Displays 91 (2026) 103179

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Exploiting independent query information for few-shot image segmentation[☆]

Weide Liu [a], Zhonghua Wu [a], Henghui Ding [b], Fayao Liu [c], Jie Lin [c], Guosheng Lin [a], Wei Zhou [d],[*]

[a] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[b] Institute of Big Data, Fudan University, China
[c] Institute for Infocomm Research (I^{2}R)-Agency for Science, Technology and Research, Singapore
[d] Cardiff University, UK

ARTICLE INFO

Keywords:
Few-shot
Image segmentation
Independent query

ABSTRACT

This work addresses the challenging task of few-shot segmentation. Previous few-shot segmentation methods mainly employ the information of support images as guidance for query image segmentation. Although some works propose to build a cross-reference between support and query images, their extraction of query information still depends on the support images. In this paper, we propose to extract the information from the query itself independently to benefit the few-shot segmentation task. To this end, we first propose a prior extractor to learn the query information from the unlabeled images with our proposed global–local contrastive learning. Then, we extract a set of predetermined priors via this prior extractor. With the obtained priors, we generate the prior region maps for query images, which locate the objects, as guidance to perform cross-interaction with support features. In such a way, the extraction of query information is detached from the support branch, overcoming the limitation by support, and could obtain more informative query clues to achieve better interaction. Without bells and whistles, the proposed approach achieves new state-of-the-art performance for the few-shot segmentation task on public datasets.

1. Introduction

Semantic segmentation is a fundamental task that involves classifying each pixel into a particular class. Deep learning has achieved remarkable success in fully supervised semantic segmentation [1]. However, this approach has some intrinsic limitations, such as requiring a large number of pixel-level annotated image samples for model training and abundant annotations for novel classes when extending the current segmentation model to new classes.

Few-shot segmentation has been proposed to address these issues by training a network to predict segmentation masks for novel classes with only a few annotated novel class training samples. Currently, state-of-the-art few-shot segmentation methods [2–4] utilize support images as guidance for query image segmentation with a two-branch architecture. However, they only use support images' information as guidance for query mask prediction and do not consider clues from the query images themselves. To use the query information, CRNet [5] proposes a cross-reference mechanism that enables interaction between support and query image features. Similarly, PANet [3] uses prototype alignment regularization to align the query and support prototypes. Although these methods build cross-guidance between query and support images, they are limited to labeled support images since all information extraction/propagation, such as support-to-query and query-to-support, essentially rely on support masks.

In this work, we argue that detaching query information extraction from the support branch and generating query clues independently can enhance few-shot segmentation performance and generalization. To this end, we propose a prior extractor to learn query information from the unlabeled query images themselves with self-supervised learning.

Contrastive learning is among the most promising directions in self-supervised learning methods [6]. The process involves transforming images into different variants, using contrastive loss to minimize the feature distances between the variants from the same images, and maximizing the feature distances obtained from different images. The objective of contrastive learning is to learn a predetermined prior of the objects without labeled data, which can bridge the gap between fully and less supervised classification. However, the current state-of-the-art design [7] is sub-optimal for segmentation tasks for two reasons. Firstly, image segmentation is a pixel-level classification task where both local and global representations are crucial, but [7] is only designed for global image-level representation. Secondly, there are often multiple objects co-existing in one image, such as a keyboard, desk,

[☆] This paper was recommended for publication by Guangtao Zhai.

[*] Corresponding author.
E-mail address: wei.zhou@uwaterloo.ca (W. Zhou).

https://doi.org/10.1016/j.displa.2025.103179
Received 13 April 2025; Received in revised form 27 June 2025; Accepted 6 August 2025
Available online 20 August 2025

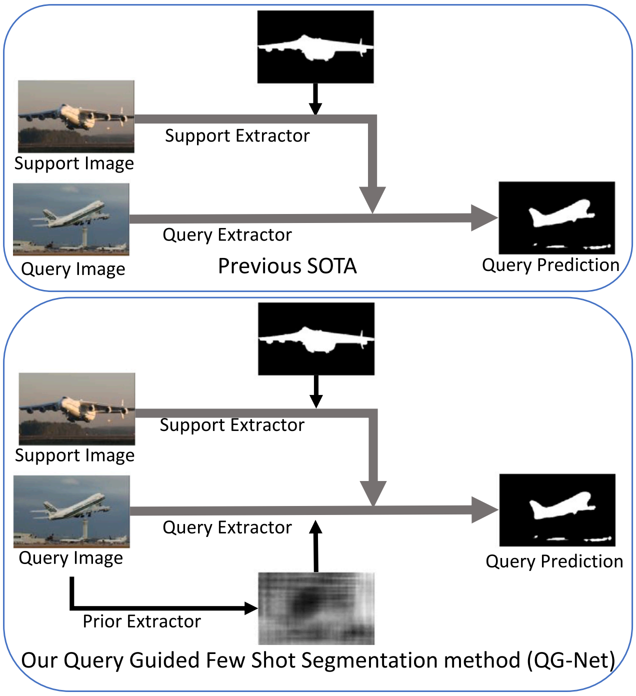

Fig. 1. Comparison between the pipeline of our proposed Query Guided Network with previous state-of-the-art (SOTA) few-shot segmentation methods. Previous works (upper part) only employ support images' information as guidance for query mask estimation, while our QGNet (lower part) utilizes the clues of query images with query extractor as guidance for final query mask prediction.

and computer. However, the global contrastive loss cannot distinguish between these different objects within the same image.

To address these issues, we propose a predetermined prior learning method to obtain more distinguishable image features for the few-shot semantic segmentation task. We leverage both global and local contrastive losses to learn a prior extractor for few-shot image segmentation. A global contrastive loss is applied to the global representations to minimize the feature distances obtained from different variants of an identical image. To further differentiate different objects within the same image, we obtain local representations by dividing the image into local patches, with each patch containing similar features and contexts. We then apply a local contrastive loss between the local patches to learn a local predetermined prior. Similar to the global contrastive loss, the local contrastive loss aims to maximize the feature distance between different patches, such as computer patches and keyboard patches. We use both global and local contrastive losses to train our prior extractor to leverage both advantages.

We utilized the previously mentioned self-supervised learning method to train a prior extractor on the available unlabeled data, which holds the predetermined prior for the query images. Subsequently, we developed a new few-shot segmentation architecture, as illustrated in Fig. 1, to extract semantic segmentation information of the query category from the priors. In this approach, the given query images undergo encoding using the prior and feature extractors, resulting in prior features and bridge features, respectively. The target objects are then identified by computing pixel-wise similarity maps between the prior and bridge features. This method enables us to obtain the prior region maps of the query images utilizing their own information, thereby addressing the aforementioned limitations. Furthermore, the support images undergo encoding by the same feature extractor to be projected onto the same feature space as the bridge features. Subsequently, we establish cross-interactions between the support and bridge features to enhance the segmentation performance.

Our main contributions can be summarized as follows:

- To the best of our knowledge, we are the first to utilize self-supervised feature learning methods on unlabeled query images to benefit the few-shot segmentation task. Specifically, we propose a prior extractor to generate maps of prior regions from the query images themselves to guide the final query mask prediction.
- We propose a global and local contrastive loss to train the prior extractor, making contrastive learning more suitable for few-shot segmentation tasks. With our proposed global–local learning, the query branch independently extracts informative clues from the query images themselves, which greatly enhances the cross-interaction between the query and support.
- Our network achieves state-of-the-art results on the COCO datasets.

2. Related work

2.1. Fully supervised semantic segmentation

Semantic segmentation is a crucial computer vision task that involves predicting a predefined class for each pixel in an image. Recent state-of-the-art methods rely heavily on fully convolutional networks (FCN) [1]. The encoder–decoder [8] architecture is a popular choice for generating high-resolution prediction maps for semantic segmentation. The encoder gradually downsamples the feature maps to obtain a large field-of-view, while the decoder recovers the fine-grained information. To avoid losing the resolution of the feature maps while keeping the field-of-view large, dilated convolution [8] increases the field-of-view without decreasing the feature map size or increasing the number of parameters. In our network, we follow the encoder–decoder structure with dilated convolution to produce the query segmentation masks.

2.2. Few-shot segmentation

Early approaches for few-shot segmentation adopt a two-branch architecture to condition on a given support example. For instance, Shaban et al. [9] introduced the first such model (OSLSM) with a conditioning branch that generates parameters for a query segmentation network. Subsequent methods like co-FCN [10] and CANet [2] improved this paradigm by embedding support features into the query branch (via feature concatenation and attention) and by iteratively refining the predicted masks. Another prominent direction is to use metric learning with class prototypes. Dong and Xing [11] first proposed to compute a prototype representation from the support mask and match query features to this embedding for segmentation, an idea also exploited by the similarity guidance network SG-One [12].

