Remember the Difference: Cross-Domain Few-Shot Semantic Segmentation via Meta-Memory Transfer

Wenjian Wang,[1,2,3] Lijuan Duan,[1,2,3] Yuxi Wang,[5] Qing En,[4] Junsong Fan,[5] Zhaoxiang Zhang[5,6*]
[1]Faculty of Information Technology, Beijing University of Technology,
[2]Beijing Key Laboratory of Trusted Computing,
[3]National Engineering Laboratory for Key Technologies of Information Security Level Protection, China
[4]School of Computer Science, Carleton University, Canada
[5]Centre for Artificial Intelligence and Robotics, (HKISI_CAS)
[6]Institute of Automation, Chinese Academy of Sciences (NLPR, CASIA, UCAS)
wangwj@emails.bjut.edu.cn, ljduan@bjut.edu.cn, yuxiwang93@gmail.com
QingEn@cunet.carleton.ca, {fanjunsong2016, zhaoxiang.zhang}@ia.ac.cn

2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-6946-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPR52688.2022.00693

Abstract

Few-shot semantic segmentation intends to predict pixel-level categories using only a few labeled samples. Existing few-shot methods focus primarily on the categories sampled from the same distribution. Nevertheless, this assumption cannot always be ensured. The actual domain shift problem significantly reduces the performance of few-shot learning. To remedy this problem, we propose an interesting and challenging cross-domain few-shot semantic segmentation task, where the training and test tasks perform on different domains. Specifically, we first propose a meta-memory bank to improve the generalization of the segmentation network by bridging the domain gap between source and target domains. The meta-memory stores the intra-domain style information from source domain instances and transfers it to target samples. Subsequently, we adopt a new contrastive learning strategy to explore the knowledge of different categories during the training stage. The negative and positive pairs are obtained from the proposed memory-based style augmentation. Comprehensive experiments demonstrate that our proposed method achieves promising results on cross-domain few-shot semantic segmentation tasks on COCO-20^i , PASCAL-5^i, FSS-1000, and SUIM datasets.

1. Introduction

Recently, semantic segmentation [2, 19, 45] has made remarkable progress benefiting from the large amounts of human-annotated datasets and deep convolutional neural networks [11]. However, obtaining these large annotated datasets is time-consuming and labor-intensive, and the trained model always fails to segment novel unseen categories. To tackle this problem, few-shot semantic segmentation methods [6, 42, 47] have received a lot of attention in recent years, aiming to produce pixel-level predictions for the novel categories when given only few (one) training images.

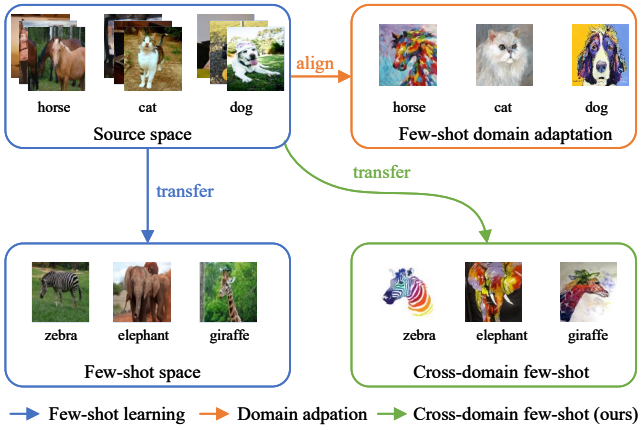

Figure 1. Comparison with few-shot segmentation and few-shot domain adaptation segmentation. In few-shot learning, the training set and testing sets come from the same domain. In few-shot domain adaptation, the training set and testing sets are from separate domains, but the label space is the same. In our cross-domain few-shot segmentation task, the training set and testing sets come from different domains, and the categories in the testing set are unseen.

Few-shot semantic segmentation methods [1, 16] address the challenge of adapting a segmentation network to a new

*Corresponding Author

978-1-6654-6946-3/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPR52688.2022.00693

category represented by few support images via a meta-learning [9, 43], which enables the model to transfer the knowledge from the support set to the query set. Existing approaches assume that the base training set is sampled from the same domain as the testing set. However, collecting sufficient examples in specific areas, such as dermatology or satellite imagery, is infeasible or impossible. Alternatively, we train a satisfactory few-shot semantic segmentation network for the test dataset (target domain) by transferring the knowledge from the existing base training dataset (source domain). Consequently, we formulate a new but essential problem called cross-domain few-shot semantic segmentation.

Recently, a tiny number of cross-domain few-shot methods [21, 31, 34] have been developed to tackle the similar issue on image classification. However, the proposed methods are hard to apply to our scenario because the pixel-level segmentation is fundamentally different from the image-level classification. Another close work is domain adaptation [18, 26, 40] or few-shot domain adaptation semantic segmentation [46, 49], which eliminates domain shift issue with few-shot labeled target images. But our setting is more challenging because we have disjoint categories between the source and target domains. The comparison is shown in Figure 1.

