IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 August 2023, accepted 6 September 2023, date of publication 14 September 2023,
date of current version 19 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3315327

RESEARCH ARTICLE

Class-Prompting Transformer for Incremental Semantic Segmentation

ZICHEN SONG, (Student Member, IEEE), ZHAOFENG SHI, CHAO SHANG, (Member, IEEE), FANMAN MENG, (Member, IEEE), AND LINFENG XU, (Member, IEEE)
School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: Zichen Song (szc@std.uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62071086 and Grant 62271119, and in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2021YGLH210.

ABSTRACT Class-incremental Semantic Segmentation (CISS) aims to learn new tasks sequentially that assign a specific category to each pixel of a given image while preserving the original capability to segment the old classes even if the labels of old tasks are absent. Most existing CISS methods suppress catastrophic forgetting by directly distilling on specific layers, which ignores the semantic gap between training data from the old and new classes with different distributions and leads to distillation errors, thus affecting segmentation performance. In this paper, we propose a Class-prompting Transformer (CPT) to introduce external prior knowledge provided by a pre-trained vision-language encoder into CISS pipelines for bridging the old and new classes and performing more generalized initialization and distillation. Specifically, we proposed a Prompt-guided Initialization Module (PIM), which measures the relationships between the class prompts and old query parameters to initialize the new query parameters for relocating the previous knowledge to the learning of new tasks. Then, a Semantic-aligned Distillation Module (SDM) is proposed to incorporate class prompt information with the class-aware embeddings extracted from the decoder to prevent the semantic gap problem between distinct class data and conduct adaptive knowledge transfer to suppress catastrophic forgetting. Extensive experiments on Pascal VOC and ADE20K datasets for the CISS task demonstrate the superiority of the proposed method, which achieves state-of-the-art performance.

INDEX TERMS Incremental semantic segmentation, knowledge distillation, class prompt learning.

I. INTRODUCTION

Semantic segmentation [1], [2], [3] is a fundamental computer vision task that aims to classify each pixel of the given image and assign the corresponding class label. Despite the remarkable achievement in the traditional semantic segmentation field where samples for all classes are available, continuous learning of data streams in real-world scenarios [4], [5] remains challenging. In recent years, researchers are widely interested in Class-incremental Semantic Segmentation (CISS), which means the semantic segmentation model has to learn newly emerged classes incrementally while preserving its capabilities of segmenting object(s) of old classes without any old examples. As a representative incremental learning task, the main challenge of CISS lies in preventing catastrophic forgetting that represents a significant degradation in the performance of the previous tasks.

An intuitive approach is directly fine-tuning [6] the trained old model on the new task to adjust parameters for fitting the distribution of new data, which causes a dramatic degradation of the model's performance on the old tasks, i.e. catastrophic forgetting [7], [8]. To alleviate this problem, various knowledge distillation-based methods [9], [10], [11], [12], [13] are carried out to preserve the original capability by transferring knowledge of the old classes to the new model. Despite the knowledge distillation-based methods have contributed significantly to the development of CISS, two common limitations still exist. On the one hand, in most of the methods, the newly input class parameters are randomly initialized. However, the learning of new tasks relies heavily on the originally learned knowledge of the model during the class incremental learning process. Simple random initialization strategies cannot efficiently transfer

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyu Zhou.

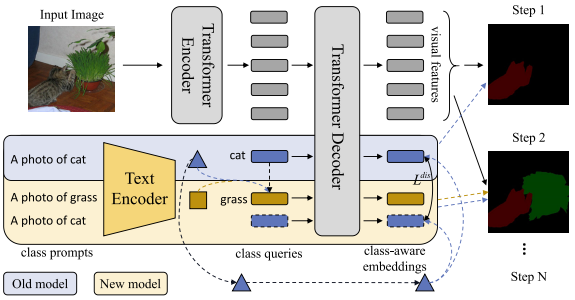

FIGURE 1. Illustration of the proposed Class-prompting Transformer (CPT), the class prompts are introduced into the initialization of class queries and distillation of class-aware embeddings of the class-incremental semantic segmentation (CISS) framework.

the learned knowledge for the learning of the next step. On the other hand, existing CISS methods typically transfer knowledge by distilling the visual features from single or multiple layers of the segmentation framework. Nevertheless, the different distributions between the old and new class data lead to the semantic gap, which causes errors during the plain distillation methods without prior semantic knowledge.

To overcome the aforementioned problems, we propose a novel Class-prompting Transformer (CPT) for the class-incremental semantic segmentation task. As shown in Fig. 1, different from other Transformer-based CISS methods [14], [15], our CPT leverages external prior knowledge provided by the large-scale vision-language pre-trained model. Specifically, the class prompts are fed into the CLIP [16] to obtain the semantic guidance, which is injected into the CISS framework for bridging the semantic gap between different classes and performing more generalized initialization and distillation. Firstly, in terms of the initialization, we propose a Prompt-guided Initialization Module (PIM), which measures the relationship between prompts generated by the text encoder of CLIP corresponding to the old and new classes for guiding the initialization of the newly emerged class queries. Such a prompt-based initialization strategy transfers the learned knowledge of the old CISS model to the new model for facilitating the learning of the new classes. Next, for the distillation processing, the proposed Semantic-aligned Distillation Module (SDM) integrates the old class-aware embeddings generated by the Transformer decoder with prompt-based semantic guidance for filling the semantic gap between distinct class data. With the help of external prior knowledge provided by the pre-trained model, we bridge the semantic gap and perform knowledge distillation between the old and new class-aware embeddings to prevent the model from catastrophic forgetting.

The major contributions of this paper are concluded as follows:

- The Prompt-guided Initialization Module (PIM) is proposed to combine the class prompts and old query features and injects them into the initialization procedure of the new queries to relocate the learned knowledge to the new task learning processing.
- We design a novel Semantic-aligned Distillation Module (SDM), which computes correlations between the semantic guidance and the old and new class-aware embeddings, respectively. The prior semantic information based on class prompts prevents the model from forgetting the previously learned knowledge.
- We propose a Class-prompting Transformer (CPT) for the CISS task, whose key modules are SDM and PIM. Extensive quantitative and qualitative experiments on Pascal VOC and ADE20K datasets show that our CPT outperforms other methods and achieves state-of-the-art performance.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

In recent years, researchers have proposed multiple remarkable methods for semantic segmentation task, which aims to assign semantic categories to each pixel within the given image. Benefit from high-quality semantic segmentation datasets [17], [18], [19], [20] with pixel-level annotations, several deep architectures [1], [2], [21], [22], [23] have been designed and achieve significant performance. FCN [1] is an end-to-end fully convolutional network, which predicts pixel-level segmentation masks. Deeplab series [2], [24], [25], [26] introduce atrous convolution layers to enlarge the visual receptive field and capture more comprehensive contexts. Some researchers attempt to adopt Encoder-Decoder structures [21], [26], [27], [28] for semantic segmentation for capturing abundant spatial information of the given image. With the emergence and development of the attention mechanism, several works [29], [30], [31], [32], [33] try to integrate various attention-based modules for extracting dense visual correlations between image patches. Recently, several Transformer-based methods [3], [23], [34] have been proposed to facilitate the capture of long-range dependencies and yield unprecedented segmentation precision.

