SCIENCE CHINA
Information Sciences

November 2023, Vol. 66 210103:1–210103:19
https://doi.org/10.1007/s11432-022-3748-5

• RESEARCH PAPER •
Special Topic: Deep Learning for Computer Vision

A meaningful learning method for zero-shot semantic segmentation

Xianglong LIU[1], Shihao BAI[1], Shan AN[2], Shuo WANG[1], Wei LIU[1], Xiaowei ZHAO[1] & Yuqing MA[1*]

[1]*State Key Lab of Software Development Environment, Beihang University, Beijing* 100191*, China;*
[2]*JD Health International Inc., Beijing* 100191*, China*

Received 16 November 2022/Revised 7 February 2023/Accepted 3 April 2023/Published online 17 October 2023

Abstract Zero-shot semantic segmentation, which is developed to segment unseen categories, has attracted increasing attention due to its strong practicability. Previous approaches usually applied semantic-visual mapping based on seen categories to unseen categories, and thus failed to generate meaningful unseen visual representations and struggled to balance the seen and unseen concepts in the classifier. To overcome the above limitations, we propose a novel meaningful learning method that could be embedded into any generation-based zero-shot semantic segmentation model, borrowing the idea from the educational psychology field. The proposed meaningful learning method refers to the process that the new concepts could be learned by relating to existing comprehensible concepts and harmoniously incorporated into the concept schema. Specifically, we introduce a generator with conjugate conceptual correlation (G3C) which generates meaningful unseen visual information through anchoring into existing concepts. Moreover, simulating the rational thinking mechanism, we introduce a fast-slow concept modulator to alleviate the noisy over-correlation problem introduced by G3C and further construct a comprehensive concept schema. Extensive experiments conducted on three benchmarks demonstrate the superior performance of our method, especially according to the commonly-acknowledged h-mIoU (e.g., 4% improvement on the Pascal-VOC dataset).

Keywords meaningful learning, zero-shot learning, semantic segmentation, conjugate conceptual correlation, fast-slow conceptual modulator

Citation Liu X L, Bai S H, An S, et al. A meaningful learning method for zero-shot semantic segmentation. Sci China Inf Sci, 2023, 66(11): 210103, https://doi.org/10.1007/s11432-022-3748-5

1 Introduction

Semantic segmentation [1–5] is one of the most fundamental problems in computer vision and has been widely applied in many real-world scenarios, such as geographic information systems [6], self-driving [7,8], and medical image analysis [9–11]. However, it heavily depends on dense pixel-wise annotations which are time-consuming and labor-expensive, and struggles to transfer previous knowledge to unseen classes. On the other hand, humans could easily recognize an unseen class by connecting it with the concepts they knew before.

To bridge the gap between humans and machine visual intelligence, zero-shot semantic segmentation has been developed to segment objects of unseen categories based on previously seen categories with semantic word embeddings, as shown in Figure 1(a). There have been a few attempts on zero-shot multi-class semantic segmentation [12–14]. SPNet [13] learns to align each image pixel of seen classes with a semantic word embedding vector to accomplish classification during training, and extend the existing seen classifier with semantic embeddings of unseen classes. However, the classifier of SPNet is only trained on seen classes, resulting in poor performance on unseen classes. In contrast, generation-based methods, such as ZS3Net [14] and CaGNet [15] which utilize a generator to establish semantic-visual mapping and generate the visual representations for both seen and unseen classes, could provide the visual representations to train a comprehensive classifier.

* Corresponding author (email: mayuqing@buaa.edu.cn)

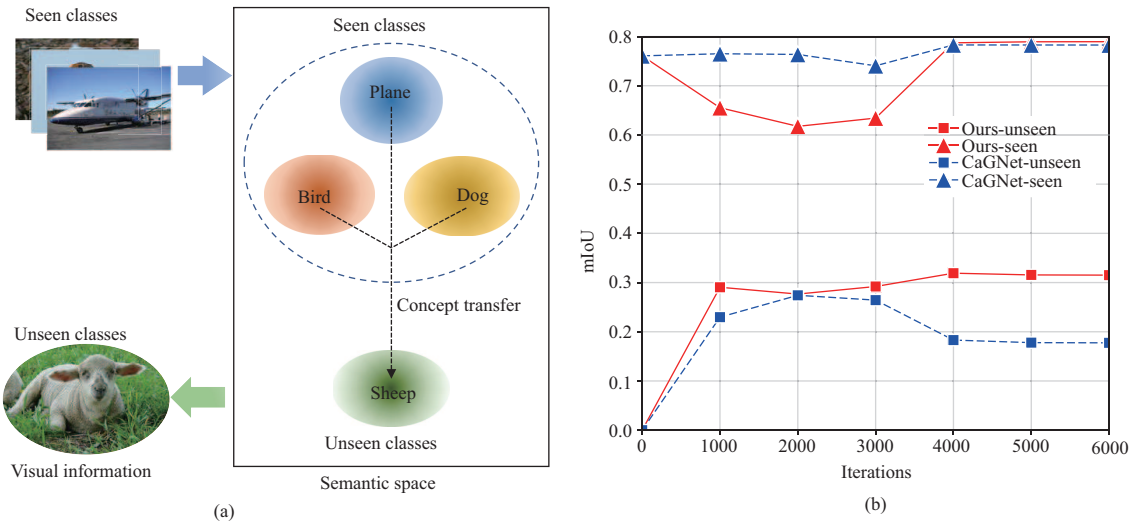

Figure 1 (Color online) The zero-shot semantic segmentation task and the underlying catastrophic forgetting problem. (a) The zero-shot segmentation task which transfers the concepts from seen classes to unseen classes to accomplish segmentation on both seen and unseen categories. (b) The mIoU value on seen classes (s-mIoU) and unseen classes (u-mIoU) of the previous state-of-the-art method CaGNet and our method every 1000 iterations, respectively. It manifests that CaGNet suffers from catastrophic forgetting while our method could evenly handle the previous experience and novel concepts.

Although previous generation-based zero-shot segmentation approaches have achieved promising performance, they fail to generate meaningful unseen visual representations and struggle to balance the seen and unseen concepts in the classifier. For one thing, the semantic-visual mapping trained on seen categories is hard to apply to unseen categories, leading to unreasonable unseen visual representation which further affects the training of the final classifier. For another, it is hard to prevent newly-learned information from interfering with existing concepts in the classifier, also known as catastrophic forgetting [16].

What is worse, since the seen categories account for the largest share of the overall classes and thus have a strong impact on evaluating the overall performance according to m-IoU metric, previous studies tend to dampen the classification performance of unseen classes to maintain the performance for seen categories. Figure 1(b) shows the mean intersection-over-union (mIoU) performance on seen classes (s-mIoU) and unseen classes (u-mIoU) every 1000 iterations of previous zero-shot semantic segmentation method CaGNet and ours. The u-mIoU of CaGNet increases at the first 2000 iterations, while the s-mIoU is falling. To maintain the overall performance, CaGNet suppresses the learning of unseen classifier, with the u-mIoU dropping more than 0.1 until 4000 iterations while s-mIoU is slightly rising. This is against the intention of zero-shot learning.

To address the above problems, we propose a novel meaningful learning method for zero-shot semantic segmentation, which is inspired by the relevant theory from the educational psychology field [17], to generate meaningful unseen visual representation and construct a comprehensive concept schema of previous seen classes and novel unseen classes. The term "meaningful learning" refers to the process of learning the new information through anchoring into existing comprehensible concepts, thus making the new-learned concept meaningful to the student, and integrating the existing and new concepts together to eventually promote a comprehensive understanding. Our meaningful learning method correspondingly consists of two novel components: a generator with conjugate conceptual correlation (G3C) which generates meaningful unseen visual information through anchoring into existing concepts, and a fast-slow concept modulator (FSCM) which alleviates the noisy over-correlation problem introduced by G3C and further constructs a comprehensive concept schema for segmentation. The G3C completes the process of learning new information through existing comprehensible concepts of meaningful learning and the FSCM accomplishes the process of the integrating the existing and new information. With the meaningful learning and the establishment of such valuable correlations between new and existing concepts, it is possible to better understand new knowledge and integrate it effectively with the previous knowledge.

Specifically, G3C generates the unseen visual representation through anchoring the existing seen visual representations, which could be guided by their semantic correlations through word embeddings. In other words, G3C models the valuable correlations between the new and existing concepts respectively

from semantic and visual perspectives, and puts constraints on the semantic-visual mapping to pursue correlation consistency. Connecting with the comprehensible existing concepts, the generated unseen visual representations are more meaningful and authentic. Moreover, a spatially adaptive margin is introduced to handle the semantic-visual gap and encourage visual diversity.

Although G3C could generate meaningful unseen visual representations through the correlations, it would also cause the over-correlation problem due to the noisy semantic word embeddings pre-trained with large-scale irrelevant corpus. Thus, the model cannot grasp the discriminative characteristics of a category and struggles to construct a comprehensive concept schema, exacerbating the catastrophic forgetting. Hence, we propose an FSCM consisting of a fast classifier, a slow classifier, and a recognition controller, to alleviate the over-correlation and form a comprehensive concept schema. The recognition controller allows the highly distinguishable samples obtained by the fast classifier to build the initial discriminative unseen concept, and then slowly control the concept integration in the slow classifier step by step.

To the best of our knowledge, this is the first endeavor to adopt the meaningful learning theory for the zero-shot semantic segmentation problem. Figure 1(b) also exhibits that our solution harmoniously improves the performance of both seen classes and unseen classes during the training phase. We hope our attempt could provide a new insight into the community and promote a wide breadth of applications. In summary, our main contributions are listed as follows.

(1) We propose a novel and general meaningful learning method for zero-shot segmentation, borrowing the idea from the educational psychology field, which well learns novel concepts without visual instances and handles conflicts between new concepts and the previous ones. It could be embedded into general zero-shot semantic segmentation models and significantly improve the segmentation performance.

(2) We design a novel generator with conjugate conceptual correlation to generate meaningful unseen visual representations through anchoring into existing comprehensible concepts.

(3) A novel fast-slow concept modulator is designed to alleviate the noisy over-correlation problem introduced by G3C and further constructs a comprehensive concept schema for segmentation.

(4) Extensive experiments demonstrate that the proposed meaningful learning method outperforms the state-of-the-art approaches on Pascal-VOC [18], Pascal-Context [19], and COCO-stuff [20] datasets. Our model can achieve superior performance (e.g., 4% improvement of h-mIoU on Pascal-VOC dataset) compared with state-of-the-art baseline methods.

2 Related work

2.1 Semantic segmentation

Semantic segmentation is a long-standing and challenging task in computer vision, aiming to accurately predict pixel-wise semantic labels in an image. Most recent state-of-the-art models [21–26] are based on FCN [27], the first framework adopting fully convolutional networks to address the task in an end-to-end manner. Several studies [21, 28] introduce a post-processing module such as conditional random fields (CRFs) on the predicted results of the network to improve the performance. Besides, models such as PSPNet [23] or DeepLab [21] attempt to obtain multi-scale contextual information by performing spatial pyramid pooling at multiples scales or applying several parallel atrous convolutions with different rates. Moreover, the encoder-decoder networks [29, 30] have been successfully applied to semantic segmentation tasks, owing to their strong ability to capture higher semantic information and visual contextual information. For instance, U-Net [29] consists of an encoder which down-samples the input image to a feature map and a decoder that up-samples the feature map to input image size. Deeplabv3+ [30] integrates the spatial pyramid pooling module and the encoder-decoder structure into a network, targeting to encode multi-scale contextual information and capture sharper object boundaries. Although these state-of-the-art methods have achieved good performance, they still need accurate annotations for each category, limiting them to adapt to novel categories in new tasks quickly.