The fundamental challenges for cross-domain few-shot semantic segmentation lie in two aspects. (1) There is a domain shift problem between the training and testing tasks due to sampling from different domains. It leads to performance degrades significantly for conventional few-shot semantic segmentation methods. (2) The labeled data of new categories is scarce, so the fine-tuning or distribution alignment methods are challenging. In this paper, we propose a novel cross-domain few-shot semantic segmentation framework, named as **CDFSS**. Our framework mainly focuses on reducing the domain gap by domain generalization and exploring the discriminative information from the few-shot novel categories. Specifically, we first construct a meta-memory to collect domain-specific information among source domain instances. The source instances continuously register the stylized domain information into the meta-memory as the training progresses. Aggregated style information can effectively describe the source data distribution and enhance the target domain features. Then, we load the stored memory into both the source and target domain to enhance the model's generalization. For the meta-training stage on the source domain, the loaded memory is mainly responsible for augmenting the features stylization, and the model is encouraged to produce consistent representations despite the style differences. In the test stage on the target domain, the model loads the source meta-knowledge to guide the feature enhancement and alleviate the domain gap in cross-domain problems. Memory-enhanced features help novel categories generate more diversified prototypes so that the model can provide robust predictions. Moreover, we adopt the contrastive loss using the memory enhanced features to further constrain the prototypes between categories to improve the model's adaptability in few-shot learning.

The contributions of this article are summarized as follows:

1. We propose a novel framework to solve the cross-domain problem in few-shot semantic segmentation. Compared to the standard few-shot segmentation network, we use the most primitive feature transfer to solve the cross-domain problem and effectively broaden the use scenarios of few-shot segmentation tasks.

2. We propose a plug-and-play meta-knowledge module to transfer the prior source distribution to the target domain. Our model can effectively alleviate the influence of domain shift in few-shot segmentation with exclusive contrastive loss.

3. We demonstrate the effectiveness of our framework on four different cross-domain few-shot segmentation scenarios. In particular, it can achieve state-of-the-art performance under the cross-domain setting.

2. Related Work

2.1. Few-Shot Segmentation

Prototype-based [30, 39] few-shot segmentation network extracts category prototype to matching novel samples. However, the single prototype cannot accurately describe the category. ASGNet [16] and RPMMs [33] obtain muti-prototype through clustering and EM algorithm. In recent related research, RePRI [1] abandoned the meta-learning strategy, used more effective transduction reasoning to solve the few-shot segmentation problem. SCL [47] focuses on solving the problem of information lost and uses self-information to achieve the optimized effect. Different from [1], CWT [22] uses meta-learning to give the network an adaptive classification weight and complete the segmentation of new categories. HSNet [24] makes full use of the features of the intermediate convolution to adequately improve the accuracy of segmentation to a new level.

These solutions mainly focus on the accurate segmentation of new categories. It is assumed that the data distribution of the new categories is consistent with the training data distribution, but this setting is not always guaranteed. The model segmentation effect will be significantly reduced when a domain gap exists between the data distribution. We explore the few-shot segmentation in cross-domain scenarios and use the memory mechanism to reduce the domain gap, improve the model's generalization.

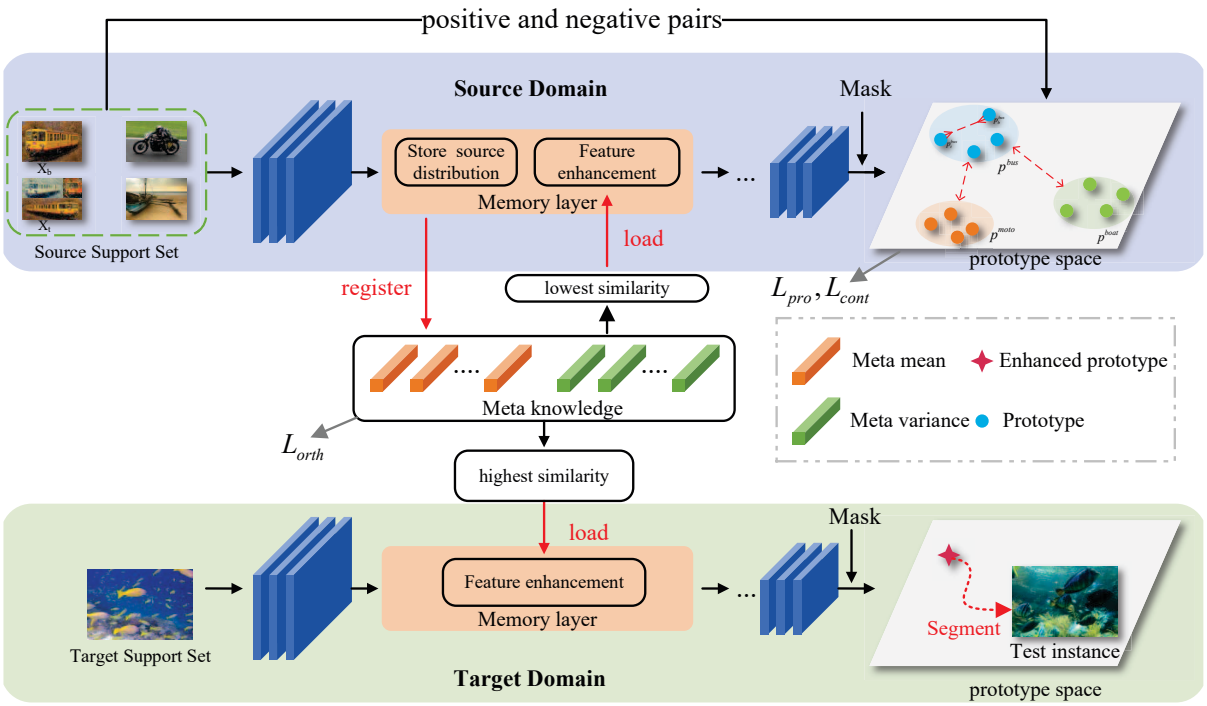

Figure 2. Overview of our proposed cross-domain few-shot semantic segmentation framework. We first conduct the Memory module during the training process on the source domain, which stores the stylized domain information. Then we use the Memory to strengthen the stylization of source and target features during the meta-training and adaptation process. Moreover, we adopt the contrastive loss to constrain the prototypes using the augmented features from the memory information,