Despite the remarkable achievement of semantic segmentation, it requires the whole dataset including images of all the pre-defined categories. Traditional semantic segmentation methods are not capable of incrementally learning the newly emerged classes, which is common in real-world scenarios.

B. INCREMENTAL LEARNING

Incremental learning refers to continually learning new class data while preserving the original knowledge learned previously. The pioneering work is LwF [6], which proposes an incremental learning method without catastrophic forgetting and conducts extensive experiments. In follow-up studies, researchers try various incremental learning strategies including replay-based methods [35], [36], [37], architecture-based methods [38], [39], [40], and regularization-based methods [41], [42], [43]. Replay-based incremental learning means retaining the exemplars of the old classes in the new task learning. Old exemplars can be divided into three forms, i.e. raw data [35], [37], synthetic data [36], [44], and prototype representations [45]. The introduction of old class samples in the learning of new tasks facilitates the model's retention of prior knowledge. Architecture-based methods [38], [39], [40] mean dynamically changing the

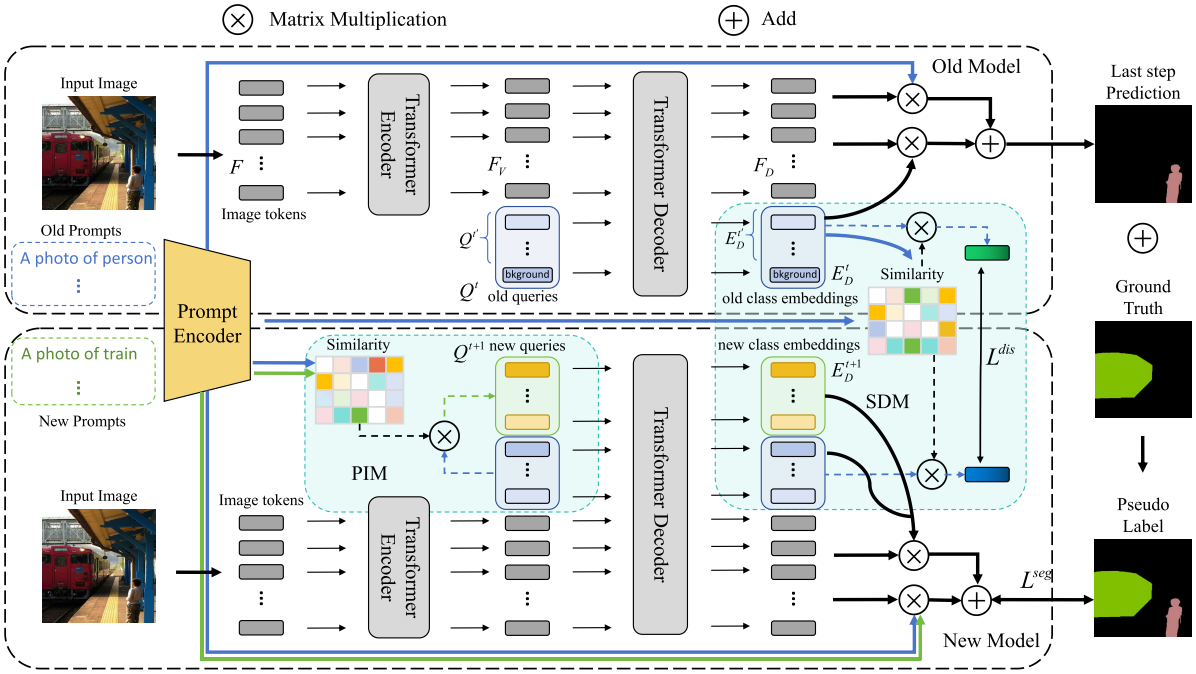

FIGURE 2. The overall architecture of the proposed CPT. The input image is first converted into visual tokens and fed into the Transformer encoder to obtain the visual features, and we input the pre-defined prompts into the pre-trained text encoder to get the prompt features. Then, we combine the encoded visual features with class queries, which are initialized by the Prompt-guided Initialization Module (PIM) for transferring the previous knowledge into the learning of the new task. Next, we feed the combined features into the Transformer decoder and get the decoded visual features and class-aware embeddings. A Semantic-aligned Distillation Module (SDM) is applied to build the semantic-level linkage between the old and new models by incorporating prompt guidance with class-aware embeddings. Finally, we use the prediction of the old model and the current ground truth to get the pseudo label for supervising the training of the new model.

structure of the network to learn new knowledge while preserving the original capabilities. Regularization-based methods perform regularization to the parameters of specific layers to prevent bias in the network parameters and maintain the previously learned knowledge. The specific operation includes knowledge distillation [46], [47], adversarial learning [48], and vanilla regularization methods [41], [49].

With the development of the computer vision field, incremental learning technology has been applied to multiple complicated visual tasks such as object detection, semantic segmentation, and instance segmentation.

C. CLASS-INCREMENTAL SEMANTIC SEGMENTATION

Class-incremental Semantic Segmentation (CISS) is a more challenging computer vision task with the difficulty of preventing catastrophic forgetting. ILT [9] is the pioneering work that proposes the CISS task and designs a CISS baseline based on the DeepLab [2] framework. MiB [10] draws out the problem of the semantic shift of background and proposes an unbiased distillation method to avoid forgetting the learned knowledge. SDR [12] utilizes a prototype learning strategy to improve the model's learning on the new tasks and prevent forgetting. RECALL [50] is a replay-based method that introduces real or generated additional data. Distllation-based methods PLOP [11] and REMINDER [51] suppress forgetting by distillating features from multiple scales or assigning weights during distillation. For the architecture-based method, RCN [52] designs a two-stream network architecture to accommodate both old and new knowledge. In recent years, RBC [53] points out that context in the CISS task is important and decouples different classes through context-rectified image-duplet learning. SPPA [54] alleviates forgetting by measuring and constraining inter-class and intra-class relationships. Incrementer [15] proposes a full Transformer framework and designs brand-new distillation and class de-confusion strategies.

Nevertheless, the aforementioned methods are difficult to extract generalized information between the old and new task data due to the absence of external semantic-level guidance. We propose a Class-prompting Transformer (CPT) to inject semantic prompts into the CISS framework.

III. METHOD

The overall framework of the proposed Class-prompting Transformer (CPT) is illustrated in Fig. 2. We first input the given image into the trained old model in the last step to get the segmentation masks of the old classes. Then, using a pre-trained text encoder of CLIP [16] to convert the old and new prompts into semantic guidance features and measure the relationships between them, and re-weight the learned classes queries to initialize the visual queries of the new model via the Prompt-guided Initialization Module (PIM) for transferring the previous knowledge into the newly emerged learning step.