2.2 Zero-shot learning

Zero-shot learning has attracted significant attention in recent years, due to its great potential in generalizing to unseen classes in classification tasks. These zero-shot learning methods can be roughly divided into two classes of methods, namely (a) projection-based methods and (b) generation-based methods.

The former class of methods [31–36] tend to learn a projection function which projects the semantic embedding and visual features into a public latent space, and then directly apply the learned projection to unseen classes in the testing process. Moreover, for the class-imbalance issue in zero-shot learning, Ref. [36] designed a sample-balanced training process to encourage all training classes to contribute equally to the learned model. However, such projection-based methods always have poor performance in the generalized zero-shot learning which is more appropriate for real-life applications. Because there is no extra supervised learning for unseen classes and the model trained on seen classes has a strong bias for the seen classes. Generation-based methods mainly utilize generative adversarial networks [37] (GANs) to generate unseen class features, which transfers the zero-shot learning into a fully supervised problem. Ref. [38] firstly adopted a conditional Wasserstein GAN architecture for feature generation, consistently improving the classification performance in both zero-shot learning and generalized zero-shot learning settings, which is followed by later studies [34, 39–45]. Such as, Ref. [39] proposed a cycle consistent loss to constrain the generated visual features to reconstruct its original semantic features. Ref. [34] borrowed the knowledge distillation idea to ensure the generated classifiers are discriminative to the visual features. While these GAN-based methods have achieved good performance in zero-shot classification tasks, they can only handle images with one object and can not directly guide zero-shot segmentation task, where the input images always consist of many objects. Therefore, an explicit correlation constraint inside the image is needed to guide the generator to generate more realistic feature maps for zero-shot semantic segmentation.

In addition to classification, there are also some studies focusing on detection tasks. Ref. [46] proposed an intra-class semantic diverging component and an inter-class structure preserving component to help synthesize robust region features for zero-shot detection task. Ref. [47] designed an end-to-end framework with class-sensitive modeling, semantic-agnostic modeling, and content-aware modeling.

2.3 Zero-shot semantic segmentation

Prior studies on zero-shot semantic segmentation [12, 48] mainly focused on single-class segmentation which has less helpful to solving practical problems in real life. Recently, there are some methods [13–15, 49, 50] proposed to solve multi-class semantic segmentation task. These methods can be roughly divided into two categories, discriminative-based methods and generative-based methods. The discriminative-based methods utilize the seen classes to explicitly build the relations between the semantic information and vision information. SPNet [13] learned a visual-semantic embedding module with seen classes to produce intermediate feature maps in the word embedding space and then projected those feature maps into class probabilities via a fixed word embedding projection matrix. Finally, it segmented unseen categories by replacing the projection matrix with word embeddings of novel classes. Ref. [50] introduced the uncertainty aware losses for zero-shot semantic segmentation to learn at image level and pixel level. Ref. [51] devised a self-training pipeline to obtain strong supervision for unseen classes. Generation-based methods utilize the generative adversarial network to generate training examples for unseen classes, and retrain the network with both seen classes and generated unseen classes. ZS3Net [14] learned the semantic-visual mapping through a generative model which allowed the generation of visual representations from semantic word embeddings of unseen classes. The generated visual representations were then used to fine-tune the classifier of the network, making it compatible with seen and unseen categories. Based on ZS3Net, CaGNet [15] introduced a contextual module after the feature embedding network to capture the pixel-wise contextual information, which encouraged the diversity of visual representations and improved the segmentation performance. CSRL proposed a consistent structural relation learning approach to harness the similarity of category-level relations on the semantic word embedding space and to learn a better visual feature generator. Ref. [52] introduced the spatial information module to incorporate spatial information in semantic segmentation and proposed an annealed self-training to generate pseudo-annotations for unlabeled samples.

3 Approach

Previous methods [12–14] fail to generate meaningful unseen visual representations and cannot accommodate the newly learned concepts with the existing concepts, resulting in inferior performance. To address the above problems, we propose a meaningful learning method, borrowing the idea from the educational

Table 1 Notations

Symbol	Meaning
$\mathcal{C}^s$	The set of seen classes
$\mathcal{C}^u$	The set of unseen classes
$\boldsymbol{I}^s$	The images containing seen classes
$\boldsymbol{I}^{s\cup u}$	The images containing seen and unseen classes
$\boldsymbol{W}_n^s$	Semantic word embedding map according to the segmentation annotations
$\boldsymbol{W}_n^{s\cup u}$	Semantic word embedding map sampled by randomly stacking word embedding of both seen and unseen classes
$\mathbb{E}$	Feature extractor
$\mathbb{G}$	Conjugate consistent generator
$\mathbb{D}$	Discriminator
$\mathbb{RC}$	Recognition controller
$\mathbb{FC}$	Fast classifier
$\mathbb{SC}$	Slow classifier
$\boldsymbol{X}_n^s$	The real visual feature map for seen classes
$\widetilde{\boldsymbol{X}}_n^{s\cup u}$	The generated visual feature map for seen and unseen classes
$\boldsymbol{Z}_n$	The sampled noise map
α	The adaptive semantic-visual margin of seen classes
$\boldsymbol{P}_{m,i}$	The compatibility of $\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u}$ with the prior knowledge
$\boldsymbol{R}_{m,i}$	The predicted logits of slow classifier for $\boldsymbol{X}_{m,i}^{s\cup u}$

psychology field, to learn the novel concepts by relating to the previous experience and incorporating them into the existing concept schema.

The following content is organized as follows: we will introduce the preliminary of zero-shot segmentation first. Then we will depict the overall method and present the two crucial operations in meaningful learning: the generator with conjugate conceptual correlation and the fast-slow concept modulator in Subsections 3.3 and 3.4. Finally, we will introduce the learning and inference process.

3.1 Preliminary

In this part, we first present the definitions of all symbols in Table 1, in order to facilitate the understanding of the formula in the paper. Then we give the formal definition of zero-shot semantic segmentation. Specifically, let $\mathcal{C}^s$ and $\mathcal{C}^u$ respectively denote the set of seen classes and unseen classes, where $\mathcal{C}^s \cap \mathcal{C}^u = \emptyset$. The goal of zero-shot semantic segmentation model is to segment images $\boldsymbol{I}^{s\cup u}$ of $\mathcal{C}^s \cup \mathcal{C}^u$, with the concepts of the images only containing seen classes and their corresponding ground-truth segmentation maps $\{\boldsymbol{I}^s, \boldsymbol{Y}^s\}_{n=1}^N$, and the semantic word vectors of all the classes $\mathcal{C}^s \cup \mathcal{C}^u$.

Previous generation-based zero-shot semantic segmentation architectures usually consist of a feature embedding network $\mathbb{E}$, a generator $\mathbb{G}$, and a pixel-wise classifier $\mathbb{C}$. The learning pipeline usually consists of two stages: a basic learning stage for obtaining informative seen information and an extended learning stage for learning and integrating unseen concepts. In the basic learning stage, the generator $\mathbb{G}$ mainly constructs the semantic-visual mapping and generates fake seen visual representation $\widetilde{\boldsymbol{X}}_n^s = \mathbb{G}([\boldsymbol{W}_n^s, \boldsymbol{Z}_n])$, close to the real seen visual representation $\boldsymbol{X}_n^s = \mathbb{E}(\boldsymbol{I}_n^s)$. $\boldsymbol{W}_n^s$ is the word embedding map, and $\boldsymbol{Z}_n$ is the noise map. $\mathbb{C}$ is trained with both the real and generated seen visual representations. CaGNet [15] also introduces a discriminator $\mathbb{D}$, sharing the first layer with a classifier, to further regularize the mapping. In the extended learning stage, the semantic word embedding of unseen categories is involved, and we could obtain the generated unseen visual representations $\widetilde{\boldsymbol{X}}_m^{s\cup u} = \mathbb{G}([\boldsymbol{W}_m^{s\cup u}, \boldsymbol{Z}_m])$, and train the extended $\mathbb{C}$ with both seen and unseen representations. Thus, in the inference phase, the classifier could process the real unseen visual representation for prediction.

It can be seen that the semantic-visual mapping of the generator is only trained with seen samples, which is inappropriate for unseen categories. In other words, the generator cannot provide meaningful unseen visual representations for the classifier, which will affect the segmentation performance. Moreover, simply extending and fine-tuning the classifier is hard to balance the learning of the seen and unseen concepts, causing the interference for both the unseen concepts construction and seen concepts adjustment.

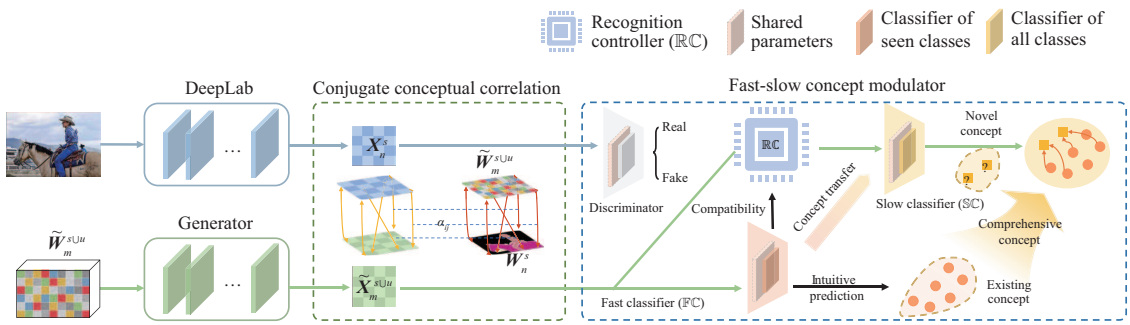

Figure 2 (Color online) The architecture of the proposed meaningful learning method for zero-shot semantic segmentation. The conjugate conceptual correlation (G3C) encourages the diversity and reality of unseen representations via establishing correlations with meaningful seen representations, guiding the learning of the unseen concept schema in the classifier. The fast-slow concept modulator (FSCM), consisting of the fast classifier $\mathbb{FC}$, the recognition controller $\mathbb{RC}$, and the slow classifier $\mathbb{SC}$, alleviates the noisy over-correlation problem introduced by G3C and further constructs a comprehensive concept schema eventually.

3.2 The overall

In this part, we will introduce our meaningful learning method for zero-shot semantic segmentation. We first constrain the generator $\mathbb{G}$ with a conjugate conceptual correlation to generate the meaningful unseen visual representations, and then replace the pixel-wise classifier $\mathbb{C}$ with our fast-slow concept modulator to alleviate the over-correlation and accomplish concept integration. Figure 2 illustrates the architecture of the proposed meaningful learning method.

