Few-shot Learning via Saliency-guided Hallucination of Samples

Hongguang Zhang[1,2]   Jing Zhang[1,2]   Piotr Koniusz[2,1]
[1]Australian National University,   [2]Data61/CSIRO
firstname.lastname@{anu.edu.au[1], data61.csiro.au[2]}

Abstract

Learning new concepts from a few of samples is a standard challenge in computer vision. The main directions to improve the learning ability of few-shot training models include (i) a robust similarity learning and (ii) generating or hallucinating additional data from the limited existing samples. In this paper, we follow the latter direction and present a novel data hallucination model. Currently, most datapoint generators contain a specialized network (i.e., GAN) tasked with hallucinating new datapoints, thus requiring large numbers of annotated data for their training in the first place. In this paper, we propose a novel less-costly hallucination method for few-shot learning which utilizes saliency maps. To this end, we employ a saliency network to obtain the foregrounds and backgrounds of available image samples and feed the resulting maps into a two-stream network to hallucinate datapoints directly in the feature space from viable foreground-background combinations. To the best of our knowledge, we are the first to leverage saliency maps for such a task and we demonstrate their usefulness in hallucinating additional datapoints for few-shot learning. Our proposed network achieves the state of the art on publicly available datasets.

1. Introduction

Convolutional Neural Networks (CNN) have demonstrated their usefulness in numerous computer vision tasks *e.g.*, image classification and scene recognition. However, training CNNs on these tasks requires large numbers of labeled data. In contrast to CNNs, human ability to learn novel concepts from a few of samples remains unrivalled. Inspired by this observation, researchers [8] proposed the one- and few-shot learning tasks with the goal of training algorithms with low numbers of datapoints.

Recently, the concept of learning relations with deep learning has been explored in several papers [36, 33, 34, 32] which can be viewed as a variant of metric learning [39, 21, 11] adapted to the few-shot learning scenario. In these works, a neural network extracts convolutional descriptors, and another learning mechanism (*e.g.*, a relation network)
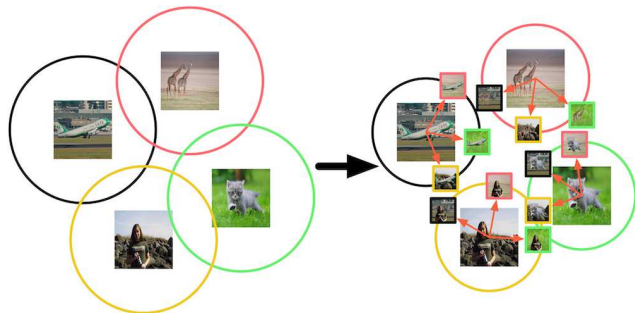

Figure 1: Illustration of saliency-based data generation for one-shot case. The foreground objects are combined with different backgrounds in attempt to refine the classification boundaries.

captures relationship between descriptors. Most papers in this category propose improvements to relationship modeling for the purpose of similarity learning. In contrast, [12] employs a separate Multilayer Perceptron (MLP) to hallucinate additional image descriptors by modeling foreground-background relationships in feature space to obtain implicitly augmented new samples. To train the feature generator, MLP uses manually labelled features clustered into 100 clusters, which highlights the need for extra labelling. Another approach [38] generates data in a meta-learning scenario, which means the network has to be pre-trained on several datasets, thus increasing the cost of training.

In this paper, we adopt the data hallucination strategy and propose a saliency-guided data hallucination network dubbed as *Salient Network* (*SalNet*). Figure 1 shows a simple motivation for our work. Compared with previous feature hallucinating approaches, we employ a readily available saliency network [46] pre-trained on MSRA Salient Object Database (MSRA-B) [25] to segment foregrounds and backgrounds from given images, followed by a two-stream network which mixes foregrounds with backgrounds (we call it the *Mixing Network*) in the feature space of an encoder (*c.f.* image space). As we obtain spatial feature maps from this process, we embed mixed feature vectors into a second-order representation which aggregates over the spatial dimension of feature maps. Then, we capture the similarity between final co-occurrence descriptors of a so-called training query sample and hallucinated support matrices via