Next, the Semantic-aligned Distillation Module (SDM) is utilized to fill the semantic gap between different class data by introducing the correlations between the old class prompts and old class-aware embeddings to suppress the errors during distillation processing. Finally, a cross-entropy loss and a distillation loss are adopted as the supervision during the training process. In the following sections, the basic settings and preliminaries of CISS are introduced in section III-A, the overall Transformer-based framework is illustrated in section III-B, and the proposed PIM and SDM are detailed presented in section III-C and section III-D, respectively.

A. PRELIMINARIES

Different from the semantic segmentation task with the annotations of all the classes available, Class-incremental Semantic Segmentation (CISS) aims to continuously learn the newly emerged classes and divide the whole dataset into multiple subsets. The dataset is divided based on class and each of the subsets contains a portion of class labels. There are no intersections between subsets and the CISS model learns one subset at each learning step without the previous data.

We assume there are T learning steps and the whole dataset is split into T subsets, which are denotes as $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^T\}$, and the corresponding classes are denoted as $\mathcal{C} = \{\mathcal{C}^1, \mathcal{C}^2, \ldots, \mathcal{C}^T\}$. The sample pairs of the subset $\mathcal{D}^t$ are represented as $\{\mathcal{I}_i^t, \mathcal{Y}_i^t\}$, where $\mathcal{I}_i^t \in \mathbb{R}^{3\times H\times W}$ is the *i-th* input image and $\mathcal{Y}_i^t$ is the segmentation annotation. $|\mathcal{D}^t|$ denotes the number of samples of the subset $\mathcal{D}^t$ and $|\mathcal{C}^t|$ means the number of classes contained in the subset $\mathcal{D}^t$. During the *t-th* learning step, only the data of subset $\mathcal{D}^t$ and classes of $\mathcal{C}^t$ are available, where the labels of previous steps $\mathcal{C}^{1:t-1}$ are not contained. Due to the disjoint nature of the subsets of different learning steps, thus $\mathcal{C}^{1:t-1} \cap \mathcal{C}^t = \emptyset$. The labels of $\mathcal{C}^t$ and samples of $\mathcal{D}^t$ are unavailable after the *t-th* learning step and the CISS model will fit the data of $\mathcal{D}^{t+1}$ and forget the knowledge of old classes, i.e. catastrophic forgetting.

To address the problem, we propose a novel Class-prompting Transformer (CPT), which takes the human-defined class prompt as the semantic guidance and introduces it into a Transformer-based CISS framework for more generalized initialization and distillation.

B. CLASS-PROMPTING TRANSFORMER FRAMEWORK

In this section, we introduce the basic Transformer-based structure of the proposed method.

Our Transformer-based method consists of an encoder network and a decoder network, which are denoted as *Enc* and *Dec*, respectively. As with ViT [55], we first divide the given input image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$ into several patches. The size of each patch is set to $P \times P$ and the number of patches is $N_p = HW/P^2$. Then, each patch is flattened and projected into representational vector by a linear layer. The obtained feature vectors are denoted as $F = \{f_1, f_2, \ldots, f_{N_p}\} \in \mathbb{R}^{N_p\times D}$, where f_i denotes the feature of *i-th* image patch. Next, we feed the feature vectors into the vision Transformer encoder, which consists of multiple self-attention layers. The deep Transformer encoder extracts abundant long-range dependencies and contextual information between image patches and outputs the representational visual feature sequence, which is denoted as $F_V = \{f_{v,1}, f_{v,2}, \ldots, f_{v,N_p}\} \in \mathbb{R}^{N_p\times D}$. The encoding process can be formulated as:

$$F_V = Enc(F) \tag{1}$$

For the Transformer decoder, which takes the visual features $F_V = \{f_{v,1}, f_{v,2}, \ldots, f_{v,N_p}\} \in \mathbb{R}^{N_p\times D}$ extracted by the encoder as one of the input. Inspired by the ideology of the pioneering method [3], [15], we add a series of extra learnable queries to represent each introduced class. However, unlike the previous work where the class queries are randomly initialized, we develop a novel Prompt-guided Initialization Module (PIM) to incorporate the class prompts and old queries into the initialization of the new queries for transferring the original knowledge to the learning of new tasks. The class query sequence are denoted as $Q = \{q_0, q_1, \ldots, q_M\} \in \mathbb{R}^{(M+1)\times D}$, where M denotes the number of classes and $M = |\mathcal{C}^{1:t}|$. q_0 represents the query of the background class. We concatenate the visual features and class queries and feed them into the decoder to obtain the decoded visual features $F_D = \{f_{d,1}, f_{d,2}, \ldots, f_{d,N_p}\} \in \mathbb{R}^{N_p\times D}$ and class-aware embeddings $E = \{e_0, e_1, \ldots, e_M\} \in \mathbb{R}^{(M+1)\times D}$. The decoder framework can be formulated as:

$$[F_D; E_D] = Dec(F_V; E) \tag{2}$$

Finally, we compute the similarity between the decoded visual features F^D and the decoded class-aware embedding $E_D = \{e_{d,0}, e_{d,1}, \ldots, e_{d,M}\} \in \mathbb{R}^{(M+1)\times D}$. Moreover, the similarity between the F^D and the prompt features of the current learning step $P = \{p_0, p_1, \ldots, p_M\} \in \mathbb{R}^{(M+1)\times D}$ is also computed for incorporating prior semantic information provided by the large-scale pre-trained model. The two similarity matrices are combined to get the final probability distribution prediction. After getting the probability map S of every class, we reshape and unsample it to the size of the input image and perform *Softmax* to get the final prediction S'. To address the semantic shift problem [10] of the background class, we generate the segmentation prediction using the trained old model and re-label the ground truth $\mathcal{Y}^t$ of the present sample to obtain the pseudo label $\hat{\mathcal{Y}}^t$. We use a cross-entropy loss as the segmentation loss $\mathcal{L}^{seg}$ for training the current model, and the formulation is shown as follows:

$$\mathcal{L}^{seg} = \frac{1}{HW}\sum_{i=1}^{HW}\sum_{c\in\mathcal{C}^{0:t}} \hat{\mathcal{Y}}_{c,i}^t log\mathcal{S}'^t_{c,i} \tag{3}$$

C. PROMPT-GUIDED INITIALIZATION MODULE

Previous Transformer-based CISS method [15] typically randomly initializes the new class queries, which prevents retaining knowledge of old classes and learning new ones. To address this problem, we propose a novel Prompt-guided

Initialization Module (PIM) to initialize the new class queries based on the old query features and class prompts. In this paper, we utilize the pre-trained text encoder of CLIP [16] model to extract the features of the input human-defined class prompts. And the class prompt features of the *t-th* learning step can be denoted as $P^t = \{p_0^t, p_1^t, \ldots, p_M^t\} \in \mathbb{R}^{(M+1)\times D}$, and the features of the class queries during the *t-th* learning step are denoted as $Q^t = \{q_0^t, q_1^t, \ldots, q_M^t\} \in \mathbb{R}^{(M+1)\times D}$.

