SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

October 2024, Vol. 67, Iss. 10, 202103:1–202103:22
https://doi.org/10.1007/s11432-023-4113-1

Saliency-guided meta-hallucinator for few-shot learning

Hongguang ZHANG[1*†], Chun LIU[1†], Jiandong WANG[1†], Linru MA[1], Piotr KONIUSZ[2,3], Philip H. S. TORR[4] & Lin YANG[1*]

[1]*Systems Engineering Institute, Academy of Military Sciences, Beijing* 100141*, China;*
[2]*Data61, Commonwealth Scientific and Industrial Research Organisation, Canberra* ACT 2601*, Australia;*
[3]*Australian National University, Canberra* ACT 2601*, Australia;*
[4]*Oxford University, Oxford* OX3 6PJ*, UK*

Received 16 May 2023/Revised 19 October 2023/Accepted 15 December 2023/Published online 26 September 2024

Abstract Learning novel object concepts from limited samples remains a considerable challenge in deep learning. The main directions for improving the few-shot learning models include (i) designing a stronger backbone, (ii) designing a powerful (dynamic) meta-classifier, and (iii) using a larger pre-training set obtained by generating or hallucinating additional samples from the small scale dataset. In this paper, we focus on item (iii) and present a novel meta-hallucination strategy. Presently, most image generators are based on a generative network (i.e., GAN) that generates new samples from the captured distribution of images. However, such networks require numerous annotated samples for training. In contrast, we propose a novel saliency-based end-to-end meta-hallucinator, where a saliency detector produces foregrounds and backgrounds of support images. Such images are fed into a two-stream network to hallucinate feature samples directly in the feature space by mixing foreground and background feature samples. Then, we propose several novel mixing strategies that improve the quality and diversity of hallucinated feature samples. Moreover, as not all saliency maps are meaningful or high quality, we further introduce a meta-hallucination controller that decides which foreground feature samples should participate in mixing with backgrounds. To our knowledge, we are the first to leverage saliency detection for few-shot learning. Our proposed network achieves state-of-the-art results on publicly available few-shot image classification and anomaly detection benchmarks, and outperforms competing sample mixing strategies such as the so-called Manifold Mixup.

Keywords few-shot learning, saliency detection, object recognition, anomaly detection, computer vision

1 Introduction

Convolutional neural network (CNN) models perform well on various computer vision tasks, such as image classification, scene recognition, and object detection. However, CNNs require many labeled training samples, whereas humans learn novel concepts well, even from a few samples. Inspired by the notion of transfer of practice [1], many researchers study the so-called one- and few-shot learning paradigm [2, 3] to adapt models to new concepts given a few of samples.

The concept of learning object relations has recently been explored in several studies [4–8], which can be considered a form of metric learning [9–11] adapted to few-shot learning. In the above studies, a neural network extracts convolutional representations, and a learner, such as a relation module, logistic classifier, or nearest-neighbor classifier, is used to capture the relationship between the so-called support and query images [12–15]. Many recent studies in this area improve relationship modeling [5, 6, 8]. In contrast, a low-shot feature generating approach [16] mines image analogies in the feature space. However, the authors noted that "generated examples are unlikely to be as good as real examples" and they added the generated samples mainly to categories with the fewest samples. This strategy is beneficial to long-tail distributed classification datasets but not viable in one-shot learning where each class may access only

* Corresponding author (email: zhang.hongguang@outlook.com, yanglin61s@126.com)
† Zhang H G, Liu C, and Wang J D have the same contribution to this work.

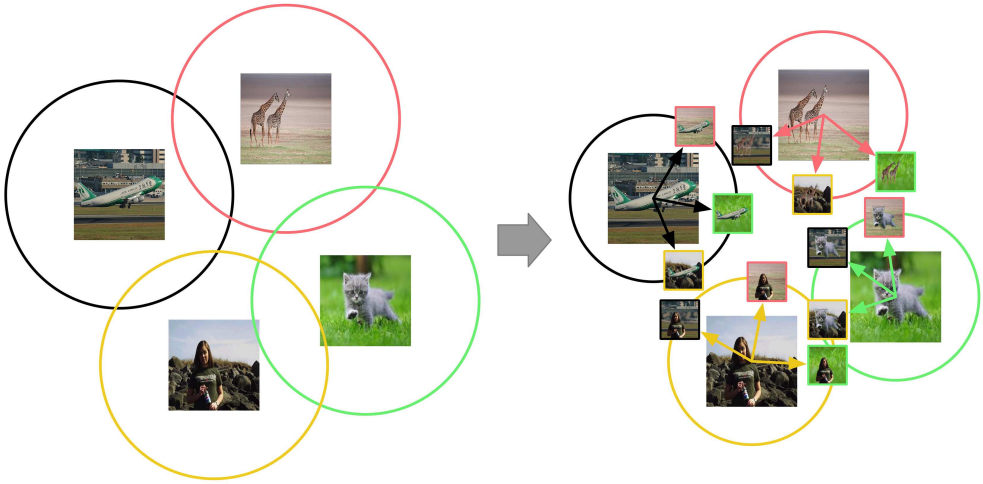

Figure 1 Saliency-based data generation (one-shot case). The foreground objects are combined with different backgrounds to form auxiliary training samples which help refine the classification boundary of similarity learner.

one sample. Low-shot learning from imaginary data [17] is pretrained on numerous tasks and classes, thus increasing the cost of training. As approaches [16, 17] are formulated as low-shot learning, which differs from few-shot learning, and the code is not available, we report results of related pipelines, i.e., MetaGAN [18] which uses an adversarial generator conditioned on tasks to augment vanilla few-shot classification models, and AFHN [19] which uses feature generator to synthesize fake features. Finally, as some recent pipelines [20,21] use Manifold Mixup [22,23], which performs implicit sample augmentation in the feature space by a convex combination of pairs of samples and corresponding labels, we also compare our method to Manifold Mixup.

In this paper, we extend our saliency-based few-shot learning pipeline [24], which combines our earlier few-shot learning pipeline [8] with a saliency detector [25] pretrained on the MSRA salient object database (MSRA-B) [26]. By removing backgrounds from images, one can model and learn relations between foreground objects alone. This strategy should improve the performance if the saliency maps capture objects of interest accurately. However, as salient detectors are not oracles, direct learning of similarity in foregrounds is also suboptimal. Moreover, as object recognition benefits from a meaningful visual context (i.e., backgrounds) associated with objects, backgrounds should not be completely discarded.

To this end, we adopt the data hallucination strategy and propose a novel saliency-guided meta-hallucination pipeline, dubbed the saliency-guided meta-hallucination (SMH). Figure 1 illustrates the key principle underpinning our SMH, whereas Figure 2 shows its pipeline. In contrast to standard feature hallucination approaches, we employ a salient detector to explicitly segment out foregrounds and backgrounds from given images, followed by a two-stream Mixing Nets, which extracts image representations to mix foreground and background representations in the feature space. As we obtain spatial feature maps from this process, we mix foreground and background feature vectors into a combined second-order representation, aggregated over the spatial modes of feature maps. We investigate two basic mixing operators such as a linear combination and concatenation. Subsequently, we employ a base learner such as a relation network (RN), a nearest-neighbor (NN) classifier, or a linear classifier (LC) to learn to make predictions based on the final co-occurrence descriptors of a so-called training query sample and hallucinated support matrices.

We propose several strategies for selecting meaningful backgrounds for mixing with a given foreground. To this end, we perform (i) intra-class hallucination (foregrounds/backgrounds of the same class) or (ii) inter-class hallucination (for any given foreground. We take its corresponding background, retrieve its NN backgrounds from various classes, and use a simple retrieval distance to express the likelihood of how valid the mixed pair is before using or discarding it).

Moreover, we address the negative impact of non-salient images on our hallucination strategy. In some images, class-related objects may not coincide with salient objects (or the saliency map may be heavily corrupted). Thus, we introduce a simple controller that decides if an image should participate in the foreground-background mixing step. This strategy results in a conditional meta-hallucination process in which only images with valid saliency maps may mix their foregrounds with alternative backgrounds.

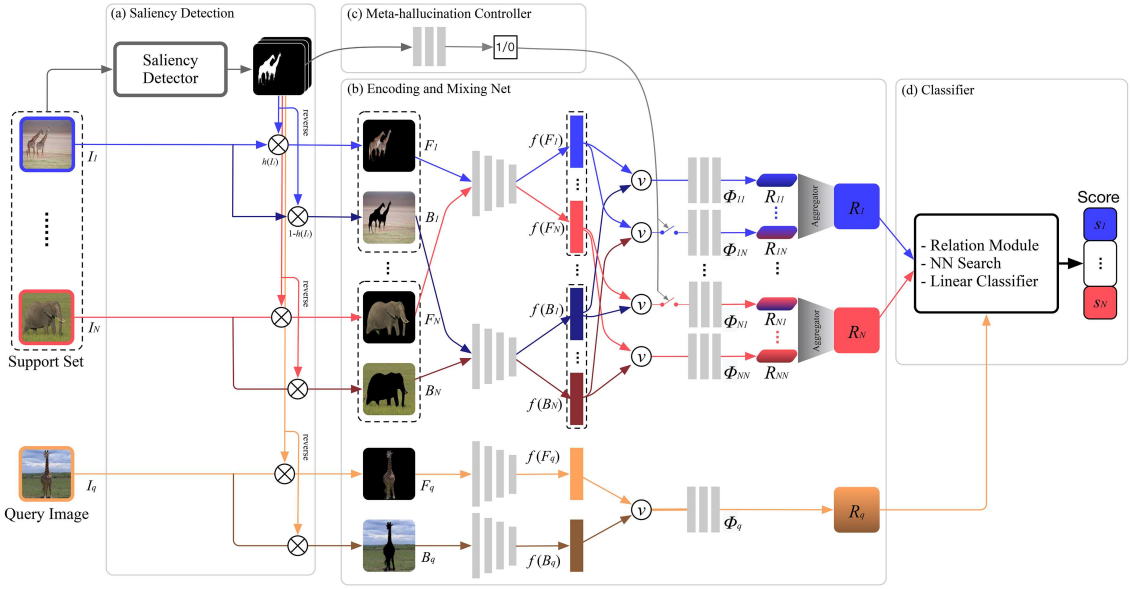

Figure 2 Our pipeline comprises four units: (a) pretrained saliency net, (b) foreground-background encoding and Mixing Nets (FEMN), (c) meta-hallucination controller, and (d) base learner. The FEMN block comprises two streams that take foreground/background images as inputs, respectively, and a Mixing Nets which combines foreground-background pairs via $\oplus$ and refines them via a single-stream network before aggregation of the resulting feature maps via the second-order encoder. The network receives support images $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_N$ (where N is the number of classes, say 5, in an episode) and a query image $\boldsymbol{I}_q$. Support foregrounds and backgrounds are given by $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_N$ and $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_N$, respectively, where $\boldsymbol{F}_q$ and $\boldsymbol{B}_q$ are the foreground and background of $\boldsymbol{I}_q$, and $\boldsymbol{\Phi}_{11}, \ldots, \boldsymbol{\Phi}_{NN}$ are N^2 support samples after mixing.

We argue that our foreground-background mixing strategies (including the above controller) increase intra-class variability while limiting unrealistic mixtures. In contrast, generative adversarial networks (GAN)-based hallucinators cannot directly control whether generated samples are diverse and meaningful for few-shot learning, and Manifold Mixup cannot guarantee that plausible class-wise mixtures are created, e.g., mixing class whale with bedroom is obviously a bad choice. As Manifold Mixup does not mix foregrounds and backgrounds of images separately like our intra- and inter-class hallucination strategies, one could consider our saliency-driven mixing strategies to be related to the general principles of Manifold Mixup but specifically designed to model foreground-background feature mixing for few-shot learning (how to mix, what to mix, and what not to mix), rather (c.f., interpolated mixing of image features and class labels).

Furthermore, we propose several strategies for regularizing our Mixing Nets and promoting the hallucination of realistic blends of foreground and background representations, which also improves the diversity of hallucinated samples. To this end, we introduce representation distillation (RD), salient dilation (SD), and generative modeling (GM). We believe GM-based modeling is especially novel/interesting, as our generator admits two noise inputs for generating foreground-background mixtures according to learned distributions.

To summarize the above strategies, (i) when a foreground-background pair is extracted from the same image (c.f., two separate images), the RD strategy constrains the resulting blended representation via the ℓ_2-norm to be close to a representation from a supervising network that, by its design, is trained only on real foreground-background pairs (not real and thus potentially infeasible combinations are never used by the supervising network). (ii) The SD strategy applies a convolution with a Gaussian kernel to soften the saliency boundary, e.g., enlarge it to filter out the shape of foregrounds from background candidates (as the shape itself provides some degree of discriminative information). (iii) The GM replaces the meta-hallucination step with a simple generative model to hallucinate samples based on learned distribution parameters, thus improving the diversity of hallucinated samples.