In terms of enhanced feature matching, HDMNet [13] introduce hierarchical dense correlation distillation using transformer-based multi-scale matching to mitigate overfitting. Xu et al. [14] propose SCCAN, employing self-calibrated cross-attention to better align query-support patches, and further improve segmentation accuracy via their ambiguity elimination strategy (AENet) [15]. Moon et al. [16] present MSI, maximizing support-set information by enriching correlation maps to handle small or ambiguous query targets.

Another promising trend involves leveraging external knowledge and pretrained models. MIANet [17] aggregate instance-specific and general class prototype embeddings guided by textual semantics to overcome class bias. Zhu et al. [18] propose LLaFS, uniquely utilizing large language models (LLMs) to guide segmentation with language-driven region attributes, demonstrating substantial improvements in few-shot generalization. Meanwhile, visual prompting techniques have also emerged, such as the multi-scale visual prompts introduced by Hossain et al. [19], designed for generalized few-shot segmentation tasks without altering transformer weights.

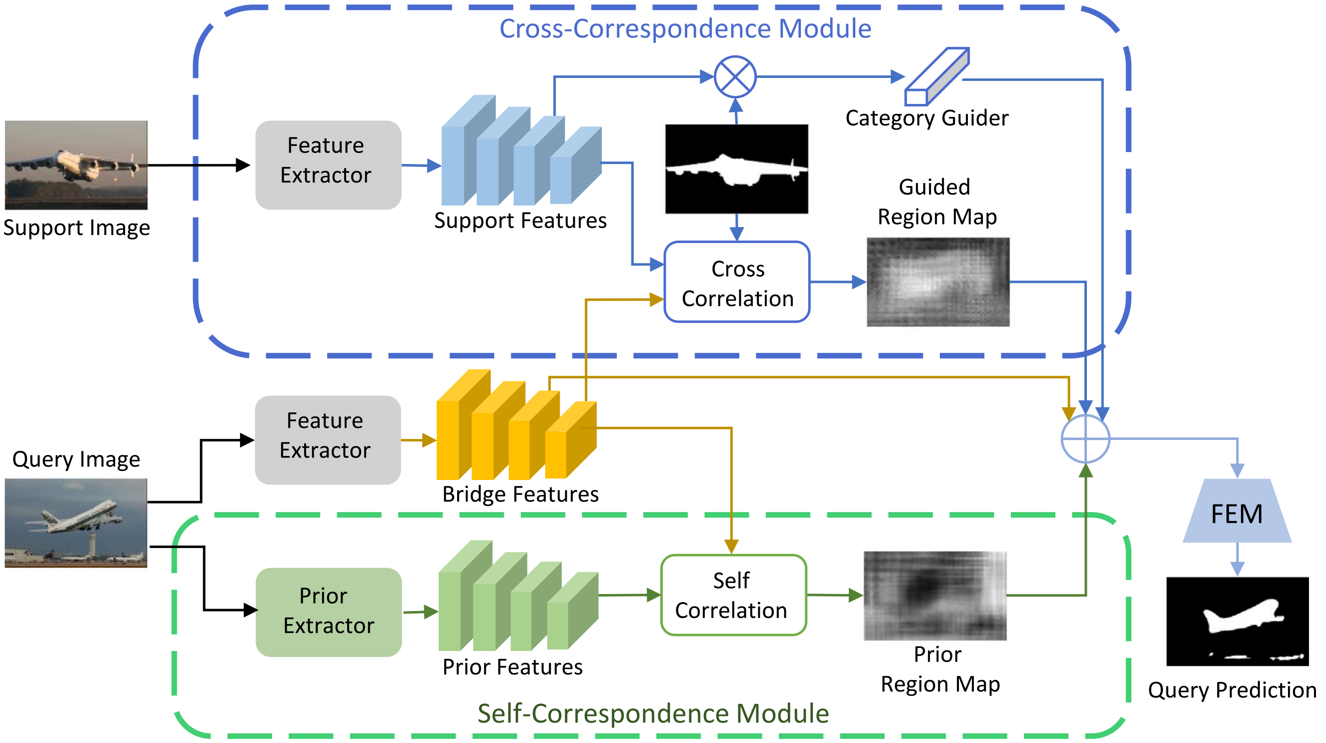

Fig. 2. Our proposed method consists of a self-correspondence module and a cross-correspondence module. Unlike the previous SOTA, a self-correspondence module (green) is proposed to extract prior features and generate a prior region map to locate the target object regions from the query image itself. A Cross-correspondence module (blue) is proposed to generate a guided region map to identify the query object region with the category guide from the masked support feature. Finally, the prior region map and guided region map are concatenated with category and bridge features for the final query mask prediction (FEM).

The two-branch network architecture is a common approach in few-shot segmentation tasks. Various prior studies [4,5,10,12,20–28] have employed methods to encode category features from labeled support images for guiding query mask prediction. However, these methods primarily rely on support images for guidance and do not fully exploit information from query images. In pursuit of better leveraging query image information, CRNet [5] introduces a cross-reference mechanism to facilitate interaction between query and support images, enhancing model training. PANet [3] incorporates a prototype alignment network to guide segmentation for both query and support images. SSNet [29] introduces a self-supervised approach, enhancing performance by introducing random pseudo-classes in the background of query images through superpixel segmentation. PFENet [20] employs a prior-guided feature enhancement network for comparing query and support features through pyramid feature comparison. Building upon these advancements, methods like ASGNet [30], RePRI [31], CWT [32], and ABPNet [33] have aimed to mitigate model bias and adapt more effectively to novel classes. ASGNet dynamically determines the number of prototypes and their spatial dimensions. RePRI fine-tunes models on support images to enhance adaptation to novel classes. In contrast, CWT adjusts classifier weights using episodic training within a self-attention block, and ABPNet employs an attention mechanism with a meta-training strategy to predict task-specific backgrounds. Additionally, SCL [34] introduces self-guided learning to recover lost critical information during the encoding process. This method employs masked Global Average Pooling (GAP) to encode both covered and uncovered foreground regions, leading to improved segmentation performance on query images. Nevertheless, despite these methods aligning both support and query clues and guiding each other, the query mask prediction is still limited to the labeled support images as their information extraction of query images still depends on support. In this work, we contend that training an image prior extractor using unlabeled query images could enhance the query segmentation mask prediction.

2.3. *Contrastive learning*

Self-supervised contrastive learning methods, such as MoCo [7,35] and Bootstrap [36], have gained popularity in computer vision for learning a feature extractor from unlabeled images as pre-training for downstream tasks. These methods aim to pull representations of different views of the same image closer together while pushing away representations of different images. SimCLR [37] proposes a straightforward framework for self-supervised learning by applying contrastive learning to representations of the same image with different data augmentations. MoCo [7,35] increases the negative memory bank size with a moving average network, known as a momentum encoder. RegionContrast [38] improves similarity between semantically similar pixels while maintaining discrimination. It leverages a memory bank and region centers for efficient feature storage, enabling region-level contrastive learning, which is more memory-efficient than pixel-level methods. IIC [39] focuses on maximizing mutual information between class assignments in each pair, a method designed for unsupervised image classification and segmentation. Achanta [40] propose to identify salient regions in images by utilizing low-level features related to brightness (luminance) and color, while Krishna et al. [41] propose strategies to enhance semi-supervised segmentation of volumetric medical images, incorporating domain-specific and problem-specific cues. These strategies lead to improved performance, as demonstrated on three MRI datasets with limited annotations, surpassing other self-supervised and semi-supervised methods. DetCo [42] introduces multi-level supervision for intermediate representations and employs contrastive learning between the entire image and local patches. This design ensures consistent and discriminative global and local representations across feature pyramid levels, benefiting both detection and classification.

In this work, we leverage contrastive learning to train a prior extractor from the unlabeled query images to enhance few-shot segmentation. By using contrastive loss, our prior extractor can capture informative features from unlabeled query images to support query segmentation without relying solely on labeled support images (see Fig. 3).