a similarity-learning network. Moreover, we regularize our mixing network to promote hallucination of realistically blended foreground-background representations. To this end, whenever a foreground-background pair is extracted from the same image (*c.f.* two separate images), we constrain the resulting blended representation via the ℓ_2-norm to be close to a representation from a supervising network which, by its design, is trained only on real foreground-background pairs (*c.f.* infeasible combinations). We refer to this strategy as *Real Representation Regularization* (*TriR*). Lastly, we propose the similarity-based strategies regarding how to choose backgrounds for mixing with a given foreground. To this end, we perform either (i) intra-class mixing (foregrounds/backgrounds of the same class) or (ii) inter-class mixing (for any given foreground, we take its corresponding background, retrieve its nearest-neighbour backgrounds from various classes, and use the retrieval distance to express the likelihood how valid the mixed pair is). Below, we list our contributions:

I. We propose a novel saliency-guided data hallucination network for few-shot learning.

II. We investigate various hallucination strategies. We propose a simple but effective regularization and two strategies to prevent substandard hallucinated samples.

III. We investigate the effects of different saliency map generators on the few-shot learning performance.

To the best of our knowledge, we are the first to employ saliency maps for datapoints hallucination for few-shot learning. Our experiments achieve the state of the art on two challenging publicly available few-shot learning datasets.

2. Related Work

In what follows, we describe popular zero-, one- and few-shot learning algorithms followed by the saliency detection methods and a discussion on second-order statistics.

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"* [28]. Learning from scarce data poses a challenge to typical CNN-based classification systems [31] which have to learn millions of parameters. Current trends in computer vision highlight 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 was introduced in 1901 under a notion of *"transfer of particle"* [40] and is closely related to zero-shot learning [23, 7, 1] which can be defined as an ability to generalize to unseen class categories from categories seen during training. For one- and few-shot learning, some

"*transfer of particle*" is also a desired mechanism as generalizing from one or few datapoints to account for intra-class variability of thousands images is a formidable task.

One- and Few-shot Learning has been studied widely in computer vision in both shallow [26, 24, 9, 2, 8, 22] and deep learning scenarios [15, 36, 33, 10, 33, 34].

Early works [8, 22] propose generative models with an iterative inference for transfer. In contrast, a recent Siamese Network [15] uses a two-stream convolutional neural network which performs simple metric learning. Matching Network [36] introduces the concept of support set and N-way W-shot learning protocols. It captures the similarity between one query and several support images, and also implicitly performs metric learning. Prototypical Networks [33] learn a model that computes distances between a datapoint and prototype representations of each class. Model-Agnostic Meta-Learning (MAML) [10] is a meta-learning model which can be seen a form of transfer learning. Relation Net [34] is similar to Matching Network [36], but uses an additional network to learn similarity between images. Second-order Similarity Network (SoSN) [45] leverages second-order descriptors and power normalization which help infer rich relation statistics. SoSN descriptors are more effective than the first-order Relation Net [34].

Hallucination-based approaches [12] and [38] use descriptors manually assigned into 100 clusters to generate plausible combinations of datapoints. Mixup network [42] applies a convex combination of pairs of datapoints and labels. In contrast, we decompose images into foreground and background representations via saliency maps and we propose several strategies for mixing foreground-background pairs to hallucinate meaningful auxiliary training samples.

Zero-shot Learning can be implemented within few-shot learning frameworks [15, 36, 33, 34]. Attribute Label Embedding (ALE) [1], Zero-shot Kernel Learning (ZSKL) [44] all use so-called compatibility mapping (linear/non-linear) and some form of regularization to associate feature vectors with attributes (class descriptors). Recent methods such as Feature Generating Networks [41] and Model Selection Network [43] hallucinate the training data for unseen classes via Generative Adversarial Networks (GAN).