Finally, we go beyond the few-shot object classification and extend our pipeline to few-shot anomaly detection, a challenging task in which only a few positive samples are available for training, whereas negative samples are not given. Applying our SMH pipeline to this task simultaneously provides additional positive training samples and negative adversarial samples, which effectively refines the decision boundary of a few-shot anomaly classifier via contrastive loss.

Below, we list our contributions:

(i) We propose a novel end-to-end conditional saliency-guided meta-hallucination for few-shot image classification, including a controller that excludes poor quality saliency foregrounds for foreground-background coupling, and investigates different types of mixing and hallucination strategies.

(ii) We additionally propose several auxiliary strategies such as RD, SD, and GM to prevent the creation of substandard hallucinated samples, which improves the overall quality and diversity of hallucination.

(iii) We extend our pipeline to a challenging task of few-shot anomaly detection to further demonstrate the effectiveness of our saliency-guided meta-hallucination model.

In this paper, we build on our few-shot saliency-guided model [24]. To our knowledge, this model is the first that employs the saliency maps for data hallucination in the few-shot learning scenario. In this manuscript, we extend our conference paper [24] by (i) adding new backbones, (ii) introducing a new controller that judges the quality of saliency-based foregrounds for conditional foreground and background mixing in the feature space, with the inclusion of (iii) salient dilation and generative modeling, and (iv) we extend our ideas to a new task of few-shot anomaly detection.

Our experiments demonstrate state-of-the-art performance on (1) publicly available few-shot benchmarks, namely, mini-ImageNet, tiered-ImageNet, and Open MIC, (2) three fine-grained image classification datasets, namely, Flower102, CUB-200-2011, and Food-101, and (3) the few-shot anomaly detection benchmark on CIFAR-10.

2 Related work

Below, we discuss one- and few-shot learning models, saliency detectors, second-order statistical models, and anomaly detectors.

2.1 Learning from few samples

For deep learning algorithms, the ability of "learning quickly from only a few examples is definitely the desired characteristic to emulate in any brain-like system" [27]. Learning from the limited number of samples poses a challenge to typical CNN approaches [28], which must learn millions of parameters. The current trend in computer vision highlights the need for "an ability of a system to recognize and apply knowledge and skills learned in previous tasks to novel tasks or new domains, which share some commonality." This problem, introduced in 1901 under the notion of "transfer of particle" [1], is closely related to zero-shot learning [29–32], which can generalize to unseen classes from samples of categories seen during training. For one- and few-shot learning, some "transfer of particle" is also desirable as generalizing from one or few samples to account for the intra-class variability of visually diverse object classes is a formidable task.

One- and few-shot learning has been studied in computer vision in shallow [2, 3, 33–36] and deep learning [4–6, 12, 37] settings.

Early studies [3, 36] proposed generative models with an iterative inference for the transfer step. The recent Siamese network [37] uses a two-stream CNN that performs simple metric learning to distinguish whether two samples share the same class, which is simply a form of similarity learning. Matching Nets [4] introduces the concept of a support set and N-way W-shot learning protocols. It captures the similarity between one query and several support images and implicitly performs metric learning. Prototypical Nets [5] learn a model that computes distances between a sample and prototype representations of each class. Model-agnostic meta-learning (MAML) [12] is a meta-learning model that "learns to learn" by adapting gradients of individual tasks to then update the global gradient of an encoder. Relation Nets [6] is similar to Matching Nets [4], but uses an additional network to learn the similarity between images. The second-order similarity network (SoSN) [8] leverages second-order descriptors and power normalization which help capture rich relation statistics. SoSN descriptors are more effective than the first-order Relation Nets [6].

A hallucination-based approach [16] manually assigns descriptors into 100 clusters to generate so-called analogies. The approach of [17] employs a generative model to hallucinate additional samples to supplement the support classes; however, it requires a large-scale benchmark for training, which considerably limits its application in realistic settings. In contrast, MetaGAN [18] incorporates GAN into existing meta-learning pipelines to refine the class boundaries by learning to discriminate the fake samples. AFHN [19] is also a GAN-based model and uses support sample features as conditional cues

to generate new samples. Manifold Mixup [22, 23] is a different type of data augmentation strategy that applies a convex combination of pairs of samples in the feature space and corresponding labels. Inspired by the mixup strategy, few-shot learning approaches [20, 21] leverage it to boost their performance.

Our method clearly differs from the above hallucination models as we decompose images into foreground and background representations via saliency maps, and we propose several strategies for pairing up and mixing foreground and background features to hallucinate meaningful auxiliary training samples.

Recent studies [14, 15, 38] revisited few-shot learning and proposed conventional fine-tuning strategies, additional regularization terms, and extra supervisory cues, and often used deeper backbones, such as ResNet-12 or ResNet-18 (c.f., Conv-4-64). DeepEMD [39] focuses on the matching strategy between sets of patches rather than entire images, using the optimal transportation plan.

Because we address whole images, not patch sets, our performance should not be directly compared with the above orthogonal few-shot strategies, as our method and the above models are complementary rather than "competing" with each other; i.e., we would expect a boost if we were to use patch sets/optimal transport.

2.2 Generative approaches

Generative models aim to capture the inherent data distribution within a training dataset by mapping it into a latent space. GAN [40] stands out as the predominant approach in this domain. The GAN framework is grounded in a zero-sum game between a generator and a discriminator, wherein the generator's objective is to produce data that are virtually indistinguishable from real data. Numerous advancements have been made in refining the GAN architecture, as well as enhancing training and regularization techniques [41–43]. Other approaches include variational autoencoders (VAEs) [44], which model the data distribution explicitly via invertible transformations.

In few-shot learning, a generative model can be trained on base-class data to augment real novel-class data with synthetic data. An early concept in this realm is feature hallucination [16]. Various strategies built on GANs have been explored, such as MetaGAN [18], which combines MAML [12] with a conditional GAN to generate instances in the input space; AFHN [19], an approach employing feature hallucination and wGAN [45]; and FUNIT [46], a GAN-driven method designed for few-shot image-to-image translation.

Alternative approaches to GANs include IDeMe-Net [47], which combines novel-class support with similar base-class images; VI-Net [48], which employs a class-conditional VAE as a feature hallucinator; and TFH [49], which uses VAE and class prototypes to augment original data.

However, the generative technique (e.g., GAN) requires numerous annotated samples for training, which violates the few-shot setting. Moreover, GAN-based hallucinators cannot directly control whether generated samples are diverse and meaningful for few-shot learning. The proposed saliency-guided meta-hallucinator will address these issues.

2.3 Saliency detection

Saliency detectors capture image regions with foreground objects that correlate with human visual perception. They produce a dense likelihood saliency map with relevance scores in the range of $[0, 1]$. Conventional salient detectors underperform on complex scenes because of computations based on human-defined priors [50]. In contrast, deep saliency models [51, 52] perform well, but they require laborious pixel-wise labels. The combination of contrastive learning and self-supervised learning [53–56] has recently achieved some success in discovering semantic groupings in real-world images, but fails to separate semantically similar object instances nearby. As we use saliency maps as a guiding cue, we adopt highly-efficient weakly-supervised deep convolutional salient detectors, i.e., RFCN [51], RBD [50], MNL [25], UC-Net [57], SelfMask [58] FOUND [59], and MOVE [60].

2.4 Second-order statistics

Second-order statistics have been studied in the context of texture recognition [61, 62] via so-called region covariance descriptors (RCDs), often applied to semantic segmentation [63] and object category recognition [64, 65]. Using correlation patterns in CNNs, similar in spirit to RCDs, is a popular direction. Approach [66] fuses two CNN streams via the outer product in the context of fine-grained image recognition. A face recognition algorithm [67] uses co-occurrences of CNN feature vectors and facial attribute vectors to provide state-of-the-art face recognition. Second-order statistics must address the

so-called burstiness, which is "the property that a given visual element appears more times in an image than a statistically independent model would predict" [68]. Power normalization [64, 69], used with Bag-of-Words [64, 65, 69, 70], was shown to limit this burstiness. A survey [69] showed that so-called MaxExp feature pooling [71] is a detector of "at least one particular visual word being present in an image". MaxExp on second-order matrices was shown in [70] to coincide with the sigmoid function, and it performed well in few-shot learning [8]. Recent studies [72, 73] further demonstrated the usefulness of global covariance pooling in capturing richer statistics of deep features. Thus, we employ second-order pooling with the sigmoid because of its simplicity and established track record. Note that we do not claim second-order pooling as a contribution.

2.5 Anomaly detection

Anomaly detection (AD) has a wide range of applications, such as bank fraud detection, surveillance, and airport security. AD detects outliers or anomalies that differ from the dataset distribution. AD can be divided into supervised, semi-supervised, and unsupervised variants based on the availability of labels. In what follows, we focus on unsupervised few-shot anomaly detection, in which a few training samples are available but labels are not.

Shallow AD models, based on classification, often rely on classical machine learning, such as decision trees [74], or support vector machine [75]. Clustering-based AD leverages clustering algorithms such as k-means and kNN [76] to detect anomalies from samples. These algorithms perform well in low-dimensional spaces on human-selected features. On real-world datasets with intricate, nonlinear, and heterogeneous samples, deep learning approaches are more robust [77].

Deep AD uses neural networks to model anomaly detection. Semi-supervised deep AD approaches are label-free or label-dependent [78]. Label-free approaches often use generative models such as AE [79] and GAN [80], whereas label-dependent AD combines traditional AD with deep learning, such as DeepSVDD [81] or leverages geometric transformations [82].

Yet, only a few few-shot anomaly detection (FSAD) models exist [83–85]. Based on the mutual information maximization [86], a framework [83] is applied to polyp detection in colonoscopy. However, this approach relies on numerous images. Prototypical Nets were adapted to a one-way setting in [84] by introducing a "null class" via a batch normalization (BN) layer. The MAML algorithm was also adapted to FSAD in [85], but this model leverages the embeddings of a model trained on auxiliary datasets.

To the best of our knowledge, we introduce the first saliency-guided contrastive FSAD framework, which learns and detects anomalies given only a few samples with no labels.

3 Approach

Our pipeline builds on the generic few-shot Relation Nets [6], which implicitly learns a metric for so-called query and support images. To this end, an encoding network encodes images into feature vectors. Subsequently, the so-called episodes with query and support images are formed. Each query-support pair is forwarded to a so-called RN (a.k.a., similarity network), and a loss function learns a binary label in $\{0, 1\}$, indicating whether a query-support pair is of the same class. However, such methods suffer from scarce training data, which we address below.

3.1 Network

Figure 3(a) shows that the standard few-shot learning pipeline (e.g., Relation Nets or SoSN) encodes support and query images with the feature encoder. Subsequently, it employs the similarity network to learn the relations between features of support-query pairs. Specifically, one can opt for a deep similarity network, a linear classifier, or a nearest-neighbor match, etc.