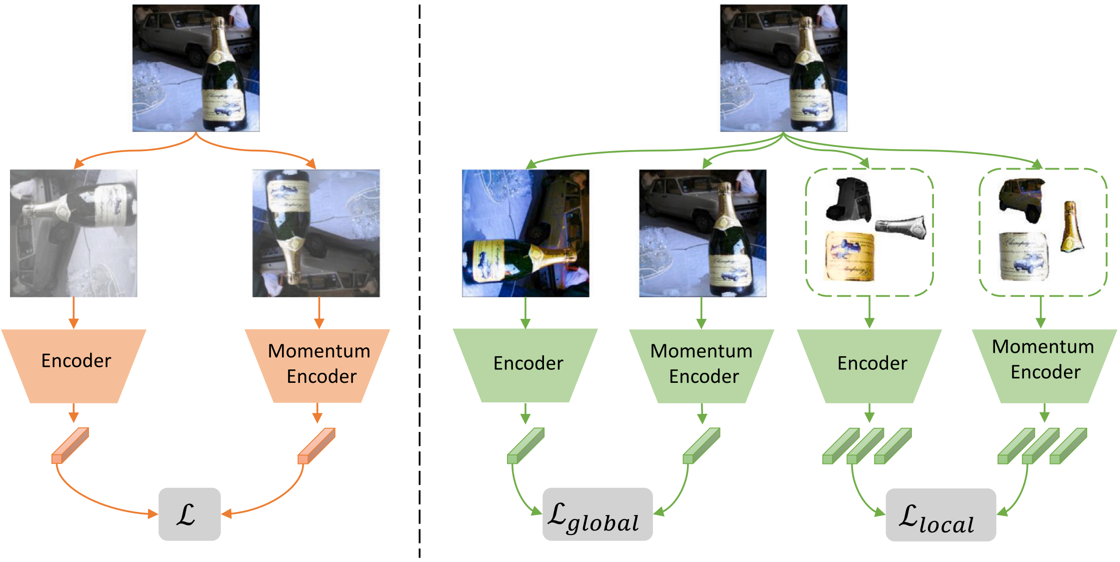

Fig. 3. The difference between conventional contrastive learning and our proposed global–local contrastive learning. Conventional contrastive learning methods (left) only learn contrast from a global perspective. Our global–local contrastive learning learns the contrast with two additional local patches as input and builds a contrastive loss across the global and local representation.

3. Task definition

Few-shot semantic segmentation aims to perform segmentation with only a few annotated examples on novel classes. We divide the images into two sets, the target images are called query set $\mathcal{Q}$ and the annotated images serve as support set $\mathcal{S}$. Provided K annotated samples from the support set, the model aims to predict binary masks for the query images, while the annotated support images should provide the foreground categories. The categories in the training image set have no overlap with the test image set.

Given a network $\mathcal{R}_\theta$ parameterized by θ, in each episode, we first sample a target category c from the dataset C. Based on the sampled classes, we then collect $K+1$ labeled images $\{(x_s^1, y_s^1), (x_s^2, y_s^2), \dots (x_s^k, y_s^k), (x_q, y_q)\}$ from the dataset, which contains at least one object belongs to the sampled category c. Among them, we sample K labeled images constitute the support set $\mathcal{S}$, and the last one serves as the query set $\mathcal{Q}$. After that, we make predictions on the query images by inputting the support set and the query image into the model $\hat{y}_q = \mathcal{R}_\theta(\mathcal{S}, x_q)$. At training time, we learn the model parameters θ by optimizing the cross-entropy loss $\mathcal{L}(\hat{y}_q, y_q)$, and repeat such procedures until convergence.

4. Method

Few-shot semantic segmentation task aims to train a network to predict novel classes with only a few annotated data. Previous works [2,4,5,20,28] solve this task by learning a network with annotated support images. However, we aim to extract information from query images themselves to improve the prediction of query masks. Our proposed method comprises two main modules: the self-correspondence module and the cross-correspondence module, as shown in Fig. 2. In the self-correspondence module, we train a prior extractor from unlabeled query images using global and local contrastive losses. We store prior features obtained with this method to propagate stored prior information to bridge features, obtaining object regions as prior region maps. The cross-correspondence module generates guided region maps and category guider to determine query category information with annotated support images, allowing for the final prediction of query masks.

This section introduces the global contrastive loss to differentiate dissimilar image features and the proposed local contrastive loss to differentiate different object patches. The quality of patches is crucial for local contrastive learning, which impacts prior feature generation. We analyze different ways of generating patches to calculate the local contrastive loss and optimize the prior extractor. We then describe how the prior region map and guided region map are generated from our trained prior extractor and feature extractor. Finally, we present an interesting discovery: the prior region maps have low correspondence values at object regions, while conventional guided region maps have high correspondence values at object regions.

4.1. *Prior extractor*

4.1.1. *Global contrastive learning*

The global contrastive loss [6] has been widely used for pre-training by distinguishing features obtained from different images. In this paper, we also use the global contrastive loss to learn a global level prior to distinguishing the image level features.

Following [35], we random sample one image from the dataset and transform the image into two different variants as a positive pair. For the negative pairs I_k, we sample one image from the queue. Consider we encode one query image I_q to vector q, and encode a set of reference images as keys to vectors $k_{0,1,2\dots}$. We use the InfoNCE [43] as our contrastive loss:

$$\mathcal{L}_{global} = -\log \frac{\exp(q\cdot k_0/\tau)}{\sum_{i=1}^{K}\exp(q\cdot k_i/\tau) + exp(q\cdot k_0/\tau)}. \qquad (1)$$

In line with previous approaches [7,35], we adopt $K = 65\,536$ negative samples for our method. k_0 denotes the positive pair from the same image, and we set the temperature hyper-parameter τ [44]. The contrastive loss's value is low when the feature similarity is high for positive pairs and low for negative pairs.

To maintain a large dictionary size, we build a dynamic dictionary following MoCo [7]. The keys of the dictionary are updated by replacing the oldest mini-batch with the new one in a momentum-based manner.

4.1.2. *Local contrastive learning*

The global contrastive loss aims to learn a global-prior feature by teaching the model to distinguish dissimilar images. However, several limitations still exist in applying global contrastive loss for segmentation task pre-training, (1) The global contrastive loss only discriminates between dissimilar images and does not consider different objects within a single image. (2) Instances that commonly coexist might have a negative effect with only global contrastive loss. For instance, the keyboard and computer frequently appear together in the same image, but they typically need to be segmented into different classes.

To address these limitations, we propose a local contrastive loss to differentiate between different objects and improve the few-shot segmentation task. We divide the images into local patches and pretrain the model using the local contrastive loss on these patches. After introducing the local contrastive loss, we will explain the methods used to generate the patches and demonstrate how they can be used to improve few-shot segmentation tasks.

Local contrastive loss. We aim to distinguish the patches by applying the local contrastive loss on the patches generated with the above methods. Specifically, we randomly sample one query patch and encode it to q_p, and a set of reference patches as our reference patches key, $k_{p0,p1,\dots}$. A local contrastive loss aims to minimize the distance between similar patches and maximize the distance between dissimilar patches. Thus, our local contrastive loss can be formulated as:

$$\mathcal{L}_{local} = -\log \frac{\exp(q_p \cdot k_{p0}/\tau)}{\sum_{i=1}^{M} \exp(q_p \cdot k_{pi}/\tau) + exp(q_p \cdot k_{p0}/\tau)}. \tag{2}$$

Here, we keep $M = 65\,536$ negative sample patches in the queue, and the k_{p0} is the positive patch pairs with different views from the same patch, the k_{pi} denotes the negative patch pairs, and the τ is a temperature hyper-parameter [44].

We use both local and global contrastive loss to generate our query prior:

$$\mathcal{L} = \mathcal{L}_{local} + \mathcal{L}_{global}. \tag{3}$$

Patch generation method 1: Felzenszwalb's efficient graph based segmentation. To guarantee that the generated patches align with the same semantic categories, follow [45], we segment the images by assigning the local patches with the same contexts. As shown in algorithm 1, at first, we build an undirected graph $G = (V, E)$, the initial elements v_i in V are pixels, the edges E are the edges connecting the pixels. The weight w of the edges is the dissimilarity of two pixels connected by the edge. We define the weight of edge $e_{i,j}$ with their pixel intensity $w_{i,j} = I_i - I_j$. In the beginning, we treat each pixel as a component and merge the components connected with low weight as one component until all the edges have been calculated. MInt(C_i, C_j) is the internal difference in the components C_i and C_j. Each final component serves as a local patch for our prior generation.

Algorithm 1 Generate local patches with Felzenszwalb's efficient graph based segmentation [45]

```
Input: One image
Output: $S = (C_1, C_2, ..., C_r)$
Initialization: We build an undirected graph based on the image color $G = (V, E)$ with $n$ vertices and $m$ edges, where $v_i \in V$, $V$ denotes the image pixels, $e \in E$, $E$ denotes the correspondences of the neighboring vertices, and $w$ denotes the weight of $e$.
Sort the edge $E$ with their edge weight.
$S^0$: Where each $v_i$ is each pixel.
for $q = 1, 2, .., m$ do
    Construct $S^q$ based on $S^{q-1}$
    Let $o_q = (v_i, v_j)$
    Let $v_i \in C_i$ and $v_j \in C_j$
    if $C_i^{q-1} \neq C_j^{q-1}$ and $w(o_q) \leq MInt(C_i^{q-1}, C_j^{q-1})$ then
        $S^q = S^{q-1} \cup C_i^{q-1} \cup C_j^{q-1})$
    else
        $S^q = S^{q-1}$
    end if
end for
```

Patch generation method 2: Simple Linear Iterative Clustering (SLIC). We generate the image patches by clustering pixels based on their color similarity and proximity in the image plane. Following [46], we fuse both the color and position information of each pixel into a five-dimension format ($labxy$). Here lab denotes the color information and xy denotes the spatial position. To cluster the 5D spaces pixels, new distance measures between pixels need to be used. As shown in Algorithm 2, for every input image, we choose K cluster centers C_k with regular grid intervals $S = sqrt(N/K)$, where N denotes the number of pixels of the input image. We assume that the cluster center connects to the associated pixels within a $2S \times 2S$. We calculate the 5D distance (D_s) between the pixels k, i with the following equation:

$$D_s = D_{lab} + D_{xy}/S \times m. \tag{4}$$

In Eq. (4), D_{lab} denotes the color distance and D_{xy} denotes the spatial distance:

$$D_{lab} = \sqrt{((l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2)}, \tag{5}$$

$$D_{xy} = \sqrt{((x_k - x_i)^2 + (y_k - y_i)^2)}. \tag{6}$$

In Eq. (4), we normalize the spatial distance by dividing the interval S and the variable m aims to control the compactness of the superpixel.

