CGMGM: A Cross-Gaussian Mixture Generative Model for Few-Shot Semantic Segmentation

Junao Shen[1], Kun Kuang[2], Jiaheng Wang[1], Xinyu Wang[1], Tian Feng[1*], Wei Zhang[1, 3]

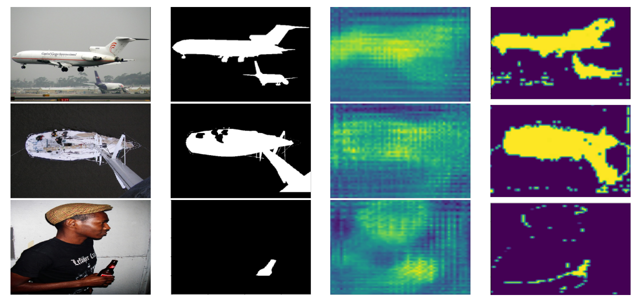
[1]School of Software Technology, Zhejiang University
[2]College of Computer Science and Technology, Zhejiang University
[3]Innovation Center of Yangtze River Delta, Zhejiang University
{jashen, kunkuang, jiahengwang, xinyu.w, t.feng, cstzhangwei}@zju.edu.cn

Abstract

Few-shot semantic segmentation (FSS) aims to segment unseen objects in a query image using a few pixel-wise annotated support images, thus expanding the capabilities of semantic segmentation. The main challenge lies in extracting sufficient information from the limited support images to guide the segmentation process. Conventional methods typically address this problem by generating single or multiple prototypes from the support images and calculating their cosine similarity to the query image. However, these methods often fail to capture meaningful information for modeling the de facto joint distribution of pixel and category. Consequently, they result in incomplete segmentation of foreground objects and mis-segmentation of the complex background. To overcome this issue, we propose the Cross Gaussian Mixture Generative Model (CGMGM), a novel Gaussian Mixture Models (GMMs)-based FSS method, which establishes the joint distribution of pixel and category in both the support and query images. Specifically, our method initially matches the feature representations of the query image with those of the support images to generate and refine an initial segmentation mask. It then employs GMMs to accurately model the joint distribution of foreground and background using the support masks and the initial segmentation mask. Subsequently, a parametric decoder utilizes the posterior probability of pixels in the query image, by applying the Bayesian theorem, to the joint distribution, to generate the final segmentation mask. Experimental results on PASCAL-5^i and COCO-20^i datasets demonstrate our CGMGM's effectiveness and superior performance compared to the state-of-the-art methods.

Introduction

Semantic segmentation is a fundamental task in the field of computer vision and can formulated as a pixel-wise classification problem. Recent significant improvements have occurred to semantic segmentation given the considerable development of deep neural networks (DNNs) (Chen et al. 2018; Huang et al. 2019; Zhao et al. 2017; Yuan et al. 2019; Xie et al. 2021). Conventional methods for semantic segmentation rely heavily on large amount annotated datasets (Everingham et al. 2010; Nguyen and Todorovic 2019), whereas collecting high-quality data is time-consuming and laborious. In the case of extremely limited data, the performances of these method may degrade notably. To address this challenge, few-shot semantic segmentation (FSS) (Shaban et al. 2017; Cheng et al. 2021), which aims to segment the objects of a novel category in a *query* image with a few annotated *support* images for training, has emerged as a noteworthy subfield of semantic segmentation. The setting of FSS is closer than that of the general semantic segmentation to the way humans recognize unseen objects in the real word with limited supporting information or knowledge. Obviously, the key difficulty of FSS is to extract adequate guidance information from the support images.



(a) Query (b) GT (c) CS (d) GP

Figure 1: Examples of cosine similarity (CS) and GMMs posterior (GP), compared to Ground Truth (GT), on Pascal-5^i (Shaban et al. 2017) under 1-shot setting.

Existent FSS methods (Lang et al. 2022a; Tian et al. 2020; Li et al. 2021; Zhang et al. 2019) counteract the above-mentioned difficulty via extracting semantic-level prototypes from the feature representations of the support images and adopting a metric learning pipeline between these prototypes and the feature representation of the query image for guidance information. Among these methods, the prototypical ones (Lang et al. 2022a; Li et al. 2021; Zhang et al. 2019, 2021) have achieved the state-of-the-art performance, while they undergo specific limitations. First, the conventional process can be summarized as guiding the feature representation of the query image using prototypes, no matter category-wise (Zhang et al. 2019; Lang et al. 2022a), cluster-wise (Li et al. 2021) or pixel-wise (Zhang et al. 2021), and then adopting simplistic cosine similarity for measurement. As shown in Figure 1, calculating cosine similarity relies

*Corresponding author (t.feng@zju.edu.cn).

Guidance Method	Similarity Measure	GmIoU	FmIoU
Prototype with CS (Lang et al. 2022a)	$(x_q^T x_s)/(\parallel x_q \parallel\parallel x_s \parallel)$	24.11	67.81
CGMGM (Ours)	$\sum \pi_{cm}^{s,q}\mathcal{N}(x_q;\mu_{cm}^{s,q},\Sigma_{cm}^{s,q})$	58.05	69.85
CGMGM w/ GT	$\sum \pi_{cm}^{q}\mathcal{N}(x_q;\mu_{cm}^{q},\Sigma_{cm}^{q})$	64.89	82.17

Table 1: Comparison between our CGMGM and prototype with cosine similarity, *i.e.*, BAM (Lang et al. 2022a), on Pascal-5^i under 1-shot setting. The mIoU between the similarity map and the ground truth as *Guidance mIoU* to quantify the guidance information from the feature representations of the support image x_s to that of the query image x_q, where CS, GT, GmIoU, and FmIoU denote Cosine Similarity, Ground Truth, Guidance mIoU, and Final mIoU.

heavily on generalizing the feature representations of the support images and can hardly model the *de facto* joint distribution of pixel and category. Second, the guidance information extracted from the support images lacks of the cues on the novel category in the query image, which results in category bias during the generalization of feature representations. Third, the information on the scene (*i.e.*, background) has been so far only considered from the perspective of the support images (Lang et al. 2022b), whereas the scene of the query image can differ significantly. It may aggravate the performance for FSS without introducing the information on the scene of the query image.