Figure 2 presents our complete model, the saliency-guided two-stream similarity network with hallucination (SMH wHal). For this model, we leverage saliency maps to isolate foreground and background image representations. The network comprises (i) a saliency detector whose role is to generate foreground hypotheses, (ii) foreground-background encoding and Mixing Nets (FEMN), whose role is to combine foreground-background image pairs into episodes, and (iii) base learner (BL), e.g., linear classifier, kNN classifier, or the relation module, which learns whether the support and query samples belong to the same class or different classes. In addition to mixing original foreground-background pairs in the

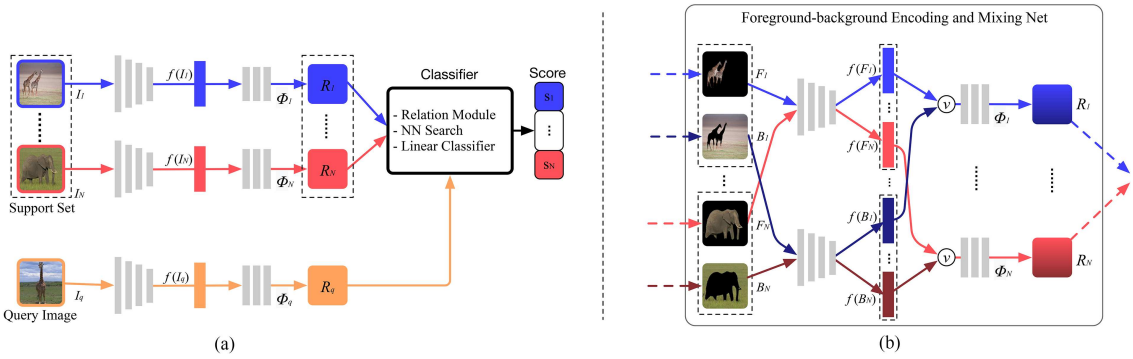

Figure 3 (a) Pipeline without saliency-guided foreground-background segmentations. This simplified diagram coincides with the design of SoSN [8]. In this scenario, saliency is not used on support images $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_N$ (where N is the number of classes e.g., 5 in an episode) and a query image $\boldsymbol{I}_q$. (b) "Sanity check" pipeline without hallucination of auxiliary samples. We use saliency maps to segment the foregrounds and backgrounds of an image, then mix them "back" in feature space. Here, we show the process only for support images $\boldsymbol{I}_1, \ldots, \boldsymbol{I}_N$ whose foregrounds and backgrounds are given by $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_N$ and $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_N$, respectively, and N is the number of classes in an episode (e.g., 5).

feature space, this network also mixes every foreground object in the support set with available background candidates from other images in the support set. For instance, Figure 2 shows that we may mix features representing the giraffe with the features of its own background or the background of an image containing the elephant; thus, we obtain auxiliary training samples. Hallucinating additional training samples from existing samples is an easy way to improve the performance of few-shot learning.

Figure 3(b) is a "sanity check" variant of SMH that, if substituted into Figure 2, results in the saliency-guided two-stream similarity network without hallucination (SMH w/o Hal). The SMH units from Figure 3(b) and the model in Figure 3(a) differ in that support/query images are segmented into foregrounds/backgrounds by the saliency maps and fed into the foreground/background encoder, respectively. We mix the foreground/background features from the same image in the feature space and then apply second-order pooling on the mixed features.

To illustrate how our network works, consider the i-th image $\boldsymbol{I}_i$, which is passed through saliency network $\boldsymbol{h}$ to extract the corresponding saliency map $\boldsymbol{h}(\boldsymbol{I}_i)$, the foreground $\boldsymbol{F}_i$, and the background $\boldsymbol{B}_i$ of $\boldsymbol{I}_i$:

$$\boldsymbol{F}_i = \boldsymbol{h}(\boldsymbol{I}_i) \odot \boldsymbol{I}_i, \tag{1}$$

$$\boldsymbol{B}_i = (1 - \boldsymbol{h}(\boldsymbol{I}_i)) \odot \boldsymbol{I}_i, \tag{2}$$

where $\odot$ is the Hadamart product. The feature encoding network consists of two parts, f and g. For images $\boldsymbol{I}_i \in \mathbb{R}^{3\times M\times M}$ and $\boldsymbol{I}_j \in \mathbb{R}^{3\times M\times M}$ ($\boldsymbol{I}_i = \boldsymbol{I}_j$ or $\boldsymbol{I}_i \neq \boldsymbol{I}_j$), we proceed by encoding their foreground $\boldsymbol{F}_i \in \mathbb{R}^{3\times M\times M}$ and background $\boldsymbol{B}_j \in \mathbb{R}^{3\times M\times M}$ via feature encoder $f : \mathbb{R}^{3\times M\times M} \rightarrow \mathbb{R}^{K\times Z^2}$, where $M \times M$ denotes the spatial size of an image, K is the feature size, and Z^2 refers to the vectorized spatial dimension of map f of size $Z \times Z$. Then, the encoded foreground and background are mixed via summation and refined in encoder $g : \mathbb{R}^{K\times Z^2} \rightarrow \mathbb{R}^{K'\times Z'^2}$, where K' is the feature size and Z'^2 corresponds to the vectorized spatial dimension of map g of size $Z' \times Z'$. As in the SoSN approach [8], we apply the outer-product on $g(\cdot)$ to obtain an autocorrelation of features, and we perform pooling via sigmoid $\boldsymbol{\psi}$ to tackle the burstiness in our representation. Thus, we have

$$\boldsymbol{\Phi}_{ij} = g(\vartheta(f(\boldsymbol{F}_i), f(\boldsymbol{B}_j))), \tag{3}$$

$$\boldsymbol{R}_{ij} = \boldsymbol{\psi}(\boldsymbol{\Phi}_{ij}\boldsymbol{\Phi}_{ij}^{\mathrm{T}}, \sigma), \tag{4}$$

where ϑ refers to the mixing operator, e.g., sum, mean, concatenation, inner-product, and outer-product, and $\boldsymbol{\psi}$ is a zero-centered sigmoid function with σ (the value is set to 0.4) as the parameter that controls the slope of its curve:

$$\boldsymbol{\psi}(\boldsymbol{X}; \sigma) = \frac{1 - \mathrm{e}^{-\sigma\boldsymbol{X}}}{1 + \mathrm{e}^{-\sigma\boldsymbol{X}}} = \tanh(2\sigma\boldsymbol{X}). \tag{5}$$

Descriptors $\boldsymbol{R}_{ii} \in \mathbb{R}^{K'\times K'}$ represent a given image $\boldsymbol{I}_i$ while $\boldsymbol{R}_{ij} \in \mathbb{R}^{K'\times K'}$ represents a combined pair of the foreground of image $\boldsymbol{I}_i$ and the background of $\boldsymbol{I}_j$, and its class label is given as l_i.

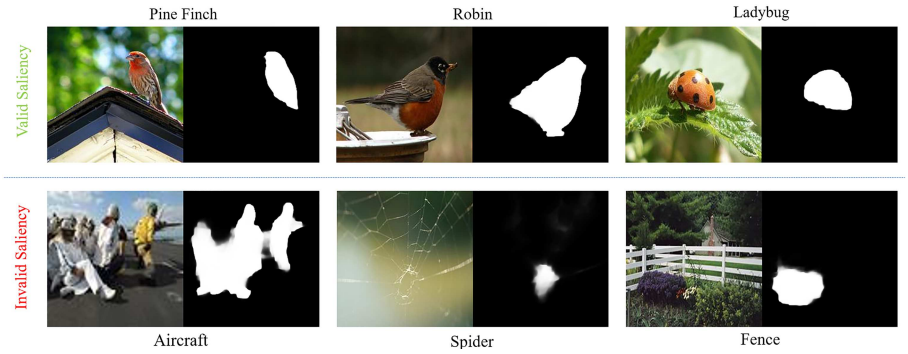

Figure 4 Examples of salient and non-salient images from mini-ImageNet. The saliency detector generates invalid saliency maps for non-salient samples, e.g., the detector focuses on the "cow" rather than the "fence" for an image (2nd row, 3rd column) labeled as "fence." Similarly, humans selected as foregrounds for hallucinating samples (2nd row, 1st column) are invalid foreground choices given that the image is labeled as "aircraft." Thus, we introduce to the text a meta-hallucination controller that filters out non-salient images.

A loss example. From the above equations, one can devise various mixing/hallucination strategies. Let us first explain how to recover a basic SoSN pipeline. Taking the relation module-based base learner as an example, we form the query-support pairs (e.g., we concatenate their representations), and we pass episodes to the relation module followed by the mean square error (MSE) loss to train our network without any hallucination strategy:

$$\mathcal{L} = \frac{1}{NW}\sum_{i=1}^{NW}(r(\boldsymbol{R}_{ii}, \boldsymbol{R}_q) - \delta(l_i - l_q))^2, \tag{6}$$

where index i indexes samples from originally sampled support images, q chooses the query image, r is the similarity network, l_i is the label of the i-th image, N is the number of classes in an episode, W is the shot number per support class (so in total, we have NW support images per episode), and $\delta(0) = 1$ (0 elsewhere). Note that Eq. (6) does not form foreground and background hallucinated pairs. In (6), we merely decompose each support image into the foreground and background pair and stitch them back via (3), which results in a pipeline similar to SoSN. We describe the hallucination process in Subsection 3.3.

3.2 Saliency maps

For brevity, we consider three approaches: deep supervised saliency approaches MNL [25], RFCN [51], and an unsupervised shallow method RBD [50]. In this paper, we use saliency maps as a prior to generate foreground and background hypotheses.

Because of its superior performance, we use the deep weakly-supervised saliency detector MNL [25] in our main experiments. Moreover, we investigate the deep supervised RFCN approach [51] pretrained on the THUS10K dataset [87], which has no intersection with our few-shot learning datasets. We also investigate the cheap RBD model [50], which outperformed other unsupervised models [88].

Figure 4 shows the difference in results between saliency maps obtained with all three methods. In the top row of Figure 4, the foreground and background have distinct textures. Thus, conventional and deep models isolate the foreground well. However, for the scenes whose foreground and background share color and texture composition (bottom row), the unsupervised method fails to detect the correct foreground. As our dataset contains simple and complex scenes, the performance of our method depends, to some extent, on the salient detector. For example, results based on RBD [50] are expected to be worse in comparison to RFCN [51] and MNL [25]. The performance of few-shot learning combined with different salient detectors will be presented in Subsection 5.3. Moreover, Figure 4 shows that saliency masks are not always valid; therefore we provide several advanced foreground and background mixing strategies and a saliency controller to exclude troublesome images from mixing. Initially, we detail our strategies for hallucinating additional training data for few-shot learning.

3.3 Saliency-guided meta-hallucination

The additional samples are hallucinated by summing foreground and background feature vector pairs obtained from the feature encoder $f(\cdot)$ and refined by the encoder $g(\cdot)$. Taking the N-way W-shot

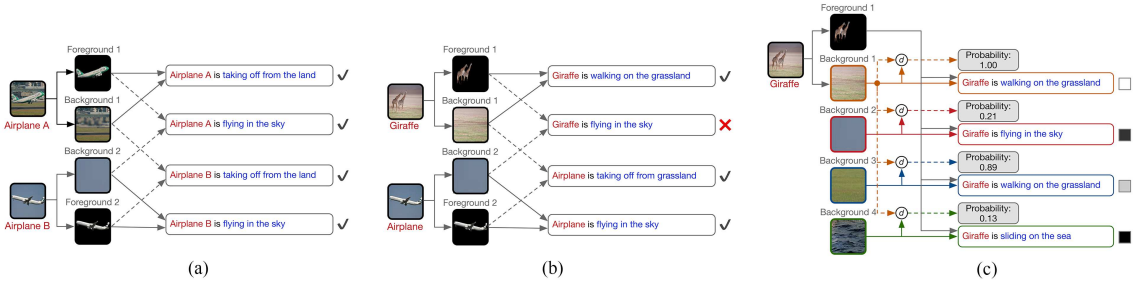

Figure 5 (a) Intra-class meta-hallucination strategy: most samples generated in this way are statistically plausible. (b) Inter-class meta-hallucination may generate impossible instances, e.g., "giraffe in the sky" is an unlikely concept (except for a giraffe falling off a helicopter during transportation?) (c) Inter-class meta-hallucination strategy guided by the similarity prior. We assign likelihoods to generated samples according to the similarity of a background of a given image to other backgrounds.

problem as an example (see Relation Nets [6] or SoSN [8] for the detailed definition of this protocol), we will randomly sample W images from each of N training classes. Let i, j be the index selecting the $N \times W$ support images and q be the index selecting the query image. Where required, assume the foreground and background descriptors for images are extracted. Then, the following strategies for the meta-hallucination of auxiliary samples can be formulated.

3.3.1 *Strategy I: intra-class meta-hallucination*

For this strategy, given a support image index i, a corresponding foreground is only mixed with the backgrounds of images from the same class c. Thus, we can generate $W-1$ extra samples for every image. Figure 5(a) shows that the intra-class hallucination produces plausible new samples. Note that the image class l_i typically correlates with foreground objects, and such objects appear on backgrounds which, statistically speaking, if swapped, will produce plausible object-background combinations. However, the above strategy cannot work in one-shot setting as only one support image per class is given. The loss function of using intra-class meta-hallucination based on the relation module is given as

$$\mathcal{L} = \frac{1}{NW}\sum_{i=1}^{NW}\frac{1}{W}\sum_{j:l_j=l_i}(r(\boldsymbol{R}_{ij}, \boldsymbol{R}_q) - \delta(l_i - l_q))^2, \tag{7}$$

where W denotes the number of support samples with the class label l_i. Although our intra-class hallucination presents a promising direction, our results will show that sometimes the performance may sit below baseline few-shot learning because of a very simple mixing foreground-background strategy that includes the foreground-background feature vector summation followed by the refining encoder $\boldsymbol{g}(\cdot)$. This strategy incurs possible noises from (i) the substandard saliency maps and/or (ii) mixing incompatible foreground-background pairs.

