02983, 2019. [Online]. Available: <http://arxiv.org/abs/1908.02983>
- [16] A. Maracani, U. Michieli, M. Toldo, and P. Zanuttigh, "RECALL: Replay-based Continual Learning in Semantic Segmentation," *CoRR*, vol. abs/2108.03673, 2021. [Online]. Available: <https://arxiv.org/abs/2108.03673>
- [17] U. Michieli and P. Zanuttigh, "Continual Semantic Segmentation via Repulsion-attraction of Sparse and Disentangled Latent Representations," *CoRR*, vol. abs/2103.06342, 2021. [Online]. Available: <https://arxiv.org/abs/2103.06342>
- [18] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," *CoRR*, vol. abs/1210.5644, 2012. [Online]. Available: <http://arxiv.org/abs/1210.5644>
- [19] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional Random Fields as Recurrent Neural Networks," *CoRR*, vol. abs/1502.03240, 2015. [Online]. Available: <http://arxiv.org/abs/1502.03240>
- [20] A. Arnab, S. Jayasumana, S. Zheng, and P. H. S. Torr, "Higher Order Conditional Random Fields in Deep Neural Networks," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 524–540. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_33
- [21] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: <http://arxiv.org/abs/1411.4038>
- [22] M. Goyal and M. H. Yap, "Multi-class Semantic Segmentation of Skin Lesions via Fully Convolutional Networks," *CoRR*, vol. abs/1711.10449, 2017. [Online]. Available: <http://arxiv.org/abs/1711.10449>
- [23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation," *CoRR*, vol. abs/1511.00561, 2015. [Online]. Available: <http://arxiv.org/abs/1511.00561>
- [25] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik, "Hypercolumns for Object Segmentation and Fine-grained Localization," *CoRR*, vol. abs/1411.5752, 2014. [Online]. Available: <http://arxiv.org/abs/1411.5752>

- [26] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale Context Intertwining for Semantic Segmentation," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11207. Springer, 2018, pp. 622–638. [Online]. Available: https://doi.org/10.1007/978-3-030-01219-9_37
- [27] G. Lin, A. Milan, C. Shen, and I. D. Reid, "RefineNet: Multi-path Refinement Networks for High-resolution Semantic Segmentation," *CoRR*, vol. abs/1611.06612, 2016. [Online]. Available: <http://arxiv.org/abs/1611.06612>
- [28] H. Noh, S. Hong, and B. Han, "Learning Deconvolution Network for Semantic Segmentation," *CoRR*, vol. abs/1505.04366, 2015. [Online]. Available: <http://arxiv.org/abs/1505.04366>
- [29] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network," *CoRR*, vol. abs/1703.02719, 2017. [Online]. Available: <http://arxiv.org/abs/1703.02719>
- [30] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders Matter for Semantic Segmentation: Data-dependent Decoding Enables Flexible Feature Aggregation," *CoRR*, vol. abs/1903.02120, 2019. [Online]. Available: <http://arxiv.org/abs/1903.02120>
- [31] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to Scale: Scale-aware Semantic Image Segmentation," *CoRR*, vol. abs/1511.03339, 2015. [Online]. Available: <http://arxiv.org/abs/1511.03339>
- [32] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context Contrast Feature and Gated Multi-scale Aggregation for Scene Segmentation," 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. 2393–2402. [Online]. Available: http://openaccess.thecvf.com/content/_cvpr/_2018/html/Ding/_Context/_Contrasted/_Feature/_CVPR/_2018/_paper.html
- [33] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual Attention Network for Scene Segmentation," *CoRR*, vol. abs/1809.02983, 2018. [Online]. Available: <http://arxiv.org/abs/1809.02983>
- [34] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization Attention Networks for Semantic Segmentation," *CoRR*, vol. abs/1907.13426, 2019. [Online]. Available: <http://arxiv.org/abs/1907.13426>
- [35] Z. Niu, W. Liu, J. Zhao, and G. Jiang, "DeepLab-based Spatial Feature Extraction for Hyperspectral Image Classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 16, no. 2, pp. 251–255, 2019. [Online]. Available: <https://doi.org/10.1109/LGRS.2018.2871507>
- [36] S. Mehta, M. Rastegari, A. Caspi, L. G. Shapiro, and H. Hajishirzi, "ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation," *CoRR*, vol. abs/1803.06815, 2018. [Online]. Available: <http://arxiv.org/abs/1803.06815>