We propose a novel generative method for FSS, that is, the Cross Gaussian Mixture Generative Model (CGMGM), to address above-discussed limitations via modeling the joint distribution of pixel and category in the support images and the query image. Normally, a generative model (Bernardo et al. 2007) establishes the joint distribution $p(x, c)$ between high-dimensional feature corresponding to pixel x and category c, and use $p(x, c)$ to evaluate the category conditional probability $p(x|c)$, which is to model the input data itself (Liang et al. 2022) and thus has the potential to overcome the shortcomings of previous methods. Our CGMGM models the de facto joint distribution of pixel and category to provide high-quality results for FSS. Specifically, an initial segmentation mask is generated via matching the feature representations of the support images and the query image, and is further refined using a cycle-consistency strategy; Separate mixtures of Gaussians are then adopted to model the joint distribution of pixel and category for the novel category (*i.e.*, *foreground*) and others (*i.e.*, *background*); Afterwards, the category posterior probability is evaluated for each pixel in the query image, serving as the guidance information of the support images for the query image; Finally, a parametric decoder takes as input both the posterior probability and the feature representation of the query image to predict a high-quality segmentation mask for the query image. In particular, the proposed method optimizes a generative Gaussians mixture model (GMM) (Reynolds et al. 2009) uisng the Expectation Maximization (EM) algorithm (Dempter 1977) during training. Besides, we employ an end-to-end design for the proposed method and the cross-entropy loss function to maximize the joint distribution and the representation learning parameters.

The proposed method differs fundamentally from previous methods in the following ways. **First**, we exploit the de facto joint distribution of pixel and category to guide the segmentation, instead of relying on prototypes and cosine similarity, as shown in Figure 1 and Table 1. **Second**, we use the information on both background and foreground from the support images and the query image to overcome the limitation of only involving the support images, which enables a significant improvement in the performance under the 1-shot setting. **Third**, our distribution modeling process is entirely parameter-free enhancing the generalization of the proposed method and preventing the overfitting to base categories, which is particularly crucial to FSS. We conduct extensive experiments on two datasets (*i.e.*, PASCAL-5^i(Shaban et al. 2017) and COCO-20^i(Nguyen and Todorovic 2019)), where our CGMGM can achieve the state-of-the-art performances.

The primary contributions of this paper are three-fold: (1) To the best of our knowledge, we for the first time introduce the GMMs, which establish the joint distribution of pixel and category, to FSS; (2) We propose a novel method to exploit the information on foreground and background in the support images and the query image, which guides the segmentation; (3) The proposed method can achieve the state-of-the-art performances on PASCAL-5^i and COCO-20^i under both 1-shot and 5-shot settings.

Related Work

Few-Shot Segmentation

FSS aims to predict the query image's pixel target in the condition of a few annotated samples (Shaban et al. 2017). The primary challenge lies in extracting abundant information from the support images to guide the query image segment. Subsequent studies (Li et al. 2021; Lang et al. 2022a; Zhang et al. 2019; Tian et al. 2020; Yang et al. 2020) extract the typical information, called 'prototype'. For example, ASGNet (Li et al. 2021) and PMM (Yang et al. 2020) construct multiple prototypes using parameter-free methods such as superpixel-guided clustering or EM algorithm. PFENet (Tian et al. 2020) uses high-level features with cosine similarity to generate prior mask and introduces a feature-enrichment module. BAM (Lang et al. 2022a) proposes a base learner to predict the base category in the query image and uses the results of base learner to suppress false segmentation in the meta learner. However, relying solely on category-wise prototypes and cosine similarity may result in spatial structural loss. To address this issue, HSNet (Min, Kang, and Cho 2021) exploits pixel-wise cosine similarity information from support and query features and constructs 4D correlation tensors to represent dense correspondences. However, all these methods neglect the joint distribution of pixel and category in the support images and the query image. In this paper, we propose GMMs to model the de facto joint distribution to alleviate these issues.

Gaussian Mixture Generative Model

Generative models and discriminative models can be seem as two contrasting ways of solving deep learning classification tasks (Bernardo et al. 2007). Generative models (*i.e.*, naive Bayes) learn the category conditional probability $p(x|c)$, while the discriminative models (*i.e.*, MLP with softmax) learn the category posterior $p(c|x)$ without consider about the underlying joint distribution. MLP-softmax is widely used in classification models due to its simplicity and efficiency. However, generative models still have great potential in fields where accurate modeling of data distribution and high generalization ability are required. Consequently, recent Trusty AI-related fields have focused on generative models, *i.e.*, adversarial defense, explainable AI (Mackowiak et al. 2021; Schott et al. 2018; Serrà et al. 2019), out-of-distribution recognition (Lu et al. 2022), and semi-supervised learning (Izmailov et al. 2020).

Gaussian Mixture Models (GMMs) have been combined with neural networks in the standard supervised classification tasks in early studies given their capability to model arbitrary continuous distributions. However, they almost focus on training the GMMs in a *discriminatively* way (*i.e.*, maximizing the category posterior $p(c|x)$) (Hayashi and Uchida 2019; Variani, McDermott, and Heigold 2015; Tüske et al. 2015; Klautau, Jevtic, and Orlitsky 2003). Recent studies(Liang et al. 2022; Lu et al. 2022) have instead exploited the nature of generative models and adopted GMMs to evaluate the category conditional probability $p(x|c)$. GMM-Seg (Liang et al. 2022), a semantic segmentation method, optimizes the GMMs via EM, while the deep representations are obtained via gradient backpropagation of the discriminative loss. In domain adaptive segmentation tasks, BiSMAP (Lu et al. 2022) employs GMMs to estimate category condition between source and target domains to fit the *de facto* distribution of the source domain and estimate the likelihood of target samples based on probability densities. However, these methods model the de facto joint distribution based on massive annotated data and are difficult to adopt with limited annotations. It is noteworthy that several recent studies, including PMM (Yang et al. 2020), DG-PNet (Johnander et al. 2022), demonstrate somewhat relevance to GMMs but are still essential not GMMs-based methods. To the best of our knowledge, our work for the first time introduces GMMs to FSS, especially for modeling the joint distribution of pixel and category in the support images and the query image.