Mixing operators. Below, we illustrate the mixing operators we use. Let $\boldsymbol{F}_i$ and $\boldsymbol{B}_j$ denote the foreground of the i-th support image and the background of j-th support sample, respectively, and $f(\cdot)$ refers to the feature encoder. We investigate the following mixing operators in end-to-end meta-hallucination.

• **Linear Comb.:** $\vartheta(f(\boldsymbol{F}_i), f(\boldsymbol{B}_j)) = \lambda f(\boldsymbol{F}_i) + (1-\lambda) f(\boldsymbol{B}_j)$. This strategy enables us to linearly mix foreground and background representations according to λ.

• **Concat.:** $\vartheta(f(\boldsymbol{F}_i), f(\boldsymbol{B}_j)) = \text{cat}(f(\boldsymbol{F}_i), f(\boldsymbol{B}_j))$. Concatenation relies on the mixing ability of the subsequent network.

3.3.2 *Strategy II: inter-class meta-hallucination*

This strategy permits mixing the foregrounds of support images with all available backgrounds (e.g., between-class mixing is allowed) in the support set. Therefore, the inter-class hallucination can generate $W-1+W(N-1)$ samples for each support image, which is $W(N-1)$ more than the intra-class generator. However, many foreground-background pairs will be statistically implausible, as shown in Figure 5(b), which would cause the degradation of the classification accuracy. The loss function for the inter-class

meta-hallucination based on the relation module is

$$\mathcal{L} = \frac{1}{NW}\sum_{i=1}^{NW}\frac{1}{NW}\sum_{j=1}^{NW}(r(\boldsymbol{R}_{ij}, \boldsymbol{R}_q) - \delta(l_i - l_q))^2. \tag{8}$$

Eq. (8) generates $(NW)^2$ samples from NW support images during the inter-class meta-hallucination process, which is a large number compared with standard few-shot learning. However, this strategy introduces many invalid foreground-background couplings. To eliminate the unlikely couplings from the inter-class hallucination step, we introduce a similarity prior, which assigns probabilities to backgrounds in terms of their compatibility with a given foreground. Figure 5(c) illustrates such a setting.

Scoring foreground-background couplings. Numerous similarity priors can be proposed; e.g., one can use the label information to specify some similarity between two given classes. Intuitively, the backgrounds of images containing dogs and cats should be more correlated than the backgrounds of images of dogs and radios. However, explicitly modeling such relations may be cumbersome and has shortcomings; e.g., the backgrounds of images containing cars may also be suitable for rendering animals on the road or sidewalk despite an apparent lack of correlation between cat and car classes. Thus, we ignore strategies based on class labels and perform a background retrieval instead. Specifically, once all backgrounds of support images are extracted, we measure the distance between the background of a chosen image of index i vs. all other backgrounds to assign a probability score of how similar two backgrounds are

$$d(\boldsymbol{B}_i, \boldsymbol{B}_j) = \big\|f(\boldsymbol{B}_i) - f(\boldsymbol{B}_j)\big\|_2^2, \tag{9}$$

$$p(\boldsymbol{B}_j|\,\boldsymbol{B}_i) = \frac{2\mathrm{e}^{-\alpha d(\boldsymbol{B}_i,\,\boldsymbol{B}_j)}}{1+\mathrm{e}^{-\alpha d(\boldsymbol{B}_i,\,\boldsymbol{B}_j)}}, \tag{10}$$

where α is a hyper-parameter to control our probability profile function $p(d)$.

We apply the profile p to hallucinated outputs of $\boldsymbol{g}$ to obtain $\boldsymbol{g}'$. We show this strategy in Figure 5(c), and we refer to it as a soft similarity prior (SSP):

$$\boldsymbol{g}'(\boldsymbol{F}_i, \boldsymbol{B}_j) = p(\boldsymbol{B}_j|\,\boldsymbol{B}_i)\, g(f(\boldsymbol{F}_i), f(\boldsymbol{B}_j)). \tag{11}$$

In addition, we propose a hard similarity prior (HSP) according to which we combine a given foreground with the most relevant retrieved backgrounds whose p is above a certain τ (the empirical value is 0.5):

$$\boldsymbol{g}'(\boldsymbol{F}_i, \boldsymbol{B}_j) = \begin{cases} 0, & \textbf{if } p(\boldsymbol{B}_j|\boldsymbol{B}_i) \leqslant \tau, \\ \boldsymbol{g}(f(\boldsymbol{F}_i), f(\boldsymbol{B}_j)), & \text{otherwise.} \end{cases} \tag{12}$$

We will show in our experiments that the use of priors considerably enhances the performance of the inter-class hallucination, particularly for the 1-shot protocol, to which the intra-class hallucination is not applicable. We will show in Section 5 that HSP and SSP improve the performance of few-shot learning; SSP is a consistent performer on all protocols.

3.3.3 *Conditional meta-hallucination*

Although we investigate different mixing/hallucination strategies to address the mismatch issue between foregrounds and backgrounds, many images contain invalid saliency maps or no salient objects in the first place. Thus, such images will introduce considerable noise in the meta-hallucination process and lead to poorer performance.

To verify the impact of such non-salient images, we first manually draw non-salient and salient samples from mini-ImageNet to construct two subsets and perform few-shot learning ablation studies on them to show the impact on our meta-hallucination strategy. Specifically, the salient subset contains 20–40 samples per class in 30 classes, whereas the non-salient subset contains 10–15 samples per class in 20 classes. The samples from two subsets are shown in Figure 4. The experimental results of the 5-way 1-shot setting are given in Table 1, which shows that our SMH pipeline clearly performs worse on the non-salient subset, which motivates us to develop a meta-hallucination controller that decides which images participate in generating auxiliary samples.

Table 1 Results for salient vs. non-salient subsets with intra- and inter-class hallucination (5-way accuracy, Conv-4 backbone, MNL detector). Notably, SMH performs worse than SoSN on the challenging non-salient subset, but better on the salient subset.

Model	Non-salient set		Salient set	
	1-shot	5-shot	1-shot	5-shot
RN [6]	60.67	81.35	53.38	70.92
SoSN [8]	62.73	84.21	55.21	72.26
SMH Intra-class	59.98	81.07	57.53	74.31
SMH Inter-class	58.09	79.32	58.96	76.05

To address this fundamental issue of our SMH pipeline, we introduce a meta-hallucination controller $c(\cdot)$ that decides if a given sample participates in the process of meta-hallucination; thus, $c(\cdot)$ is called a conditional SMH (conSMH). Specifically, we annotate the salient and non-salient subsets (detailed earlier) with non-salient/salient {0,1} labels $l_i^{(s)}$. Subsequently, we train a two-class linear classifier on the concatenation of original images and their saliency maps. This meta-hallucination controller detects which saliency maps contain valid/invalid saliency regions and is given as

$$\mathcal{L}^{(c)} = -\sum_i l_i^{(s)} \log c(\text{cat}(\boldsymbol{I}_i, \boldsymbol{h}(\boldsymbol{I}_i))). \tag{13}$$

If the controller predicts 1 for $c(\text{cat}(\boldsymbol{I}_i, \boldsymbol{h}(\boldsymbol{I}_i)))$, then $\boldsymbol{I}_i$ will participate in the meta-hallucination. Otherwise, if the controller predicts 0, $\boldsymbol{I}_i$ will merely be used by combining its own foreground and background (the original pair) to obtain the non-hallucinated representation. The controller is trained standalone before the training of SMH (not end-to-end) in our implementation, but it could also be trained end-to-end for further improvements if needed. The above process can be considered a type of conditional meta-hallucination strategy.

3.4 Techniques for improving the quality and diversity of hallucinated samples

In this section, we propose several techniques for further improving the quality and diversity of hallucinated samples, therefore helping improve the few-shot learning performance.

3.4.1 *Representation distillation*

To further refine the hallucinated samples, we propose to exploit foreground-background mixed pairs $\boldsymbol{F}_i$ and $\boldsymbol{B}_i$ that come from the same image (e.g., their mixing should produce the original image) and enforce their feature vectors to be close in the MSE sense to some baseline teacher network that does not perform hallucination. Specifically, we take $\boldsymbol{\Phi}_{ii} = \boldsymbol{g}(\vartheta(\boldsymbol{F}_i, \boldsymbol{B}_i))$ and encourage its proximity to some teacher representation $\boldsymbol{\Phi}^*_{ii} = \boldsymbol{g}^*(\vartheta(\boldsymbol{F}_i, \boldsymbol{B}_i))$:

$$\Omega_{rd} = \frac{1}{NW}\sum_{i=1}^{NW} \left\| g(\vartheta(f(\boldsymbol{F}_i), f(\boldsymbol{B}_i))) - g^*(\vartheta(f^*(\boldsymbol{F}_i), f^*(\boldsymbol{B}_i))) \right\|_2^2, \tag{14}$$

$$\mathcal{L}' = \mathcal{L} + \beta_1 \Omega_{rd}, \tag{15}$$

where β_1 adjusts the impact of Ω, $\mathcal{L}'$ is the combined loss, and $f^*(\cdot), g^*(\cdot)$ are pretrained.

We investigate $\boldsymbol{g}^*(\cdot)$, which encodes (i) the original images only i.e., $g^*(f(\boldsymbol{I}_i))$ or (ii) foreground-background pairs from original images, i.e., $g^*(f(\boldsymbol{F}_i)+f(\boldsymbol{B}_i))$. We refer to Ω as the RD. Our experiments will demonstrate that RD improves the final results.

3.4.2 *Saliency dilation*

As backgrounds extracted via a salient detector contain silhouettes of objects (shapes), they unintentionally carry foreground information. Figure 6 shows that high quality saliency maps contain sharp boundary information (background maps then contain the information about object foregrounds). This attribute leads to an undesired influence during pairing foregrounds with backgrounds. Thus, we propose to dilate the saliency maps with a Gaussian kernel $k(x, y; \sigma)$ with the standard deviation given by σ_d. The size of the Gaussian kernel is set to 5×5 and 7×7 depending on σ_d. Figure 6 shows that applying the Gaussian kernel and a threshold over saliency maps eliminates the shape of foreground objects. This result prevents leaking the foreground information to background representations.

Figure 6 Examples of gradual dilation of the foreground mask. The larger the dilation effect, the less information about the foreground class is contained in the saliency map.

3.4.3 *Generative modeling*

The above-demonstrated hallucination processes combine specific foregrounds and backgrounds, which results in a deterministic hallucination pattern learned from the foreground-background pairs. However, we expect that learning a generative model might improve the diversity of hallucinating patterns. In what follows, we employ a VAE to implement the generative variant of foreground-background mixing meta-hallucination.

To this end, let $f_e(\cdot)$ and $f_d(\cdot)$ denote the encoder and decoder of the VAE, respectively. For given sample $\boldsymbol{I}_i$, we first segment it into a foreground and a background, given by $\boldsymbol{F}_i$ and $\boldsymbol{B}_i$, respectively. Subsequently, we forward them into the foreground and background variational encoder, respectively:

$$(\boldsymbol{\mu}_{\boldsymbol{F}_i}, \boldsymbol{\sigma}_{\boldsymbol{F}_i}) = f_e(\boldsymbol{F}_i), \quad (\boldsymbol{\mu}_{\boldsymbol{B}_i}, \boldsymbol{\sigma}_{\boldsymbol{B}_i}) = f_e(\boldsymbol{B}_i), \tag{16}$$

$$\boldsymbol{\Phi}_{ii} = g(\vartheta(\boldsymbol{\sigma}_{\boldsymbol{F}_i} \odot \boldsymbol{z} + \boldsymbol{\mu}_{\boldsymbol{F}_i}, \boldsymbol{\sigma}_{\boldsymbol{B}_i} \odot \boldsymbol{z}' + \boldsymbol{\mu}_{\boldsymbol{B}_i})), \text{ and} \tag{17}$$

$$\boldsymbol{R}_{ii} = \boldsymbol{\psi}(\boldsymbol{\Phi}_{ii}\boldsymbol{\Phi}_{ii}^{\mathrm{T}}, \sigma), \tag{18}$$

where $\boldsymbol{z} \sim \mathcal{N}(0,1)$ and $\boldsymbol{z}' \sim \mathcal{N}(0,1)$ are the random vectors sampled from the normal distribution.