In the basic stage, we follow the previous learning pipeline, while during the extended learning stage, we first propose a conjugate conceptual correlation to regularize the mapping $\mathbb{G}$, to encourage the diversity and reality of unseen representations via establishing correlations with meaningful seen representations. Moreover, we obtain the $\widetilde{\boldsymbol{X}}_n^{s\cup u}$ and feed it into the FSCM. FSCM is composed of a fast classifier $\mathbb{FC}$, a recognition controller $\mathbb{RC}$, and a slow classifier $\mathbb{SC}$. We first put it in the $\mathbb{FC}$. The recognition controller $\mathbb{RC}$ screens out samples containing significant characteristics for unseen concepts according to the output of $\mathbb{FC}$, and allows such samples to pass to the slow classifier $\mathbb{SC}$ to make the final prediction, guiding the model to establish a good initial concept for unseen classes in the comprehensive concept schema. At the end of the training, the slow classifier $\mathbb{SC}$ evenly learns the meaningful novel concepts and fine-tunes the existing concept schema, and eventually incorporates them into the entire concepts. We will further elaborate on them in Subsections 3.3 and 3.4.

3.3 Generator with conjugate conceptual correlation

The generator in previous studies [14,15] constructed semantic-visual mapping only through seen samples and cannot generate meaningful unseen visual representations. To overcome this problem, we design a generator with G3C. The proposed G3C encourages the generated unseen visual representations to connect with the comprehensible existing concepts, and the visual connecting should be consistent with the semantic counterparts. Figure 3(a) depicts the idea.

Formally, given the m-th semantic word embedding map $\boldsymbol{W}_m^{s\cup u}\in\mathcal{R}^{c\times h\times w}$, we reshape the map to the size of $c\times n$, where $n=h\times w$. Correspondingly, $\boldsymbol{W}_{m,i}^{s\cup u}$ of length c is the i-th column in the reshaped map, where $i=1,\ldots,n$. In the meanwhile, we randomly pick a sample $\{\boldsymbol{I}^s,\boldsymbol{Y}^s\}$ of seen classes, and could also obtain its visual representation $\boldsymbol{X}_n^s$, its semantic word embedding map $\boldsymbol{W}_n^s$, and its generated visual representation $\widetilde{\boldsymbol{X}}_n^s$.

The semantic correlations between the i-th patch of $\boldsymbol{W}_m^{s\cup u}$ and the j-th patch of $\boldsymbol{W}_n^s$ could be calculated by $(\boldsymbol{W}_{m,i}^{s\cup u})^{\mathrm{T}}\boldsymbol{W}_{n,j}^s$. The visual connections could be calculated through $(\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u})^{\mathrm{T}}\boldsymbol{X}_{n,j}^s$ likewise. Intuitively, the connections from both semantic and visual perspectives should be consistent. Hence, G3C could generate the meaningful unseen visual representation $\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u}$ through the following constraints:

$$\begin{aligned}\ell_{\mathrm{CON}}=&\sum_{i,j}\Big(\max(0,(\boldsymbol{W}_{m,i}^{s\cup u})^{\mathrm{T}}\boldsymbol{W}_{n,j}^s-(\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u})^{\mathrm{T}}\boldsymbol{X}_{n,j}^s-\alpha)\Big)^2\\&+\sum_{i,j}\Big(\max(0,(\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u})^{\mathrm{T}}\boldsymbol{X}_{n,j}^s-(\boldsymbol{W}_{m,i}^{s\cup u})^{\mathrm{T}}\boldsymbol{W}_{n,j}^s-\alpha)\Big)^2,\end{aligned}\tag{1}$$

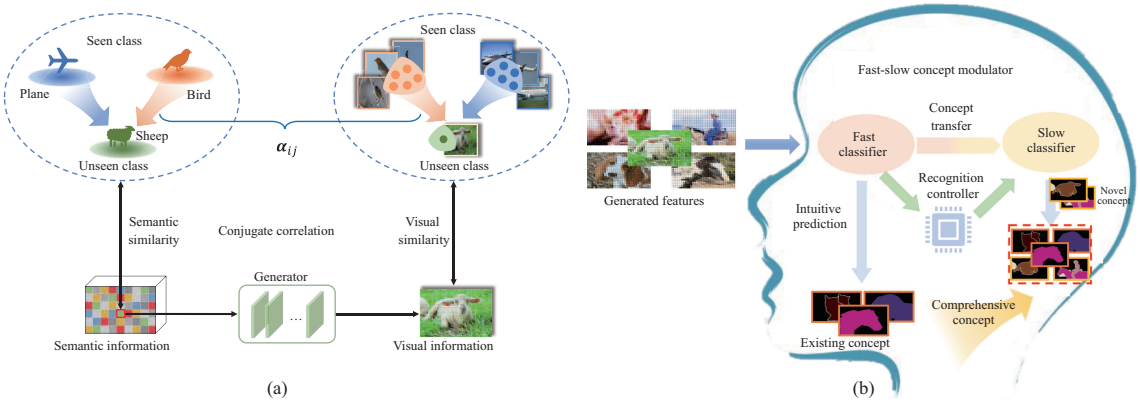

Figure 3 (Color online) Two novel components of the proposed meaningful learning method. (a) The generator with conjugate conceptual correlation (G3C). The proposed G3C encourages the generated unseen visual representations to connect with the comprehensible existing concepts, and the visual connecting should be consistent with the semantic counterparts. The learning process of fast-slow concept modulator (FSCM). (b) The FSCM consists of a fast classifier $\mathbb{FC}$, a recognition controller $\mathbb{RC}$, and a slow classifier $\mathbb{SC}$. $\mathbb{RC}$ justifies the intuitive decision made by the fast thinking of $\mathbb{FC}$, and thus controls the rational thinking for concept integration in the $\mathbb{SC}$.

where α is the margin indicating the semantic-visual gap and $\max(\cdot,\cdot)$ refers to the element-wise maximum value.

In practice, to encourage visual diversity at different spatial positions, a pixel-wise adaptive margin is introduced to ℓ_{CON} to explicitly transfer the spatial diversity of the seen classes to the unseen classes. Specifically, we use the semantic-visual gap of real seen samples as the margin instead of a fixed one:

$$\boldsymbol{\alpha}_{ij} = \text{abs}\left\{(\boldsymbol{W}^s_{n,i})^{\text{T}}\boldsymbol{W}^s_{n,j} - (\boldsymbol{X}^s_{n,i})^{\text{T}}\boldsymbol{X}^s_{n,j}\right\}, \tag{2}$$

where $\text{abs}\{\cdot\}$ represents the element-wise absolute value.

In summary, our conjugate conceptual correlation loss can be formulated as

$$\begin{aligned}\ell_{\text{CON}} = &\sum_{i,j}\left(\max(0,(\boldsymbol{W}^{s\cup u}_{m,i})^{\text{T}}\boldsymbol{W}^s_{n,j} - (\widetilde{\boldsymbol{X}}^{s\cup u}_{m,i})^{\text{T}}\boldsymbol{X}^s_{n,j} - \boldsymbol{\alpha}_{ij})\right)^2 \\ &+\sum_{i,j}\left(\max(0,(\widetilde{\boldsymbol{X}}^{s\cup u}_{m,i})^{\text{T}}\boldsymbol{X}^s_{n,j} - (\boldsymbol{W}^{s\cup u}_{m,i})^{\text{T}}\boldsymbol{W}^s_{n,j} - \boldsymbol{\alpha}_{ij})\right)^2.\end{aligned} \tag{3}$$

The adaptive margin takes the spatial information into account and flexibly adjusts the semantic-visual gap, thus ensuring the visual diversity and improving the generalization ability.

G3C generates the unseen visual representation through the conjugate conceptual correlation, connecting the unseen visual representation with the comprehensible seen ones and thus making them meaningful and authentic. With the meaningful unseen representations, it could guide the learning of the unseen concept schema in the classifier and the harmonious integration of the overall concepts.

3.4 Fast-slow concept modulator

Although G3C could generate meaningful unseen visual representations through the correlations, it would also cause the over-correlation problem due to the noisy semantic word embeddings. Thus, the model cannot grasp the discriminative characteristics of a category and struggles to construct a comprehensive concept schema, exacerbating the catastrophic forgetting. Contrastively, humans could effortlessly integrate the new concepts to existing concepts, filtering out the noisy information. Therefore, we follow the two types of human thinking mechanism [17] proposed by Kahneman, namely fast thinking and slow thinking, and propose the FSCM to handle the over-correlation problem and construct a comprehensive concept schema, harmoniously integrating the existing seen concepts and the novel unseen concepts for the final classification.

Figure 3(b) depicts the learning process of the proposed FSCM. Specifically, FSCM consists of a fast classifier $\mathbb{FC}$, a recognition controller $\mathbb{RC}$, and a slow classifier $\mathbb{SC}$. $\mathbb{RC}$ justifies the intuitive decision made by the fast thinking of $\mathbb{FC}$, and thus controls the rational thinking for concept integration in the $\mathbb{SC}$. $\mathbb{SC}$ replicates the parameters of $\mathbb{FC}$ as an initialization and extends it with randomly initialized parameters

for unseen classes. The parameters of the prediction layer in the $\mathbb{FC}$ are fixed in the extended learning stage, while the parameters of the first layer will be updated during the extended learning stage.

Formally, the recognition controller $\mathbb{RC}$ separates the highly discriminative samples of unseen classes according to the output of $\mathbb{FC}$, which has been already trained in the basic learning stage with existing concepts:

$$\boldsymbol{p}_{m,i} = \mathbb{RC}(\mathbb{FC}(\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u})), \tag{4}$$

where $\boldsymbol{p}_{m,i} \in [0,1]^2$ is a two-dimensional vector respectively indicating the compatibility of $\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u}$ with the prior concepts of seen classes and the possibility of it being out of the distribution.

We use a common cross entropy loss to guide the learning of the recognition controller:

$$\ell_{\mathrm{BCE}} = \sum_{k=0}^{1} \overline{\boldsymbol{p}}_{m,i}[k] \log \boldsymbol{p}_{m,i}[k], \tag{5}$$

where the value of $\overline{\boldsymbol{p}}_{m,i}$ depends on its exact class label $\boldsymbol{y}_{m,i}^{s\cup u}$ at the i-th column of the reshape feature map $\boldsymbol{X}_{m,i}^{x\cup u}$:

$$\overline{\boldsymbol{p}}_{m,i} = \begin{cases} (0,1)^{\mathrm{T}}, & \text{if } \boldsymbol{y}_{m,i}^{s\cup u} \notin \mathcal{C}^s, \\ (1,0)^{\mathrm{T}}, & \text{if } \boldsymbol{y}_{m,i}^{s\cup u} \in \mathcal{C}^s. \end{cases} \tag{6}$$

Intuitively, if $\boldsymbol{p}_{m,i}[0] < \boldsymbol{p}_{m,i}[1]$, it means the possibility that $\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u}$ falls into the seen classes is lighter than it does not. That is to say, $\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u}$ is a discriminative representation and should be processed first in the $\mathbb{SC}$ to build a discriminative initial unseen concept in the overall concept schema:

$$\boldsymbol{R}_{m,i} = \mathcal{I}\left(\boldsymbol{p}_{m,i}[0] < \boldsymbol{p}_{m,i}[1]\right) \mathbb{SC}\left(\boldsymbol{X}_{m,i}^{s\cup u}\right), \tag{7}$$

where $\mathcal{I}\left(\boldsymbol{p}_{m,i}[0] < \boldsymbol{p}_{m,i}[1]\right)$ is the Kronecker delta function that is equal to 1 when $\boldsymbol{p}_{m,i}[0] < \boldsymbol{p}_{m,i}[1]$ and 0 otherwise. $\boldsymbol{R}_{m,i} \in \mathbb{R}^{|\mathcal{C}^s \cup \mathcal{C}^u|}$ is the predicted probability of the slow classifier for $\boldsymbol{X}_{m,i}^{s\cup u}$.