We re-center the cluster center to the lowest gradient position within $n \times n$ distance to the existing center (we choose $n = 3$) for every cluster center. The gradient generated with :

$$\begin{aligned} G_{xy} = {} & \|I(x+1, y) - I(x-1, y)\|^2 + \\ & \|I(x, y+1) - I(x, y-1)\|^2. \end{aligned} \tag{7}$$

Where x, y denotes the pixel positional and the $I_{x,y}$ denotes the corresponding color feature vector—each pixel connected to its nearest cluster center within the search distance. Furthermore, a new center is generated by averaging all the pixels within the same cluster. The process of associating pixels with the nearest cluster center and re-centering the cluster is repeated until convergence. When there remain unconnected pixels E less than the $threshold$, we enforce them to connect to the largest neighboring cluster and stop the process of clustering. Each superpixel (cluster) serves as a local patch for our prior generation.

Algorithm 2 Generate local patches with Simple Linear Iterative Clustering

```
Initialize cluster centers $C_k = [C_1, C_2, C_3, ...]$ with a grid distance S
Re-center the cluster within 3 × 3 regions to the lowest gradient position with equation (7)
repeat
    for Each cluster do
        Assign the pixels to the cluster with the distance measure in equation (4)
    end for
    Compute new cluster centers and residual error $E$.
until $E <= threshold$
return Each cluster
```

4.2. Generate region maps for few-shot segmentation tasks

As shown in Fig. 2, in this section, we show how to generate the prior region map and guided region map from our pretrained prior extractor and the feature extractor to benefit the few-shot segmentation tasks.

We set I_q denotes the query images, I_s denotes the support images, and M_s denotes the support annotations. F_p denotes our pretrained prior extractor, and F denotes the feature extractor. We encode the images with:

$$X_q = F(I_q), X_s = F(I_s), P_q = F_p(I_q). \tag{8}$$

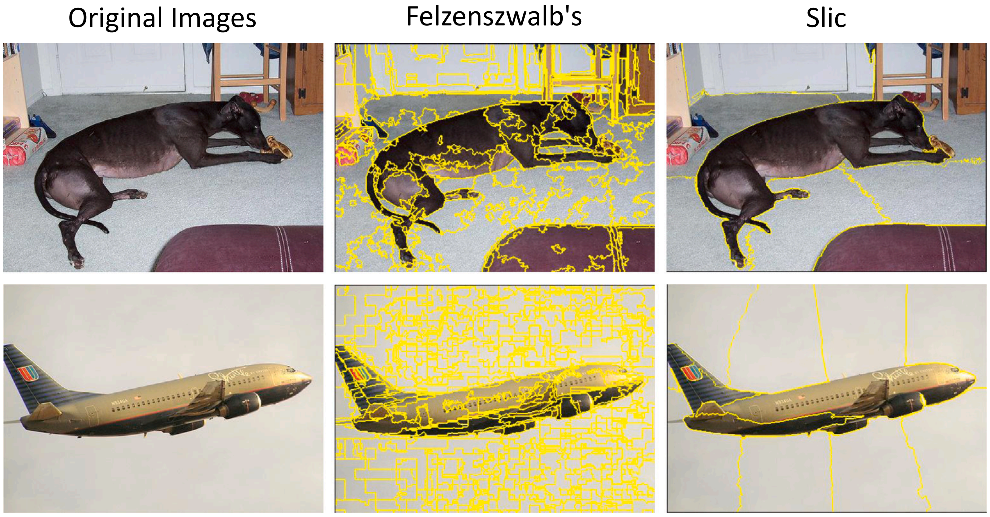

Fig. 4. Visualization of the local patches generated by Felzenszwalb's method and Slic.

4.2.1. Generate prior region maps

For every query image, we retrieve the query information from our prior features by calculating the correlation between the bridge features X_q and P_q. Specifically, we generate the prior region maps with the following steps: we first calculate the pixel-wise cosine similarity $cos(x_q, p_{q'})$ between the bridge feature $x_q \in X_q$, and the prior feature $p_{q'} \in P_q$:

$$cos(x_q, p_{q'}) = \frac{x_q^T p_{q'}}{\left\| x_q \right\| \left\| p_{q'} \right\|} \quad . \tag{9}$$

where $q, q' \in \{1, 2, 3, \dots, hw\}$ and h, w denote the feature size. For the prior region map, we locate the object by defining the confident background with the maximum similarity value alone the all prior features and combine them as the similarity map:

$$s_q = \max_{q' \in \{1,2,\dots,hw\}} (cos(x_q, p_{q'})), \tag{10}$$

$$S_Q = [s_1, s_2, \dots, s_{hw}]. \tag{11}$$

We normalize the correspondence value to $(0, 1)$ with:

$$S_Q = \frac{S_Q - \min(S_Q)}{\max(S_Q) - \min(S_Q) + \epsilon}. \tag{12}$$

where ϵ is set to $1e-7$ in our experiments. Finally, we reshape S_Q into $h \times w \times 1$ to the same size as our final prior region map to locate the objects. To be noticed, as shown in Fig. 2, the prior features are obtained directly from our global–local contrastive learning using the query image. These features are extracted without any labeled information, relying solely on unsupervised learning.

In contrast, the bridge features are derived from the query image, yet they are generated using a shared Feature Extractor with the support image. The parameters of this Feature Extractor have been fine-tuned using labeled training images, which provide domain knowledge.

4.2.2. Generate guided region maps

Following the same step as 4.2.1, we generate the guided region maps with the support features X_s and bridge features X_q within the cross-correspondence module. We filter out the irrelevant support features by multiplying the support features with the annotated support mask $M_s \times X_s$. In this way, the bridge feature pixels yield no correspondence with the background and only correlate with the target object area. Confident query object regions are then identified by selecting the highest cosine similarity for each bridge feature across all filtered support features.

Subsequently, we combine the prior and guided region maps to produce our final correspondence maps, guiding our network's predictions. Additionally, to directly provide category information without unrelated information, we derive the category guider by multiplying the support features with the annotated support mask. We then concatenate the bridge features, the category guider, the prior region maps, and the guided region maps into a final feature group, which is processed by FEM [20] as a decoder to generate our final query masks.

4.3. Opposite region maps

As shown in Fig. 5, we discover that the prior region map generated by bridge features and prior features has a low correspondence value at the object regions while having a high value at the background regions. Our conjecture is that the prior learns the object features from the unlabeled images and stores the information within the prior. The learned object features have been transferred to a new feature space, which is dissimilar to the features generated from the bridge encoder. However, the irrelevant objects (*e.g.*, the background) remain in a similar feature space, which indicates high similarity to the features from the bridge encoder. So the prior region maps locate the target objects with the low similarity value. On the other hand, the support feature and bridge feature stay in the same feature space and are generated by different images; a high similarity score indicates high confidence that the regions contain similar objects.

5. Experiment

5.1. Implementation details

We trained the prior extractor using SGD optimizer with InfoNCE loss [43] and followed MOCO-v2 [35] to apply data augmentation techniques such as blur augmentation, color distortion, random crop, and random flip. The entire dataset was used to train the prior extractor for the COCO dataset with 200 epochs and for the PASCAL VOC dataset with 1000 epochs.

To ensure a fair comparison with previous methods [2–5,20], we selected multiple backbones, including Resnet101, Resnet50, and VGG, and multiple image training sizes (321 × 321, 473 × 473). We used dilated convolution as the backbone to encode the features, maintaining the feature resolution, as done in previous works [2,5,20,47,48]. During the training process, we used data augmentation techniques such as random crop, random scale, and random flip.

We conducted ablation experiments with a Resnet-50 backbone and 473 × 437 training image size, unless specified otherwise.