At the beginning of the *t*+1-*th* learning step, the learnable class queries corresponding to the new classes $\mathcal{C}^{t+1}$ are initialized by the PIM. Specifically, we first get the prompt features of the newly emerged classes $P^{t+1} = \{p_{M+1}^{t+1}, p_{M+2}^{t+1}, \ldots, p_{M+|\mathcal{C}^{t+1}|}^{t+1}\} \in \mathbb{R}^{|\mathcal{C}^{t+1}|\times D}$ and remove the "background" class within the old prompt features to get $P^{t'} = \{p_1^t, \ldots, p_M^t\} \in \mathbb{R}^{M\times D}$. The similarities between features of the old prompts and newly emerged prompts are calculated to measure the semantic correlations between the old and new tasks, which can be formulated as follow:

$$\mathcal{M}_{PIM} = P^{t+1} \cdot (P^{t'})^T \tag{4}$$

where $\mathcal{M}_{PIM} \in \mathbb{R}^{|\mathcal{C}^{t+1}|\times M}$ denotes the similarity matrix, which measures the semantic correlations between the old and new task classes. Then, we perform the *Softmax* operation on the similarity matrix and calculate the weighted sum of the old class queries with the "background" class removed, which can be formulated as follow:

$$Q^{t+1} = \mathcal{M}_{PIM} \cdot Q^{t'} \tag{5}$$

where $Q^{t'} \in \mathbb{R}^{M\times D}$ denotes the old class query features with the "background" class removed, and $Q^{t+1} \in \mathbb{R}^{|\mathcal{C}^{t+1}|\times D}$ denotes the initialized query features of the next learning step. Finally, the model learns the new task with the initialized class query features Q^{t+1} as the additional initial Transformer decoder inputs, which facilitates the model to transfer the original knowledge to the new learning process.

D. SEMANTIC-ALIGNED DISTILLATION MODULE

We propose a Semantic-aligned Distillation Module (SDM) to introduce the semantic guidance information provided by the class prompts for filling the semantic gap between different class data and alleviating the distillation errors. Details of the proposed SDM are illustrated below:

As shown in Fig. 2, we denote the old class prompts and old decoded class-aware embeddings (remove the "background" class) as $P^{t'} = \{p_1^t, \ldots, p_M^t\} \in \mathbb{R}^{M\times D}$ and $E_D^t = \{e_{d,1}^t, e_{d,2}^t \ldots, e_{d,M}^t\} \in \mathbb{R}^{M\times D}$, respectively. In the next learning step, the newly decoded class-aware embeddings are denoted as $E_D^{t+1} = \{e_{d,1}^{t+1}, e_{d,2}^{t+1} \ldots, e_{d,M+|\mathcal{C}^{t+1}|,new}^{t+1}\} \in \mathbb{R}^{(M+|\mathcal{C}^{t+1}|)\times D}$. We first calculate the similarities between the old class-aware embeddings and the old class prompts, which can be formulated as:

$$\mathcal{M}_{SDM} = E_D^t \cdot (P^{t'})^T \tag{6}$$

where $\mathcal{M}_{SDM} \in \mathbb{R}^{M\times M}$ denotes the similarity matrix calculated in SDM. Then, a *Softmax* function is applied to $\mathcal{M}_{SDM}$ for normalization. To build a unified semantic-level linkage across the old model and the new model, we utilize the calculated $\mathcal{M}_{SDM}$ to re-weight the old and new class-aware embeddings. Note that since only the old class information should be distilled, we keep the old class embeddings within the newly decoded E_D^{t+1}, which can be denoted as $E_D^{t+1'} = \{e_{d,1}^{t+1}, e_{d,2}^{t+1} \ldots, e_{d,M}^{t+1}\} \in \mathbb{R}^{M\times D}$. Next, the weighted old and new class-aware embeddings can be computed as follows:

$$(E_R^t)^T = (E_D^t)^T \cdot \mathcal{M}_{SDM} \tag{7}$$

$$(E_R^{t+1})^T = (E_D^{t+1'})^T \cdot \mathcal{M}_{SDM} \tag{8}$$

where $E_R^{t+1} = \{e_{r,1}^{t+1}, e_{r,2}^{t+1} \ldots, e_{r,M}^{t+1}\} \in \mathbb{R}^{M\times D}$ and $E_R^t = \{e_{r,1}^t, e_{r,2}^t \ldots, e_{r,M}^t\} \in \mathbb{R}^{M\times D}$ denote the weighted class-aware embeddings of the old and new model. Finally, after establishing semantic-level linkage by class prompt features extracted by the text encoder, more accurate and generalized knowledge distillation can be achieved and the distillation loss function can be formulated as:

$$\mathcal{L}^{dis} = \frac{1}{M}\sum_{i=1}^{M} ||e_{r,i}^{t+1} - e_{r,i}^t||^2 \tag{9}$$

The total loss for training the proposed CPT is illustrated as follows:

$$\mathcal{L}^T = \mathcal{L}^{seg} + \mathcal{L}^{dis} \tag{10}$$

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

1) DATASETS

We conduct extensive experiments on Pascal VOC [17] and ADE20K [18] datasets.

The Pascal VOC dataset [17] is a mainstream dataset for semantic segmentation task. It contains 20 foreground classes and one background class with 10,852 images for training and 1,449 images for testing.

The ADE20K [18] is a large-scale visual scene understanding dataset and is widely adopted to evaluate the effectiveness of semantic segmentation method. This dataset contains 150 semantic classes with 20,210 images for training and 2,000 images for testing.

2) PROTOCOLS

There are two settings for evaluating the effectiveness of the CISS model: Overlapped and Disjoint. For the disjoint setting, the data in each step just contains the objects of the classes $\mathcal{C}^{1:t-1}$ learned in the previous steps and the current classes $\mathcal{C}^t$, without the objects of future classes. For the overlapped settings, the data of each step contains visual objects of future classes, which is more compatible with realistic scenarios. In this paper, we conduct experiments under Overlapped setting to evaluate the proposed method.

Following the existing settings [11], [15], for Pascal VOC, we evaluate our method with similar protocols. The

TABLE 1. The mIoU(%) of the last step for different incremental class learning scenarios on Pascal VOC dataset.