Except for the controlling of training samples for $\mathbb{SC}$, we introduce a balancing factor to restrict the severity of the penalty, concentrating mostly on discriminative unseen samples and indiscriminative seen samples. Formally, the balancing factor is defined as

$$\gamma_i = \begin{cases} \beta_1, & \text{if } \overline{\boldsymbol{y}}_{m,i}^{s\cup u} = \boldsymbol{y}_{m,i}^{s\cup u}, \\ \beta_2, & \text{if } \overline{\boldsymbol{y}}_{m,i}^{s\cup u} \neq \boldsymbol{y}_{m,i}^{s\cup u}, \end{cases} \tag{8}$$

where $\overline{\boldsymbol{y}}_{m,i}^{s\cup u}$ is the predicted class according to $\boldsymbol{R}_{m,i}$.

The weighted classification loss can be written as

$$\ell_{\mathrm{CLS}} = -\sum_{i=1}^{n} \sum_{k=0}^{|\mathcal{C}^s \cup \mathcal{C}^u|-1} \gamma_i \boldsymbol{Y}_{m,i}^{s\cup u}[k] \log \boldsymbol{R}_{m,i}[k]. \tag{9}$$

As the distinguishing ability of $\mathbb{SC}$ grows, some seen samples are inappropriate for the fast classifier to be predicted due to the adjustment of the shared layer and will gradually be passed through the slow classifier for further identification. At this time, the slow classifier simultaneously fine-tuned the existing concepts and the newly-learned concepts, attempting to integrate them into the entire comprehensive concept schema, and thus is capable of accomplishing the classification task for all the categories.

3.5 Learning and inference

In this subsection, we will elaborate on the training procedure. In the basic learning stage, the modules $\{\mathbb{E}, \mathbb{G}, \mathbb{D}, \mathbb{FC}\}$ are trained with seen samples and corresponding annotations, guided by the loss terms similar as follows:

$$\min_{\mathbb{E},\mathbb{G},\mathbb{FC}} \max_{\mathbb{D}} \bar{\ell}_{\mathrm{CLS}} + \ell_{\mathrm{ADV}} + \lambda \ell_{\mathrm{REC}} + \ell_{\mathrm{KL}}, \tag{10}$$

where $\bar{\ell}_{\mathrm{CLS}}$ is the cross entropy loss to guide the learning of $\mathbb{FC}$, ℓ_{ADV} is the adversarial loss for constraining the generator $\mathbb{G}$ and the discriminator $\mathbb{D}$, ℓ_{REC} is applied to mapping the visual embedding and the semantic embedding, and ℓ_{KL} is utilized to project the visual context information to unit Gaussian

Algorithm 1 Extended learning stage

Input: Seen samples $\{\boldsymbol{I}^s, \boldsymbol{Y}^s\}_{n=1}^N$, semantic word embedding map $\boldsymbol{W}_m^{s\cup u} \in \mathcal{R}^{c\times h\times w}$, $\boldsymbol{W}_n^s \in \mathcal{R}^{c\times h\times w}$.
Output: The generator $\mathbb{G}$, discriminator $\mathbb{D}$, slow classifier $\mathbb{SC}$, and recognition controller $\mathbb{RC}$.
Initialize $\mathbb{SC}$ with parameters of $\mathbb{FC}$:
while iterations $t < T^{\text{step}}$ **do**
  Sample batch images $\boldsymbol{I}_n^s$;
  Construct semantic word embedding map $\boldsymbol{W}_n^s$;
  Calculate the real visual representation $\boldsymbol{X}_n^s = \mathbb{E}(\boldsymbol{I}_n^s)$;
  Randomly sample the semantic word embedding map $\boldsymbol{W}_m^{s\cup u}$ and the noise map $\boldsymbol{Z}_m$;
  Produce the visual representation $\widetilde{\boldsymbol{X}}_m^{s\cup u} = \mathbb{G}([\boldsymbol{W}_m^{s\cup u}, \boldsymbol{Z}_m])$;
  Generate the determination $\boldsymbol{p}_{m,i} = \mathbb{RC}(\mathbb{FC}(\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u}))$;
  Predict the probability $\boldsymbol{R}_{m,i}$ of the slow classifier for $\widetilde{\boldsymbol{X}}_{m,i}^{s\cup u}$;
  Calculate ℓ_{ADV} by $\boldsymbol{X}_n^s$ and $\widetilde{\boldsymbol{X}}_n^s$, ℓ_{CON} by $\boldsymbol{X}_n^s, \widetilde{\boldsymbol{X}}_n^s, \boldsymbol{W}_n^s$, and $\boldsymbol{W}_n^{s\cup u}$, ℓ_{BCE} by $\boldsymbol{p}_{m,i}$, ℓ_{CLS} by $\boldsymbol{R}_{m,i}$;
  Update $\{\mathbb{G}, \mathbb{D}, \mathbb{SC}, \mathbb{RC}\}$ with ℓ_{CON}, ℓ_{ADV}, ℓ_{CLS}, and ℓ_{BCE}.
end while

distribution $\mathcal{N}(0,1)$ which ensures the generator can sample the visual context information of unseen classes directly from $\mathcal{N}(0,1)$ and generate the visual representations of unseen classes in the extended learning stage. The meaningful learning method is a general learning strategy, which could be inserted in any zero-shot semantic segmentation model. We implement the proposed model based on the recent state-of-the-art approach CaGNet [15] and inherit its loss function. Therefore, detailed definitions of the above loss terms can be referred to as CaGNet [15].

In the extended learning stage, we update $\{\mathbb{G}, \mathbb{D}, \mathbb{SC}, \mathbb{RC}\}$ with parameters of other modules fixed:

$$\min_{\mathbb{G},\mathbb{SC},\mathbb{RC}} \max_{\mathbb{D}} \ell_{\text{CON}} + \ell_{\text{ADV}} + \ell_{\text{CLS}} + \ell_{\text{BCE}}, \tag{11}$$

where ℓ_{CON} deriving from (3) is the proposed conjugate conceptual correlation loss to generate meaningful unseen visual information, ℓ_{BCE} deriving from (5) is a binary cross entropy loss to guide the learning of the recognition controller, and ℓ_{CLS} deriving from (9) is the weighted classification loss to guide the learning of the slow classifier. ℓ_{ADV} is the adversarial loss for constraining the generator $\mathbb{G}$ and the discriminator $\mathbb{D}$, which is the same as ℓ_{ADV} in (10). The detailed optimization process is described in Algorithm 1. To better construct a comprehensive concept schema, we perform the basic learning stage and the extended learning stage alternately every 100 steps.

In the test phase, our model has constructed a comprehensive concept schema of the seen objects and unseen objects and can directly predict the segmentation results via $\mathbb{SC}$, with the fast classifier and the recognition controller removed. Given an image $\boldsymbol{I}^{s\cup u}$ which may contain seen classes and unseen classes as input, the final predicted segmentation map is formulated as $\boldsymbol{R} = \mathbb{SC}\left(\mathbb{E}\left(\boldsymbol{I}^{s\cup u}\right)\right)$.

3.6 Discussion

The meaningful learning method comprises two main operations, learning the new information through anchoring into existing comprehensible concepts, thus making the new-learned concept meaningful, and integrating the existing and new concepts to eventually promote a comprehensive understanding. We implement the meaningful learning method with the proposed conjugate conceptual correlation and the fast-slow concept modulator. To the best of our knowledge, this is the first endeavor to adopt the meaningful learning method to tackle the zero-shot learning problem. As for anchoring the new information into existing comprehensible concepts, most studies [53, 54] simply utilized the semantic-visual mapping trained on seen categories which are hard to apply to unseen categories to adapt to novel concepts. A recent study [40] for zero-shot classification attempted to regularize the correlation between existing concepts and novel ones which can be formulated as follows:

$$\begin{aligned}\ell = &\sum_{i,j} \left(\max(0, (\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_j - (\widetilde{\boldsymbol{\mu}}_i^{\mathrm{T}}\boldsymbol{\mu}_j - \alpha)))\right)^2 \\ &+ \sum_{i,j} \left(\max(0, (\widetilde{\boldsymbol{\mu}}_i^{\mathrm{T}}\boldsymbol{\mu}_j - (\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_j - \alpha)))\right)^2,\end{aligned} \tag{12}$$

where $\boldsymbol{w}_i, \boldsymbol{w}_j$ represent the semantic word embedding for the i-th and j-th class, while $\boldsymbol{\mu}_j$ refers to the prototype of j-th seen class and $\widetilde{\boldsymbol{\mu}}_i$ refers to the generated prototype for class i. Although Ref. [40] also could encourage the generated unseen visual representations connecting with the existing concepts,

they also have the over-correlation problem due to the noisy semantic word embeddings pre-trained with large-scale irrelevant corpus. Moreover, they merely utilized correlations of the prototype of each class to constrain the generator and ignored the diversity of visual instances in different spatial positions, and the fixed margin greatly limits the diversity of generated features. Therefore, it cannot well handle the zero-shot segmentation tasks. Recent CSRL [49] which is similar with [40] also merely utilizes correlations of the prototype of each class to constrain the generator and ignores the diversity of visual instances in different spatial positions, still facing the same problem. Contrastively, we introduce an FSCM to alleviate the noisy over-correlation problem and adopt an adaptive margin to take the spatial information into account and flexibly adjust the semantic-visual gap, thus ensuring visual diversity and improving the generalization ability. We also carry out experiments on ablation study with conjugate conceptual correlation to verify our point.

As for concepts integration of meaningful learning, some zero-shot studies [55,56] attempted to preserve existing concepts of seen classes via knowledge distillation when integrating the existing seen concepts and the novel unseen concepts for the final classification which can be formulated as follows:

$$\ell = \|\boldsymbol{\mu}_i^{s\cup u} - \boldsymbol{\mu}_i^{s}\|_1 + \|\boldsymbol{\sigma}_i^{s\cup u} - \boldsymbol{\sigma}_i^{s}\|_1 , \tag{13}$$

where $\boldsymbol{\mu}_i^s, \boldsymbol{\sigma}_i^s$ refer to the distribution of the i-th seen class predicted by origin model trained on seen classes, and $\boldsymbol{\mu}_i^{s\cup u}, \boldsymbol{\sigma}_i^{s\cup u}$ are the distribution of the i-th class predicted by the model fine-tuned with novel classes. These methods are computationally expensive and more likely to be biased towards the seen classes with a large number of annotated images, leading to unbalanced seen and unseen concepts in the classifier. Ref. [57] tried to directly map the novel concepts to the existing concepts learned from ImageNet, which also restricts the learning of novel concepts and cannot balance the seen and unseen concepts well. Contrastively, our fast-slow concept modulator simulates the thinking mechanism of the human brain and well balances the learning of novel concepts and adapting of existing concepts, leading to a new comprehensive concept schema. Combined with the G3C and FSCM, the proposed method is a completely meaningful learning method that is simple yet effective, and achieves superior performance on zero-shot semantic segmentation tasks.