- [37] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for Semantic Segmentation in Street Scenes," 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. 3684–3692. [Online]. Available: http://openaccess.thecvf.com/content/_cvpr/_2018/html/Yang/_DenseASPP/_for/_Semantic/_CVPR/_2018/_paper.html
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [40] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end Video Instance Segmentation with Transformers," *CoRR*, vol. abs/2011.14503, 2020. [Online]. Available: <https://arxiv.org/abs/2011.14503>
- [41] Y. Zeng, J. Fu, and H. Chao, "Learning Joint Spatial-temporal Transformations for Video Inpainting," *CoRR*, vol. abs/2007.10247, 2020. [Online]. Available: <https://arxiv.org/abs/2007.10247>
- [42] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking Semantic Segmentation from a Sequence-to-sequence Perspective with Transformers," *CoRR*, vol. abs/2012.15840, 2020. [Online]. Available: <https://arxiv.org/abs/2012.15840>
- [43] C. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention Multi-scale Vision Transformer for Image Classification," *CoRR*, vol. abs/2103.14899, 2021. [Online]. Available: <https://arxiv.org/abs/2103.14899>
- [44] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," *CoRR*, vol. abs/2102.12122, 2021. [Online]. Available: <https://arxiv.org/abs/2102.12122>
- [45] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *CoRR*, vol. abs/2105.15203, 2021. [Online]. Available: <https://arxiv.org/abs/2105.15203>
- [46] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature Pyramid Transformer," *CoRR*, vol. abs/2007.09451, 2020. [Online]. Available: <https://arxiv.org/abs/2007.09451>
- [47] 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," *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: <https://arxiv.org/abs/2103.14030>

- [48] R. Strudel, R. G. Pinel, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," *CoRR*, vol. abs/2105.05633, 2021. [Online]. Available: <https://arxiv.org/abs/2105.05633>
- [49] Z. Chen and B. Liu, *Lifelong Machine Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2016. [Online]. Available: <https://doi.org/10.2200/S00737ED1V01Y201610AIM033>
- [50] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert Gate: Lifelong Learning with a Network of Experts," *CoRR*, vol. abs/1611.06194, 2016. [Online]. Available: <http://arxiv.org/abs/1611.06194>
- [51] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient Lifelong Learning with A-GEM," *CoRR*, vol. abs/1812.00420, 2018. [Online]. Available: <http://arxiv.org/abs/1812.00420>
- [52] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual Lifelong Learning with Neural Networks: A Review," *CoRR*, vol. abs/1802.07569, 2018. [Online]. Available: <http://arxiv.org/abs/1802.07569>
- [53] R. Aljundi, M. Rohrbach, and T. Tuytelaars, "Selfless Sequential Learning," *CoRR*, vol. abs/1806.05421, 2018. [Online]. Available: <http://arxiv.org/abs/1806.05421>
- [54] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.
- [55] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual Learning with Deep Generative Replay," *CoRR*, vol. abs/1705.08690, 2017. [Online]. Available: <http://arxiv.org/abs/1705.08690>
- [56] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory Aware Synapses: Learning what (not) to forget," *CoRR*, vol. abs/1711.09601, 2017. [Online]. Available: <http://arxiv.org/abs/1711.09601>
- [57] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. S. Torr, "Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence," *CoRR*, vol. abs/1801.10112, 2018. [Online]. Available: <http://arxiv.org/abs/1801.10112>
- [58] A. Gepperth and C. Karaoguz, "A Bio-inspired Incremental Learning Architecture for Applied Perceptual Problems," *Cogn. Comput.*, vol. 8, no. 5, pp. 924–934, 2016. [Online]. Available: <https://doi.org/10.1007/s12559-016-9389-5>
- [59] S. Rebuffi, A. Kolesnikov, and C. H. Lampert, "iCaRL: Incremental Classifier and Representation Learning," *CoRR*, vol. abs/1611.07725, 2016. [Online]. Available: <http://arxiv.org/abs/1611.07725>
- [60] V. Dobson and R. Gregory, "Reviews: studies of mind and brain: neural principles of learning, perception, development, cognition and motor control, organization in vision: essays on gestalt perception," 1984.