2.2. Domain Generalization

Domain generalization (DG) has been widely used in many fields [12, 15, 28, 29]. It focuses on training a robust generalization model to achieve better results in unknown domains. Since the target domain data cannot be accessed during training, DG is challenging and practical. Data enhancement is a vital strategy to solve the DG problem. Image transformations [25, 28, 36], feature enhancement [23, 53], and generation strategies [51, 52] can all effectively improve the generalization of the model. At the same time, the DG framework based on the adversarial network hopes to decouple the domain-invariant features from the domain-special features and smoothly transfer the decoupled features to the target domain [4, 17, 37]. Using meta-learning to solve the DG problem has gradually become a promising research direction. The effect of DG can be effectively improved by learning highly adaptive parameters [3, 38, 50] and setting a more reasonable meta-learning framework [5, 7, 29].

When DG faces the few-shot scene, since the categories in the new domain have not been fitted before and the amount of labeling is tiny, this poses a significant challenge to the model's generalization. In order to solve this problem, we use the difference between the instances to enhance the features so that the few-shot model can have more abundant features for fitting and cooperate with the memory to enhance the generalization of the model.

3. Problem Setting

In this work, we pay attention to the problem of the few-shot segmentation in the cross-domain scenario. We define a training domain D_s. Each category $c_s \in D_s$ has sufficient pixel labels. At the same time, there is a test domain D_t, and each category $c_t \in D_t$ has only limited pixel labels. There is no category intersection between C_s and C_t, i.e., $C_s \cap C_t = \varnothing$. Moreover, there exists strong domain gaps between D_s and D_t. The task is to train the model in D_s, then apply the limited samples to segment the new categories in D_t.

We follow the standard practice to formalize the few-shot segmentation problem [30,32]. We first sample a large number of episodes [35] in the D_s. Each episode is composed of a query sample $Q = (X_q, Y_q)$ and K support samples $S = \{(X_s^i, Y_s^i)\}_{i=1}^K$, where X is the image, and Y is the corresponding pixel label. All the support and query samples in the episode belong to the same category. Then, the model extracts prototype based on S and performs inference and loss evaluation on the Q. After completing the episodes training on the D_s, the model adopts support samples of novel categories in the target domain D_t to extract

prototypes [30] and complete the segmentation on the corresponding new categories.

4. Methodology

Our primary motivation is to apply the Meta-Memory module as an intermediary to bridge the gap between the source and the target domains. The Memory represents stylized domain information accumulated on the source domain during the training process. We apply Memory in both the source and the target domain to improve the model's generalization.

During the meta-training stage on the source domain, the Memory is employed to strengthen the features stylization, and the model is encouraged to produce consistent representations ignoring the style differences. Moreover, we adopt the contrastive loss to further constrain the prototypes between categories to improve the model's adaptability in few-shot learning.

During the adaptation stage on the target domain, we load the Memory to help generate prototypes on novel categories so that the prototypes can cover more source information to provide robust predictions. The overall framework of our approach is shown in Figure 2.

4.1. Meta-Memory Module

The function of the memory module includes storing the source data distribution, Figure 3 (a), and using the meta-knowledge for feature enhancement, Figure 3 (b). During the training process, the mean and variance of channel features can reflect the domain information [53]. Inspired by this, we construct a Meta-Memory to represent the domain-specific information by storing the mean and variance of source instances. Memory is designed as a plug-and-play layer that can be placed behind any network layer. We put the memory module behind the shallow backbone layer in our network to collect more generalized domain information. First, we randomly initialize the meta-knowledge for each memory module as $Memory = \{M = (m_j \in R^{1\times C})_{j=1}^{N}, E = (e_j \in R^{1\times C})_{j=1}^{N}\}$ before our training. M and E are meta means and meta variance, responsible for collecting source instances feature information. C is the number of feature channels of the previous layer output, and N represents the number of meta-knowledge pairs.

Given an instance feature $f_b \in \mathcal{R}^{C\times H\times W}$ from the S, with C, H, and W denoting the dimension of the channel, height and width of the corresponding feature, respectively. We first calculate mean μ_b and variance v_b within each channel of each instance as follows:

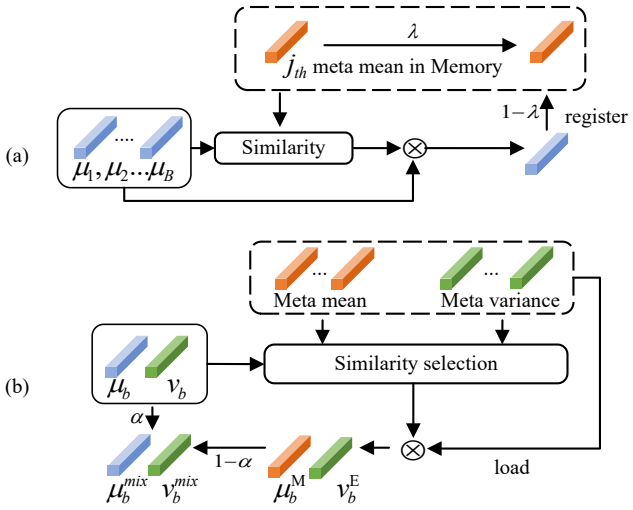

Figure 3. The specific implementation of the memory module. (a) Update the memory domain distribution during source training. (b) Use historical knowledge in the memory module to reconstruct and enhance features.