Problem Definition

In a FSS task, the dataset can be divided into a train set D_{train} with *base* categories C_{base}, and a test set D_{test} with *novel* categories C_{novel}, where the categories in C_{base} and those in C_{novel} are completely disjoint (i.e., $C_{base} \cap C_{novel} = \emptyset$). Conventional methods (Zhang et al. 2019; Tian et al. 2020; Lang et al. 2022a) usually adopt meta-learning with episodic training that enables the learning of transferable knowledge on D_{train} for high-quality generalization on D_{test}. Specifically, each episode in episodic training works with a support set $\mathcal{S} = \{(\mathrm{x}_i^s, \mathrm{m}_i^s)\}_{i=1}^k$ and a query set $\mathcal{Q} = \{(\mathrm{x}^q, \mathrm{m}^q)\}$ for k-shot semantic segmentation (*i.e.*, $k \in \{1, 5\}$ in our setting), where x^* and m^* represent an image and its corresponding foreground mask on category c, respectively. Trained with episodes on D_{train}, a FSS method aims to segment the objects of a novel category in a query image x^q according to the knowledge of k support images and support masks from D_{test}.

Overview of Generative Models

We provide a brief comparison between discriminative and generative models theoretically to emphasize the advantages of adopting generative models for FSS. As discussed above, recent deep learning-based methods for FSS usually employ a parametric network for representation learning (*i.e.*, $f_\theta : \mathbb{R}^3 \to \mathbb{R}^d$, where $\mathbb{R}^3$ and $\mathbb{R}^d$ denote the 3 channels of a RGB image and its corresponding d-dimensional feature representation, respectively), and the Softmax function for label prediction (*i.e.*, $p(C|x) = h_\theta : \mathbb{R}^d \to \mathbb{R}^{|C|}$, where $\mathbb{R}^{|C|}$ represents the prediction for categories $C = \{c_i\}$).

Generative models achieve the predictive results with the Bayes theorem (Rish et al. 2001), rather than directly obtaining the posterior probability $p(C|x)$ over a dataset D. Specifically, a generative model begins with building the joint distribution $p(x, c)$ of pixel x and category c. The category conditional probability $p(x|c)$ is then estimated with the category prior probability $p(c)$, and $p(c|x)$ is reached as,

$$p(c|x) = \frac{p(c)p(x|c)}{\sum_{i=1}^{|C|} p(c_i)p(x|c_i)}, \tag{1}$$

where the category prior probability $p(c)$ is normally set to a uniform prior (*i.e.*, $p(c) = 1/|C|$).

Generative models focus on estimating and optimizing the data distribution $\prod_{(x,c)\in D} p(x|c)$ (*i.e.*, generative training (Bernardo et al. 2007)). Extensive studies have explored the optimization of generative training, among which a representative one is Gaussian mixture models (GMMs) (Reynolds et al. 2009). In this paper, we introduce GMMs using the Expectation Maximization (EM) algorithm (Dempter 1977) to the optimization of evaluating $p(x|c)$. Generative models are able to capture the intrinsic characteristics of categories and then model the de facto distribution over the unseen data, which suggests their capability of excellent generalization towards the goal of FSS. Hence, adopting generative models may become an alternative to the current FSS paradigm.

Cross Gaussian Mixture Generative Model

In this paper, we present the Cross Gaussian Mixture Generative Model (CGMGM) as a novel FSS method. As illustrated in Figure 2, the proposed method comprises three modules on Initial Mask Generating and Refining (IMGR), Cross Data Gaussian Mixture Generating (CDGMG), and Query Category Posterior Extracting (QCPE). Suppose k is set to 1, a shared-weight backbone first extracts the high-level feature representations X_q^h and X_s^h, and the mid-level feature representations X_q and X_s, respectively, from a query image I_q and a support image I_s. To begin with, the IMGR module takes as input X_q^h, X_s^h, X_q, and X_s to generate and refine the initial segmentation mask M_q on a novel category c. The CDGMG module then takes as input M_q and the support mask M_s, X_q, and X_s as input, and adopts

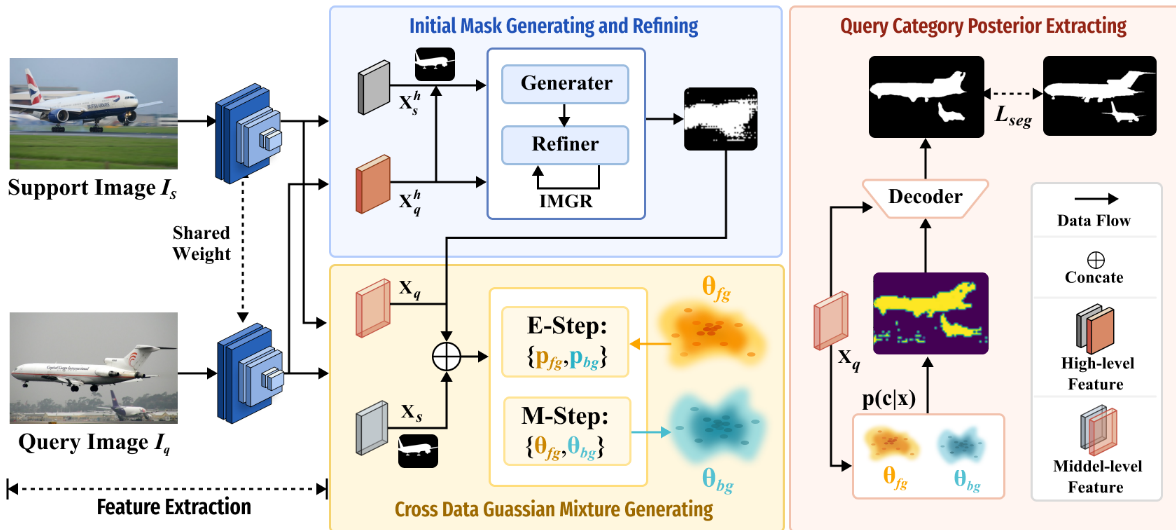


Figure 2: Architecture of the proposed CGMGM. Following a shared backbone, the IMGR module generates and refines the initial segmentation mask M_q for the target objects in the query image. The CDGMG module then models the joint distribution of pixel and category for the novel category regarding foreground θ_{fg} and background θ_{bg}, using M_q, the support mask M_s, mid-level feature representations X_s and X_q. Afterwards, the QCPE module evaluates the category posterior $p(c|x)$ to be used by a parametric decoder as the guidance information to generate the output segmentation mask with X_q.