4 Experiments

In this section, we conduct extensive experiments on three widely used datasets. First, we introduce the experimental setup. Then we compare our method with state-of-the-art zero-shot semantic segmentation approaches, after which we evaluate the effectiveness of the two crucial operations in meaningful learning. Subsequently, we provide the performance of each component in FSCM during extended learning. Eventually, we analyze the effect of conceptual correlations of semantic word embeddings.

4.1 Experimental setup

Datasets. We employ the widely used datasets in prior studies, including Pascal-VOC 2012 [18], Pascal-Context [19], and COCO-stuff [20]. The Pascal-VOC has 1464 training images with dense pixel-level annotations from 20 categories. The Pascal-Context dataset consists of 33 categories, a total of 4998 training images, and 5105 validation images. The COCO-stuff has a larger data size than Pascal-Context and Pascal-VOC, which contains 182 classes and 164k annotated images, and is more challenging. We adopt the data split setting of the recent work CaGNet [15] for all models, and retrain the state-of-the-art comparison methods with their published codes to guarantee a fair comparison. Specifically, we treat "potted plant, sheep, sofa, train, tv monitor" as 5 unseen categories and the other 15 categories as seen classes on Pascal-VOC. On Pascal-Context, we treat "cow, motorbike, sofa, cat" as 4 unseen categories. And "frisbee, skateboard, cardboard, carrot, scissors, suitcase, giraffe, cow, road, wallconcrete, tree, grass, river, clouds, playingfield" are treated as 15 unseen categories on COCO-stuff. We use two different word embedding models, word2vec [58] ($d = 600$) trained on Google News and fast-Text [59] ($d = 300$) trained on Common Crawl, and concatenate the two embeddings as the semantic supervision information, following previous work [13]. For categories with multiple words, we directly average the embeddings of each individual word. We adopt a generalized zero-shot semantic segmentation setting that both seen and unseen categories can appear in test samples, but we only use the pixel-wise annotations of seen categories and ignore the unseen pixels by marking the annotations of these pixels as 'ignored' during training.

Implementation details. As mentioned above, we implement the proposed model based on the recent state-of-the-art approach CaGNet [15] and inherit its original network modules. Because CaGNet has a more complete experimental setup, which is easy to reproduce the results in the paper. In particular, our model is built on the DeepLab-v2 [23] which adopts ResNet-101 as the backbone to extract features and applies an atrous spatial pyramid pooling layer to produce the visual features. Following SPNet [13] and CaGNet [15], the weights pre-trained on ImageNet are utilized to initialize the backbone network, but the unseen classes of the three datasets we employ have been specifically excluded in the pretraining phase, which does not violate the zero-shot constraint. The generator $\mathbb{G}$ is implemented by a multi-layer perceptron with 512 hidden neurons, Leaky ReLU, and dropout for each layer. The discriminator $\mathbb{D}$, the fast classifier $\mathbb{FC}$, and the slow classifier $\mathbb{SC}$ are implemented with two 1×1 convolutional layers, respectively, and share the parameters of the first layer. The recognition controller $\mathbb{RC}$ is implemented with one convolutional layer. The number of iterations in the basic and extended learning stages is set to 40k and 6k, respectively. We adopt the SGD optimizer with an initial learning rate 2.5e^{-4} for the feature embedding network, and two Adam optimizers for the generator and the discriminator with an initial learning rate 2e^{-4} and 2.5e^{-4}, respectively. The weight decay is set to 5e^{-4} and the mini-batch size is set to 8 for all experiments. As for the hyper-parameters, we set $\lambda = 10, \beta_1 = 0.5, \beta_2 = 2$ via cross-validation. We randomly crop the input images to 368×368 during training, and resize them to 513×513 when testing. Following CaGNet [15], the images are first resized to 513×513, and then randomly cropped to the size of 368×368 during training, which is a commonly-used data augmentation method, and no data augmentation is used in testing, keeping the image size at 513×513. For fair comparison, we keep the same setting on all comparison methods (SPNet and ZS3Net) with published code.

Evaluation metrics. We investigate the performance of different methods using the standard semantic segmentation metrics: pixel accuracy (PA), mean accuracy (MA), and mean intersection-over-union (mIoU). Moreover, following the previous work, we also report harmonic metrics (h), namely harmonic pixel accuracy (h-PA), harmonic mean accuracy (h-MA), and harmonic mean Intersection-over-Union (h-mIoU). PA is the proportion of correctly marked pixels to total pixels, while MA is the average of the proportion of correctly classified pixels in each class. mIoU calculates the degree of overlap between the predicted segmentation map and the ground truth. Although PA, MA, and mIoU are commonly used to evaluate semantic segmentation models, it should be emphasized that they tend to be usually dominated by seen classes and neglect the improvement of unseen classes because there is a serious imbalance between the number of seen and unseen classes in zero-shot semantic segmentation. Thus, these metrics may not fully reflect the overall improvement of zero-shot semantic segmentation methods objectively and fairly. On the contrary, we should pay more attention to the results of the harmonic metrics for the overall classes. The harmonic metric eliminates the impacts brought by the imbalance class number, via respectively considering the seen classes and unseen classes, which is defined as below:

$$\text{h-metric} = \frac{2 \times \text{s-metric} \times \text{u-metric}}{\text{s-metric} + \text{u-metric}}, \tag{14}$$

where "metric" could be mIoU, PA, or MA.

4.2 Comparison with state-of-the-arts

We investigate the performance of our model compared with state-of-the-art zero-shot semantic segmentation approaches: SPNet, ZS3Net, CaGNet, and SIGN. We carefully reproduce the methods which publish their code and provide both quantitative results and quality results. As for the latest SIGN which did not publish their code, we directly list the results they provided in their paper for fair comparison. We also report the results of SPNet-c which deducts the prediction scores of seen categories by a calibration factor. Moreover, we list the results with self-training (ST) strategy for all approaches following [14,15].

Table 2 lists the quantitative results obtained with all methods on COCO-stuff, Pascal-Context, and Pascal-VOC. Owing to the effective meaningful learning method, we can find that our method almost achieves peak performance in all cases, especially according to harmonic metrics for overall classes and metrics for unseen classes. According to harmonic evaluations, the proposed model reaches 0.2115, 0.2207, and 0.4540 h-mIoU, respectively on COCO-stuff, Pascal-Context, and Pascal-VOC datasets. Besides, on COCO-stuff dataset, our method almost brings 9% gains compared with the baseline model CaGNet on unseen classes according to PA and MA. Compared with the latest SIGN, our method can also achieve better performance on all datasets according to the most meaningful metric harmonic IoU. On the Pascal-VOC dataset, our method could even get a 4% performance improvement. It is worth noting that our

Table 2 Quantitative results of all zero-shot semantic segmentation methods on COCO-stuff, Pascal-Context, and Pascal-VOC datasets[a)]

Method	Overall-harmonic			Overall-mean		Seen			Unseen		
	h-mIoU	h-PA	h-MA	mIoU	MA	mIoU	PA	MA	mIoU	PA	MA
COCO-stuff											
SPNet	0.0145	0.0353	0.0148	0.3169	0.4650	0.3491	**0.6667**	**0.5126**	0.0074	0.0181	0.0075
SPNet-c	0.1607	0.3890	0.2655	**0.3324**	0.4450	**0.3563**	0.6237	0.4721	0.1037	0.2827	0.1847
ZS3Net	0.1670	0.3399	0.2228	0.3111	0.4627	0.3302	0.6326	0.4627	0.1118	0.2324	0.1467
CaGNet	0.1929	0.3960	0.3535	0.3224	**0.4643**	0.3405	0.6336	0.4821	0.1346	0.2880	0.2790
SIGN	0.2093	–	–	–	–	0.3231	–	–	**0.1547**	–	–
Ours	**0.2115**	**0.4653**	**0.4102**	0.3251	0.4617	0.3416	0.5956	0.4712	0.1532	**0.3818**	**0.3632**
ZS3Net + ST	0.1708	0.4589	0.2217	0.3124	0.4315	0.3314	**0.6425**	0.4589	0.1151	0.3506	0.1462
CaGNet + ST	0.1947	0.4265	0.3784	0.3130	0.4481	0.3299	0.6162	0.4604	0.1381	0.3261	0.3213
SIGN + ST	**0.2139**	–	–	–	–	**0.3639**	–	–	0.1515	–	–
Ours + ST	0.2131	**0.4761**	**0.4066**	**0.3241**	**0.4663**	0.3404	0.5990	**0.4771**	**0.1551**	**0.3950**	**0.3543**
Pascal-Context											
SPNet	0.0013	0.0012	0.0013	0.2962	0.4414	0.3370	**0.6281**	**0.5022**	0.0006	0.0006	0.0006
SPNet-c	0.1134	0.3300	0.3400	0.3055	0.4156	0.3382	0.5670	0.4419	0.0682	0.2328	0.2246
ZS3Net	0.1430	0.2761	0.1927	0.3068	0.4540	0.3360	0.6167	0.5002	0.0908	0.1779	0.1194
CaGNet	0.2028	0.4671	0.4080	0.3176	0.4745	0.3415	0.5883	0.4917	0.1443	0.3873	0.3493
SIGN	0.2067	–	–	–	–	0.3367	–	–	0.1493	–	–
Ours	**0.2207**	**0.5131**	**0.4378**	**0.3202**	**0.4678**	**0.3419**	0.5845	0.4764	**0.1629**	**0.4572**	**0.4050**
ZS3Net + ST	0.1247	0.3026	0.1752	0.2995	0.4293	0.3302	**0.6133**	0.4736	0.0768	0.2008	0.1075
CaGNet + ST	0.2266	0.4795	0.4214	0.3212	**0.4809**	0.3421	0.5996	**0.4968**	0.1694	0.3995	0.3659
SIGN + ST	0.2260	–	–	–	–	**0.3491**	–	-	0.1671	–	–
Ours + ST	**0.2399**	**0.5291**	**0.4430**	**0.3231**	0.4710	0.3422	0.5919	0.4792	**0.1847**	**0.4782**	**0.4119**
Pascal-VOC											
SPNet	0.00001	0.00002	0.00001	0.5703	0.7043	0.7603	0.9476	0.9391	0	0	0
SPNet-c	0.0977	0.1316	0.1091	0.5467	0.6928	0.7527	**0.9588**	**0.9574**	0.0523	0.0706	0.0579
ZS3Net	0.2825	0.3665	0.3668	0.6143	0.7307	0.7613	0.9128	0.8974	0.1734	0.2293	0.2305
CaGNet	0.4036	0.6275	0.5921	0.6423	0.7571	0.7650	0.8639	0.8588	0.2741	0.4927	0.4518
SIGN	0.4174	–	–	–	–	0.7540	–	–	0.2886	–	–
Ours	**0.4540**	**0.6602**	**0.6225**	**0.6702**	**0.7689**	**0.7872**	0.8791	0.8629	**0.3190**	**0.5286**	**0.4868**
ZS3Net + ST	0.3312	0.4074	0.4220	0.6336	0.7503	0.7781	**0.9185**	**0.9088**	0.2104	0.2617	0.2748
CaGNet + ST	0.4547	0.7091	0.6678	0.6621	0.7675	0.7756	0.8482	0.8384	0.3217	0.6091	0.5549
SIGN + ST	0.4661	–	–	–	–	**0.7862**	–	–	0.3312	–	–
Ours + ST	**0.4791**	**0.7241**	**0.6712**	**0.6734**	**0.7804**	0.7829	0.8675	0.8557	**0.3451**	**0.6213**	**0.5517**

a) The bold indicates the best performance, and the underline indicates the second best performance. "ST" stands for self-training.