$$\mu_b = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} f_{c,h,w},$$
$$v_b = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(f_{c,h,w} - \mu_b)^2} \tag{1}$$

Where $c \in C$, notice that μ_b and v_b perform the channel normalization of each instance with $\mu_b, v_b \in \mathcal{R}^{1\times C}$. According this, we can obtain the normalized instance feature f_b^{norm}:

$$f_b^{norm} = \frac{f_b - \mu_b}{v_b} \tag{2}$$

To achieve the collection of source information, we first calculate the similarity between (μ_b, v_b) and each meta-knowledge in memory (m_j, e_j), respectively:

$$s_M^{jb} = \frac{sim(m_j, \mu_b)}{\sum_{b=1}^{B} sim(m_j, \mu_b)}, s_E^{jb} = \frac{sim(e_j, v_b)}{\sum_{b=1}^{B} sim(e_j, v_b)} \tag{3}$$

where B denotes the number of batch size during training.

Then, M and E are updated by aggregating the mean and variance information in the whole batch as below:

$$m_j = \lambda m_j + (1-\lambda)\sum_{b=1}^{B} s_M^{jb}\mu_b$$
$$e_j = \lambda e_j + (1-\lambda)\sum_{b=1}^{B} s_E^{jb} v_b \tag{4}$$

λ is the aggregation weight and we set it to 0.9 in all experiments. From Eq. 4, we can assign the style information obtained from an instance to the most similar meta-memory. Our memory bank represents all kinds of domain-specific style information for the source domain.

We can observe that each meta-knowledge pairs in our memory bank represents a style pattern for the source domain. To characterize the distribution of source domains, we hope that each meta in the memory bank is as independent as possible from each other. Therefore, we develop an orthogonal loss to constrain our memory module during training as follows:

$$L_{orth} = \frac{1}{2N^2}(\sum_{i=1}^{N}\sum_{j=1}^{N} h_M^{ij} + \sum_{i=1}^{N}\sum_{j=1}^{N} h_E^{ij}), \quad (5)$$

where $h_M^{i,j}$ indicates the similarity of the m_i and the m_j and $h_E^{i,j}$ indicates the similarity of the e_i and the e_j in $Memory$.

4.2. Memory-based Feature Enhancement

After constructing Meta-Memory on the source domain, we can obtain a large range of style information of the source data, which is feasible to enhance feature representation and improve the generalization of the segmentation model. Although previous works [14, 29] achieve this goal by adopting low-level data enhancement strategies, such as cropping, rotation, and contrast enhancement, they only provide low-level data transformations and lacks the augmentation on high-level features. To tackle this, we propose the memory-based feature enhancement method for both source data and target data.

Source Data Enhancement. The motivation of this process is to generate various source features with different styles, which can significantly improve the generalization of the segmentation model. Specifically, for a given instance's feature f_b, we first calculate the similarity $s_b^M \in R^{1\times N}$ and $s_b^E \in R^{1\times N}$ between the corresponding (μ_b, v_b) and each meta pair in $Memory$. Then we use the meta element with the lowest similarity to enhance the feature f_b, because our model utilizes the difference information among instance. The enhance feature is obtained as below:

$$f_b^{enh} = f_b^{norm} v_b^{mix} + \mu_b^{mix}, \quad (6)$$

and

$$\mu_b^{mix} = \alpha\mu_b + (1-\alpha)\mu_b^M, v_b^{mix} = \alpha v_b + (1-\alpha)v_b^E, \quad (7)$$

where μ_b^M and v_b^E are the elements with the lowest similarity selected from the $Memory$ using s_b^M and s_b^E respectively, α mainly controls the retention ratio of original features, we set it to 0.1 referring to Mixstyle [53]. Finally, the obtained enhanced support features f_b^{enh} pass through the whole backbone getting the category prototype p_b following [16].

Target Data Enhancement. The process of target data enhancement is similar to source data enhancement. The only difference is we select the highest similarity meta element (μ_b^M, v_b^E) through (s_b^M, s_b^E) in the $Memory$ to calculate the enhanced feature, due to the large gap between target style and the source meta.

4.3. Memory Enhanced Contrastive Learning

Basically, the model is trained by the prototype-based loss:

$$L_{pro} = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} Y_q^{ij}\log(R(f_q^{ij}, p_b)), \quad (8)$$

where R is the ASGNet [16] to achieve predicting, f_q is query features and $Y_q^{i,j}$ is the corresponding mask. To further empower the representation of the model, we apply the contrastive loss to prevent interference between categories, utilizing the enhanced features obtained from Sec 4.2. We first perform data enhancement and domain enhancement [14, 29] on input image (X_b, Y_b) getting transformed input (X_t, Y_t). According to our memory-based feature enhancement, we can get a transformed prototype p_t for a transformed image. Obviously, for all samples, the initial prototype p_b has the same semantic information as the corresponding transformed prototype p_t, while it should be distinguished from other categories' prototypes. Therefore, we adopt contrastive losses to increase the distance between different category prototypes. regarding the other categories' prototypes as negative pairs and the transformed prototypes as positive pairs for each p_b on batch.

$$L_{cont} = \frac{1}{2|B|}\sum_{i=1}^{|B|} -\log\frac{pos(i)}{pos(i)+neg(i)} \quad (9)$$

where

$$pos(i) = exp(sim(p_b^i, p_t^i)) \quad (10)$$

$$neg(i) = \sum_{j=1,j\neq i}^{|B|}[exp(sim(p_b^i, p_b^j)) + exp(sim(p_b^i, p_t^j))], \quad (11)$$

where $|B|$ is the categories in each batch, i,j are specific categories.