2.2. Saliency Detection

A saliency detector highlights image regions containing foreground objects which correlate with human visual attention, thus producing a dense likelihood saliency map which assigns some relevance score in range $[0, 1]$ to each pixel. Conventional saliency detectors underperform on complex scenes due to computations based on human-defined priors [47]. In contrast, deep saliency models [37, 13] outperform conventional saliency detectors but they require laborious pixel-wise labels. In this paper, we use saliency maps as a guiding signal, thus we adopt a highly-efficient weakly-
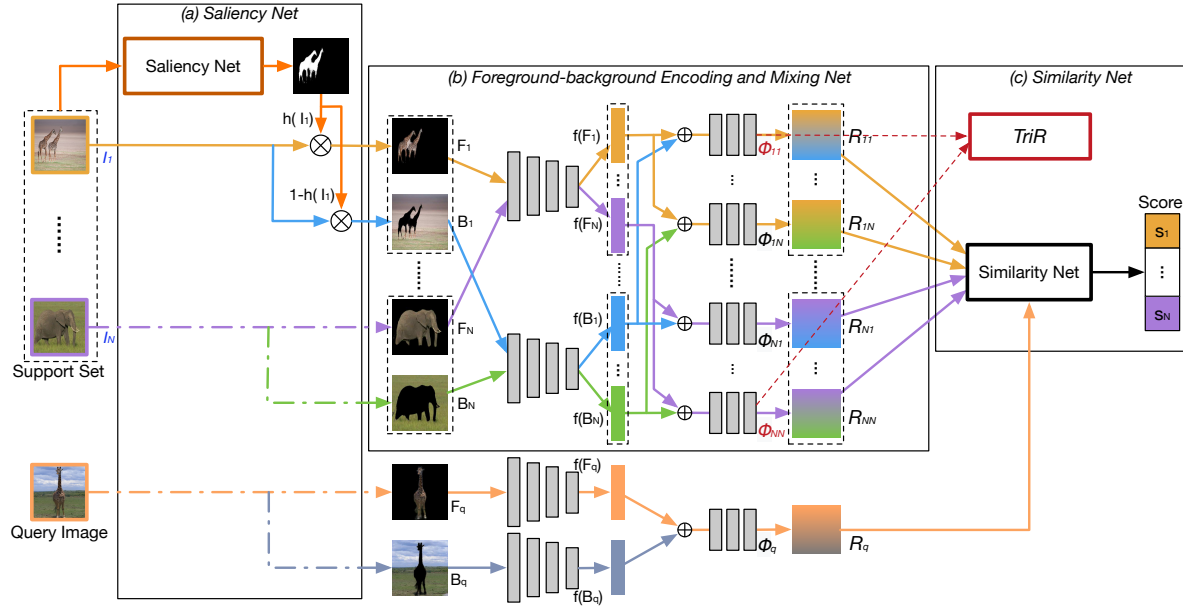
$$\beta \sum_{i=1}^{N} ||g^*(F_i, B_i) - g(F_i, g_i)||^2$$

Figure 2: Our pipeline consists of three units: (a) pre-trained *Saliency Net*, (b) *Foreground-background Encoding and Mixing Net* (*FEMN*), and (c) *Similarity Net*. The FEMN block consists of two streams which take foreground/background images as inputs, respectively, and a *Mixing Net* which combines foreground-background pairs via ⊕ and refines them via a single-stream network prior to aggregation of the resulting feature maps via the *Second-order Encoder*.

supervised deep convolutional saliency detector MNL [46]. We compare the performance of MNL with (i) RFCN [37], a fully-supervised deep model, and (ii) a cheap non-CNN Robust Background Detector (RBD) [47], one of the best unsupervised saliency detectors according to evaluation [3].

2.3. Second-order Statistics

Below we discuss briefly the role of second-order statistics and related shallow and CNN-based approaches. **Second-order statistics** have been studied in the context of texture recognition [35, 30] via so-called Region Covariance Descriptors (RCD), often applied to semantic segmentation [5] and object category recognition [17, 18].