Method	**19-1** (2 steps)			**15-5** (2 steps)			**15-1** (6 steps)			**10-1** (11 steps)		
	0-19	20	*all*	0-15	16-20	*all*	0-15	16-20	*all*	0-10	11-20	*all*
ILT [9]	67.75	10.88	65.05	67.08	39.23	60.45	8.75	7.99	8.56	7.15	3.67	5.50
MiB [10]	71.43	23.59	69.15	76.37	49.97	70.08	34.22	13.50	29.29	12.25	13.09	12.65
PLOP [11]	75.35	37.35	73.54	75.73	51.71	70.09	65.12	21.11	54.64	44.03	15.51	30.45
RECALL [50]	67.90	53.50	68.40	66.60	50.90	64.00	65.70	47.80	62.70	59.50	46.70	54.80
UCD [13]	71.40	47.30	70.00	77.50	53.10	71.30	49.00	19.50	41.90	42.30	28.30	35.30
CAF [56]	75.50	34.80	73.40	77.20	49.90	70.40	55.70	14.10	45.30	-	-	-
RCN [52]	-	-	-	78.80	52.00	72.40	70.60	23.70	59.40	55.40	15.10	34.30
RBC [53]	77.26	55.60	76.23	76.59	52.78	70.92	69.54	38.44	62.14	-	-	-
SPPFA [54]	76.50	36.20	74.60	78.10	52.90	72.10	66.20	23.30	56.00	-	-	-
EWF [57]	77.80	12.20	74.70	-	-	-	77.70	32.70	67.00	71.50	30.30	51.90
AMSS [58]	79.40	42.80	77.66	79.31	55.88	73.73	78.54	50.82	71.94	-	-	-
Incrementer [15]	82.54	60.95	82.14	82.53	69.25	79.93	79.60	59.56	75.55	**77.62**	60.33	70.16
Ours	**83.49**	**71.45**	**82.92**	**84.63**	**71.68**	**81.54**	**84.19**	**62.06**	**78.92**	77.47	**67.71**	**72.82**
Joint(no clip)	83.69	78.90	83.46	84.52	80.10	83.46	84.52	80.10	83.46	83.79	83.09	83.46
Joint	83.94	81.83	83.84	84.88	80.51	83.84	84.88	80.51	83.84	84.32	83.30	83.84

experimental settings including VOC-19-1 (2 steps, first training on 19 classes and then on 1 new class), VOC-15-5 (2 steps, first training on 15 classes and then on 5 new classes), VOC-15-1 (6 steps, first training on 15 classes and then on 1 new class in each of the next 5 steps), and VOC-10-1 (11 steps, first training on 10 classes and then on 1 new class in each of the next 10 steps).

For ADE20K [18], we evaluate our method on multiple divisions including: ADE-100-50 (2 steps in total, first training on 100 classes and then on 50 new classes), ADE-50-50 (3 steps, first training on 50 classes and then on 50 new classes on each of the following 2 steps), ADE-100-10 (6 steps, first training on 100 classes and then on 10 new classes in each of the next 5 steps), and ADE-100-5 (11 steps, first training on 100 classes and then on 5 new classes in each of the next 10 steps).

3) METRIC

We adopt mean Intersection over Union (mIoU) to quantitatively evaluate the performance of the proposed method. In detail, after T training steps, we first compute the mIoU of the initial classes to evaluate the model's stability. Then, we compute the mIoU of subsets of the following classes. Finally, we compute the mIoU of all classes to evaluate the overall performance.

4) IMPLEMENTATION DETAILS

The proposed CPT is based on Transformer [59] architecture. We adopt ViT-B/16 [55] pre-trained on the ImageNet [60] as visual encoder. And the decoder is a two-layer Transformer network. For the text encoder that converts prompts into representational vectors, we adopt the text encoder of CLIP [16]. The input images are cropped into 512×512 following the common setting. For the training strategy, we use the SGD optimizer with a momentum of 0.9 to train our model. We train our model on the Pascal VOC dataset with a learning rate of 1e-4 for 30 epochs every step, and on ADE20K with a learning rate of 1e-3 for 60 epochs every step, and the learning rate is multiplied by 0.5 in the incremental steps. In the training stage, we also use the loss function proposed by MiB [10] following RCN [52].

B. COMPARISONS WITH THE STATE-OF-THE-ARTS

1) PASCAL VOC

In Table 1, the quantitative results of the proposed method and other state-of-the-art class-incremental semantic segmentation methods on the Pascal VOC dataset are shown. Our method is comprehensively evaluated under four different settings: VOC-19-1, VOC-15-5, VOC-15-1, and VOC-10-1. The results demonstrate that our method outperforms current state-of-the-art methods in most settings by considerable points.

Compared with the most recent CISS method Incrementer [15], our CPT outperforms it by 0.78%, 1.61%, 3.37%, and 2.66% under the VOC-19-1, VOC-15-5, VOC-15-1, and VOC-10-1 settings for all classes, respectively. Notice that under the VOC-15-1, and VOC-10-1 settings whose number of learning steps are more than that under the VOC-19-1, and VOC-15-5 settings, our method outperforms Incrementer by a larger margin. Under the VOC-15-1 setting, our method leads by Incrementer as much as 4.59% mIoU. It demonstrates the effectiveness of our method under multi-step incremental learning settings, which are more consistent with real-world scenes.

2) ADE20K

In Table 2, we show the performance on the more challenging ADE20K dataset and compare our performance with other CISS methods under the ADE-100-50, ADE-50-50, ADE-100-10, and ADE-100-5 settings. The quantitative results show that our method outperforms the most recent Incrementer significantly under all of the learning and class settings.

Under the ADE-100-50 setting, our method outperforms Incrementer by 1.28% for 1-100 classes, 1.16% for 101-150 classes, and 1.24% for all classes. Under the ADE-50-50

TABLE 2. The mIoU(%) of the last step for different incremental class learning scenarios on ADE20K dataset.

Method	**100-50** (2 steps)			**50-50** (3 steps)			**100-10** (6 steps)			**100-5** (11 steps)		
	1-100	101-150	*all*	1-50	51-150	*all*	1-100	101-150	*all*	1-100	101-150	*all*
ILT [9]	18.29	14.40	17.00	3.53	12.85	9.70	0.11	3.06	1.09	0.08	1.31	0.49
MiB [10]	40.52	17.17	32.79	45.57	21.01	29.31	38.21	11.12	29.24	36.01	5.66	25.96
PLOP [11]	41.87	14.89	32.94	48.83	20.99	30.40	40.48	13.61	31.59	39.11	7.81	28.75
UCD [13]	42.12	15.84	33.31	47.12	24.12	31.79	40.80	15.23	32.29	-	-	-
CAF [56]	37.30	31.90	35.50	47.50	26.80	33.70	39.00	17.44	31.80	-	-	-
RCN [52]	42.30	18.80	34.50	48.30	25.00	32.50	39.30	17.60	32.10	38.50	11.50	29.60
RBC [53]	42.90	21.49	35.81	49.59	26.32	34.18	39.01	21.67	33.27	-	-	-
SPPFA [54]	42.90	19.90	35.20	49.80	23.90	32.50	41.00	12.50	31.50	-	-	-
EWF [57]	41.20	21.30	34.60	-	-	-	41.50	16.34	33.20	41.40	13.40	32.10
AMSS [58]	44.06	24.96	37.74	-	-	-	43.88	25.14	37.67	43.35	18.53	35.13
CoMFormer [14]	44.70	26.20	38.40	-	-	-	40.60	15.60	32.30	39.50	13.60	30.90
Incrementer [15]	49.42	35.62	44.82	56.15	37.81	43.92	48.47	34.62	43.85	46.93	31.31	41.72
Ours	**50.70**	**36.78**	**46.06**	**57.70**	**40.17**	**46.01**	**49.96**	**36.07**	**45.33**	**47.96**	**32.18**	**42.70**
Joint(no clip)	51.09	39.06	47.09	58.25	41.50	47.09	51.09	39.06	47.09	51.09	39.06	47.09
Joint	51.18	39.15	47.17	58.62	41.45	47.17	51.18	39.15	47.17	51.18	39.15	47.17