Finally, the overall objective considers all constraints we introduced above as:

$$L_{all} = L_{orth} + L_{cont} + L_{pro} \quad (12)$$

Table 1. Cross-domain few-shot semantic segmentation results on COCO-20^i to PASCAL-5^i task.

COCO-20^i to PASCAL-5^i											
Backbone	Method	1-shot					5-shot				
		split0	split1	split2	split3	Mean	split0	split1	split2	split3	Mean
ResNet50	RPMMs [44]$_{(ECCV20)}$	36.3	55.0	52.5	54.6	49.6	40.2	58.0	55.2	61.8	53.8
	RePRI [1]$_{(CVPR21)}$	52.4	**64.3**	65.3	71.5	63.3	57.0	68.0	70.4	76.2	67.9
	ASGNet [16]$_{(CVPR21)}$	42.5	58.7	65.5	63.0	57.4	53.7	**69.8**	67.1	75.9	66.6
	PFENet [33]$_{(TPAMI)}$	-	-	-	-	60.8	-	-	-	-	61.9
	CWT [22]$_{(ICCV21)}$	53.5	59.2	60.2	64.9	59.4	60.3	65.8	67.1	72.8	66.5
	HSNet [24]$_{(ICCV21)}$	48.7	61.5	63.0	72.8	61.5	58.2	65.9	**71.8**	**77.9**	68.4
	Ours	**57.4**	62.2	**68.0**	**74.8**	**65.6**	**65.7**	69.2	70.8	75.0	**70.1**
ResNet101	SCL [47]$_{(CVPR21)}$	43.1	60.3	66.1	68.1	59.4	43.3	61.2	66.5	70.4	60.3
	HSNet [24]$_{(ICCV21)}$	46.3	**64.7**	67.7	**74.2**	63.2	59.1	69.0	**73.4**	78.7	70.0
	Ours	**59.4**	64.3	**70.8**	72.0	**66.6**	**67.2**	**72.7**	72.0	**78.9**	**72.7**

5. Experiments

5.1. Datasets

COCO-20^i [20] is the most extensive and challenging few-shot segmentation dataset at present. It contains 82,081 training images and 40,137 validation images, covering 80 common categories in life. Following to [16], we divide 80 categories into four non-intersecting subsets. We select three subsets as the training set and the other subset as the test set. Therefore, the model will get four sub-datasets containing 60 categories, and we use these four sub-datasets as the source domain training model to select the non-intersecting categories in other datasets as the target domain test.

PASCAL-5^i [27] is an expanded version of PASCAL VOC 2012 [8], with additional annotation enhancement information of the SDS dataset [10]. We refer [1] to divide pascal-5^i into four subsets, and each subset category does not intersect with the corresponding subset category in COCO-20^i. We regard PASCAL-5^i as a target domain for testing.

FSS-1000 [41] contains 1000 categories, and each category has ten images with annotations. The categories cover daily necessities to cartoons and small parts. We divide it into 20 subsets. Each subset contains nearly 50 categories, and each COCO-20^i split is tested on 5 FSS-1000 subsets. And we regard FSS-1000 as a target domain for testing.

SUIM [13] acting as a benchmark dataset for underwater image segmentation, is an exclusive segmentation dataset. This dataset contains over 1500 images with pixel annotations for eight object categories. SUIM can test the effect of our model against typical cross-domain scenarios. We regard SUIM as a target domain for testing.

5.2. Implementation Details

Source training We choose ASGNet [16] as the baseline model, the corresponding backbone is selected as ResNet50 and ResNet101. For the source domain COCO-20^i, we train 20 epochs with batch size 12. The image size is resized to 473×473. As for the optimizer, we choose SGD with a momentum of 0.9 and weight decay of 1e-4. The initial learning rate is set to 0.0025. During the training process, we insert two memory modules into the first two layers of ResNet respectively and initialize ten pairs of distributions in each memory module.

Target testing For PASCAL-5^i, we set 4 subsets, FSS-1000 set 20 subsets, and SUIM one subset. We remove all categories that intersect with COCO-split in all subsets. During the testing process, we removed the domain enhancement.

5.3. Comparisons with State-of-the-art

In Table 1, we compare our method with several state-of-the-art few-shot semantic segmentation methods on COCO-20^i to PASCAL-5^i task. Recent works CWT [22] and RePRI [1] have revealed a domain discrepancy between COCO-20^i and PASCAL-5^i datasets. Although these previous works provide promising results on few-shot segmentation for intra-domain, they usually ignore the bridging domain gap between source and target domains. The results show that our approach significantly improves the performance on cross-domain semantic segmentation for 1-shot and 5-shot tasks. Specifically, we surpass traditional state-of-the-art few-shot semantic segmentation method HSNet [24] based on ResNet-50 by 4.1% and 1.7% mIoU, respectively. It also outperforms the method RPMMs [44] by a large margin over 16.0 mIoU for 1-shot segmentation and 16.3 mIoU for 5-shot segmentation. Furthermore, for the backbone of ResNet-101, our method performs mean re-

Table 2. Cross-domain few-shot semantic segmentation results on COCO-20^i to FSS-1000 task with the ResNet50 as backbone.