Gaussian mixtures to model the joint distribution $p(x, c)$ of pixel x and category c; It outputs the distribution parameters $\theta_{fg} = \{\pi_{fg}, \mu_{fg}, \Sigma_{fg}\}$ and $\theta_{bg} = \{\pi_{bg}, \mu_{bg}, \Sigma_{bg}\}$. Afterwards, the QCPE module takes as input $\theta_{fg}\&\theta_{bg}$ and X_q to extract the guidance information via evaluating the posterior probability $p(c|x)$ for pixel x in X_q; It employs a parametric decoder to predict the output segmentation mask $\hat{M}_q$ on c for X_q using both features and guidance information.

Initial Mask Generating and Refining

Inspired by PFENet (Tian et al. 2020), we propose the Initial Mask Generating and Refining (IMGR) module to extract information from the query image's feature representation to accurately model the joint distribution of pixel and category. In particular, the IMGR module leverages high-level feature representations (*e.g.*, Conv5 of ResNet50) to generate and refine the initial segmentation mask.

In the generating step, a double-branch structure is adopted to separately establish global and local similarities for the improvement of the accuracy of the initial segmentation mask, which is different from the previous methods. As shown in Fig. 3, the IMGR module takes input the high-level feature representation $X_s^h \in \mathbb{R}^{C_h \times H_s \times W_s}$ of the support image I_s, the high-level feature representation $X_q^h \in \mathbb{R}^{C_h \times H_q \times W_q}$ of the query image I_q, and the segmentation mask $M_s \in \mathbb{R}^{1 \times H_s \times W_s}$ of I_s. In the global branch, the global similarity S_G is obtained by calculating the cosine similarity between X_q^h and the masked high-level feature representation of I_s as,

$$S_G = MatMul(X_s^h \otimes M_s, X_q^h), \tag{2}$$

where $\otimes$ denotes the element-wise multiplication operation, and *MatMul* represents the cosine similarity operation. In the local branch, we extract the patch representations R_s and R_q of I_s and I_q as,

$$\begin{aligned} R_s &= \mathcal{PS}(x_s^h \otimes M_s) \in \mathbb{R}^{hw \times C_h \times \frac{H_s W_s}{hw}}, \\ R_q &= \mathcal{PS}(x_q^h) \in \mathbb{R}^{hw \times C_h \times \frac{H_q W_q}{hw}}, \end{aligned} \tag{3}$$

where $\mathcal{PS}$ denotes the patch split operation, (h, w) are the patch height and width. In the experiments, we set h and w to 2 . The local similarity S_L is obtained by calculating the cosine similarity between R_s and R_q as,

$$S_L = MatMul(R_s, R_q) \in \mathbb{R}^{hw \times \frac{H_s W_s}{hw} \times \frac{H_q W_q}{hw}}. \tag{4}$$

Each pixel is represented by the mean of all pixels in its corresponding patch to maintain consistency in local features and reduce the incorrect pixel matching. Afterwards, S_L is reshaped and upsampled, followed by the concatenation with S_G. The concatenated global and local similarities are averaged to reach the coarse segmentation mask $M_q^0 \in \mathbb{R}^{1 \times H_q \times W_q}$.

In the refining step, we improve the confusing feature representation of I_q in the way similar to the use of cycle-consistency by CyCTR (Zhang et al. 2021). As shown in Figure 3, X_q^h is first masked with M_q^0, and the affinity map $\mathcal{A} \in \mathbb{R}^{H_s W_s \times H_q W_q}$ between X_s^h and the masked high-level feature representation of I_q is then calculated as,

$$\mathcal{A} = MatMul(X_q^h \otimes M_q^0, X_s^h). \tag{5}$$

For a query pixel *j*, its most similar support pixel i^* and most similar query pixel j^* are obtained by Argmax operation as,

$$i^* = Argmax\mathcal{A}_{(i,j)}, \quad j^* = Argmax\mathcal{A}_{(i^*,j)}. \tag{6}$$

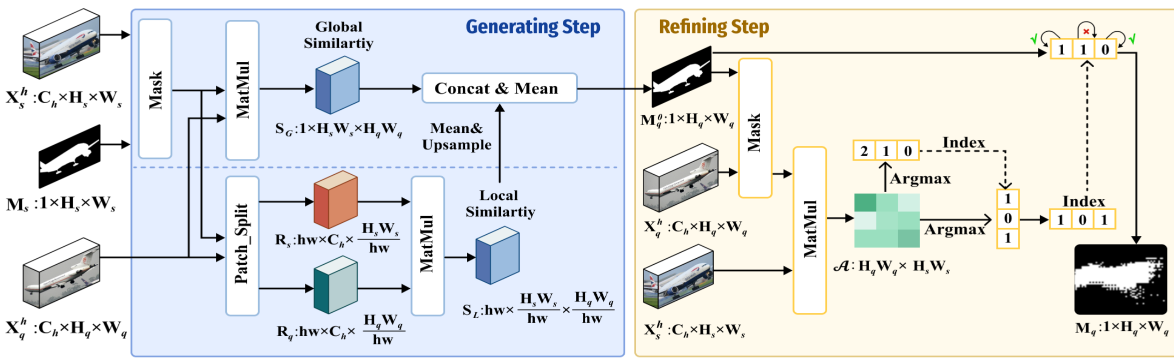


Figure 3: Structure of the IMGR module that comprises a generating step (left) and a refining step (right).