method achieves a similar performance with SIGN on the unseen class of the COCO-stuff dataset. It is because SIGN cannot balance the knowledge of seen and unseen classes well. Therefore, although it achieves the same performance as ours in unseen classes, it sacrifices the performance of seen classes, so it is inferior to ours in h-mIoU. It fully illustrates that the proposed model could well learn the novel concepts and generate meaningful unseen visual representations due to the valuable guidance by the meaningful existing concepts provided by conjugate conceptual correlations. In the meanwhile, it can obtain competitive results on seen classes, due to the fast-slow concept modulator that integrates the existing and new concepts entirely. Moreover, the "ST" is a useful strategy in semi-supervised learning that leverages a model's own predictions on unlabelled data to heuristically obtain additional pseudo annotated training data. It is a relaxed zero-shot setup where unlabelled pixels from unseen classes are already available at training time and can always improve the performance for zero-shot segmentation task. With the useful "ST" strategy, all three models could enhance their segmentation performance according to the h-mIoU metric, and the proposed model still outperforms the others on the most cases, proving the robustness of our model. However, with "ST" strategy, the mIoU performance for overall classes on COCO-stuff is even worse, we believe that the core reason is that COCO-stuff has a long-tail distribution and the model's prediction for unseen classes is not precise. Therefore, the obtained pseudo-labels have a serious long-tail problem, and the bias towards unseen classes further

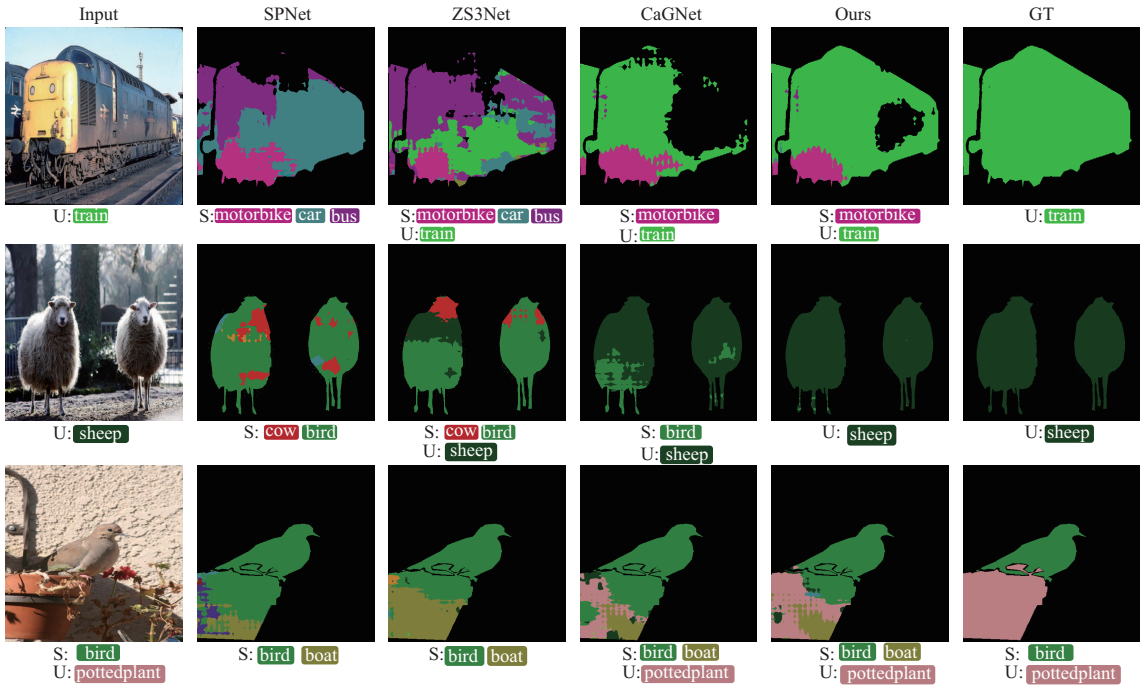

Figure 4 (Color online) Visualization of zero-shot semantic segmentation results on Pascal-VOC. "GT" is ground-truth segmentation mask. "U" means unseen class, and "S" means seen class.

worsens the results. But our meaningful learning method not only does not lead to worse results of self-training, on the contrary, but it has also greatly alleviated the problem compared to the baseline method by better integrating the concepts of seen and unseen classes to balance the predictions. As seen in Table 2, Ours+ST improves the h-mIoU by 1.84% on COCO-stuff dataset relative to the baseline method CaGNet+ST. In particular, SPNet and SPNet-c usually show favorable performance on seen classes due to the abundant data of seen classes, while they barely recognize unseen classes, which is against the intention of zero-shot learning.

Figure 4 shows the visualization of zero-shot segmentation results on Pascal-VOC. For fair comparison, we only provide the results of methods with a published code. The first two lines only contain unseen classes ("sheep" and "train"), while the last row involves both unseen classes ("pottedplant") and seen classes ("bird"). It can be observed that our model delivers the closest segmentation performance to the ground-truth, proving the superiority of the proposed meaningful learning method. Besides, it achieves the lowest recognition error, compared with the other approaches where the pixels inside a sheep are classified to "bird" and "cow". Moreover, the zero-shot segmentation approaches tend to categorize the pixels into seen classes due to the imbalanced data. For example, in the third row, previous methods classify the unseen class "pottedplant" as the seen classes "bird" and "boat". The proposed meaningful learning method somehow alleviates that problem and thus yields competitive results.

From these observations, we can conclude that our meaningful learning method adequately generates meaningful unseen visual representations and constructs a comprehensive concept schema of previous seen classes and novel unseen classes, finally achieving decent zero-shot semantic segmentation performance. Moreover, since all methods use DeepLab-v2 as the segmentation network, they have the same inference cost.

4.3 Ablation study

In this subsection, we first study the effects of different parts in the proposed model and then provide a detailed analysis of the proposed conceptual correlation and the fast-slow concept modulator. We report the common-used evaluation metrics: h-mIoU, mIoU of all classes (o-mIoU), mIoU of seen classes (s-mIoU), and mIoU of unseen Classes (u-mIoU).

Validation of network modules. To demonstrate the effectiveness of different components in our method, Figure 5(a) respectively shows the evaluation of our modules from different perspectives. Specifically, the four sub-figures sequentially present the results of the baseline model (I), the baseline model

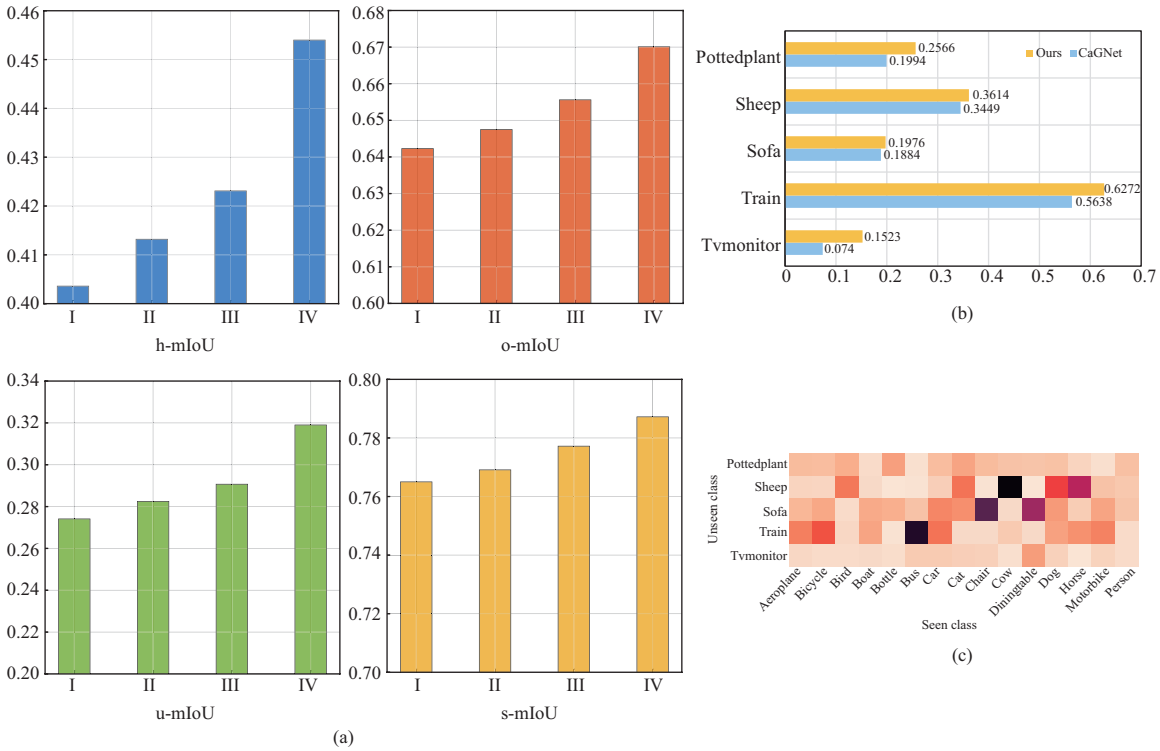

Figure 5 (Color online) (a) Ablation study on Pascal-VOC. It respectively shows the mIoU metrics of the baseline model (I), the baseline model with the conjugate conceptual correlation (II), the baseline model with the fast-slow concept modulator (III), and the full model (IV). (b) The mIoU performance of CaGNet and ours on each unseen class. (c) The semantic similarity between the unseen and seen word embeddings. The darker color indicates a higher similarity.

Table 3 Ablation study of different implementations of conjugate conceptual correlation loss on Pascal-VOC. It shows the results using the method without conjugate conceptual correlation (w/o ℓ_{CON}), the naive conjugate conceptual correlation without a margin (ℓ_{CON} w/o $\boldsymbol{\alpha}$), the conjugate conceptual correlation with a fixed margin (ℓ_{CON} with a fixed α), and the conjugate conceptual correlation with the adaptive margin (G3C), respectively[a)]

	h-mIoU	o-mIoU	s-mIoU	u-mIoU
w/o ℓ_{CON}	0.4036	0.6423	0.7650	0.2741
ℓ_{CON} w/o $\boldsymbol{\alpha}$	0.4026	0.6419	0.7642	0.2733
ℓ_{CON} with a fixed α	0.3934	0.6332	0.7556	0.2659
G3C	**0.4132**	**0.6475**	**0.7691**	**0.2826**

a) The bold indicates the best performance.

with the conjugate conceptual correlation (II), the baseline model with the fast-slow concept modulator (III), and the full model (IV).

From Figure 5(a), we can see that each component brings a performance gain compared with the baseline model. The proposed model achieves the best performance, according to all evaluation protocols, which validates the effectiveness of the meaningful learning method. According to s-mIoU and u-mIoU, we can see that, although building the correlations would benefit the comprehension of concepts for both seen classes and unseen classes, without the fast-slow concept modulator, the novel concepts learned by model II cannot handle the potential over-correlation or well accommodate the concept schema and thus could not achieve peak performance. Similarly, even though model III discreetly controls the integration of existing concepts and new ones, without establishing the valuable correlations, the model just conducts rote learning where novel concepts are hardly comprehensible and meaningful to such a model. Only combining the two contributions can the model learn in a meaningful way and form a compatible concept schema. Moreover, it can be observed that, the increases of h-mIoU and u-mIoU, nearly 5%, are more obvious compared with those of o-mIoU and s-mIoU, about 2%, indicating that the proposed model mainly devotes to the learning of unseen classes. However, the new concept schema also improves its distinguishable ability for seen classes from 0.7650 to 0.7872, due to the comprehensive understanding on both seen and unseen classes.