COCO-20^i to FSS-1000						
Backbone	Method	split-0	split-1	split-2	split-3	mean
ResNet50	ASGNet [16]$_{(CVPR21)}$	76.2	72.2	72.7	71.6	73.2
	HSNet [24]$_{(ICCV21)}$	79.9	80.5	**81.1**	82.1	80.8
	SCL [47]$_{(CVPR21)}$	81.6	78.3	77.5	74.4	78
	Ours	**82.2**	**82.6**	79.6	**83.4**	**81.9**

Table 3. Cross-domain few-shot semantic segmentation results on COCO-20^i to SUIM and PASCAL-5^i to SUIM tasks with ResNet50 as backbone.

COCO-20^i to SUIM						
Backbone	Methods	split-0	split-1	split-2	split-3	mean
ResNet50	ASGNet [16]$_{(CVPR21)}$	28.1	27.5	26.1	32.3	28.5
	HSNet [24]$_{(ICCV21)}$	**33.8**	35.9	35.3	35.4	35.1
	SCL [47]$_{(CVPR21)}$	27.3	28.8	26.5	25.3	27.0
	Ours	30.5	**38.6**	**42.5**	**36.6**	**37.1**
PASCAL-5^i to SUIM						
Backbone	Methods	split-0	split-1	split-2	split-3	mean
ResNet50	ASGNet [16]$_{(CVPR21)}$	32.4	30.9	28.9	35.2	31.9
	HSNet [24]$_{(ICCV21)}$	30.7	30.0	27.3	27.0	28.8
	SCL [47]$_{(CVPR21)}$	31.3	31.2	32.2	32.5	31.8
	Ours	**35.2**	**33.4**	**34.3**	**36**	**34.7**

sults of 66.6% and 72.7% for 1-shot and 5-shot tasks, significantly improved by 7.2 and 12.4 mIoU compared to SCL [47]. These results demonstrate the effectiveness of our method and indicate the essential of bridging the domain gap across domains.

FSS-1000 [41] is another challenging cross-domain dataset for COCO-20^i, which contains abundant 1000 categories that can effectively test the model's stability. Therefore we conduct a cross-domain few-shot semantic segmentation task on COCO-20^i to FSS-1000 based on ResNet-50, and the results are shown in Table 2. We compare our method with previous state-of-the-art approaches, including ASGNet [16], HSNet [24], and SCL [47]. The results show that our method is stable when transferred to large-scale new categories and achieves the best performance on three COCO-20^i splits to FSS-1000 tasks.

We also construct a cross-domain task with a large margin domain gap between source and target domains from COCO-20^i and PASCAL-5^i to SUIM [13]. SUIM is collected from underwater scenario, which is completely different from COCO-20^i and PASCAL-5^i. The segmentation results in Table 3 indicate that conventional few-shot methods perform poorly due to the large domain gap, with SCL [47] achieving 27.0% mIoU and 31.8% mIoU on COCO-20^i and PASCAL-5^i to SUIM, respectively. Benefiting from the our technique to eliminate domain discrepancy, our method provides 37.1% and 34.7% mIoU results for COCO-20^i and PASCAL-5^i to SUIM tasks, which outperforms SCL [47] by 10.1% and 2.9%.

Table 4. Comparison to domain generalization approaches on COCO-20^i to PASCAL-5^i, "BS" denotes the ASGNet [16]$_{(CVPR21)}$ baseline.

	BS	MixUp [48]$_{(ICLR18)}$	MixStyle [53]$_{(ICLR21)}$	ours
mean-IoU	57.4	58.6	61.5	**65.6**

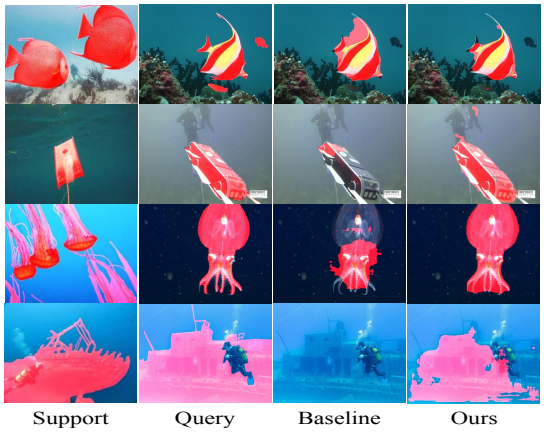

Figure 4. Visualization for predicted masks for COCO-20^i to SUIM task.

Morever as shown in Table 4, our approach outperforms previous SOTA domain generalization approaches MixUp [48] and MixStyle [53]. We attribute the improvement to our memory strategy that can explicitly transfer domain information to novel categories, while previous DG approaches tend only to fit seen categories. We also compare our method with state-of-the-art methods on the intra-domain few-shot segmentation in our Supplementary Material.