With the coarse segmentation mask M_q^0, cycle-consistency is satisfied if $M_{q(j)}^0 = M_{q(j^*)}^0$ and $M_{q(j)}^0 = 1$. Due to the complexity of the background pixels, we only focus on the consistency of the foreground pixels in the mask and obtain the initial segmentation mask M_q as,

$$M_q = \begin{cases} 1, & \text{if } M_{q(j)}^0 = M_{q(j^*)}^0 \& M_{q(j)}^0 = 1, \\ 0, & \text{otherwise .} \end{cases} \tag{7}$$

Cross Data Gaussian Mixture Generating

We devise the Cross Data Gaussian Mixture Generating (CDGMG) module to model the joint distribution $p(x, c)$ of pixel x and category c for the novel category (*i.e.*, *foreground*) and others (*i.e.*, *background*). As illustrated in Figure 2, each episode adopts GMMs using the EM algorithm taking as input the mid-level feature representations X_s and X_q of I_s and I_q along with M_s and M_q. Specifically, each GMM employs a weighted mixture of M multivariate Gaussians θ_c to model the conditional probability of $x \in \{X_s, X_q\}$ for c in the d-dimensional embedding space as,

$$\begin{aligned} p(x|c;\theta_c) &= \textstyle\sum_{m=1}^{M} p(m|c;\pi_c)p(x|c,m;\mu_c,\Sigma_c) \\ &= \textstyle\sum_{m=1}^{M} \pi_{cm}\mathcal{N}(x;\mu_{cm},\Sigma_{cm}), \end{aligned} \tag{8}$$

where $m|c = \pi_{cm}$ is the prior probability, $\mu_{cm} \in \mathbb{R}^d$ and $\Sigma_{cm} \in \mathbb{R}^{d\times d}$ are the mean vector and the covariance matrix of Gaussian component m for c, and $\theta_c = \{\pi_c, \mu_c, \Sigma_c\}$. The mixture design of the model enables accurate approximation of data densities. Notably, each Gaussian component m has an independent covariance structure, which allows flexibly measuring the importance across feature dimensions.

Then, the EM algorithm maximizes the log likelihood over the feature representations with label masks $(x_n, c_n)_{n=1}^N$ using the initial parameters θ_c^0. The optimal parameters θ_c^* can be reached as,

$$\theta_c^* = \underset{\theta_c}{argmax} \sum_{x_n:c_n=c} \log p = \underset{\theta_c}{argmax} \sum_{x_n:c_n=c} \log \sum_{m=1}^{M} p(x_n, m|c;\theta_c). \tag{9}$$

The EM algorithm iteratively calculates the intermediate parameters θ_c^t. In each iteration t, the probability of x that belongs to m is repeatedly optimized ($\mathrm{p}[\mathrm{m}] = p(m|x, c;\theta_c)$) in the **E-step** as,

$$\mathrm{p}_{\mathrm{cn}}^{\mathrm{t}}[\mathrm{m}] = \frac{\pi_{cm}^{t-1}\mathcal{N}(x_n|\mu_{cm}^{(t-1)}, \Sigma_{cm}^{(t-1)})}{\sum^M \pi_{cm}^{(t-1)}\mathcal{N}(x_n|\mu_{cm}^{(t-1)}, \Sigma_{cm}^{(t-1)})}, \tag{10}$$

and the parameters are then updated in the **M-step** as follows:

$$\begin{aligned} \pi_{cm}^t &= \frac{N_{cm}^t}{N_c}, \mu_{cm}^t = \frac{\sum_{x_n:c_n=c} \mathrm{p}_{\mathrm{cn}}^{\mathrm{t}}[\mathrm{m}] x_n}{N_{cm}^t}, \\ \Sigma_{cm}^t &= \frac{\sum_{x_n:c_n=c} \mathrm{p}_{\mathrm{cn}}^{\mathrm{t}}[\mathrm{m}](x_n - \mu_{cm}^t)(x_n - \mu_{cm}^t)^T}{N_{cm}^t}, \end{aligned} \tag{11}$$

where N_c denotes the number of the training samples labeled c and $N_{cm} = \sum_{n:c_n=c} \mathrm{p}_{\mathrm{c_n}}[\mathrm{m}]$. Because of to the setting of FSS, we adopt foreground and background as the labels regarding the novel category. Therefore, the final distribution parameters $\theta_{fg} = \{\pi_{fg}, \mu_{fg}, \Sigma_{fg}\}$ and $\theta_{bg} = \{\pi_{bg}, \mu_{bg}, \Sigma_{bg}\}$ are obtained at the end of the loop to represent the joint distribution of the novel category.

Query Category Posterior Extracting

The proposed Query Category Posterior Extracting (QCPE) module takes as input the mid-level feature representation X_q of I_q and distribution parameters θ_{fg} and θ_{bg} to extract the category posterior $p(c|x)$ that serves as guidance information. To begin with, the conditional probability of pixel $x_q \in X_q$ for category c is calculated as,

$$\begin{aligned} p(x_q|c;\theta_c) &= \log(\textstyle\sum_{m=1}^{M} \pi_{cm}\mathcal{N}(x_q;\mu_{cm},\Sigma_{cm})), \\ \mathcal{N}(x_q;\mu_c,\Sigma_c) &= \frac{\exp\{-\frac{1}{2}(x_q-\mu_{cm})^T\Sigma_{cm}^{-1}(x_q-\mu_{cm})\}}{(2\pi)^{d/2}\|\ \Sigma_{cm}\ \|^{1/2}}. \end{aligned} \tag{12}$$

As mentioned in CDGMG, the category posterior of x_q pixels can be regarded as the foreground posterior of the novel category in FSS. Based on the Bayes theorem, the guidance information is obtained as,

$$p(c|x) = p(fg|x_q) = \frac{p(x_q|c_{fg};\theta_{fg})}{p(x_q|c_{fg};\theta_{fg}) + p(x_q|c_{bg};\theta_{bg})}, \tag{13}$$

where c_{fg} and c_{bg} denote the labels of fore- and background.

Finally, both X_q and $p(c|x)$ are fed to a parameter decoder, which consists of an ASPP module to obtain multi-scale information, and a series of convolutional layers followed by the ReLU function to generate the output segmentation mask $\hat{M}_q$. The parameter decoder network is optimized with the cross-entropy loss function.