We note that the extracted generative representation is simultaneously fed into the base learner to learn the object relations and to the VAE encoder to reconstruct the original image $\boldsymbol{I}_i$. Thus, we additionally have

$$\hat{\boldsymbol{I}}_i = f_d(\vartheta(\boldsymbol{\sigma}_{\boldsymbol{F}_i} \odot \boldsymbol{z} + \boldsymbol{\mu}_{\boldsymbol{F}_i}, \boldsymbol{\sigma}_{\boldsymbol{B}_i} \odot \boldsymbol{z}' + \boldsymbol{\mu}_{\boldsymbol{B}_i})), \tag{19}$$

$$\Omega_{\text{vae}} = \sum_{i=1}^{NW} \|\boldsymbol{I}_i - \hat{\boldsymbol{I}}_i\|_F^2, \text{ and} \tag{20}$$

$$\mathcal{L}' = \mathcal{L} + \beta_1 \Omega_{rd} + \beta_2 \Omega_{\text{vae}}. \tag{21}$$

To conclude, by introducing GM into the pipeline, we encourage the meta-hallucination step to mix-up foregrounds and backgrounds more diversely.

4 Saliency-guided few-shot anomaly detection

Below, we consider a challenging scenario for our saliency-related model, namely, anomaly detection. We further evaluate the usefulness of our proposed saliency-guided meta-hallucination strategy in the setting described below.

Anomaly detection, a.k.a., outlier detection, uses positive samples to train the classifier to detect positive and unseen negative samples. Figure 7 shows an extension of our proposed SMH to FSAD. For pairs of positive samples $\boldsymbol{I}_i$ and $\boldsymbol{I}_j$, one could attempt to minimize the following objective:

$$\mathcal{L} = \frac{1}{W^2} \sum_{i=1}^{W} \sum_{j=1}^{W} (r(\boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j) - 1)^2, \tag{22}$$

where $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Phi}_j$ represent second-order matrices obtained from encoded feature vectors of samples $\boldsymbol{I}_i$ and $\boldsymbol{I}_j$, and W is the shot number.

However, the decision boundary achieved in such a way is unlikely to be accurate. Therefore, we apply the salient detector $\boldsymbol{h}(\cdot)$ on the image pairs to extract foregrounds and backgrounds. Subsequently, we apply $\boldsymbol{h}(\cdot)$ to generate K background patches per image. These patches are first filtered according

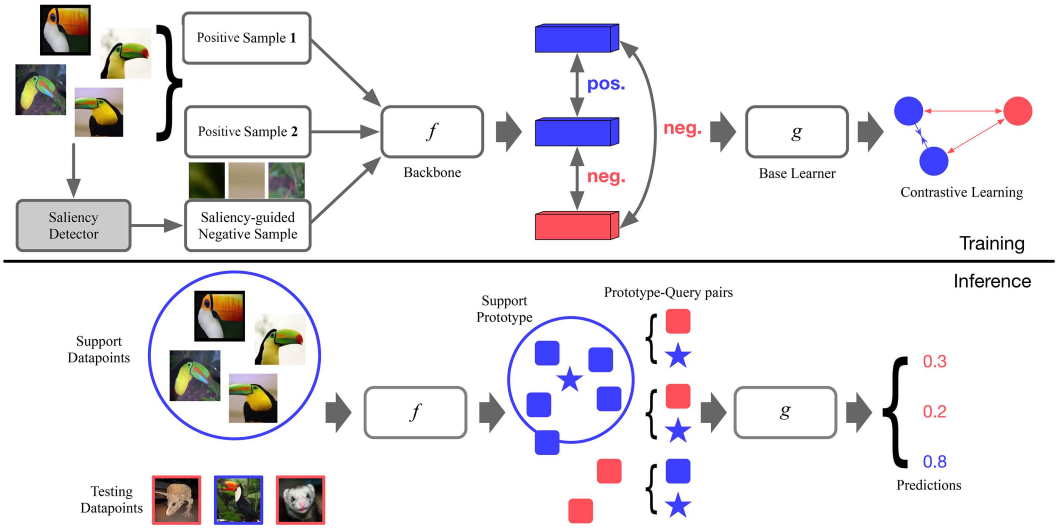

Figure 7 Proposed saliency-guided few-shot anomaly detection pipeline. We apply the salient detection to segment out foregrounds and backgrounds from given images. Foreground patches form positive samples whose feature representations are pushed closer to each other. Background patches are used as negative samples for contrastive learning (their feature representations are pushed away from the feature representations of positive samples). At the test time, feature representations of several foregrounds form a prototype. Test samples that are far from the prototype are anomalous.

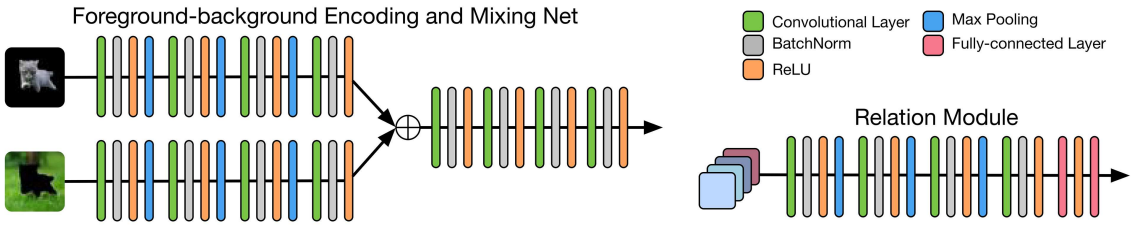

Figure 8 Detailed architecture of foreground-background encoding and Mixing Nets and the relation module. Best viewed in color.

to their pixel density. We use such background patches and the VAE to obtain M negative samples. Subsequently, we perform learning in the contrastive setting as follows:

$$\mathcal{L} = \frac{1}{W^2}\sum_{i=1}^{W}\sum_{j=1}^{W}(r(\boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j) - 1)^2 + \frac{1}{WM}\sum_{i=1}^{W}\sum_{k=1}^{M} r^2(\boldsymbol{\Phi}_i, \boldsymbol{\Phi}'_{ik}) + \frac{1}{WM}\sum_{j=1}^{W}\sum_{k=1}^{M} r^2(\boldsymbol{\Phi}_j, \boldsymbol{\Phi}'_{jk}), \tag{23}$$

where $\boldsymbol{\Phi}$ are foreground representations and $\boldsymbol{\Phi}'$ are background representations.

5 Experiments

Our network is evaluated in the few-shot learning scenario on mini-ImageNet [4], tiered-ImageNet [89], CUB-200-2011 [90], Flower102 [91], Food-101 [92], and Open MIC [93] datasets. For the anomaly detection, we evaluate our model on CIFAR-10. Our implementation is based on PyTorch, and the models are trained on a Nvidia V100 via the Adam solver. The architecture of our saliency-guided hallucination network is shown in Figures 2 and 8. The runtime of saliency detector is approximately 30–50 ms, which leads to very limited overhead w.r.t. GPU running memory and training time. The results are compared with several state-of-the-art methods for one- and few-shot learning.

5.1 Datasets

Below, we describe our setup, datasets, and evaluations.

mini-ImageNet [4] comprises 60000 RGB images from 100 classes. We follow the standard protocol [4] and use 80 classes for training (including 16 classes for validation) and 20 classes for testing. All images are resized to 84×84 pixels for fair comparison to other methods. We also investigate larger sizes, e.g., 224×224, as our SMH model can use richer spatial information from larger images to obtain high-rank autocorrelation matrices without needing to modify the similarity network to larger feature maps.

tiered-ImageNet [89] comprises 608 classes from ImageNet. We follow the protocol that uses 351 base classes, 96 validation classes, and 160 novel test classes.

Caltech-UCSD-Birds 200-2011 (CUB-200-2011) [90] has 11788 images of 200 bird species, described by 312 attributes. A total of 100/50/50 classes are randomly selected for training, validation, and testing, respectively.

Flower102 [91] is a fine-grained category recognition dataset with 102 flower classes, each with 40–258 images. We randomly select 60 training classes, 20 validation classes, and 22 testing classes.

Food-101 [92] comprises 101000 images in total and 1000 images per category. We choose 80 classes for training and 21 classes for testing.

For the above three fine-grained classification datasets, we follow the evaluation protocol and settings proposed in [94].

Open MIC [93], Open Museum Identification Challenge (Open MIC) dataset, contains photos of various exhibits, e.g., paintings, timepieces, sculptures, glassware, relics, science exhibits, natural history pieces, ceramics, pottery, tools, and indigenous crafts, captured from 10 museum exhibition spaces, according to which it is divided into 10 subproblems. In total, Open MIC has 866 diverse classes and 1–20 images per class. The within-class images undergo various geometric and photometric distortions as the data were captured with wearable cameras. Consequently, Open MIC is a perfect candidate for testing one-shot learning algorithms. Following the setup in SoSN [8], we combine (shn+hon+clv), (clk+gls+scl), (sci+nat), and (shx+rlc) splits into subproblems $p1, \ldots, p4$. We randomly select 4 out of 12 possible pairs in which subproblem x is used for training and y for testing ($x \to y$).

CIFAR-10 contains 60000 natural images of 10 classes. We evaluate the 100-shot anomaly detection performance in every class and calculate the average AUC-ROC as the final performance.

5.2 Experimental setup

For the mini-ImageNet, we resize images to 84×84 and perform 1- to 10-shot experiments in a 5-way scenario to demonstrate the improvements obtained with our SMH on different numbers of W-shot images. For each training and testing episode, we randomly select 5 and 15 query samples per class, respectively. We average the final results over 600 episodes. The initial learning rate is set to 1E−3. We train the model with 150000 episodes. For fine-grained classification datasets, e.g., Flower102, CUB-200-2011, and Food-101, we follow the similar setup with mini-ImageNet, but we use the evaluation protocol proposed in [94].

For the Open MIC dataset, we select 4 out of 12 possible subproblems, that is, $p1 \to p2$, $p2 \to p3$, $p3 \to p4$, and $p4 \to p1$. First, we apply the mean extraction on patch images (Open MIC provides three large crops per image) and resize them to 84×84 pixels. As some classes of Open MIC contain fewer than 3 images, we apply 5-way 1-shot to the 3-shot learning protocol. During training, to form an episode, we randomly select 1–3 patch images for the support set and another 2 patch images for the query set for each class. During testing, we use the same number of support and query samples in every episode, and we average the accuracy of over 1000 episodes for the final score. The initial learning rate is set to 1E−4. The models are trained with 50000 episodes.

For anomaly detection experiments on the CIFAR-10 dataset, we use a standard few-shot setup that performs training and testing on every class; then, we average the per-class AUC-ROC.

5.3 Results

For the mini-ImageNet and tiered-ImageNet datasets, Table 2 [4–6, 8, 12–14, 18–20, 38, 94–105] shows that our proposed SMH outperforms all other state-of-the-art methods on standard 5-way 1- and 5-shot protocols. Compared with Conv-4-based methods, our SMH Inter-class model achieves ∼5.4% and ∼4.8% higher accuracy than SoSN on 1- and 5-shot protocols on mini-ImageNet, respectively, while our SMH Intra-class model yields improvements of ∼4.4% and ∼3.6% accuracy over the baseline model SoSN. Moreover, the same SMH Inter-class and SMH Intra-class models outperform Manifold Mixup by 4%–5%

Table 2 Evaluations on the mini-ImageNet and tiered-ImageNet datasets[a)]. See [6,8] for details of baselines. Note that intra-class hallucination has no effect on one-shot learning, so the scores of without (w/o Hal.) and with intra-class hallucination (Intra-class Hal.) on 1-shot are identical. The 5-way accuracy is reported. For comparisons of our saliency-guided hallucination strategy with other hallucination methods, we include MetaGAN and AFHN (generative hallucination), and Manifold Mixup (results on ResNet-18 are from [20]. For results on Conv-4-64, we equipped [20] with the Conv-4-64 backbone and ran evaluations ourselves).