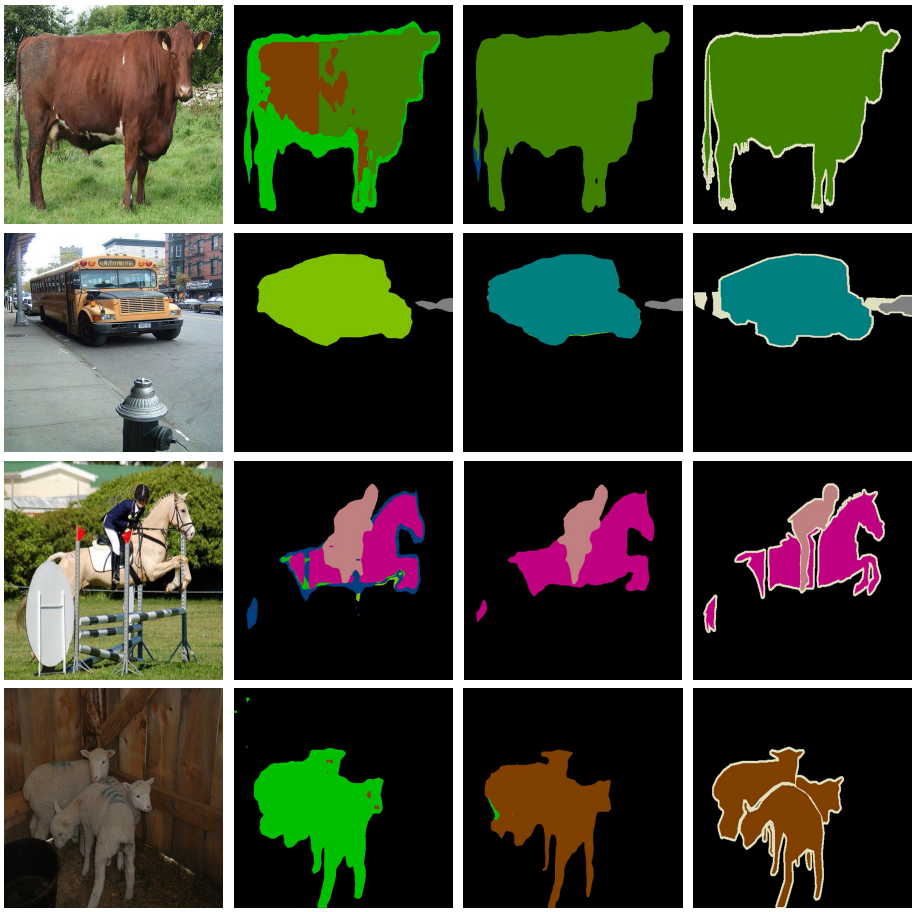

FIGURE 3. Qualitative results on Pascal VOC dataset [17] from the last step of 15-1 setting. 'GT' denotes the ground truth.

setting, our method yields 1.55%, 2.36%, and 2.09% lead for 1-50, 51-150, and all classes respectively. For the ADE-100-10 setting with 6 learning steps, our method outperforms by 1.49%, 1.45%, and 1.48% for 1-100, 101-150, and all classes respectively. For the ADE-100-5 setting with 11 learning steps, our method is 1.03%, 0.87%, and 0.98% mIoU ahead. The above results demonstrate the superiority of our proposed CPT.

C. ABLATION STUDIES

To verify the effectiveness of the proposed Class-prompt Transformer (CPT). We first conduct several ablation

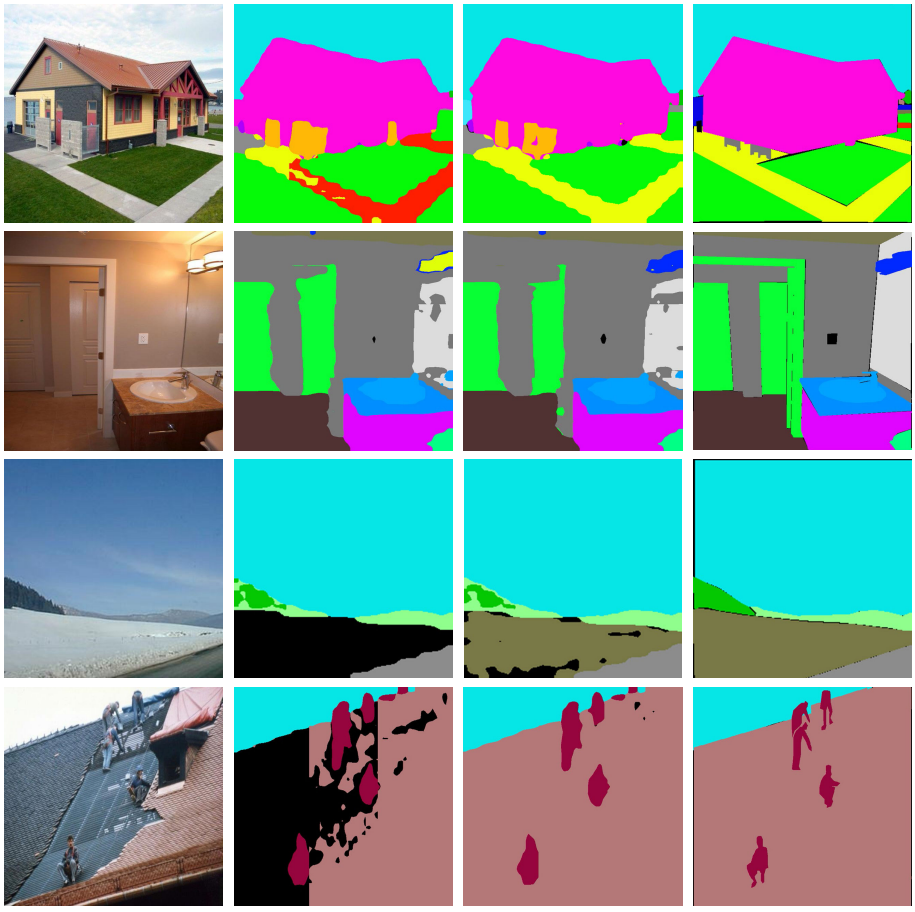

FIGURE 4. Qualitative results on ADE20K dataset [18] from the last step of 100-10 setting. 'GT' denotes the ground truth.

experiments to evaluate the effectiveness of the key components (i.e. Class prompting, Prompt-guided Initialization, and Semantic-aligned Distillation) in Section IV-C1. Then, in Section IV-C2, we tried several different query initialization strategies to confirm the validity of our prompt-guided strategy. Finally, we conduct a series of ablation experiments toward the distillation strategy in Section IV-C3. All of the ablation experiments are conducted under the setting of VOC-15-1 and the detailed analysis is as follows:

TABLE 3. Component ablation results on Pascal VOC 15-1 setting. The mIoU(%) of the last step are reported. CP denotes class prompting, PIM denotes Prompt-guided Initialization Module, and SDM denotes Semantic-aligned Distillation Module.