Methods	Backbone	1-shot					5-shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
PFENet (Tian et al. 2020)	VGG-16	56.90	68.20	54.40	52.40	58.00	59.00	69.10	54.80	52.90	59.00
HSNet (Min, Kang, and Cho 2021)		59.60	65.70	59.60	54.00	59.70	64.90	69.00	64.10	58.60	64.10
DPCN (Liu et al. 2022a)		58.90	69.10	63.20	55.70	61.70	63.40	70.70	68.10	59.00	65.30
NTRENet (Liu et al. 2022b)		57.70	67.60	57.10	53.70	59.00	60.30	68.00	55.20	57.10	60.20
BAM †(Lang et al. 2022a)		*63.18*	**70.77**	*66.14*	*57.53*	*64.41*	*67.36*	*73.05*	*70.61*	*64.00*	*68.76*
CGMGM (**Ours**)		**66.54**	*69.79*	**68.01**	**59.87**	**66.04**	**68.03**	**73.95**	**71.84**	**64.79**	**69.65**
PFENet (Tian et al. 2020)	ResNet50	61.70	69.50	55.40	56.30	60.80	63.10	70.70	55.80	57.90	61.90
ASGNet (Li et al. 2021)		58.84	67.86	56.79	53.66	59.29	63.66	70.55	64.17	57.38	63.94
CyCTR (Zhang et al. 2021)		65.70	71.00	59.50	59.70	64.00	69.30	73.50	63.80	63.50	67.50
HSNet (Min, Kang, and Cho 2021)		64.30	70.70	60.30	60.50	64.00	70.30	73.20	67.40	67.10	69.50
DCAMA (Shi et al. 2022)		67.50	72.30	59.60	59.00	64.60	70.50	73.90	63.70	65.80	68.50
NTRENet (Liu et al. 2022b)		65.40	72.30	59.40	59.80	64.20	66.20	72.80	61.70	62.20	65.70
DPCN (Liu et al. 2022a)		65.70	71.60	69.10	60.60	66.70	70.00	73.20	70.90	65.50	69.90
IPMT (Liu et al. 2022c)		**72.80**	*73.70*	59.20	61.60	66.80	**73.10**	74.70	61.60	63.40	68.20
BAM †(Lang et al. 2022a)		68.97	73.59	*67.55*	*61.13*	*67.81*	70.59	*75.05*	**70.09**	*67.20*	*70.91*
CGMGM (**Ours**)		*71.14*	**74.99**	**69.62**	**63.65**	**69.85**	*71.77*	**78.89**	*69.11*	**68.59**	**72.09**

Table 2: Comparison of our CGMGM and other FSS methods in mIoU (%) on PASCAL-5^i under 1-shot and 5-shot settings. Best scores are in bold and second best scores are in *italics*. †: baseline method.

Experiments

Datasets, Metrics, and Implementation Details

We evaluated the performance of our CGMGM for FSS on two benchmark datasets: PASCAL-5^i (Shaban et al. 2017) and COCO-5^i (Nguyen and Todorovic 2019). PASCAL-5^i is generated from the PASCAL VOC 2012 (Everingham et al. 2010) dataset with external annotation from SDS (Hariharan et al. 2014), which consists of 20 categories. COCO-20^i is constructed based on the MSCOCO (Lin et al. 2014) dataset, which consists of 80 categories. Following previous studies(Shaban et al. 2017; Tian et al. 2020; Yang et al. 2021), we grouped the categories in both datasets into four folds for cross-validation. During training, three folds were used for training and the remaining one for validation.

For the metrics, we adopted mean intersection over union (mIoU) and foreground-background IoU (FB-IoU) to conduct the evaluation under 1-shot and 5-shot settings.

In our experiments, we used VGG-16 (Simonyan and Zisserman 2014) and ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) as the backbones. Using BAM (Lang et al. 2022a) as the baseline, we adopted its initialization weights and the dataset setting. It is noteworthy that the BAM dataset setting can refine the mask output from the IMGR module in our CGMGM by reducing the probability of novel categories being categorized as background. Besides, we froze the weights of the backbones and the base learner in BAM. For fine-tuning parameters, we used the SGD optimizer with cosine learning rate decay, where the learning rate, momentum, and weight decay were set to 0.05, 0.9, and 0.0001, respectively. our method was trained for 200 epochs with the batch size of 8 and the image size of 473×473 on PASCAL-5^i, and for 50 epochs with the batch size of 8 and the image size of 641×641 on COCO-20^i. The number of Gaussian components M was set to 3 on PASCAL-5^i, and to 6 on COCO-20^i. In our evaluation, 1000 support-query pairs were randomly sampled from each of both datasets.

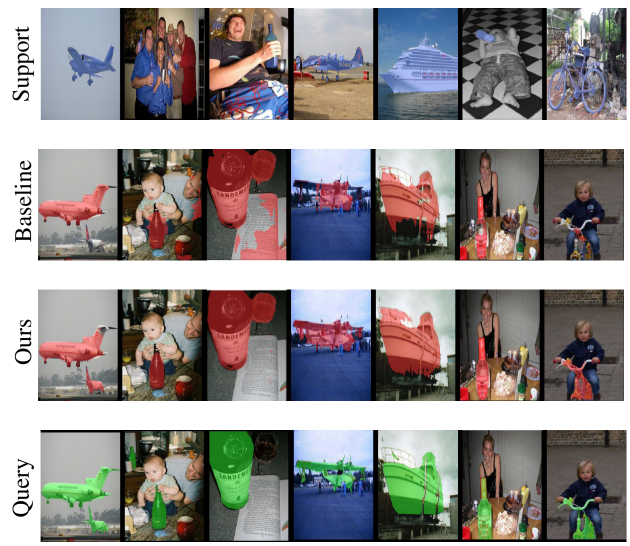


Figure 4: Qualitative comparison between our CGMGM and baseline under 1-shot setting on Pascal-5^i. From top to bottom: support images, baseline segmentation, CGMGM segmentation, query images.