- [61] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end Incremental Learning," *CoRR*, vol. abs/1807.09536, 2018. [Online]. Available: <http://arxiv.org/abs/1807.09536>
- [62] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark Experience for General Continual Learning: a Strong, Simple Baseline," *CoRR*, vol. abs/2004.07211, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07211>
- [63] D. Lopez-Paz and M. Ranzato, "Gradient Episodic Memory for Continual Learning," 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. 6467–6476. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddeb-Abstract.html>
- [64] D. L. Silver and R. E. Mercer, "The task rehearsal method of life-long learning: Overcoming impoverished data," in *Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 2002, pp. 90–101.
- [65] Z. Li and D. Hoiem, "Learning without Forgetting," *CoRR*, vol. abs/1606.09282, 2016. [Online]. Available: <http://arxiv.org/abs/1606.09282>
- [66] A. R. Triki, R. Aljundi, M. B. Blaschko, and T. Tuytelaars, "Encoder Based Lifelong Learning," *CoRR*, vol. abs/1704.01920, 2017. [Online]. Available: <http://arxiv.org/abs/1704.01920>
- [67] D. Baek, Y. Oh, S. Lee, J. Lee, and B. Ham, "Decomposed Knowledge Distillation for Class-incremental Semantic Segmentation," *CoRR*, vol. abs/2210.05941, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2210.05941>
- [68] M. Phan, T. Ta, S. L. Phung, L. Tran-Thanh, and A. Bouzerdoum, "Class Similarity Weighted Knowledge Distillation for Continual Semantic Segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 16 845–16 854. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01636>
- [69] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, M. Nabi, X. Alameda-Pineda, and E. Ricci, "Uncertainty-aware Contrastive Distillation for Incremental Semantic Segmentation," *CoRR*, vol. abs/2203.14098, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2203.14098>
- [70] C. Zhang, J. Xiao, X. Liu, Y. Chen, and M. Cheng, "Representation Compensation Networks for Continual Semantic Segmentation," *CoRR*, vol. abs/2203.05402, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2203.05402>
- [71] Z. Huang, W. Hao, X. Wang, M. Tao, J. Huang, W. Liu, and X. Hua, "Half-real Half-fake Distillation for Class-incremental Semantic Segmentation," *CoRR*, vol. abs/2104.00875, 2021. [Online]. Available: <https://arxiv.org/abs/2104.00875>
- [72] S. Yan, J. Zhou, J. Xie, S. Zhang, and X. He, "An EM Framework for Online Incremental Learning of Semantic Segmentation," *CoRR*, vol. abs/2108.03613, 2021. [Online]. Available: <https://arxiv.org/abs/2108.03613>

- [73] S. Cha, B. Kim, Y. Yoo, and T. Moon, "SSUL: Semantic Segmentation with Unknown Label for Exemplar-based Class-incremental Learning," *CoRR*, vol. abs/2106.11562, 2021. [Online]. Available: <https://arxiv.org/abs/2106.11562>
- [74] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015, p. 0.
- [75] V. V. Ramasesh, E. Dyer, and M. Raghu, "Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics," *CoRR*, vol. abs/2007.07400, 2020. [Online]. Available: <https://arxiv.org/abs/2007.07400>
- [76] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *CoRR*, vol. abs/1708.02002, 2017. [Online]. Available: <http://arxiv.org/abs/1708.02002>
- [77] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a Unified Classifier Incrementally via Rebalancing," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 831–839. [Online]. Available: http://openaccess.thecvf.com/content/_CVPR/_2019/html/Hou/_Learning/_a/_Unified/_Classifier/_Incrementally/_via/_Rebalancing/_CVPR/_2019/_paper.html
- [78] P. Buzzega, M. Boschini, A. Porrello, and S. Calderara, "Rethinking Experience Replay: a Bag of Tricks for Continual Learning," *CoRR*, vol. abs/2010.05595, 2020. [Online]. Available: <https://arxiv.org/abs/2010.05595>
- [79] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015. [Online]. Available: <https://doi.org/10.1007/s11263-014-0733-5>
- [80] 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," *CoRR*, vol. abs/1604.01685, 2016. [Online]. Available: <http://arxiv.org/abs/1604.01685>
- [81] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-level Performance on ImageNet Classification," *CoRR*, vol. abs/1502.01852, 2015. [Online]. Available: <http://arxiv.org/abs/1502.01852>
- [82] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *CoRR*, vol. abs/1612.00796, 2016. [Online]. Available: <http://arxiv.org/abs/1612.00796>