Model	Backbone	mini-ImageNet		tiered-ImageNet	
		1-shot	5-shot	1-shot	5-shot
Matching Nets [4]	Conv-4-64	43.56 ± 0.84	55.31 ± 0.73	–	–
Meta-Learn Nets [95]	Conv-4-64	43.44 ± 0.77	60.60 ± 0.71	–	–
Prototypical Nets [5]	Conv-4-64	49.42 ± 0.78	68.20 ± 0.66	53.31 ± 0.89	72.69 ± 0.74
MAML [12]	Conv-4-64	48.70 ± 1.84	63.11 ± 0.92	51.67 ± 1.81	70.30 ± 1.75
Relation Nets [6]	Conv-4-64	51.36 ± 0.86	65.63 ± 0.72	54.48 ± 0.93	71.32 ± 0.78
MetaGAN [18]	Conv-4-64	52.71 ± 0.64	68.63 ± 0.67	–	–
SoSN [8]	Conv-4-64	52.96 ± 0.83	68.63 ± 0.68	58.62 ± 0.92	75.19 ± 0.79
SoSN + SC [94]	Conv-4-64	55.36 ± 0.72	71.23 ± 0.64	60.13 ± 0.91	77.35 ± 0.75
Manifold Mixup [20]	Conv-4-64	53.87 ± 0.70	69.98 ± 0.65	58.56 ± 0.89	78.20 ± 0.74
TADAM [96]	ResNet-12	58.50 ± 0.30	76.70 ± 0.30	–	–
MetaOpt [13]	ResNet-12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
SNAIL [97]	ResNet-12	55.71 ± 0.99	68.88 ± 0.92	–	–
MTL [98]	ResNet-12	61.20 ± 1.80	75.50 ± 0.80	–	–
Variational FSL [13]	ResNet-12	61.23 ± 0.26	77.69 ± 0.17	–	–
SimpleShot [14]	ResNet-10	60.85 ± 0.20	78.40 ± 0.15	–	–
Rethinking [38]	ResNet-12	62.02 ± 0.63	79.64 ± 0.44	69.74 ± 0.72	84.41 ± 0.55
Manifold Mixup [20]	ResNet-18	55.77 ± 0.23	71.15 ± 0.12	–	–
AFHN [19]	ResNet-18	62.38 ± 0.72	78.16 ± 0.56	–	–
BML [99]	ResNet-12	67.04 ± 0.63	83.63 ± 0.29	68.99 ± 0.50	85.49 ± 0.34
COSOC [100]	ResNet-12	69.28 ± 0.49	85.16 ± 0.42	73.57 ± 0.43	87.57 ± 0.10
TPMN [101]	ResNet-12	67.64 ± 0.63	83.44 ± 0.43	72.24 ± 0.70	86.55 ± 0.63
CNL [102]	ResNet-12	67.96 ± 0.98	83.36 ± 0.51	73.42 ± 0.95	87.72 ± 0.75
PAL [103]	ResNet-12	69.37 ± 0.64	84.40 ± 0.44	72.25 ± 0.72	86.95 ± 0.47
Setfeat [104]	SF-12	68.32 ± 0.62	82.71 ± 0.46	73.63 ± 0.88	87.59 ± 0.57
CECNet [105]	WRN-28	70.20 ± 0.46	85.00 ± 0.30	73.84 ± 0.50	87.36 ± 0.34
SMH w/o Sal. Seg. + MNL (*)	Conv-4-64	53.15 ± 0.87	68.87 ± 0.67	59.23 ± 0.89	76.10 ± 0.77
SMH Foreground Only + MNL (*)	Conv-4-64	50.32 ± 0.85	65.26 ± 0.64	57.73 ± 0.90	75.08 ± 0.79
SMH Mix w/o Hal. + MNL (*)	Conv-4-64	56.18 ± 0.86	70.35 ± 0.66	60.97 ± 0.87	77.93 ± 0.75
SMH Intra-class + MNL	Conv-4-64	57.83 ± 0.87	73.15 ± 0.72	63.43 ± 0.89	79.87 ± 0.75
SMH Inter-class + MNL	Conv-4-64	58.89 ± 0.88	74.11 ± 0.66	65.21 ± 0.88	81.35 ± 0.76
SMH Inter-class + MNL	ResNet-12	64.13 ± 0.82	80.31 ± 0.66	70.05 ± 0.87	85.11 ± 0.74
conSMH Inter-class + MNL	ResNet-12	65.83 ± 0.79	82.51 ± 0.63	72.36 ± 0.84	88.06 ± 0.71
conSMH Inter-class + SelfMask	ResNet-12	67.75 ± 0.77	85.31 ± 0.61	73.84 ± 0.84	88.93 ± 0.72
conSMH Inter-class + SelfMask	ResNet-18	68.81 ± 0.75	86.14 ± 0.63	73.98 ± 0.81	89.11 ± 0.71
conSMH Inter-class + SelfMask	WRN-28	70.35 ± 0.66	86.73 ± 0.60	74.23 ± 0.87	89.35 ± 0.74

a) The asterisk (*) denotes the "sanity check" results on our proposed pipeline for the disabled saliency segmentation and hallucination.

and 3%–4% on 1- and 5-shot protocols on mini-ImageNet, respectively, and MetaGAN by 5%–6% and 4%–5% in the above evaluation setup.

Given the ResNet-12 backbone and MNL detector, the performance of SMH increases to 64.13% and 80.31% for 1- and 5-shot protocols, respectively, outperforming previous ResNet-12 based methods. Using the conditional pipeline (conSMH), the performance can be further boosted by 1.6% and 2.2%. For tiered-ImageNet, our SMH and conSMH also achieve state-of-the-art performance for Conv-4 and ResNet-12 backbones. Moreover, our SMH Inter-class and conSMH (ResNet-12, mini-ImageNet) considerably outperform Manifold Mixup (ResNet-18 backbone) and GAN-based AFHN, which shows that the saliency-guided hallucination produces meaningful auxiliary samples. Applying the SelfMask detector further improves the overall performance by 1%–2%. Meanwhile, using a stronger backbone also considerably affects the learning accuracy. To illustrate, conSMH achieves 68.81%/86.14% for 1- and 5-shot, respectively, with the ResNet-18 backbone and achieves 70.35%/86.73% with the WRN-28 backbone.

Table 3 [4–6, 8, 12, 20, 94, 106, 107] shows the performance of SMH on three fine-grained classification

Table 3 Evaluations on fine-grained classification datasets, CUB-200-2011, Flower102, and Food-101 (5-way accuracy, Conv-4 backbone)[a)]

Model	Flower102		CUB-200-2011		Food-101	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Match. Net [4]	61.21 ± 0.91	79.24 ± 0.66	36.79 ± 0.85	51.34 ± 0.71	33.85 ± 0.73	48.21 ± 0.67
MAML [12]	61.98 ± 0.90	80.34 ± 0.65	37.83 ± 0.84	51.16 ± 0.73	34.26 ± 0.71	48.78 ± 0.65
Proto. Net [5]	62.81 ± 0.93	82.11 ± 0.65	37.42 ± 0.86	51.57 ± 0.73	34.97 ± 0.71	49.13 ± 0.66
Relation Nets [6]	68.26 ± 0.94	80.94 ± 0.66	40.62 ± 0.84	53.91 ± 0.74	36.89 ± 0.72	49.07 ± 0.65
Manifold Mixup [20]	71.35 ± 0.92	84.83 ± 0.63	41.53 ± 0.82	55.08 ± 0.71	37.96 ± 0.70	49.88 ± 0.62
SoSN [8]	73.07 ± 0.94	87.68 ± 0.62	46.72 ± 0.89	60.34 ± 0.73	41.15 ± 0.71	54.92 ± 0.64
SoSN + SC [94]	74.42 ± 0.93	89.15 ± 0.62	48.12 ± 0.86	61.79 ± 0.71	42.06 ± 0.72	56.09 ± 0.63
LRPABN [106]	76.15 ± 0.89	91.09 ± 0.60	52.37 ± 0.81	65.98 ± 0.67	44.25 ± 0.70	58.21 ± 0.65
MattML [107]	77.93 ± 0.91	92.23 ± 0.59	54.46 ± 0.80	67.83 ± 0.68	45.88 ± 0.68	59.16 ± 0.62
SMH Intra-class + MNL	75.44 ± 0.96	90.86 ± 0.56	53.86 ± 0.87	68.38 ± 0.71	44.51 ± 0.73	57.26 ± 0.66
SMH Inter-class + MNL	78.12 ± 0.91	92.16 ± 0.55	55.31 ± 0.88	70.85 ± 0.70	46.49 ± 0.73	59.35 ± 0.64
conSMH Inter-class + MNL	79.05 ± 0.89	93.08 ± 0.53	56.22 ± 0.85	72.03 ± 0.68	48.03 ± 0.70	60.87 ± 0.61
conSMH Inter-class + SelfMask	$\mathbf{80.73 \pm 0.85}$	$\mathbf{94.35 \pm 0.50}$	$\mathbf{57.74 \pm 0.81}$	$\mathbf{74.25 \pm 0.64}$	$\mathbf{49.32 \pm 0.67}$	$\mathbf{63.09 \pm 0.59}$

a) Bold indicates optimal performance.

Table 4 The 100-shot performance of our proposed saliency-guided FSAD on CIFAR-10 (Conv-4 backbone)[a)]

Class	HOG	DCAE	AnoGAN	Deep SVDD	SMH (MNL)					SMH (SelfMask)
	OC-SVM	[108]	[109]	[81]	w/o Neg.	Intra. Hal	+Gauss.	+Back.	+Back.+GM	+Back.+GM
1	55.3	56.8	59.3	63.2	61.7	63.3	65.1	76.2	**78.3**	80.2
2	58.2	46.0	55.2	56.6	55.9	58.1	62.6	62.4	**65.2**	68.1
3	43.9	64.2	40.6	41.5	61.2	64.1	64.0	62.2	**64.2**	66.3
4	46.0	58.0	53.8	54.7	61.6	**63.3**	57.7	55.0	60.8	62.1
5	60.7	74.2	52.1	51.2	69.3	70.8	**75.2**	68.8	70.9	73.2
6	52.7	53.6	60.3	61.8	57.4	58.8	59.5	62.3	**65.2**	66.9
7	63.8	70.0	56.9	57.4	65.4	67.2	71.5	70.6	**71.5**	73.3
8	53.7	48.4	60.8	62.7	55.4	58.3	58.0	61.6	**61.7**	63.0
9	56.4	62.3	72.7	74.1	62.9	64.4	62.4	72.5	**74.4**	76.0
10	38.6	38.3	70.9	**72.4**	58.1	60.1	59.6	69.7	72.3	73.5
Avg.	52.9	57.2	58.2	59.6	60.9	62.8	63.6	66.1	**68.4**	70.3

a) Bold indicates optimal performance.

datasets. SMH Inter-class and SMH Intra-class considerably improve the 1- and 5-shot classification accuracies on these datasets. Taking CUB-200-2011 as an example, our proposed SMH models outperform previous baselines by up to 9.6% and 10.5% for 1- and 5-shot, respectively, which clearly demonstrates the effectiveness of meta-hallucination in fine-grained classification tasks. Notably, applying conSMH on the Flower102 and CUB-200-2011 datasets does not achieve large improvements over SMH, as most images in these two benchmarks enjoy valid/accurate saliency maps.

Table 4 [81,108,109] reports the results of few-shot anomaly detection. Our SMH considerably outperforms baseline models, i.e., deep SVDD, by up to 8.0% AUC-ROC. Specifically, SMH w/o Hal obtains similar performance as baseline models, whereas applying intra-class hallucination brings an approximately 2% gain. Introducing contrastive samples further improves the performance. Applying random Gaussian noises to the VAE and using the background patch sampling bring improvements of 0.8% and 3.0%, respectively. According to SMH Intra-class + Back., we further apply RD and GM to enhance the performance of FSAD and achieve 67.6% AUC-ROC. The experimental results clearly demonstrate the usefulness of meta-hallucination when very few one-class samples are available.

Table 5 [6, 8] presents results on Open MIC. The improvements of SMH Intra-class and SMH Inter-class on this dataset are consistent with improvement on mini-ImageNet. However, the improvements on some splits are small (i.e., $\sim$1.1%) because of the difficult contents of these splits; for example, jewelry, fossils, complex nonlocal engine installations, and semitransparent exhibits captured with wearable cameras cannot be easily segmented out by salient detectors. Applying the conSMH pipeline improves the performance on Open MIC by 1%–2% (5-way 1-shot).