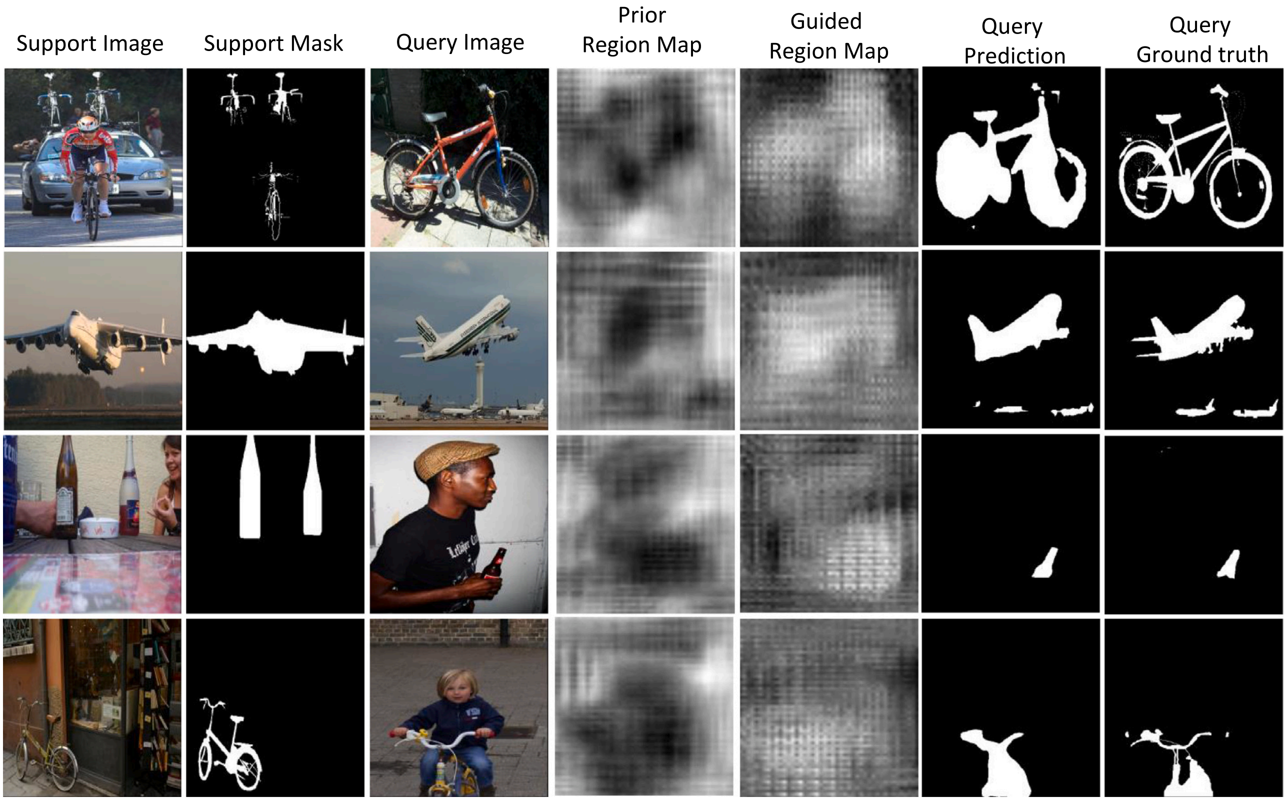

Fig. 5. Visualization results for guided region map, prior region map, and the query prediction generated by our proposed QGNet on PASCAL-5^i dataset.

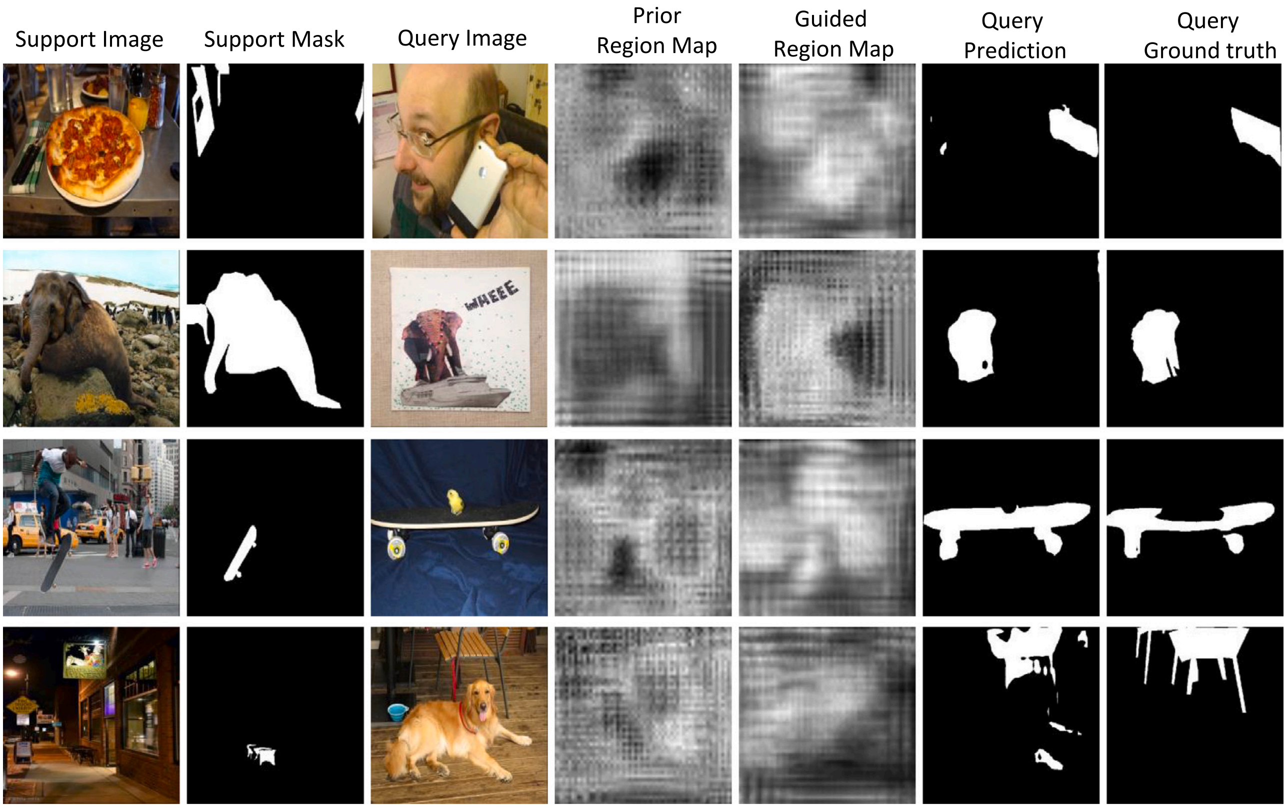

Fig. 6. Visualization results for guided region map, prior region map, and the query prediction generated by our proposed QGNet on COCO dataset.

Table 1
1-shot and 5-shot mIoU results on PASCAL-5^i dataset. The training size and backbone used by each method are listed. Our QGNet outperforms the state-of-the-art under all the experiment settings. The results reported with mIoU (%).

Methods	Training size	Backbone	1-shot					5 shot				
			Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
OSLSM	–	VGG 16	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
co-FCN	–	VGG 16	36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4
AMP-2	–	VGG 16	41.9	50.2	46.7	34.7	43.4	40.3	55.3	49.9	40.1	46.4
PFENet	473 × 473	VGG 16	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
PANet	417 × 417	VGG 16	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
FWBF	512 × 512	VGG 16	47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1
SS-PFENet	–	VGG 16	54.5	67.4	63.4	54.0	59.8	56.9	70.0	68.3	62.1	64.3
Ours	321 × 321	VGG 16	52.9	65.0	50.7	51.6	55.0	56.8	67.3	51.2	58.2	58.3
Ours	473 × 473	VGG 16	58.6	67.2	52.3	52.0	57.5	58.7	68.4	52.6	55.0	58.7
CANet	321 × 321	ResNet 50	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
PGNet	321 × 321	ResNet 50	56.0	66.9	50.6	50.4	56.0	54.9	67.4	51.8	53.0	56.8
CRNet	321 × 321	ResNet 50	–	–	–	–	55.7	–	–	–	–	58.8
PMMs	321 × 321	ResNet 50	55.2	66.9	52.6	50.7	56.3	56.3	67.3	54.5	51.0	57.3
PPNet	417 × 417	ResNet 50	47.8	58.8	53.8	45.6	51.5	58.4	67.8	64.9	56.7	62.0
PANet	417 × 417	ResNet 50	44.0	57.5	50.8	44.0	49.1	55.3	67.2	61.3	53.2	59.3
PFENet	473 × 473	ResNet 50	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
SCL (CANet)	–	ResNet 50	56.8	67.3	53.5	52.5	57.5	59.5	68.5	54.9	53.7	59.2
SCL (PFENet)	–	ResNet 50	63.0	70.0	56.5	57.7	61.8	64.5	70.9	57.3	58.7	62.9
SS-PFENet	–	ResNet 50	58.9	**69.9**	**66.4**	57.7	**63.2**	61.4	**75.0**	**70.5**	**67.7**	**68.6**
Ours	321 × 321	ResNet 50	57.9	67.2	52.4	55.5	58.2	59.2	69.4	53.0	64.5	61.5
Ours	473 × 473	ResNet 50	**63.4**	69.4	55.1	**58.4**	61.6	64.7	71.0	53.6	61.6	62.8
FWBF	512 × 512	ResNet 101	51.3	64.5	56.7	52.2	56.2	54.8	67.4	62.2	55.3	59.9
PPNet	417 × 417	ResNet 101	52.7	62.8	57.4	47.7	55.2	60.3	70.0	69.4	60.7	65.1
DAN	–	ResNet 101	54.7	68.6	57.8	51.6	58.2	57.9	69.0	60.1	54.9	60.5
PFENet	473 × 473	ResNet 101	60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
Ours	321 × 321	ResNet 101	57.6	67.5	53.0	53.6	57.9	58.6	66.5	53.3	53.6	58.0
Ours	473 × 473	ResNet 101	60.3	69.3	53.3	57.4	60.1	**67.6**	71.8	55.1	64.4	64.7

Table 2
Comparison with the state-of-the-art methods under the 1-shot and 5-shot setting. The results reported on PASCAL VOC 2012 dataset with FBIoU (%).