Analysis of conjugate conceptual correlation. The conjugate conceptual correlation establishes the valuable correlations between existing concepts and the new ones, and regularizes the generator to generate meaningful unseen visual representations. However, the gap between the semantic and visual space makes it hard to simply align the correlations in different spaces without any special designs. Therefore, in Table 3, we carefully make comparisons between (1) the generator without conjugate conceptual correlation (w/o ℓ_{CON}) and that of three margin settings which are adopted to eliminate the semantic-visual gap, (2) the naive conjugate conceptual correlation without a margin (ℓ_{CON} w/o $\boldsymbol{\alpha}$), (3) the conjugate conceptual correlation with a fixed margin (ℓ_{CON} with a fixed α), similar to the idea of [49] discussed in Subsection 3.5, and (4) the conjugate conceptual correlation with the adaptive margin (G3C).

As shown in Table 3, removing conjugate conceptual correlation has little impact on seen classes according to o-mIoU and s-mIoU, while causing about 1% drop on unseen classes. It well confirms our point that introducing conjugate conceptual correlation would help the learning of unseen classes. Without an appropriate margin, the performance nearly stays the same or becomes even worse. Moreover, with the improperly fixed margin, it witnesses 1% drop even on seen classes, meaning that it hurts the semantic-visual mapping ability of the generator. Intuitively, semantic word embeddings are the highly abstract class prototypes, while the visual instances contain both the class-specific commonness and the individual uniqueness, so the margin should be introduced to bridge the semantic-visual gap. We can observe that the variation of o-mIoU in Table 3 is essentially the same as the variation of s-mIoU, which does not reflect the larger boost in u-mIoU when the adaptive margin is used. This observation also demonstrates the weakness that the overall mIoU is usually dominated by the seen classes and ignores the boost on unseen classes. The effectiveness of the adaptive margin could be measured more objectively by comparing the h-mIoU metric and our method achieves a 1.06% improvement on h-mIoU when the adaptive margin is used. Therefore, adaptive margins should be introduced to encourage the diversity of the visual space. Only in this way can the model capture the valuable correlations and construct reasonable semantic-visual mapping, thus generating meaningful unseen visual representations.

Analysis of conceptual correlation. Figure 5(b) shows the mIoU performance of CaGNet and ours for each unseen class on the Pascal-VOC dataset and Figure 5(c) shows the similarity between seen classes and unseen classes. It can be observed that the proposed method improves the performance on each unseen class compared with CaGNet. Especially on the "tvmonitor" class, CaGNet barely segments it correctly and the mIoU of "tvmonitor" class is only 0.074. However, our method can reach 0.1523 mIoU of "tvmonitor" class which is a massive boost. Combining Figures 5(b) and (c), we can observe that if the unseen class is similar to some seen classes, its mIoU performance will be satisfactory. Such as the "train" class, on which the proposed method gains 6% mIoU performance, is highly similar to the seen class "bus". The experimental results fully validate that our method could establish the valuable correlations between classes and thus effectively anchor the existing seen visual representations and generate meaningful unseen visual representations, benefiting from the conjugate conceptual correlation in the meaningful learning method.

Analysis of fast-slow concept modulator. In this subsection, we explore some different ways of integrating the existing concepts and new ones and report the experimental results in Table 4. The first two settings drop the fast-slow mechanism and get entangled in training seen and unseen classes. The last three settings investigate different implementations for fast-slow thinking mechanism. From Table 4, we can observe that whatever implementation of fast-slow thinking mechanism performs better than dropping the mechanism, which fully demonstrates the superiority of the proposed module. As for the models abandoning the mechanism, fixing the first layer maintains the existing experience and thus delivers the best performance on seen classes. Still it cannot correct the noisy over-correlation problem caused by G3C and construct a comprehensive concept schema for segmentation. Fine-tuning all the parameters of the extended classifier could help learn the novel concepts at the cost of injuring previous information, with a nearly 2% performance drop on s-mIoU. In the fast-slow concept modulator, the recognition controller behind the first layer can also contribute to concept transfer, and increase the h-mIoU by 2% compared with the "$\mathbb{SC}$" setting. However, the generalization ability of the high-dimensional feature vector is far less than that of the low-dimensional score distribution due to the noise information contained in high-dimension vectors, which also leads to the decline of the model's performance compared with the full FSCM. Although the third experimental setting adopts the fast-slow mechanism, it witnesses a 1% drop on h-mIoU, o-mIoU, and u-mIoU, because the existing concept schema contained in the fixed fast classifier cannot be adapted to the new observations. Thus, the novel concepts

Table 4 Ablation study of fast-slow concept modulator on Pascal-VOC. It shows the results of abandoning the fast-slow thinking mechanism and fine-tuning the classification layer of the slow classifier with other parameters fixed ($\mathbb{SC}$ + 1-st layer fixed), abandoning the fast-slow thinking mechanism and simply fine-tuning all the parameters of the slow classifier ($\mathbb{SC}$), adopting fast-slow concept modulator and fine-tuning all the parameters of the slow classifier with the fast classifier fixed (FSCM + fixed $\mathbb{FC}$), adopting fast-slow concept modulator and deploying the recognition controller before the classification layer of the fast classifier (FSCM + $\mathbb{RC}_f$), and the overall fast-slow concept modulator (FSCM), respectively[a)]

	h-mIoU	o-mIoU	s-mIoU	u-mIoU
$\mathbb{SC}$ + 1-st layer fixed	0.3320	0.6432	**0.7875**	0.2103
$\mathbb{SC}$	0.4132	0.6475	0.7691	0.2825
FSCM + fixed $\mathbb{FC}$	0.4423	0.6652	0.7842	0.3081
FSCM + $\mathbb{RC}_f$	0.4322	0.6626	0.7841	0.2980
FSCM	**0.4504**	**0.6701**	0.7872	**0.3190**

a) The bold indicates the best performance.

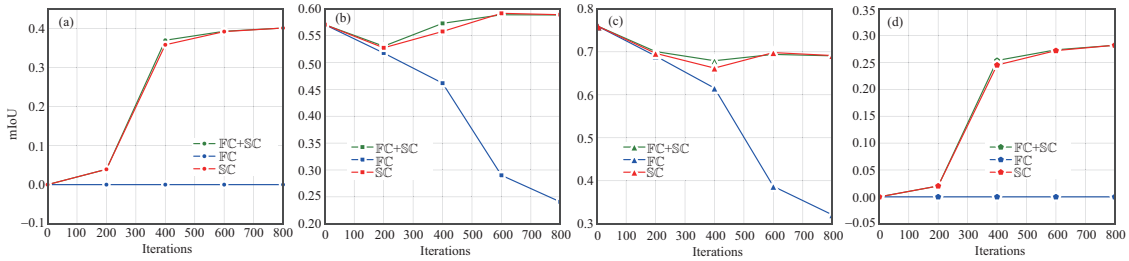

Figure 6 (Color online) Visualization of the extended learning stage on Pascal-VOC of mIoU, respectively. The model is evaluated on the test dataset every 200 iterations. (a) h-mIoU; (b) o-mIoU; (c) s-mIoU; (c) u-mIoU.

may be mistakenly classified as seen classes and cannot well accommodate in such a concept schema. Therefore, the implementation of ours is reasonable and indeed yields the favorable performance.

4.4 Analysis of the extended learning process

In this subsection, we analyze the extended learning of each module in the fast-slow concept modulator. We have presented the performance of the overall modulator in Section 1 where the full model could evenly handle the previous experience and novel concepts. This subsection looks deep into the modulator and investigates $\mathbb{FC}$, $\mathbb{SC}$, and the overall FSCM. Figure 6 provides the h-mIoU, o-mIoU, s-mIoU, and u-mIoU performance of $\mathbb{FC}$, $\mathbb{SC}$, and the overall FSCM, respectively. Since the learning of novel concepts happens rapidly, making it hard to distinguish the $\mathbb{SC}$ which becomes the final comprehensive concept schema and the overall modulator, we record the performance every 100 iterations and only provide the previous 800 iterations.

First, we can observe that the overall modulator FSCM outperforms other modules under all settings, but it is hard to distinguish it from $\mathbb{SC}$ due to the fast learning of novel concepts, except the o-mIoU and s-mIoU where seen classes play a more important role than unseen classes due to the computation of evaluation metrics. The h-mIoU and u-mIoU performance of $\mathbb{SC}$ and FSCM is continuously increasing, while s-mIoU and o-mIoU first drop and then increase. After 800 iterations, they will reach the previous level or even better as shown in Figure 1, indicating that the proposed model could preserve the existing knowledge or better understand it due to the novel observations. Moreover, the fast classifier cannot predict unseen classes, and thus the h-mIoU and u-mIoU performance remain zero. Besides, with incompatible novel observations, the whole system FSCM alters its perceiving approach, and thus the rigid $\mathbb{FC}$ even cannot be directly applied to existing concepts, leading to declining performance during extended learning. Actually, as the learning proceeds, the recognition controller $\mathbb{RC}$ gradually loses its efficacy and allows increasing seen samples to be fed in $\mathbb{SC}$. $\mathbb{SC}$ starts to accommodate the new information evenly and fine-tune the existing concept schema, eventually forming a comprehensive schema. As a result, the performance of $\mathbb{SC}$ becomes similar to the overall modulator. The existence of $\mathbb{FC}$ and $\mathbb{RC}$ effectively prevent the noisy semantic word embeddings pre-trained with large-scale irrelevant corpus from interfering the learning of novel information, and thus help the proposed model modulate the previous and current concepts.

Table 5 The mIoU performance of vanilla CaGNet/ZS3Net and CaGNet/ZS3Net applying the meaningful learning method on Pascal-VOC[a)]

	h-mIoU	o-mIoU	s-mIoU	u-mIoU
CaGNet	0.4036	0.6423	0.7650	0.2741
Ours (CaGNet)	**0.4540**	**0.6702**	**0.7827**	**0.3190**
ZS3Net	0.2825	0.6143	0.7613	0.1734
Ours (ZS3Net)	**0.3221**	**0.6432**	**0.7901**	**0.2023**

a) The bold indicates the best performance.