Comparison with the State-of-the-Art Methods

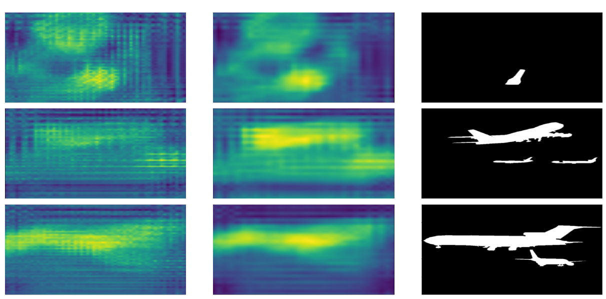
As shown in Table 2 and 3, the proposed method outperformed all the methods for comparison and achieved the state-of-the-art results on both datasets. Specifically, our CGMGM with ResNet-50 as the backbone on PASCAL-5^i achieved an increase of up to 2.04 under the 1-shot setting and 1.18 under the 5-shot setting in mIoU. Additionally, it reached an increase of up to 1.63 under the 1-shot setting and 0.89 under the 5-shot setting in mIoU when using the VGG-16 as the backbone. On COCO-20^i, our CGMGM outperformed baseline by 1.17 and 0.85 in mIoU under 1-shot and 5-shot settings, respectively, with ResNet-50 as the

Methods	Backbone	1-shot					5-shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
PFNet (Tian et al. 2020)	ResNet50	36.80	41.80	38.70	36.70	38.50	40.40	46.80	43.20	40.50	42.70
CyCTR (Zhang et al. 2021)		38.90	43.00	39.60	39.80	40.30	41.10	48.90	45.20	47.00	45.60
HSNet (Min, Kang, and Cho 2021)		36.30	43.10	38.70	38.70	39.20	43.30	51.30	48.20	45.00	46.90
DCAMA (Shi et al. 2022)		41.90	45.10	44.40	41.70	43.30	45.90	50.50	50.70	46.00	48.30
DPCN (Liu et al. 2022a)		42.00	47.00	43.20	39.70	43.00	46.00	54.90	50.80	47.40	49.80
NTRENet (Liu et al. 2022b)		36.80	42.60	39.90	37.90	39.30	38.20	44.10	40.40	38.40	40.30
IPMT (Liu et al. 2022c)		41.40	45.10	45.60	40.00	43.00	43.50	49.70	48.70	47.90	47.50
BAM †(Lang et al. 2022a)		*43.31*	**50.59**	*47.49*	*43.42*	*46.23*	*49.26*	*54.20*	**51.63**	*49.55*	*51.16*
CGMGM (**Ours**)		**47.05**	*49.34*	**48.84**	**44.35**	**47.40**	**50.33**	**54.59**	*51.28*	**51.80**	**52.01**

Table 3: Comparison of our CGMGM and other FSS methods in mIoU (%) on COCO-5^i under 1-shot and 5-shot settings. Best scores are in bold and second best scores are in *italics*. †: baseline method.

Methods	Backbone	FB-IoU	
		1-shot	5-shot
PFENet (Tian et al. 2020)	ResNet50	73.30	73.90
HSNet (Min, Kang, and Cho 2021)		76.70	80.60
DCAMA (Shi et al. 2022)		75.70	79.50
DPCN (Liu et al. 2022a)		78.00	80.70
NTRENet (Liu et al. 2022b)		77.00	78.40
IPMT (Liu et al. 2022c)		77.10	81.40
BAM †(Lang et al. 2022a)		*79.71*	*82.18*
CGMGM (**Ours**)		**80.51**	**83.05**

Table 4: Comparison of our CGMGM and other FSS methods in FB-IoU on PASCAL-5^i under 1-shot and 5-shot settings. †: baseline method.



(a) CS (b) IMGR (c) GT

Figure 5: Visualization of initial masks with Cosine Similarity (CS) and IMGR, compared to Ground Truth (GT).

backbone. These results suggest the efficacy of the proposed method. In particular, our CGMGM obtained a more significant improvement under the 1-shot setting, demonstrating that it could extract the useful information from the feature representation of the query image given a more limited support set. Table 4 shows the comparison with several other state-of-the-art methods in FB-IoU, which also validates the superiority of our CGMGM. **Quantitative Result** We visualized some segmentation examples output from our CGMGM and baseline on PASCAL-5^i in Figure 4. It is noteworthy that our method obtained more complete target objects because of the de facto distribution between query image and support images.

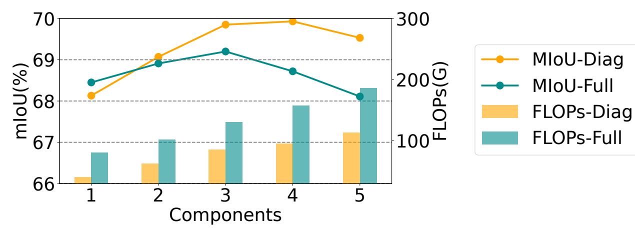


Figure 6: Effects of the number of Gaussian components and the type of covariance matrix on mIoU and floating point of operations (FLOPs) on PASCAL-5^i under 1-shot setting.

IMG	IMR	C&Q	mIoU
			67.81
		✓	66.95
✓		✓	68.27
✓	✓		68.58
✓	✓	✓	69.85

Table 5: Comparison of different configurations of parts from our CGMGM on PASCAL-5^i under 1-shot setting. C&Q denotes the coupled CDGMG and QCPE.

Ablation Studies

We conducted a series of ablation experiments under the 1-shot setting on PASCAL-5^i, and adopted BAM (Lang et al. 2022a) with ResNet-50 as the baseline.

Effectiveness of Parts As mentioned in Section , our CGMGM consists of IMGR, CDGMG, and QCPE modules. Since the QCPE module relies on the output of the CDGMG module, they were coupled as a CDGMG&QCPE (C&Q) part in the ablation experiments. Besides, we regarded the two steps in the IMGR module as an Initial Mask Generating (IMG) part and an Initial Mask Refining (IMR) part. Therefore, the ablation experiments were conducted with these three parts. As shown in Table 5, only using the C&Q part caused a decrease of 0.86 in mIoU, while adding the IMG achieved an improvement of 0.46 in mIoU, compared to the baseline. These results are regarded as reasonable since the inconsistency of the background information exists between the support images and the query image, and relying solely on the information from the support images is likely to result in incorrect distribution modeling. Overall, our CGMGM achieved the state-of-the-art performance for FSS because of combining all three modules.

Effectiveness of Double-Branch IMGR module As mentioned in Section , we innovatively proposed the double-branch initial mask generating module. We visualized the initial masks of previous methods and our double-branch IMGR module in Figure 5. It demonstrated that our IMGR can generate a higher quality initial mask by keeping the local pixels consistent.

Ablation on the CGMGM We varied the number of Gaussian components from 1 to 5 and adopted both diag-covariance and full-covariance matrices to set the GMMs to be either independent or correlated. As shown in Figure 6, using 3 multivariate Gaussians with the diag-covariance matrix led to a better accuracy-efficiency trade-off. We also performed experimental and theoretical analyses on computational complexity of our CGMGM, compared to other FSS methods.