Table 5 Evaluations on the Open MIC dataset[a)]

Model	Way	Shot	$p1 \to p2$	$p2 \to p3$	$p3 \to p4$	$p4 \to p1$
Relation Nets [6]	5	1	70.1	49.7	66.9	46.9
SoSN [8]	5	1	78.0	60.1	75.5	57.8
SMH Intra-class + MNL	5	1	78.2	60.3	75.9	58.1
SMH Inter-class + MNL	5	1	79.3	61.4	76.6	59.2
conSMH Inter-class + MNL	5	1	80.6	63.2	78.1	61.3
conSMH Inter-class + SelfMask	5	1	81.6	65.0	79.2	62.1
Relation Nets [6]	5	3	80.9	61.9	78.5	58.9
SoSN [8]	5	3	87.1	72.6	85.9	72.8
SMH Intra-class + MNL	5	3	87.5	73.9	86.5	73.6
SMH Inter-class + MNL	5	3	88.1	74.2	87.1	73.9
conSMH Inter-class + SelfMask	5	3	90.2	76.7	89.0	76.2
Relation Nets [6]	10	1	54.4	35.3	53.1	35.5
SoSN [8]	10	1	67.2	46.2	63.9	46.6
SMH Intra-class + MNL	10	1	68.7	47.9	65.2	48.1
SMH Inter-class + MNL	10	1	69.5	48.7	65.8	48.9
conSMH Inter-class + SelfMask	10	1	71.3	50.2	67.8	50.7
Relation Nets [6]	10	3	69.0	45.7	67.5	46.3
SoSN [8]	10	3	78.0	56.3	77.5	58.6
SMH Intra-class + MNL	10	3	79.2	58.3	78.3	59.1
SMH Inter-class + MNL	10	3	79.9	58.7	79.2	60.3
conSMH Inter-class + SelfMask	10	1	83.5	61.2	82.7	63.5

a) $p1$: shn+hon+clv, $p2$: clk+gls+scl, $p3$: sci+nat, $p4$: shx+rlc. Notation $x \to y$ means training on exhibition x and testing on exhibition y (5-way accuracy, Conv-4 backbone).

5.4 Ablation studies

The network proposed in our paper builds on frameworks [6,8]. However, we have added several nontrivial units/subnetworks to accomplish our goal of the sample hallucination in the feature space. Thus, we perform additional experiments to show which accuracy gains stem from which of our contributions and break down the accuracy w.r.t. various pipeline components.

Table 2 shows that if the saliency segmentation and data hallucination are disabled in our pipeline (SMH w/o Sal. Seg.), the performance on all protocols drops to the baseline level of SoSN, which demonstrates that the benefits from network modification are very limited.

Importantly, we observe that SMH outperforms SoSN even if we segment images into foregrounds and backgrounds and pass them via our network without the use of hallucinated samples (SMH w/o Hal.) We assert that such improvements stem from the ability of the salient detector to localize main objects in images. This form of spatial knowledge transfer helps our network capture the connections between foregrounds and backgrounds and thus capture the similarity between query and support images better.

Table 6 investigates base learners and shows that a linear classifier (instead of a relation learning network) is an excellent choice for our setting, likely due to the inclusion of a Mixing Nets in the pipeline. Table 7 shows that the linear combination mixing strategy is better than concatenation for mixing.

Table 8 demonstrates the impact of four diversification strategies on our meta-hallucination approach. Representation Distillation brings the most considerable improvement (0.82% for SMH Intra-class and 3.73% for SMH Inter. Hal on acc.), followed by the generative model. All strategies reported in the table consistently improve the overall performance. Our RD regularization in combination with the intra- and inter-class hallucination SMH Intra-class+RD and SMH Inter-class+RD brings gains of up to 1.6% and 1.5% accuracy on mini-ImageNet. We conclude that RD helps our end-to-end training by forcing encoder $\boldsymbol{g}(\cdot)$ to mimic teacher $\boldsymbol{g}^*(\cdot)$ for real foreground–background pairs ($\boldsymbol{g}^*(\cdot)$ is trained on such pairs only to act as a reliable supervisor). In addition, Table 9 shows that whether the teacher network encodes entire original images or stitches their foreground–background (from the same image), the use of the teacher network is in both cases, and gains are similar, as expected.

Figure 9(a) compares different baseline models and our SMH w.r.t. variation in the number of shots. The meta-hallucination brings higher gains when the shot number is small. Figure 9(b) shows the accuracy of our SMH Inter-class model on mini-ImageNet for 1-shot 5-way as a function of the SSP. The

Table 6 Ablation study (5-way accuracy, ResNet-12 backbone, MNL saliency detector) w.r.t. the choice of the base learner in the meta-hallucination pipeline on mini-ImageNet[a)]

Model	Base learner	5-way 1-shot	5-way 5-shot
SMH Inter-class	RN	63.13 ± 0.86	78.45 ± 0.68
SMH Inter-class	NN	63.81 ± 0.85	79.34 ± 0.68
SMH Inter-class	LC	$\mathbf{64.13 \pm 0.82}$	$\mathbf{80.31 \pm 0.66}$

a) RN denotes the relation learning module, NN refers to the nearest-neighbor classification, and LC denotes the linear classifier. Bold indicates optimal performance.

Table 7 Ablation study on different foreground-background mixing operators from Subsection 3.3.1 on mini-ImageNet (5-way accuracy, Conv-4 backbone, MNL detector)

Pipeline	Mix. Op.	5-way 1-shot	5-way 5-shot
SMH Intra-class	linear comb.	57.83	73.15
	concat.	57.02	72.83
SMH Inter-class	linear comb.	58.89	74.11
	concat.	56.57	71.78

Table 8 Ablation study on different diversification strategies of meta-hallucination on mini-ImageNet (5-way accuracy, Conv-4 backbone, MNL detector)

Hal. Strategy	RD	SD	GM	5-way 1-shot	5-way 5-shot
SMH Intra-class	–	–	–	55.61	70.89
	✓	–	–	56.43	71.95
	✓	✓	–	56.64	72.19
	✓	✓	✓	57.83	73.15
SMH Inter-class	–	–	–	53.72	67.76
	✓	–	–	57.45	72.01
	✓	✓	–	57.71	72.44
	✓	✓	✓	58.89	74.11

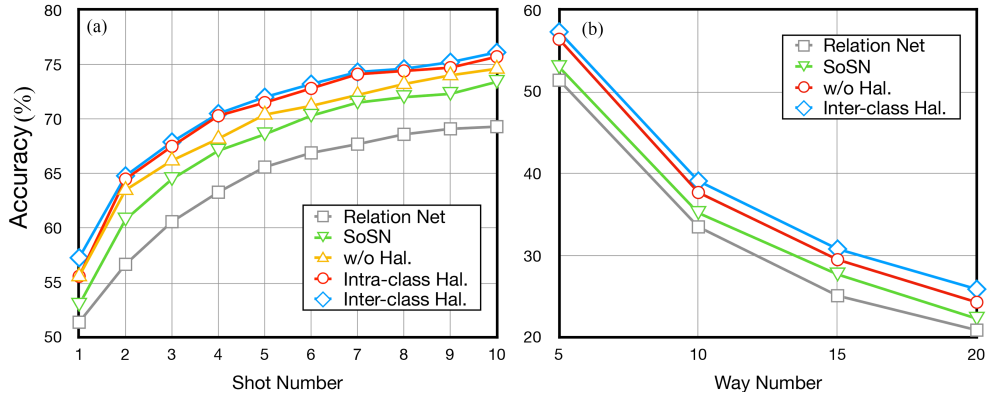

Figure 9 Accuracy w.r.t. (a) W-shot (5-way) and (b) N-way (5-shot) on mini-ImageNet on a few methods (Conv-4 backbone, MNL detector).

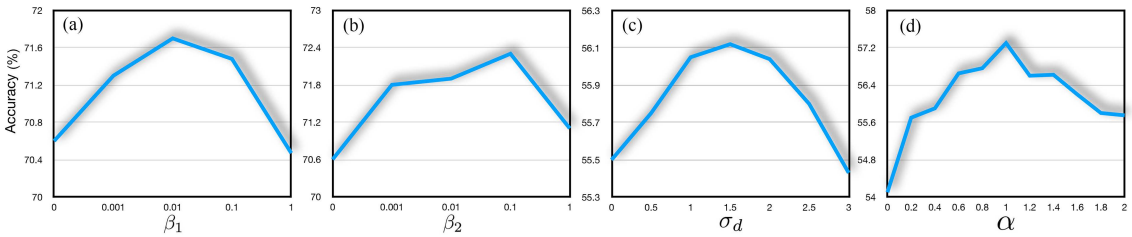

Figure 10 Accuracy on mini-ImageNet as a function of (a) β_1 of RD from Eq. (14) (5-shot 5-way), (b) β_2 of GM from Eq. (21), (c) σ_d of Saliency Dilation, and (d) α of SSP from Eq. (11) (1-shot 5-way accuracy, Conv-4 backbone, MNL detector).

maximum observed gain in accuracy is ∼3.3%.

Table 9 Evaluations on the mini-ImageNet dataset given different teacher networks (options (i) and (ii) from Subsection 3.4.1) for the RD regularization (5-way accuracy, Conv-4 backbone, MNL detector)

Model	5-way 1-shot	5-way 5-shot
baseline 1: option (i) as the teacher network in RD		
SMH Intra-class	56.11 ± 0.88	71.56 ± 0.67
SMH Inter-class	57.24 ± 0.94	72.49 ± 0.65
baseline 2: option (ii) as the teacher network in RD		
SMH Intra-class	55.57 ± 0.86	71.78 ± 0.69
SMH Inter-class	57.45 ± 0.88	72.01 ± 0.67

Table 10 Ablation study w.r.t. the choice of saliency detectors in the meta-hallucination pipeline on mini-ImageNet (5-way 5-shot accuracy, ResNet-12 backbone)

Model	Saliency detector						
	RBD [50]	RFCN [51]	MNL [25]	UC-Net [57]	SelfMask [58]	FOUND [59]	MOVE [60]
SMH Intra-class	73.23	75.96	77.23	79.23	79.68	79.02	79.13
SMH Inter-class	75.12	77.87	80.31	81.27	81.94	80.53	80.91
conSMH	78.82	80.94	82.51	83.71	84.09	83.37	83.43

Table 11 Ablation study w.r.t. the choice of backbone in the meta-hallucination pipeline on mini-ImageNet (5-way 5-shot accuracy, SelfMask detector)

Model	Backbone			
	Conv-4-64	ResNet-12	ResNet-18	WRN-28
SMH Intra-class	73.15	79.68	82.31	85.90
SMH Inter-class	75.33	81.94	84.23	87.18
conSMH	76.67	84.09	86.13	88.21

Figure 10(a) shows the accuracy of our SMH Intra-class model on mini-ImageNet for the 5-shot 5-way case as a function of the parameter β of our regularization loss RD. We observe that for $\beta = 0.01$, we gain $\sim$1% accuracy over $\beta = 0$ (RD disabled). Importantly, the gain remains stable over a large range of $0.005 \leqslant \beta \leqslant 0.5$.

Table 10 [25, 50, 51, 57–60] compares several saliency methods in terms of few-shot learning accuracy. The complex saliency methods perform equally well. Although a stronger saliency detector generally contributes to higher accuracy, the improvement is marginal when the detector is strong enough. Among all detectors, SelfMask helps SMH achieve the best performance, while the use of the RBD approach [50] results in a considerable performance loss due to its numerous failures.

Table 11 shows the impacts of various backbones. Deeper backbones generally contribute to higher performance. Using WRN-28 as the backbone achieves the best accuracy but increases the training overhead.

6 Conclusion

In this paper, we presented two novel lightweight meta-hallucination strategies for few-shot image classification and anomaly detection. In contrast to other costly hallucination methods based on GANs, these strategies leveraged a readily available saliency network to obtain foreground-background pairs on which we trained our SMH network in an end-to-end manner. We showed that mixing foreground and background features is also a better strategy than Manifold Mixup, which mixes features and corresponding labels. To cope with the noises of saliency maps and improve the diversity of hallucinated samples, we proposed several novel diversification strategies that regularize our network and help generate viable couplings. To address the mixing noise due to non-salient images, we introduced the mixing controller, which excludes non-salient images from mixing with other images. We further applied our pipeline to a challenging problem of few-shot anomaly detection and showed that it outperforms baseline models. In future work, we will investigate self-supervised variants of our SMH pipeline.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62106282) and Beijing Nova Program (Grant No. 20220484139).