5.4. Ablation Study

Influence of Each Component. We provide analysis of each proposed component on 1-shot semantic segmentation task from COCO-20^i to PASCAL-5^i in this subsection, including the contrastive loss (Cons), memory-enhanced source feature (MEnS), and memory-enhanced target feature (MEnT). The detailed results are shown in Table 5. The baseline model is a naive method for few-shot semantic segmentation following [16], without any domain alignment techniques. From the results, we can observe that the model (a) with the proposed contrastive loss can significantly improve the performance by 5.4 mIoU compared to the baseline model. After adding memory-enhanced source feature

Table 5. Ablation study of our proposed components for 1-shot semantic segmentation on COCO-20^i to PASCAL-5^i with ResNet-50 backbone. "Cons" denotes the proposed contrastive loss. "MEnS" and "MEnT" denote the memory-based feature enhancement for the source domain and target domain."BS" denotes the ASGNet [16]$_{(\text{CVPR21})}$ baseline.

COCO-20^i to PASCAL-5^i (1-shot)				
Model	Cons	MEnS	MEnT	mean-IoU
BS				57.4
(a)	✓			62.8
(b)	✓	✓		63.5
(c)		✓	✓	62.5
(d)	✓	✓	✓	65.6

Table 6. Results of the memory module placed with ResNet50 on 1-shot COCO-20^i to PASCAL-5^i.

	No memory	layer-0	layer-1	layer-2	layer-3
mean-IoU	62.8	64.0	62.7	53.3	50.8

(MEnS) in model (b), the performance is further improved with an increase of 0.7%. Furthermore, we analyze the effectiveness of the proposed meta-memory module by conducting the model with the proposed memory on source feature and target feature (model (c)). The result demonstrates the effectiveness of proposed memory module with 5.1% improvement. It reveals our cross-domain memory can reduce domain gap and improve the generalization of the segmentation model. The model (d) with both the proposed memory module and contrastive loss can further provide 3.1% gains compared to model (c). We replaced the meta-memory with random noise (same mean&var) on COCO-20^i to PASCAL-5^i, which caused 6.1% performance degrades, demonstrating that meta-memory provides valuable information learned from the source domain.

We also provide the visualization of category prototypes for COCO-20^i and PASCAL-5^i obtained from our model (d) and baseline model. The results is shown in Figure 5. We can see that our method can provide more reliable discriminative boundaries than the baseline model. It benefits from the memory module to improve the generalization and contrastive loss to enhance feature representation.

5.5. Parameters Analysis

We first analyze the influence of meta pairs N in the memory module. We set different values of N in 1-shot segmentation task from COCO-20^i to PASCAL-5^i, and results are shown in the Figure 5. As the model epoch grows, the model with the $N = 10$ reaches the best result in our experiments. We set $N = 10$ in all experiments in this work.

We also analyze the impact of memory placed in different layers in Table 6. We observe that the memory module

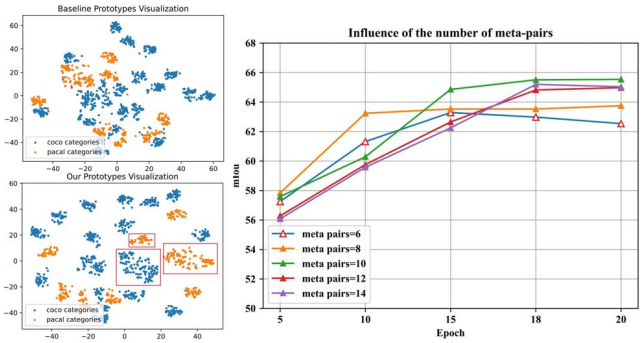

Figure 5. Left: Visualization of prototypes for COCO-20^i and PASCAL-5^i. Right: The influence of meta pairs N in the memory module.

works most effectively at the shallow layers of the backbone. This phenomenon is also in line with our understanding of deep networks, shallow features are easily transferred and shared. The memory module in the high-level layer will decrease the accuracy because of semantic noise.

5.6. Limitation

We use the lowest similarity and highest similarity to select meta-knowledge. A more reasonable solution is to learn from the network adaptively. We hope to solve this problem in subsequent research.

6. Conclusion

In this paper, we have presented a novel cross-domain few-shot framework for semantic segmentation. It aims to transfer the knowledge from the source domain to the novel classes in the target domain. To achieve this goal, a meta-memory module has been proposed to bridge the source and target domains, including reducing the domain gap and enhancing semantic feature representation. Specifically, the meta-memory stores the domain-specific information from the source data during training and transfers them to the target data to improve the generalization of the segmentation model. The memory-based feature enhancement also contributes to discriminative feature learning for the novel classes. We conduct extensive experiments and ablation studies to validate the effectiveness of the proposed framework on different cross-domain segmentation tasks.

7. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.62176009), the Project of Beijing Municipal Education Commission Project (No.KZ201910005008), the Major Project for New Generation of AI (No.2018AAA0100400), the National Natural Science Foundation of China (No. 61836014, No. U21B2042, No. 62072457, No. 62006231).