Methods	1-shot (%)	5-shot (%)
OSLM	61.3	61.5
co-fcn	60.9	60.2
sg-one	63.1	65.9
R-DFCN	60.9	66.0
PL	61.2	62.3
A-MCG	61.2	62.2
CANet	66.2	69.6
CRNet	66.8	71.5
Ours	**72.2**	**73.9**

5.2. Datasets and evaluation metric

PASCAL VOC 2012. The PASCAL-5^i dataset [49] consists of 20 categories, with 10 582 images for training, 1449 for validation, and 1456 for testing. To ensure comparability with previous work, we follow the data split used in [20] and divide the dataset into training and testing sets. Specifically, we use a cross-validation approach where we divide the 20 object categories into 4 folds, with three for training and one for testing. During training, we use a batch size of 4 and an initial learning rate of 0.0025. We train our network for 200 epochs.

MS COCO. One limitation of PASCAL-5^i is that it involves only 20 categories, which may not be enough to evaluate the model's ability to perform few-shot segmentation tasks. To address this, we conduct cross-validation experiments on the larger MS COCO dataset, which contains more categories and images. Specifically, COCO 2014 [50] includes 80 object categories with 82 783 training images and 40 504 validation images. Similar to previous works [20], we divide the 80 categories into four folds, using three folds for training and one fold for testing.

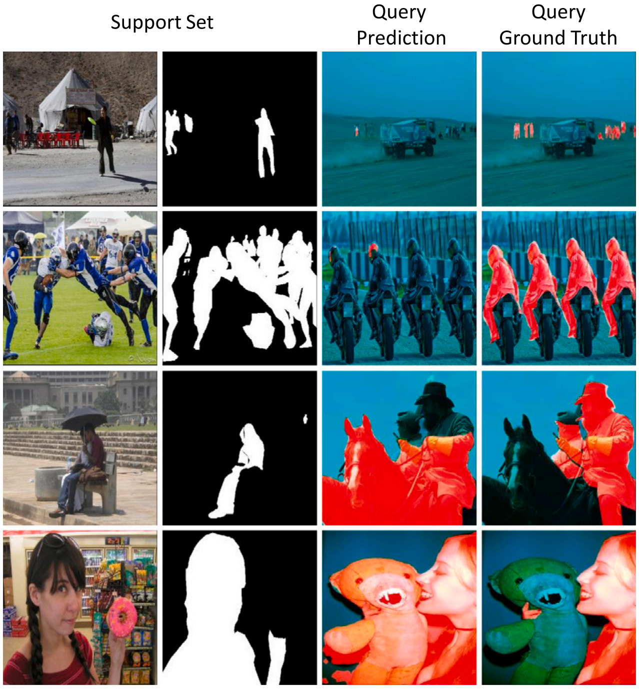

Fig. 7. The failure cases on COCO dataset.

Evaluation Metric. In previous works, there exist two evaluation metrics. Shabanet al. [9] report the results with the standard mean Intersection-Over-Union(mIoU). While [10,11] ignore the categories and report the results by averaging of foreground IoU and background IoU (FBIoU). Following the previous works [2,9,20], we choose the

Table 3
1-shot and 5-shot mIoU results on COCO dataset. The results of CANet* is obtained from [28].

Methods	Training size	Backbone	1-shot					5 shot				
			Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
PANet	417 × 417	VGG 16	–	–	–	–	20.9	–	–	–	–	29.7
FWBF	512 × 512	VGG 16	18.4	16.7	19.6	25.4	20.0	20.9	19.2	21.9	28.4	22.6
PFENet	473 × 473	VGG 16	33.4	36.0	34.1	32.8	34.1	35.9	40.7	38.1	36.1	37.7
SS-PANet	–	VGG 16	29.8	21.2	26.5	28.5	26.2	36.7	41.0	37.6	35.6	37.7
SS-PFENet	–	VGG 16	35.6	39.2	37.6	37.3	37.5	40.4	45.8	40.3	40.7	41.8
Ous	321 × 321	VGG 16	32.2	35.2	32.2	31.9	32.8	33.8	38.5	37.3	36.4	36.5
Ous	473 × 473	VGG 16	32.4	38.1	35.7	32.7	34.7	33.4	42.2	39.5	37.6	38.2
CANet*	321 × 321	ResNet 50	25.1	30.3	24.5	24.7	26.1	26.0	32.4	26.1	27.0	27.9
PMMs	321 × 321	ResNet 50	29.5	36.8	29.0	27.0	30.6	33.8	42.0	33.0	33.3	35.5
PPNet	–	ResNet 50	28.1	30.8	29.5	27.7	29.0	39.0	40.8	37.1	37.3	38.5
RPMM	321 × 321	ResNet 50	29.5	36.8	28.9	27.0	30.6	33.8	42.0	33.0	33.3	35.5
PFENet	473 × 473	ResNet 50	36.5	38.6	34.5	33.8	35.8	36.5	43.3	37.8	38.4	39.0
Ous	321 × 321	ResNet 50	34.4	38.5	34.6	33.2	35.2	38.9	46.2	39.1	38.9	40.7
Ours	473 × 473	ResNet 50	36.7	41.4	38.7	36.6	38.3	41.5	**48.1**	46.3	**43.6**	44.8
DAN	–	ResNet 101	–	–	–	–	24.4	–	–	–	–	29.6
FWBF	512 × 512	ResNet 101	19.9	18.0	21.0	28.9	21.2	19.1	21.5	23.9	30.1	23.7
PFENet	641 × 641	ResNet 101	34.3	33.0	32.3	30.1	32.4	38.5	38.6	38.2	34.3	37.4
SCL (PFENet)	–	ResNet 101	36.4	38.6	37.5	35.4	37.0	38.9	40.5	41.5	38.7	39.9
Ous	321 × 321	ResNet 101	33.0	38.4	36.8	32.8	35.2	38.9	46.7	43.1	40.4	42.3
Ous	473 × 473	ResNet 101	**39.0**	**42.6**	**40.5**	**40.2**	**40.5**	**43.8**	46.8	**47.5**	42.6	**45.2**

Table 4
Comparison with the state-of-the-art methods under the 1-shot and 5-shot setting. Our proposed network outperforms all previous methods and achieves new state-of-the-art performance. The results reported on MS COCO dataset with FBIoU (%).

Method	1-shot (%)	5-shot (%)
PANet	59.2	63.5
A-MCG	52.0	54.7
PFENet	58.6	61.9
DAN	62.3	63.9
Ours	**62.4**	**66.2**

Table 5
Ablation studies on PASCAL-5^i dataset about the training epochs and the local patch generation methods for prior extractor training. The results reported with mIoU (%).

Methods	Number of epochs	1 shot				
		fold 0	fold 1	fold 2	fold 3	mean
Felzenszwalb	1000	62.6	68.9	54.9	56.7	60.8
Felzenszwalb	2000	62.7	69.3	54.7	56.8	60.9
Slic	1000	63.4	69.4	55.1	58.4	61.6
Slic	2000	62.4	69.4	55.2	57.0	61.0

Table 6
Ablation studies on PASCAL-5^i dataset about global and local contrastive learning. The results reported with mIoU (%).

Contrastive learning type:		1 shot				
local	global	fold 0	fold 1	fold 2	fold 3	mean
√		63.2	69.3	54.9	57.4	61.2
	√	62.2	69.7	54.5	57.5	60.9
√	√	63.4	69.4	55.1	58.4	61.6

standard mIoU as our evaluation metric for the following reasons: (1) The unbalanced image distribution (*e.g.*, in the PASCAL VOC test dataset, class sheep contains 49 images while the class person contains 378 images). (2) The score of the background IoU is very high for small objects, which will fail to evaluate the model's capability. Nevertheless, we still compare the previous state-of-the-art methods with both evaluation metrics. The evaluation metrics are calculated as follows:

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP + FP + FN}, \tag{13}$$

$$mIoU = \frac{1}{n}\sum_{1}^{n}(IoU_n), \tag{14}$$

$$FBIoU = \frac{1}{2}(IoU_{fg} + IoU_{bg}). \tag{15}$$

The TP denotes true positive, FP denotes false positive, FN denotes false negative, n denotes the classes' number. The standard mIoU is calculated by averaging the IoU of all classes. The IoU_{fg} is calculated with Eq. (13), which ignores the object categories, and IoU_{bg} is calculated in the same way but reverses the foreground and background. FBIoU average the IoU_{fg} and the IoU_{bg}.