References

1 Woodworth R S, Thorndike E L. The influence of improvement in one mental function upon the efficiency of other functions. Psychological Rev, 1901, 8: 247–261
2 Miller E G, Matsakis N E, Viola P A. Learning from one example through shared densities on transforms. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2000. 464–471
3 Li F-F, Fergus R, Perona P. One-shot learning of object categories. IEEE Trans Pattern Anal Machine Intell, 2006, 28: 594–611
4 Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning. In: Proceedings of Conference on Neural Information Processing Systems, 2016. 3630–3638
5 Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proceedings of Conference on Neural Information Processing Systems, 2017. 4077–4087
6 Sung F, Yang Y, Zhang L, et al. Learning to compare: relation network for few-shot learning. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018
7 Santoro A, Raposo D, Barrett D G, et al. A simple neural network module for relational reasoning. In: Proceedings of Conference on Neural Information Processing Systems, 2017. 4967–4976
8 Zhang H, Koniusz P. Power normalizing second-order similarity network for few-shot learning. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2019. 1185–1193
9 Weinberger K Q, Blitzer J, Saul L K. Distance metric learning for large margin nearest neighbor classification. In: Proceedings of Conference on Neural Information Processing Systems, 2006. 1473–1480
10 Köstinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2012. 2288–2295
11 Harandi M, Salzmann M, Hartley R. Joint dimensionality reduction and metric learning: a geometric take. In: Proceedings of International Conference on Machine Learning, 2017. 1404–1413
12 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of International Conference on Machine Learning, 2017. 1126–1135
13 Lee K, Maji S, Ravichandran A, et al. Meta-learning with differentiable convex optimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 10657–10665
14 Wang Y, Chao W L, Weinberger K Q, et al. SimpleShot: revisiting nearest-neighbor classification for few-shot learning. 2019. ArXiv:1911.04623
15 Ziko I, Dolz J, Granger E, et al. Laplacian regularized few-shot learning. In: Proceedings of International Conference on Machine Learning, 2020. 11660–11670
16 Hariharan B, Girshick R B. Low-shot visual recognition by shrinking and hallucinating features. In: Proceedings of International Conference on Computer Vision, 2017. 3037–3046
17 Wang Y X, Girshick R, Hebert M, et al. Low-shot learning from imaginary data. 2018. ArXiv:1801.05401
18 Zhang R, Che T, Ghahramani Z, et al. MetaGAN: an adversarial approach to few-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 31
19 Li K, Zhang Y, Li K, et al. Adversarial feature hallucination networks for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 13470–13479
20 Mangla P, Kumari N, Sinha A, et al. Charting the right manifold: manifold mixup for few-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020
21 Fu Y, Fu Y, Jiang Y G. Meta-FDMixup: cross-domain few-shot learning guided by labeled target data. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021. 5326–5334
22 Zhang H, Cisse M, Dauphin Y N, et al. mixup: beyond empirical risk minimization. In: Proceedings of International Conference on Learning Representations, 2018
23 Verma V, Lamb A, Beckham C, et al. Manifold Mixup: better representations by interpolating hidden states. In: Proceedings of the 36th International Conference on Machine Learning, 2019. 6438–6447
24 Zhang H, Zhang J, Koniusz P. Few-shot learning via saliency-guided hallucination of samples. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2019. 2770–2779
25 Zhang J, Zhang T, Dai Y, et al. Deep unsupervised saliency detection: a multiple noisy labeling perspective. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018
26 Liu T, Sun J, Zheng N N, et al. Learning to detect a salient object. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2007. 1–8
27 Rajapakse J C, Wang L. Neural Information Processing: Research and Development. Berlin: Springer, 2004
28 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis, 2015, 115: 211–252
29 Larochelle H, Erhan D, Bengio Y. Zero-data learning of new tasks. In: Proceedings of AAAI Conference on Artificial Intelligence, 2008
30 Farhadi A, Endres I, Hoiem D, et al. Describing objects by their attributes. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2009. 1778–1785
31 Akata Z, Perronnin F, Harchaoui Z, et al. Label-embedding for attribute-based classification. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2013. 819–826
32 Zhang H, Koniusz P. Zero-shot kernel learning. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018. 7670–7679
33 Li F F, VanRullen R, Koch C, et al. Rapid natural scene categorization in the near absence of attention. Proc Natl Acad Sci USA, 2002, 99: 9596–9601
34 Fink M. Object classification from a single example utilizing class relevance metrics. In: Proceedings of Conference on Neural Information Processing Systems, 2005. 449–456
35 Bart E, Ullman S. Cross-generalization: learning novel classes from a single example by feature replacement. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2005. 672–679
36 Lake B M, Salakhutdinov R, Gross J, et al. One shot learning of simple visual concepts. In: Proceedings of the Annual Meeting of the Cognitive Science Society, 2011
37 Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. In: Proceedings of International Conference on Machine Learning Deep Learning Workshop, 2015
38 Tian Y, Wang Y, Krishnan D, et al. Rethinking few-shot image classification: a good embedding is all you need? 2020.

ArXiv:2003.11539
39 Zhang C, Cai Y, Lin G, et al. DeepEMD: few-shot image classification with differentiable earth mover's distance and structured classifiers. 2020. ArXiv:2003.06777
40 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM, 2020, 63: 139–144
41 Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation. 2017. ArXiv:1710.10196
42 Brock A, Donahue J, Simonyan K. Large scale gan training for high fidelity natural image synthesis. 2018. ArXiv:1809.11096
43 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 4401–4410
44 Zhang J, Zhao C, Ni B, et al. Variational few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019
45 Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of International Conference on Machine Learning, 2017. 214–223
46 Liu M Y, Huang X, Mallya A, et al. Few-shot unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 10551–10560
47 Chen Z, Fu Y, Wang Y X, et al. Image deformation meta-networks for one-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 8680–8689
48 Luo Q, Wang L, Lv J, et al. Few-shot learning via feature hallucination with variational inference. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021. 3963–3972
49 Lazarou M, Stathaki T, Avrithis Y. Tensor feature hallucination for few-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022. 3500–3510
50 Zhu W, Liang S, Wei Y, et al. Saliency optimization from robust background detection. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2014. 2814–2821
51 Wang L, Wang L, Lu H, et al. Saliency detection with recurrent fully convolutional networks. In: Proceedings of European Conference on Computer Vision, 2016. 825–841
52 Hou Q, Cheng M M, Hu X, et al. Deeply supervised salient object detection with short connections. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2017. 3203–3212
53 Yang C, Wu Z, Zhou B, et al. Instance localization for self-supervised detection pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 3987–3996
54 Hénaff O J, Koppula S, Shelhamer E, et al. Object discovery and representation networks. In: Proceedings of European Conference on Computer Vision, 2022. 123–143
55 Zhao N, Wu Z, Lau R W, et al. Distilling localization for self-supervised representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 10990–10998
56 Xie Z, Lin Y, Zhang Z, et al. Propagate yourself: exploring pixel-level consistency for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 16684–16693
57 Zhang J, Fan D-P, Dai Y C, et al. Uncertainty inspired RGB-D saliency detection. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 5761–5779
58 Shin G, Albanie S, Xie W. Unsupervised salient object detection with spectral cluster voting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 3971–3980
59 Siméoni O, Sekkat C, Puy G, et al. Unsupervised object localization: observing the background to discover objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 3176–3186
60 Bielski A, Favaro P. MOVE: unsupervised movable object segmentation and detection. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 33371–33386
61 Tuzel O, Porikli F, Meer P. Region covariance: a fast descriptor for detection and classification. In: Proceedings of European Conference on Computer Vision, 2006
62 Romero A, Terán M Y, Gouiffès M, et al. Enhanced local binary covariance matrices (ELBCM) for texture analysis and object tracking. In: Proceedings of the 6th International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, 2013
63 Carreira J, Caseiro R, Batista J, et al. Semantic segmentation with second-order pooling. In: Proceedings of European Conference on Computer Vision, 2012
64 Koniusz P, Yan F, Gosselin P, et al. Higher-order Occurrence Pooling on Mid- and Low-level Features: Visual Concept Detection. Technical Report, 2013
65 Koniusz P, Yan F, Gosselin P H, et al. Higher-order occurrence pooling for bags-of-words: visual concept detection. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 313–326
66 Lin T Y, Chowdhury A R, Maji S. Bilinear cnn models for fine-grained visual recognition. In: Proceedings of International Conference on Computer Vision, 2017
67 Hu G, Hua Y, Yuan Y, et al. Attribute-enhanced face recognition with neural tensor fusion networks. In: Proceedings of International Conference on Computer Vision, 2017
68 Jégou H, Douze M, Schmid C. On the burstiness of visual elements. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2009. 1169–1176
69 Koniusz P, Yan F, Mikolajczyk K. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. Comput Vision Image Understanding, 2013, 117: 479–492
70 Koniusz P, Zhang H, Porikli F. A deeper look at power normalizations. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2018. 5774–5783
71 Boureau Y, Ponce J, LeCun Y. A theoretical analysis of feature pooling in vision algorithms. In: Proceedings of International Conference on Machine Learning, 2010
72 Xie J, Long F, Lv J, et al. Joint distribution matters: deep Brownian distance covariance for few-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 7972–7981
73 Wang Q, Xie J, Zuo W, et al. Deep CNNs meet global covariance pooling: better representation and generalization. IEEE Trans Pattern Anal Mach Intell, 2020, 43: 2582–2597
74 Quinlan J R. Induction of decision trees. Mach Learn, 1986, 1: 81–106
75 Cortes C, Vapnik V. Support-vector networks. Mach Learn, 1995, 20: 273–297
76 Aggarwal C C. Outlier analysis. In: Proceedings of Data Mining, 2015. 237–263
77 Pang G, Shen C, Cao L, et al. Deep learning for anomaly detection: a review. 2020. ArXiv:2007.02500

78 Wang S, Zeng Y, Liu X, et al. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 5962–5975
79 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507
80 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 2672–2680
81 Ruff L, Vandermeulen R, Goernitz N, et al. Deep one-class classification. In: Proceedings of International Conference on Machine Learning, 2018. 4393–4402
82 Golan I, El-Yaniv R. Deep anomaly detection using geometric transformations. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 9758–9769
83 Tian Y, Maicas G, Pu L Z C T, et al. Few-shot anomaly detection for polyp frames from colonoscopy. 2020. ArXiv:2006.14811
84 Kruspe A. One-way prototypical networks. 2019. ArXiv:1906.00820
85 Frikha A, Krompaß D, Köpken H G, et al. Few-shot one-class classification via meta-learning. 2020. ArXiv:2007.04146
86 Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization. 2018. ArXiv:1808.06670
87 Cheng M, Zhang G, Mitra N, et al. Global contrast based salient region detection. In: Proceedings of Conference on Computer Vision and Pattern Recognition, 2011. 409–416
88 Borji A, Cheng M M, Jiang H, et al. Salient object detection: a benchmark. IEEE Trans Image Process, 2015, 24: 5706–5722
89 Ren M, Triantafillou E, Ravi S, et al. Meta-learning for semi-supervised few-shot classification. In: Proceedings of the 6th International Conference on Learning Representations, 2018
90 Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011
91 Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 2008
92 Bossard L, Guillaumin M, van Gool L. Food-101—mining discriminative components with random forests. In: Proceedings of European Conference on Computer Vision, 2014
93 Koniusz P, Tas Y, Zhang H, et al. Museum exhibit identification challenge for the supervised domain adaptation. 2018. ArXiv:1802.01093
94 Koniusz P, Zhang H. Power normalizations in fine-grained image, few-shot image and graph classification. 2020. ArXiv:2012.13975
95 Ravi S, Larochelle H. Optimization as a model for few-shot learning. In: Proceedings of International Conference on Learning Representations, 2017
96 Oreshkin B N, Rodriguez P, Lacoste A. TADAM: task dependent adaptive metric for improved few-shot learning. 2018. ArXiv:1805.10123
97 Mishra N, Rohaninejad M, Chen X, et al. A simple neural attentive meta-learner. 2017. ArXiv:1707.03141
98 Sun Q, Liu Y, Chua T S, et al. Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 403–412
99 Zhou Z, Qiu X, Xie J, et al. Binocular mutual learning for improving few-shot classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 8402–8411
100 Luo X, Wei L, Wen L, et al. Rectifying the shortcut learning of background for few-shot learning. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 13073–13085
101 Wu J, Zhang T, Zhang Y, et al. Task-aware part mining network for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 8433–8442
102 Zhao J, Yang Y, Lin X, et al. Looking wider for better adaptive representation in few-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 10981–10989
103 Ma J, Xie H, Han G, et al. Partner-assisted learning for few-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 10573–10582
104 Afrasiyabi A, Larochelle H, Lalonde J F, et al. Matching feature sets for few-shot image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 9014–9024
105 Lai J, Yang S, Zhou J, et al. Clustered-patch element connection for few-shot learning. 2023. ArXiv:2304.10093
106 Huang H, Zhang J, Zhang J, et al. Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Trans Multimedia, 2021, 23: 1666–1680
107 Zhu Y, Liu C, Jiang S. Multi-attention meta learning for few-shot fine-grained image recognition. In: Proceedings of International Joint Conferences on Artificial Intelligence, 2020. 1090–1096
108 Masci J, Meier U, Cireşan D, et al. Stacked convolutional auto-encoders for hierarchical feature extraction. In: Proceedings of International Conference on Artificial Neural Networks, 2011. 52–59
109 Schlegl T, Seeböck P, Waldstein S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Proceedings of International Conference on Information Processing in Medical Imaging, 2017. 146–157