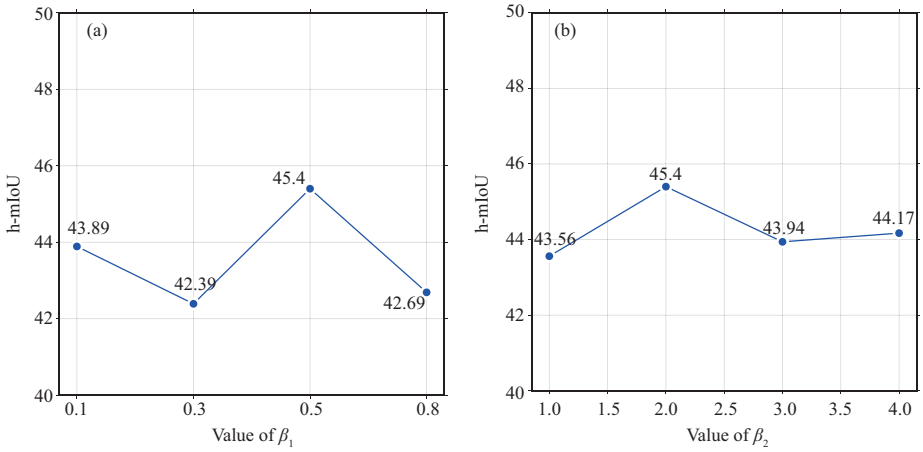

Figure 7 (Color online) Sensitivity analysis of β_1 and β_2 in Pascal-VOC dataset. β_1 varies in $[0.1, 0.8]$ and β_2 varies in $[1, 4]$.

4.5 Analysis of generality

In this subsection, we analyze the generality of the proposed meaningful learning method on two open-source generation-based methods: CaGNet and ZS3Net in Table 5. In Subsection 4.1, we have demonstrated that the proposed method could greatly improve the performance of the CaGNet model and achieve superior performance compared with the state-of-the-art methods. In this subsection, we further apply the meaningful learning method to ZS3Net and conduct additional experiments on Pascal-VOC to validate its generality. As shown in Table 5, ZS3Net applies the generator with conjugate conceptual correlation and the fast-slow concept modulator shows a 3.96% improvement on h-mIoU and a 2.89% improvement on o-mIoU, boosting the overall semantic segmentation performance of vanilla ZS3Net. Specifically, on both seen and unseen classes, the meaningful learning method could construct a comprehensive concept schema and bring significant improvement according to s-mIoU and u-mIoU in Table 5. All these results demonstrate that our method is general and effective that can be embedded into any generation-based zero-shot segmentation model, not by careful tuning to improve a particular baseline model.

4.6 Analysis of hyper-parameters

In Figure 7, we investigate the sensitivity of hyper-parameter β_1 and β_2 which are depicted in (8). As mentioned in Subsection 3.3, β_1 and β_2 are introduced as a balancing factor to restrict the severity of the penalty, concentrating mostly on discriminative unseen samples and indiscriminative seen samples. In the early time of the extend learning stage, the model has a high accuracy rate for seen classes, so it is intuitive to make $\beta_2 > \beta_1$, encouraging the model concentrating mostly on discriminative unseen samples and indiscriminative seen samples. We carry experiments on the Pascal-VOC dataset, and vary β_1 to [0.1, 0.8] and vary β_2 within the range $[1, 4]$. The h-mIoU is reported in Figure 7. It can be easily included that with the change of the our introduced hyper-parameters, although the fluctuation of h-mIoU is also around 2% with β_1 and 1% with β_2, it is relatively small compared with our 5% improvement over CaGNet. Therefore, it can be proven that our method has better robustness to the introduced hyper-parameters.

5 Conclusion

In this paper, we propose a novel meaningful learning method for zero-shot semantic segmentation borrowing the idea from the educational psychology field. Concretely, we introduce a generator with G3C to generate meaningful unseen visual representation through anchoring the novel concepts into the existing comprehensible concepts. Furthermore, simulating the rational thinking mechanism of the human brain, we design a fast-slow concept modulator to alleviate the noisy over-correlation problem introduced by G3C and further construct a comprehensive concept schema for segmentation. Extensive experiments conducted on Pascal-VOC, Pascal-Context, and COCO-stuff datasets demonstrate that our meaningful learning method could greatly improve existing zero-shot semantic segmentation methods and achieve superior segmentation performance compared to the state-of-the-art. The exhaustive ablation studies validate the rationality and effectiveness of each component. And, we also testify the generality of the proposed method on generation-based zero-shot semantic segmentation models like CaGNet and ZS3Net. Finally, we analyze the sensitivity of the hyper-parameters.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62206010, 62022009).

References

1 Feng J P, Wang X G, Liu W Y. Deep graph cut network for weakly-supervised semantic segmentation. Sci China Inf Sci, 2021, 64: 130105
2 Zhang Z J, Pang Y W. CGNet: cross-guidance network for semantic segmentation. Sci China Inf Sci, 2020, 63: 120104
3 Zhou Q, Wang Y, Liu J, et al. An open-source project for real-time image semantic segmentation. Sci China Inf Sci, 2019, 62: 227101
4 Li W X, Lin N, Zhang M Z, et al. VNet: a versatile network to train real-time semantic segmentation models on a single GPU. Sci China Inf Sci, 2022, 65: 139105
5 Peng H T, Zhou B, Yin L Y, et al. Semantic part segmentation of single-view point cloud. Sci China Inf Sci, 2020, 63: 224101
6 Chen L J, Xiao Y, Yuan X M, et al. Robust autonomous landing of UAVs in non-cooperative environments based on comprehensive terrain understanding. Sci China Inf Sci, 2022, 65: 212202
7 Wang J L, Lu Y H, Liu J B, et al. A robust three-stage approach to large-scale urban scene recognition. Sci China Inf Sci, 2017, 60: 103101
8 Chen S T, Jian Z Q, Huang Y H, et al. Autonomous driving: cognitive construction and situation understanding. Sci China Inf Sci, 2019, 62: 081101
9 Wang L F, Yu Z Y, Pan C H. A unified level set framework utilizing parameter priors for medical image segmentation. Sci China Inf Sci, 2013, 56: 110902
10 Xu Q, Xi X M, Meng X J, et al. Difficulty-aware bi-network with spatial attention constrained graph for axillary lymph node segmentation. Sci China Inf Sci, 2022, 65: 192102
11 Liu F, Li H B. Joint sparsity and fidelity regularization for segmentation-driven CT image preprocessing. Sci China Inf Sci, 2016, 59: 032112
12 Kato N, Yamasaki T, Aizawa K. Zero-shot semantic segmentation via variational mapping. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019
13 Xian Y, Choudhury S, He Y, et al. Semantic projection network for zero-and few-label semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 8256–8265
14 Bucher M, Tuan-Hung V, Cord M, et al. Zero-shot semantic segmentation. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 468–479
15 Gu Z, Zhou S, Niu L, et al. Context-aware feature generation for zero-shot semantic segmentation. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020. 1921–1929
16 McCloskey M, Cohen N J. Catastrophic interference in connectionist networks: the sequential learning problem. Psychol Learning Motiv, 1989, 24: 109–165
17 Kahneman D. Thinking, Fast and Slow. London: Macmillan, 2011
18 Everingham M, Eslami S M A, van Gool L, et al. The pascal visual object classes challenge: a retrospective. Int J Comput Vis, 2015, 111: 98–136
19 Mottaghi R, Chen X, Liu X, et al. The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, 2014. 891–898
20 Caesar H, Uijlings J, Ferrari V. COCO-stuff: thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1209–1218
21 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2017, 40: 834–848
22 Lin G, Milan A, Shen C, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1925–1934
23 Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2881–2890
24 Zhang Z, Chen A, Xie L, et al. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019. 2124–2132
25 Pei G, Shen F, Yao Y, et al. Hierarchical feature alignment network for unsupervised video object segmentation. In: Proceedings of European Conference on Computer Vision, 2022. 596–613
26 Yao Y, Chen T, Xie G S, et al. Non-salient region object mining for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 2623–2632
27 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3431–3440

28 Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. 2014. ArXiv:1412.7062
29 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015. 234–241
30 Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 801–818
31 Fu Y, Hospedales T M, Xiang T, et al. Transductive multi-view embedding for zero-shot recognition and annotation. In: Proceedings of European Conference On Computer Vision. Berlin: Springer, 2014. 584–599
32 Chen S, Hong Z, Xie G, et al. GNDAN: graph navigated dual attention network for zero-shot learning. IEEE Trans Neural Netw Learn Syst, 2022. doi: 10.1109/TNNLS.2022.3155602
33 Xu B, Zeng Z, Lian C, et al. Generative mixup networks for zero-shot learning. IEEE Trans Neural Netw Learn Syst, 2022. doi: 10.1109/TNNLS.2022.3142181
34 Yu Y, Li B, Ji Z, et al. Knowledge distillation classifier generation network for zero-shot learning. IEEE Trans Neural Netw Learn Syst, 2023, 34: 3183–3194
35 Ji Z, Sun Y, Yu Y L, et al. Attribute-guided network for cross-modal zero-shot hashing. IEEE Trans Neural Netw Learn Syst, 2020, 31: 321–330
36 Ji Z, Yu X J, Yu Y L, et al. Semantic-guided class-imbalance learning model for zero-shot image classification. IEEE Trans Cybern, 2021, 52: 6543–6554
37 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. 2014. ArXiv:1406.2661
38 Xian Y, Lorenz T, Schiele B, et al. Feature generating networks for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 5542–5551
39 Felix R, Reid I, Carneiro G, et al. Multi-modal cycle-consistent generalized zero-shot learning. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 21–37
40 Li J, Jing M, Lu K, et al. Leveraging the invariant side of generative zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 7402–7411
41 Ji Z, Yan J T, Wang Q, et al. Triple discriminator generative adversarial network for zero-shot image classification. Sci China Inf Sci, 2021, 64: 120101
42 Bai G R, He S Z, Liu K, et al. Example-guided stylized response generation in zero-shot setting. Sci China Inf Sci, 2022, 65: 149103
43 Xian Y, Sharma S, Schiele B, et al. F-VAEGAN-D2: a feature generating framework for any-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 10275–10284
44 Xie G S, Zhang Z, Liu G S, et al. Generalized zero-shot learning with multiple graph adaptive generative networks. IEEE Trans Neural Netw Learn Syst, 2022, 33: 2903–2915
45 Zou Q, Cao L, Zhang Z, et al. Transductive zero-shot hashing for multilabel image retrieval. IEEE Trans Neural Netw Learn Syst, 2022, 33: 1673–1687
46 Huang P, Han J, Cheng D, et al. Robust region feature synthesizer for zero-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 7622–7631
47 Zhang D W, Guo G Y, Zeng W Y, et al. Generalized weakly supervised object localization. IEEE Trans Neural Netw Learn Syst, 2022. doi: 10.1109/TNNLS.2022.3204337
48 Zhao H, Puig X, Zhou B, et al. Open vocabulary scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2002–2010
49 Li P, Wei Y, Yang Y. Consistent structural relation learning for zero-shot segmentation. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33
50 Hu P, Sclaroff S, Saenko K. Uncertainty-aware learning for zero-shot semantic segmentation. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 21713–21724
51 Pastore G, Cermelli F, Xian Y, et al. A closer look at self-training for zero-label semantic segmentation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021. 2687–2696
52 Cheng J, Nandi S, Natarajan P, et al. SIGN: spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 9556–9566
53 Zhu Y, Elhoseiny M, Liu B, et al. A generative adversarial approach for zero-shot learning from noisy texts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1004–1013
54 Elhoseiny M, Elfeki M. Creativity inspired zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 5784–5793
55 Wei K, Deng C, Yang X. Lifelong zero-shot learning. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020. 551–557
56 Gautam C, Parameswaran S, Mishra A, et al. Generalized continual zero-shot learning. 2020. ArXiv:2011.08508
57 Liu Q, Xie L, Wang H, et al. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 3662–3671
58 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of Advances in Neural Information Processing Systems, 2013. 3111–3119
59 Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification. 2016. ArXiv:1607.01759