5.3. Comparisons with state-of-the-art

PASCAL VOC 2012. To ensure a fair comparison with previous state-of-the-art (SOTA) methods, we conducted experiments using multiple backbone models (VGG16, Resnet50, Resnet101) and training image sizes (321 × 321 and 473 × 473). The choice of backbone and image size can significantly impact the final performance of a model.

Table 1 displays the mean intersection-over-union (mIoU) results of various methods on the PASCAL-5^i dataset. Our proposed method achieved the best results across different backbone models and image sizes. Notably, our approach outperforms the previous SOTA method, PFENet [20], by 9.4% and 3.0% mIoU in the 1-shot and 5-shot settings, respectively, when using the VGG-16 backbone. Even when compared to FWBF [51], which utilizes a larger training size, our method still outperforms it. For Resnet-101, our approach yields a 3.3 mIoU score improvement over PFENet [20] in the 5-shot setting. These results demonstrate the effectiveness and robustness of our method across different backbones and image sizes, and confirm that our approach achieves superior performance over previous SOTA methods. In addition, we compare our method with existing approaches under the FB-IoU metric, as summarized in Table 2.

MS COCO. Table 3 shows the comparison of our proposed QGNet with previous state-of-the-art methods in terms of the standard mIoU. In the large-scale MS COCO dataset experiments, PFENet used a ResNet-101 backbone with an image size of 641 × 641 for training. However, due to GPU memory constraints, we trained our model with multiple backbones (ResNet-50, ResNet-101, VGG) and two different training

Table 7
Ablation studies on PASCAL-5^i fold-0 dataset about the recall for guided region map and prior region map with ground truth query mask. The results reported with mIoU (%).

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$Recall_{g,gt}$	99.9%	99.9%	99.3%	96.0%	85.6%	65.6%	39.6%	15.5%	2.2%
$Recall_{p,gt}$	99.9%	99.9%	99.1%	96.5%	89.7%	76.2%	54.7%	28.2%	6.6%

sizes (473 × 473 and 321 × 321). Despite this limitation, our proposed QGNet outperformed the previous state-of-the-art (PFENet) by 5.9% and 7.4% in the 1-shot and 5-shot settings, respectively, using the ResNet-50 backbone.

It is worth noting that even though previous methods, such as PFENet [20] and FWBF [51], used larger training image sizes (e.g., 641 × 641 or 512 × 512), our method with the same backbone (e.g., ResNet-101) still outperforms all those methods with smaller training sizes. In particular, our method with a training size of 473 × 473 achieves a 10.1% mIoU improvement in Fold-3 with a 1-shot setting. For the 1-shot mean performance, our method outperforms PFENet by 8.1% mIoU and FWBF by more than 19.3% mIoU score. For the 5-shot setting, our method outperforms the previous state-of-the-art with a mean 7.8% mIoU with a smaller training size. These results demonstrate the superiority of our proposed QGNet. We also compare our method with existing approaches under the FB-IoU metric, as summarized in TableTable 4.

In addition, we visualize some testing results on fold 0 of the PASCAL-5^i dataset in Fig. 5 and some testing results on the COCO dataset in Fig. 6.

5.4. Ablation studies

Our method introduces a novel self-correspondence module that generates a prior region map for query images independently. To demonstrate the effectiveness of our proposed global and local contrastive losses for prior extractor training, we conducted ablation experiments. We also conducted experiments to identify the best approach for local patch generation during local contrastive loss training. To gain further insights into the impact of the generated prior region maps on query mask prediction, we analyzed the overlap between the prior region maps and the guided region maps.

Prior Extractor Training. We performed ablation experiments to analyze which patch generation method produces better performance for prior extractor training. Our results, presented in Table 5, demonstrate that we obtain improved query mask predictions when using Slic [46] as the patch generation method. Fig. 4 illustrates that the patches generated by Slic are larger than those generated by Felzenszwalb's method and each patch is likely capturing a distinct object pattern. We posit that an appropriate patch size may enhance the prior extractor's capacity to learn a more distinctive predetermined prior.

Furthermore, in Table 5, we demonstrate our ablation experiments. These results indicate that training the prior extractor for 1000 epochs with the Slic method achieves the best performance.

Effectiveness of global and local contrastive losses. We conducted ablation experiments to evaluate the effectiveness of our proposed global and local contrastive losses. The results, presented in Table 6, indicate that the combination of both losses yields the best performance. This suggests that using global and local contrastive losses during prior extractor training can help store more discriminative prior for query mask prediction.

Recall for region maps. Furthermore, we conducted additional experiments to assess the recall ability of the guided region map and prior region map with the ground truth query segmentation mask. We defined the object regions R with a threshold α, such that if the Intersection over Union (IoU) between the ground truth query mask and the object region is greater than α, the region is considered a true positive.

It is worth noting that the 'Prior' map contains more information compared to the 'Guided Region Map.' During the generation of the guided region map, we filtered out background information and retained only the target information. This filtering process may lead to the loss of some important background information, even though the background can provide valuable information for few-shot segmentation.

$$R = \begin{cases} 1, & \text{if } V(x,y) > \alpha, \\ 0, & \text{otherwise}. \end{cases} \tag{16}$$

V denotes the normalized values of the feature map.

The recall of the region maps with groundtruth mask is calculated as $Recall = (R \cap R_{GT})/R_{GT}$. Here R_{GT} indicates the groundtruth mask region. $Recall$ denotes the overlap between the region map and groundtruth mask divided by the groundtruth mask. The higher the $Recall$, the more object regions of the region map can cover the groundtruth mask. In Table 7, the recall of the prior region map (denoted as $Recall_{p,gt}$) is larger than the recall of the guided region map (denoted as $Recall_{g,gt}$), which suggests that the prior region map could cover more potential object regions.

5.5. Failure case analysis

In this section, we aim to analyze the challenging cases that our model fails to correctly identify on the COCO dataset. Fig. 7 shows some of these instances where our model is unable to distinguish between people and Teddy bears, horses, and people due to their similar patterns or colors. These scenarios are difficult to distinguish without the use of semantic information. Furthermore, our model also struggles to locate small objects in the images. When the query information is too complex, our model can only identify a small part of the human being, indicating that more advanced techniques may be required to improve performance in such challenging cases.

6. Conclusion and future work

This paper proposes the Query Guided Few Shot Segmentation method (QGNet) which employs self-supervised learning to learn query information from unlabeled images. To achieve this, the paper introduces a global–local contrastive loss to train the prior extractor, enabling the query branch to independently extract informative clues from the query image and enhancing cross-interaction between query and support. Experimental results on the PASCAL VOC 2012 and MS COCO datasets demonstrate the effectiveness of the proposed method, achieving new state-of-the-art results.

In the future, integrating techniques from perceptual quality evaluation [52–54] and image quality estimation [55–58] could establish a robust framework for objectively quantifying improvements in segmentation clarity and accuracy. Additionally, the method's capability for independent query feature extraction suggests promising potential for adaptation to 360-degree images. By leveraging visual behavior modeling and saliency prediction frameworks developed for augmented 360-degree video analysis [59–61], our approach could be effectively extended, enhancing spatial coherence and visual attention handling within immersive media applications. Furthermore, incorporating attention mechanisms and saliency-driven approaches into few-shot segmentation, as demonstrated in recent studies [62–64], presents another valuable avenue, potentially improving the model's capacity to prioritize meaningful regions and further boosting segmentation performance under limited supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[2] C. Zhang, G. Lin, F. Liu, R. Yao, C. Shen, CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5217–5226.

[3] K. Wang, J.H. Liew, Y. Zou, D. Zhou, J. Feng, PANet: Few-shot image semantic segmentation with prototype alignment, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9197–9206.

[4] Y. Liu, X. Zhang, S. Zhang, X. He, Part-aware prototype network for few-shot semantic segmentation, in: European Conference on Computer Vision, Springer, 2020, pp. 142–158.

[5] W. Liu, C. Zhang, G. Lin, F. Liu, CRNet: Cross-reference networks for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4165–4173.

[6] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, CVPR'06, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2006, pp. 1735–1742.

[7] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.

[9] A. Shaban, S. Bansal, Z. Liu, I. Essa, B. Boots, One-shot learning for semantic segmentation, 2017, arXiv preprint arXiv:1709.03410.

[10] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, S. Levine, Conditional networks for few-shot semantic segmentation, in: ICLR Workshop, 2018.