CP	PIM	SDM	0-15	16-20	*all*
			79.05	22.97	65.70
✓			80.10	55.92	74.35
✓	✓		81.45	61.80	76.77
✓		✓	83.51	60.27	77.98
✓	✓	✓	**84.19**	**62.06**	**78.92**

1) COMPONENT ABLATION

In this part, we conduct the key component ablation experiments under the overlapped VOC-15-1 setting. We comprehensively analyze the role of class prompting (CP), Prompt-guided Initialization Module (PIM), and Semantic-aligned Distillation Module (SDM). The detailed results are shown in Table 3.

The first row in Table 3 shows the performance of the baseline model. In the second row, we introduce the class prompting into the baseline framework and it leads to a 8.65% mIoU increase for all classes. It demonstrates that the incorporation of the semantic information leads to a significant performance increase. In the third row, we add the PIM, and the mIoU for all classes increases by 2.42%. In addition, the mIoU for 16-20 classes is significantly increased (i.e. 5.88%), which demonstrates that the PIM can facilitate the learning of new classes by transferring the learned knowledge into the learning process of the new tasks. In the fourth row, we add the SDM compared to the second row, which leads to a 3.63% increase in mIoU for all classes. Moreover, for the initial 0-15 classes, the SDM yields a 3.41% increase, which demonstrates the effectiveness of the SDM in preserving the original knowledge. The reason is that SDM leverages the prior information provided by the class prompts and prevents the model from forgetting the previous knowledge during the distillation processing.

TABLE 4. Ablation study of initialization strategies for visual queries of new classes. The mIoU(%) of the last step on Pascal VOC 15-1 setting are reported.

Initialization Strategy	0-15	16-20	*all*
Random	80.10	55.92	74.35
Background	79.47	56.33	73.95
Mean	80.27	56.09	74.51
Prompt-guided	**81.45**	**61.80**	**76.77**

TABLE 5. Ablation study of distillation strategies for class embeddings of old classes. The mIoU(%) of the last step on Pascal VOC 15-1 setting are reported.

Distillation Strategy	0-15	16-20	*all*
Without	80.10	55.92	74.35
Plain	81.32	61.20	76.53
Similarity	81.57	**61.75**	76.85
Plain + Similarity	82.06	61.56	77.18
Semantic-aligned	**83.51**	60.27	**77.98**

2) ANALYSIS OF INITIALIZATION STRATEGY

In this part, we try several query initialization methods and conduct the corresponding experiments to get the performance of each strategy.

In the first row of Table 4, we randomly initialize the queries. Under the VOC-15-1 setting, the model with random query initialization strategy achieves 80.10% mIoU for the 0-15 classes, 55.92% mIoU for the 16-20 classes, and 74.35% mIoU for all classes. In the second row, we use the features of the background class for initialization and the performance is 79.47% mIoU for the 0-15 classes, 56.33% mIoU for the 16-20 classes, and 73.95% mIoU for all classes. In the third row, we take the mean of the learned old classes representation for initialization and the corresponding performance is 80.27% mIoU for the 0-15 classes, 56.09% mIoU for the 16-20 classes, and 74.51% mIoU for all classes. Finally, in the fourth row, we use the proposed prompt-guided method, which increases by 2.42%, 2.82%, and 2.26% for all classes compared to Random, Background, and Mean strategies respectively. Specifically, the prompt-guided method increases by 5.88%, 5.47%, and 5.71% for 16-20 classes, which demonstrates the prompt-guided initialization facilitates the learning of newly emerged classes by injecting class prompts into the initialization of the new class queries and transferring the learned knowledge into the new learning step.

3) ANALYSIS OF DISTILLATION STRATEGY

In this part, we quantitatively analyze the effectiveness of different distillation strategies. The detailed discussions are as follows:

In Table 5, the performance corresponding to five different distillation strategies is shown. In the first row, ''Without'' means no distillation strategy is used and the model achieves 74.35% mIoU for all classes. In the second and third rows, ''Plain'' means directly distilling the old class embeddings via L2 loss, and ''Similarity'' denotes computing the similarity matrix between prompt features and class embeddings and then distilling the similarity matrix with the L2 loss function. It can be noticed that for 16-20 classes, ''Plain'' and ''Similarity'' methods increase by a large margin of 5.28% and 5.83% respectively. In the fourth row, ''Plain+Similarity'' means the combination of the ''Plain'' strategy and ''Similarity'' strategy, which leads to a relatively small increase. Finally, we evaluate the effectiveness of our SDM and the results show that the performance for 0-15 classes increases by 3.41% mIoU. For all classes, the performance of our SDM increases by 3.63% mIoU. The above results demonstrate the effectiveness of the proposed SDM, especially for the prior old classes, which verifies the role of SDM in integrating the semantic information into the distillation to prevent the model from catastrophic forgetting.

D. VISUALIZATION RESULTS

In this section, we provide some visualization results, including the segmentation prediction generated by the aforementioned baseline model and our CPT model, as shown in Fig. 3 and Fig. 4.

In Fig. 3, we show multiple results on the Pascal VOC dataset from the last learning step under the VOC-15-1 setting. It can be noticed that the segmentation masks generated by the baseline model have many misclassified areas and the fine-grained contour of visual objects can not be well depicted. However, the quality of segmentation masks generated by our CPT are significantly improved and the detailed information is well captured. In the third column of Fig. 3, we can find that the model precisely depicts the contours of each visual object with distinct semantics, accompanied by rare incorrectly segmented areas.

In Fig. 4, several visualization examples on the ADE20K dataset from the last step under the ADE-100-10 setting are shown. Despite there being up to six learning steps under this setting, the quality of the segmentation masks remains very high throughout, which indicates that with the proposed prompt-guided initialization module and the semantic-aligned distillation module, our CPT can well preserve the old class knowledge while learning the new tasks.

V. CONCLUSION

In this paper, we have proposed Class-prompting Transformer (CPT), a novel Transformer-based framework for the CISS task that leverages class prompts to bridge the semantic gap between distinct class data and achieve more generalized initialization and distillation. Firstly, a Prompt-guided Initialization Module (PIM) is developed to introduce the class prompts with prior knowledge into the initialization of the new class queries to transfer the previous knowledge into the learning of new tasks. Then, the model measures correlations between semantic guidance generated by the large pre-trained model and class-aware embeddings to fill the semantic gap between the old and new classes and reduce distillation errors via a Semantic-aligned Distillation Module (SDM). Extensive experiments on Pascal VOC and ADE20K demonstrate the effectiveness and superiority of our method.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[3] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 7262–7272.