Conclusions

The limitations of existing FSS methods are characterized by neglecting valuable information from the query image and struggling to extract effective guidance information between support and query images. In this paper, we proposed the Cross Gaussian Mixture Generative Model (CGMGM), a novel FSS method that models the de facto joint distribution of pixel and category in the support images and the query image. Our CGMGM exploits this distribution to evaluate the category posterior probability of pixels in the query image and exploits it as guidance information. Extensive experiments showed that our parameter-free generative method achieved state-of-the-art performance on two datasets, highlighting its effectiveness in pushing the boundary of FSS.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62202421, No. U23A20311, No. 62376243, and No. U20A20387); in part by Young Elite Scientists Sponsorship Program by CAST (2021QNRC001); in part by Zhejiang Provincial Natural Science Foundation of China under (Grant No. LTGS23F020001); in part by the Program of Zhejiang Province Science and Technology (2022C01044); in part by Ningbo Yongjiang Talent Introduction Program of China (Grant No. 2021A-157-G); and in part by the Key Research and Development Program of Ningbo City of China (Grant No. 2023Z130); in part by the StarryNight Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010); and in part by the Project by Shanghai AI Laboratory (P22KS00111).

References

Bernardo, J.; Bayarri, M.; Berger, J.; Dawid, A.; Heckerman, D.; Smith, A.; and West, M. 2007. Generative or discriminative? getting the best of both worlds. *Bayesian statistics*, 8(3): 3–24.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Cheng, G.; Li, R.; Lang, C.; and Han, J. 2021. Task-wise attention guided part complementary learning for few-shot image classification. *Science China Information Sciences*, 64: 1–14.

Dempter, A. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, 39: 1–22.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Hariharan, B.; Arbeláez, P.; Girshick, R.; and Malik, J. 2014. Simultaneous detection and segmentation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, 297–312. Springer.

Hayashi, H.; and Uchida, S. 2019. A discriminative gaussian mixture model with sparsity. *arXiv preprint arXiv:1911.06028*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 603–612.

Izmailov, P.; Kirichenko, P.; Finzi, M.; and Wilson, A. G. 2020. Semi-supervised learning with normalizing flows. In *International Conference on Machine Learning*, 4615–4630. PMLR.

Johnander, J.; Edstedt, J.; Felsberg, M.; Khan, F. S.; and Danelljan, M. 2022. Dense gaussian processes for few-shot segmentation. In *European Conference on Computer Vision*, 217–234. Springer.

Klautau, A.; Jevtic, N.; and Orlitsky, A. 2003. Discriminative Gaussian mixture models: A comparison with kernel classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 353–360.

Lang, C.; Cheng, G.; Tu, B.; and Han, J. 2022a. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8057–8067.

Lang, C.; Tu, B.; Cheng, G.; and Han, J. 2022b. Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation. *arXiv preprint arXiv:2204.09903*.

Li, G.; Jampani, V.; Sevilla-Lara, L.; Sun, D.; Kim, J.; and Kim, J. 2021. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8334–8343.

Liang, C.; Wang, W.; Miao, J.; and Yang, Y. 2022. GMMSeg: Gaussian Mixture based Generative Semantic Segmentation Models. arXiv:2210.02025.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, J.; Bao, Y.; Xie, G.-S.; Xiong, H.; Sonke, J.-J.; and Gavves, E. 2022a. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11553–11562.

Liu, Y.; Liu, N.; Cao, Q.; Yao, X.; Han, J.; and Shao, L. 2022b. Learning non-target knowledge for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11573–11582.

Liu, Y.; Liu, N.; Yao, X.; and Han, J. 2022c. Intermediate prototype mining transformer for few-shot semantic segmentation. *arXiv preprint arXiv:2210.06780*.

Lu, Y.; Luo, Y.; Zhang, L.; Li, Z.; Yang, Y.; and Xiao, J. 2022. Bidirectional Self-Training with Multiple Anisotropic Prototypes for Domain Adaptive Semantic Segmentation. *arXiv preprint arXiv:2204.07730*.

Mackowiak, R.; Ardizzone, L.; Kothe, U.; and Rother, C. 2021. Generative classifiers as a basis for trustworthy image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2971–2981.

Min, J.; Kang, D.; and Cho, M. 2021. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6941–6952.

Nguyen, K.; and Todorovic, S. 2019. Feature Weighting and Boosting for Few-Shot Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Reynolds, D. A.; et al. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Rish, I.; et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, 41–46.

Schott, L.; Rauber, J.; Bethge, M.; and Brendel, W. 2018. Towards the first adversarially robust neural network model on MNIST. *arXiv preprint arXiv:1805.09190*.

Serrà, J.; Álvarez, D.; Gómez, V.; Slizovskaia, O.; Núñez, J. F.; and Luque, J. 2019. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*.

Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; and Boots, B. 2017. One-Shot Learning for Semantic Segmentation.

Shi, X.; Wei, D.; Zhang, Y.; Lu, D.; Ning, M.; Chen, J.; Ma, K.; and Zheng, Y. 2022. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, 151–168. Springer.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tian, Z.; Zhao, H.; Shu, M.; Yang, Z.; Li, R.; and Jia, J. 2020. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2): 1050–1065.

Tüske, Z.; Tahir, M. A.; Schlüter, R.; and Ney, H. 2015. Integrating Gaussian mixtures into deep neural networks: Softmax layer with hidden variables. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4285–4289. IEEE.

Variani, E.; McDermott, E.; and Heigold, G. 2015. A Gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4270–4274. IEEE.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Yang, B.; Liu, C.; Li, B.; Jiao, J.; and Ye, Q. 2020. Prototype Mixture Models for Few-shot Semantic Segmentation. arXiv:2008.03898.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2021. Mining latent classes for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8721–8730.

Yuan, Y.; Chen, X.; Chen, X.; and Wang, J. 2019. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*.

Zhang, C.; Lin, G.; Liu, F.; Yao, R.; and Shen, C. 2019. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5217–5226.

Zhang, G.; Kang, G.; Yang, Y.; and Wei, Y. 2021. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems*, 34: 21984–21996.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.