References

[1] Malik Boudiaf, Hoel Kervadec, Imtiaz Masud Ziko, Pablo Piantanida, Ismail Ben Ayed, and José Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proc. CVPR*, pages 13974–13983, 2021. 1, 2, 6

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

[3] Yi-Syuan Chen and Hong-Han Shuai. Meta-transfer learning for low-resource abstractive summarization. In *Proc. AAAI*, 2021. 3

[4] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne Taery Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proc. CVPR*, pages 11575–11585, 2021. 3

[5] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *Proc. CVPR*, pages 3424–3434, 2021. 3

[6] Kaiqi Dong, Wei Yang, Zhenbo Xu, Liusheng Huang, and Zhidong Yu. Abpnet: Adaptive background modeling for generalized few shot segmentation. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 1

[7] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees G. M. Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *Proc. ECCV*, volume abs/2007.07645, 2020. 3

[8] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 6

[9] A. Frikha, Denis Krompass, Hans-Georg Koepken, and Volker Tresp. Few-shot one-class classification via meta-learning. In *Proc. AAAI*, 2021. 2

[10] Bharath Hariharan, Pablo Arbeláez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proc. ICCV*, pages 991–998, 2011. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 1

[12] Joy Hsu, Wah Chiu, and Serena Yeung. Darcnn: Domain adaptive region-based convolutional neural network for unsupervised instance segmentation in biomedical images. In *Proc. CVPR*, pages 1003–1012, 2021. 3

[13] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *Proc. IROS*. IEEE/RSJ, 2020. 6, 7

[14] Jogendra Nath Kundu, Akshay Ravindra Kulkarni, Amit Singh, V. Jampani, and R. Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proc. ICCV*, volume abs/2108.11249, 2021. 5

[15] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proc. CVPR*, pages 1482–1491, 2021. 3

[16] Gen Li, V. Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proc. CVPR*, pages 8330–8339, 2021. 1, 2, 5, 6, 7, 8

[17] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex Chichung Kot. Domain generalization with adversarial feature learning. In *Proc. CVPR*, pages 5400–5409, 2018. 3

[18] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu. Cross-domain adaptive clustering for semi-supervised domain adaptation. In *Proc. ICCV*, pages 2505–2514, 2021. 2

[19] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proc. ICCV*, pages 9166–9175, 2019. 1

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, 2014. 6

[21] Bingyu Liu, Z. Zhao, Zhenpeng Li, Jianan Jiang, Yuhong Guo, and Jieping Ye. Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification. *ArXiv*, abs/2005.08463, 2020. 2

[22] Zhihe lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *Proc. ICCV*, volume abs/2108.03032, 2021. 2, 6

[23] Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *Proc. ECCV*, volume abs/2007.12256, 2020. 3

[24] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proc. ICCV*, volume abs/2104.01538, 2021. 2, 6, 7

[25] Sebastian Otálora, Manfredo Atzori, Vincent Andrearczyk, Amjad Rehman Khan, and Henning Müller. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in Bioengineering and Biotechnology*, 7, 2019. 3

[26] Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *Proc. ICCV*, 2021. 2

[27] Amirreza Shaban, Shray Bansal, Z. Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *Proc. BMVC*, volume abs/1709.03410, 2017. 6

[28] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K. Jain. Towards universal representation learning for deep face recognition. In *Proc. CVPR*, pages 6816–6825, 2020. 3

[29] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-

augmented meta-learning. In *Proc. CVPR*, pages 9619–9628, 2021. 3, 5

[30] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Proc. NIPS*, 2017. 2, 3, 4

[31] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. Explanation-guided training for cross-domain few-shot classification. In *Proc. ICPR*, pages 7609–7616, 2021. 2

[32] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proc. CVPR*, pages 1199–1208, 2018. 3

[33] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2020. 2, 6

[34] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *Proc. ICLR*, volume abs/2001.08735, 2020. 2

[35] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proc. NIPS*, 2016. 3

[36] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proc. ICCV*, pages 7979–7988, 2019. 3

[37] Hongsong Wang, Shengcai Liao, and Ling Shao. Afan: Augmented feature alignment network for cross-domain object detection. *IEEE Transactions on Image Processing*, 30:4046–4056, 2021. 3

[38] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In *Proc. IJCAI*, 2021. 3

[39] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proc. ICCV*, pages 9196–9205, 2019. 2

[40] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proc. ICCV*, pages 9092–9101, 2021. 2

[41] Tianhan Wei, Xiang Li, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proc. CVPR*, pages 2866–2875, 2020. 6, 7

[42] Guosen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *Proc. CVPR*, pages 5471–5480, 2021. 1

[43] C. Xu, Chen Liu, Li Zhang, Chengjie Wang, Jilin Li, Feiyue Huang, X. Xue, and Yanwei Fu. Learning dynamic alignment via meta-filter for few-shot learning. In *Proc. CVPR*, pages 5178–5187, 2021. 2

[44] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *Proc. ECCV*, volume abs/2008.03898, 2020. 6

[45] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Proc. ECCV*, 2020. 1

[46] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proc. CVPR*, pages 13829–13839, 2021. 2

[47] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *Proc. CVPR*, pages 8308–8317, 2021. 1, 2, 6, 7

[48] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. ICLR*, volume abs/1710.09412, 2018. 7

[49] Junyi Zhang, Ziliang Chen, Junying Huang, Jingyu Zhuang, and Dongyu Zhang. Few-shot domain adaptation for semantic segmentation. *Proceedings of the ACM Turing Celebration Conference - China*, 2019. 2

[50] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and N. Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *Proc. CVPR*, pages 6273–6282, 2021. 3

[51] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proc. AAAI*, volume abs/2003.06054, 2020. 3

[52] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Proc. ECCV*, volume abs/2007.03304, 2020. 3

[53] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *Proc. ICLR*, volume abs/2104.02008, 2021. 3, 4, 5, 7