[11] N. Dong, E. Xing, Few-shot semantic segmentation with prototype learning, in: British Machine Vision Conference, 2018.

[12] X. Zhang, Y. Wei, Y. Yang, T.S. Huang, Sg-one: Similarity guidance network for one-shot semantic segmentation, IEEE Trans. Cybern. 50 (9) (2020) 3855–3865.

[13] B. Peng, Z. Tian, X. Wu, C. Wang, S. Liu, J. Su, J. Jia, Hierarchical dense correlation distillation for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 23641–23651.

[14] Q. Xu, W. Zhao, G. Lin, C. Long, Self-calibrated cross attention network for few-shot segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 655–665.

[15] Q. Xu, G. Lin, C.C. Loy, C. Long, Z. Li, R. Zhao, Eliminating feature ambiguity for few-shot segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2024, pp. 416–433.

[16] S. Moon, S.S. Sohn, H. Zhou, S. Yoon, V. Pavlovic, M.H. Khan, M. Kapadia, MSI: Maximize support-set information for few-shot segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 19266–19276.

[17] Y. Yang, Q. Chen, Y. Feng, T. Huang, MIANet: Aggregating unbiased instance and general information for few-shot semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 7131–7140.

[18] L. Zhu, T. Chen, D. Ji, J. Ye, J. Liu, LLaFS: When large language models meet few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 3065–3075.

[19] M.R.I. Hossain, M. Siam, L. Sigal, J.J. Little, Visual prompting for generalized few-shot segmentation: A multi-scale approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 23470–23480.

[20] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, J. Jia, Prior guided feature enrichment network for few-shot segmentation, Trans. Pattern Anal. Mach. Intell. (2020).

[21] W. Liu, C. Zhang, G. Lin, F. Liu, Crcnet: Few-shot segmentation with cross-reference and region–global conditional networks, Int. J. Comput. Vis. 130 (12) (2022) 3140–3157.

[22] W. Liu, C. Zhang, H. Ding, T.-Y. Hung, G. Lin, Few-shot segmentation with optimal transport matching and message flow, IEEE Trans. Multimed. 25 (2022) 5130–5141.

[23] W. Liu, Z. Wu, Y. Zhao, Y. Fang, C.-S. Foo, J. Cheng, G. Lin, Harmonizing base and novel classes: A class-contrastive approach for generalized few-shot segmentation, Int. J. Comput. Vis. 132 (4) (2024) 1277–1291.

[24] X. Li, T. Wei, Y.P. Chen, Y.-W. Tai, C.-K. Tang, Fss-1000: A 1000-class dataset for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2869–2878.

[25] M. Siam, B. Oreshkin, Adaptive masked weight imprinting for few-shot segmentation, 2019, arXiv preprint arXiv:1902.11123.

[26] M. Siam, B.N. Oreshkin, M. Jagersand, AMP: Adaptive masked proxies for few-shot segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5249–5258.

[27] Z. Dong, R. Zhang, X. Shao, H. Zhou, Multi-scale discriminative location-aware network for few-shot semantic segmentation, COMPSAC, in: 2019 IEEE 43rd Annual Computer Software and Applications Conference, vol. 2, IEEE, 2019, pp. 42–47.

[28] B. Yang, C. Liu, B. Li, J. Jiao, Q. Ye, Prototype mixture models for few-shot semantic segmentation, 2020, arXiv preprint arXiv:2008.03898.

[29] Y. Li, G.W.P. Data, Y. Fu, Y. Hu, V.A. Prisacariu, Few-shot semantic segmentation with self-supervision from pseudo-classes, in: British Machine Vision Conference, 2021.

[30] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, J. Kim, Adaptive prototype learning and allocation for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8334–8343.

[31] M. Boudiaf, H. Kervadec, I.M. Ziko, P. Piantanida, I.B. Ayed, J. Dolz, Few-shot segmentation without meta-learning: A good transductive inference is all you need? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13979–13988.

[32] Z. Lu, S. He, X. Zhu, L. Zhang, Y.-Z. Song, T. Xiang, Simpler is better: Few-shot semantic segmentation with classifier weight transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8721–8730.

[33] K. Dong, W. Yang, Z. Xu, L. Huang, Z. Yu, Abpnet: Adaptive background modeling for generalized few shot segmentation, in: ACM MM, 2021, pp. 2271–2280.

[34] B. Zhang, J. Xiao, T. Qin, Self-guided and cross-guided learning for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6933–6942.

[35] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, 2020, arXiv preprint arXiv:2003.04297.

[36] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P.H. Richemond, E. Buchatskaya, C. Doersch, B.A. Pires, Z.D. Guo, M.G. Azar, et al., Bootstrap your own latent: A new approach to self-supervised learning, 2020, arXiv preprint arXiv:2006.07733.

[37] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[38] H. Hu, J. Cui, L. Wang, Region-aware contrastive learning for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16291–16301.

[39] X. Ji, J.F. Henriques, A. Vedaldi, Invariant information clustering for unsupervised image classification and segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9865–9874.

[40] R. Achanta, F. Estrada, P. Wils, S. Süsstrunk, Salient region detection and segmentation, in: Computer Vision Systems: 6th International Conference, ICVS 2008 Santorini, Greece, May 12-15, 2008 Proceedings 6, Springer, 2008, pp. 66–75.

[41] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu, Contrastive learning of global and local features for medical image segmentation with limited annotations, Adv. Neural Inf. Process. Syst. 33 (2020) 12546–12558.

[42] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, P. Luo, Detco: Unsupervised contrastive learning for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8392–8401.

[43] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv preprint arXiv:1807.03748.

[44] Z. Wu, Y. Xiong, S.X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.

[45] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2) (2004) 167–181.

[46] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.

[47] W. Liu, X. Wang, L. Wang, J. Cheng, F. Liu, X. Yang, Gaussian mixture based evidential learning for stereo matching, 2024, arXiv preprint arXiv:2408.02796.

[48] W. Liu, J. Lou, X. Wang, W. Zhou, J. Cheng, X. Yang, Physically-guided open vocabulary segmentation with weighted patched alignment loss, Neurocomputing 614 (2025) 128788.

[49] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: ECCV, 2014, pp. 740–755.

[51] K. Nguyen, S. Todorovic, Feature weighting and boosting for few-shot segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 622–631.

[52] X. Min, H. Duan, W. Sun, Y. Zhu, G. Zhai, Perceptual video quality assessment: A survey, Sci. China Inf. Sci. 67 (11) (2024) 211301.

[53] Y. Zhu, Y. Li, W. Sun, X. Min, G. Zhai, X. Yang, Blind image quality assessment via cross-view consistency, IEEE Trans. Multimed. 25 (2022) 7607–7620.

[54] Y. Zhang, J. Wang, Y. Zhu, R. Xie, Subjective and objective quality evaluation of UGC video after encoding and decoding, Displays 83 (2024) 102719.

[55] X. Min, G. Zhai, K. Gu, Y. Liu, X. Yang, Blind image quality estimation via distortion aggravation, IEEE Trans. Broadcast. 64 (2) (2018) 508–517.

[56] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, C.W. Chen, Blind quality assessment based on pseudo-reference image, IEEE Trans. Multimed. 20 (8) (2017) 2049–2062.

[57] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, W. Lin, Unified blind quality assessment of compressed natural, graphic, and screen content images, IEEE Trans. Image Process. 26 (11) (2017) 5462–5474.

[58] X. Min, G. Zhai, K. Gu, X. Yang, X. Guan, Objective quality evaluation of dehazed images, IEEE Trans. Intell. Transp. Syst. 20 (8) (2018) 2879–2892.

[59] Y. Zhu, X. Min, D. Zhu, G. Zhai, X. Yang, W. Zhang, K. Gu, J. Zhou, Toward visual behavior and attention understanding for augmented 360 degree videos, ACM Trans. Multimed. Comput. Commun. Appl. 19 (2s) (2023) 1–24.

[60] Y. Zhu, G. Zhai, Y. Yang, H. Duan, X. Min, X. Yang, Viewing behavior supported visual saliency predictor for 360 degree videos, IEEE Trans. Circuits Syst. Video Technol. 32 (7) (2021) 4188–4201.

[61] Y. Zhu, G. Zhai, X. Min, J. Zhou, Learning a deep agent to predict head movement in 360-degree images, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 16 (4) (2020) 1–23.

[62] X. Min, G. Zhai, J. Zhou, M.C. Farias, A.C. Bovik, Study of subjective and objective quality assessment of audio-visual signals, IEEE Trans. Image Process. 29 (2020) 6054–6068.

[63] Y. Cao, X. Min, W. Sun, G. Zhai, Attention-guided neural networks for full-reference and no-reference audio-visual quality assessment, IEEE Trans. Image Process. 32 (2023) 1882–1896.

[64] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, X. Guan, A multimodal saliency model for videos with high audio-visual correspondence, IEEE Trans. Image Process. 29 (2020) 3805–3819.