[4] F. Tang, F. Gao, and Z. Wang, "Driving capability-based transition strategy for cooperative driving: From manual to automatic," *IEEE Access*, vol. 8, pp. 139013–139022, 2020.

[5] S. Sharmin, M. M. Hoque, S. M. R. Islam, Md. F. Kader, and I. H. Sarker, "Development of duplex eye contact framework for human-robot inter communication," *IEEE Access*, vol. 9, pp. 54435–54456, 2021.

[6] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[7] R. French, "Catastrophic forgetting in connectionist networks," *Trends Cognit. Sci.*, vol. 3, no. 4, pp. 128–135, Apr. 1999.

[8] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," 2013, *arXiv:1312.6211*.

[9] U. Michieli and P. Zanuttigh, "Incremental learning techniques for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 3205–3212.

[10] F. Cermelli, M. Mancini, S. Rota Bulo, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 9230–9239.

[11] A. Douillard, Y. Chen, A. Dapogny, and M. Cord, "PLOP: Learning without forgetting for continual semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 4040–4050.

[12] U. Michieli and P. Zanuttigh, "Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 1114–1124.

[13] 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2567–2581, Feb. 2023.

[14] F. Cermelli, M. Cord, and A. Douillard, "CoMFormer: Continual learning in semantic and panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 3010–3020.

[15] C. Shang, H. Li, F. Meng, Q. Wu, H. Qiu, and L. Wang, "Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 7214–7224.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 8748–8763.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[18] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5122–5130.

[19] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3213–3223.

[20] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1209–1218.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, 2015, pp. 234–241.

[22] Z. Jin, B. Liu, Q. Chu, and N. Yu, "ISNet: Integrate image-level and semantic-level context for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 7189–7198.

[23] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 801–818.

[27] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1520–1528.

[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," 2015, *arXiv:1511.00561*.

[29] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3640–3649.

[30] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2393–2402.

[31] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 9167–9176.

[32] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3146–3154.

[33] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 4003–4012.

[34] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 6881–6890.

[35] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2001–2010.

[36] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, vol. 30, 2017, pp. 2990–2999.

[37] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 233–248.

[38] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "PathNet: Evolution channels gradient descent in super neural networks," 2017, *arXiv:1701.08734*.

[39] X. Li, Y. Zhou, T. Wu, R. Socher, and C. Xiong, "Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 3925–3934.

[40] C.-Y. Hung, C.-H. Tu, C.-E. Wu, C.-H. Chen, Y.-M. Chan, and C.-S. Chen, "Compacting, picking and growing for unforgetting continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, vol. 32, 2019, pp. 13669–13679.

[41] J. Kirkpatrick, R. Pascanu, N. 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," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.

[42] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 532–547.
[43] K. J. Joseph, S. Khan, F. S. Khan, R. M. Anwer, and V. N. Balasubramanian, "Energy-based latent aligner for incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 7452–7461.
[44] R. Kemker and C. Kanan, "FearNet: Brain-inspired model for incremental learning," 2017, *arXiv:1711.10563*.
[45] F. Zhu, X.-Y. Zhang, C. Wang, F. Yin, and C.-L. Liu, "Prototype augmentation and self-supervision for incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 5871–5880.
[46] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 139–154.
[47] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 831–839.
[48] S. Ebrahimi, F. Meier, R. Calandra, T. Darrell, and M. Rohrbach, "Adversarial continual learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., 2020, pp. 386–402.
[49] P. Pan, S. Swaroop, A. Immer, R. Eschenhagen, R. Turner, and M. E. E. Khan, "Continual deep learning by functional regularisation of memorable past," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4453–4464.
[50] A. Maracani, U. Michieli, M. Toldo, and P. Zanuttigh, "RECALL: Replay-based continual learning in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 7026–7035.
[51] M. H. Phan, T.-A. Ta, S. L. Phung, L. Tran-Thanh, and A. Bouzerdoum, "Class similarity weighted knowledge distillation for continual semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 16866–16875.
[52] C.-B. Zhang, J.-W. Xiao, X. Liu, Y.-C. Chen, and M.-M. Cheng, "Representation compensation networks for continual semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 7053–7064.
[53] H. Zhao, F. Yang, X. Fu, and X. Li, "RBC: Rectifying the biased context in continual semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, 2022, pp. 55–72.
[54] Z. Lin, Z. Wang, and Y. Zhang, "Continual semantic segmentation via structure preserving and projected feature alignment," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, 2022, pp. 345–361.
[55] 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 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, 2021, pp. 1–21.
[56] G. Yang, E. Fini, D. Xu, P. Rota, M. Ding, T. Hao, X. Alameda-Pineda, and E. Ricci, "Continual attentive fusion for incremental learning in semantic segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 3841–3854, 2023.
[57] J. Xiao, C. Zhang, J. Feng, X. Liu, J. van de Weijer, and M. Cheng, "Endpoints weight fusion for class incremental semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Tel Aviv, Israel, Jun. 2023, pp. 7204–7213.
[58] L. Zhu, T. Chen, J. Yin, S. See, and J. Liu, "Continual semantic segmentation with automatic memory sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 3082–3092.
[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, vol. 30, 2017, pp. 5998–6008.
[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

ZICHEN SONG (Student Member, IEEE) received the B.E. degree in electronic and information engineering from the University of Electronic Science and Technology of China (UESTC), in 2016, where he is currently pursuing the Ph.D. degree in signal and information processing. His main research interests include computer vision and deep learning, especially the application of semantic segmentation and continual learning.

ZHAOFENG SHI received the B.E. degree in electronic information engineering from the University of Electronic Science and Technology of China (UESTC), in 2021, where he is currently pursuing the Ph.D. degree in information and communication engineering. His main research interests include computer vision, multi-modal processing, and incremental learning, especially the application of deep learning on image segmentation.

CHAO SHANG (Member, IEEE) received the B.E. degree in electronic information engineering from Dalian Maritime University, in 2017. He is currently pursuing the Ph.D. degree in information and communication engineering with the University of Electronic Science and Technology of China (UESTC). His main research interests include computer vision and machine learning, especially the application of deep learning on visual segmentation and multimodal analysis.

FANMAN MENG (Member, IEEE) received the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China, Chengdu, China, in 2014. From 2013 to 2014, he was a Research Assistant with the Division of Visual and Interactive Computing, Nanyang Technological University, Singapore. He is currently a Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. He has authored or coauthored numerous technical articles in well-known international journals and conferences. His current research interests include image segmentation and object detection.

LINFENG XU (Member, IEEE) received the Ph.D. degree in signal and information processing from the School of Electronic Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2014. From December 2014 to December 2015, he was with the Ubiquitous Multimedia Laboratory, The State University of New York at Buffalo, USA, as a Visiting Scholar. He is currently an Associate Professor with the School of Information and Communication Engineering, UESTC. His research interests include visual attention, saliency detection, image and video coding, visual signal processing, and multimedia communication systems.