Second-order statistics have to deal with 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*" [14]. Power Normalization [19, 17], used with Bag-of-Words [19, 17, 18, 20], was shown to limit such a burstiness. A survey [19] showed that so-called MaxExp feat. pooling [4] is in fact a detector of "*at least one particular visual word being present in an image*". MaxExp on second-order matrices was shown in [20] to be in fact the Sigmoid function. Such a pooling also performed well in few-shot learning [45]. Thus, we employ second-order pooling with Sigmoid.

3. Approach

Our pipeline builds on the generic few-shot Relation Net pipeline [34] which learns implicitly a metric for so-called query and support images. To this end, images are encoded into feature vectors by an encoding network. Then, so-called episodes with query and support images are formed. Each query-support pair is forwarded to a so-called relation network and a loss function to learn if a query-support pair is of the same class (1) or not (0). However, such methods suffer from scarce training data which we address below.

3.1. Network

Figure 2 presents a foreground-background two-stream network which leverages saliency maps to isolate foreground and background image representations in order to hallucinate additional training data to improve the few-shot learning performance. The network consists of (i) *Saliency Net* (*SalNet*) whose role is to generate foreground hypotheses, (ii) *Foreground-background Encoding and Mixing Net* (*FEMN*) whose role is to combine foreground-background image pairs into episodes, and the *Similarity Net* (*SimNet*) which learns the similarity between query-support pairs.

To illustrate how our network works, consider an image $\mathbf{I}$ which is passed through some saliency network $\mathbf{h}$ to extract the corresponding saliency map $\mathbf{h}(\mathbf{I})$, the foreground $\mathbf{F}$ and the background $\mathbf{B}$ of $\mathbf{I}$, respectively:

$$\mathbf{F}_I = \mathbf{h}(\mathbf{I}) \odot \mathbf{I}, \tag{1}$$
$$\mathbf{B}_I = (1 - \mathbf{h}(\mathbf{I})) \odot \mathbf{I}, \tag{2}$$

where $\odot$ is the Hadamart product. The feature encoding network consists of two parts, $\mathbf{f}$ and $\mathbf{g}$. For images $\mathbf{I} \in \mathbb{R}^{3 \times M \times M}$ and $\mathbf{J} \in \mathbb{R}^{3 \times M \times M}$ ($\mathbf{I} = \mathbf{J}$ or $\mathbf{I} \neq \mathbf{J}$), we proceed by encoding their foreground $\mathbf{F}_I \in \mathbb{R}^{3 \times M \times M}$ and background $\mathbf{B}_J \in \mathbb{R}^{3 \times M \times M}$ via feature encoder $\mathbf{f} : \mathbb{R}^{3 \times M \times M} \to \mathbb{R}^{K \times Z^2}$,

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 of $\mathbf{f}$ of size $Z \times Z$. Then, the encoded foreground and background are mixed via summation and refined in encoder $\mathbf{g} : \mathbb{R}^{K \times Z^2} \to \mathbb{R}^{K' \times Z'^2}$, where K' is the feature size and Z'^2 corresponds to the vectorized spatial dimension of map of $\mathbf{g}$ of size $Z' \times Z'$. As in the SoSN approach [45], we apply the outer-product on $\mathbf{g}(\cdot)$ to obtain an auto-correlation of features and we perform pooling via Sigmoid ψ to tackle the burstiness in our representation. Thus, we have:

$$\boldsymbol{\Phi}_{IJ} = \mathbf{g}(\mathbf{f}(\mathbf{F}_I) + \mathbf{f}(\mathbf{B}_J)), \tag{3}$$

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

where $\boldsymbol{\psi}$ is a zero-centered Sigmoid function with σ as the parameter that controls the slope of its curve:

$$\boldsymbol{\psi}(\mathbf{X}, \sigma) = (1 - e^{-\sigma\mathbf{X}})/(1 + e^{-\sigma\mathbf{X}}) = \tanh(2\sigma\mathbf{X}). \tag{5}$$

Descriptors $\mathbf{R}_{II} \in \mathbb{R}^{K' \times K'}$ represent a given image $\mathbf{I}$ while $\mathbf{R}_{IJ} \in \mathbb{R}^{K' \times K'}$ represent a combined foreground-background pair of images $\mathbf{I}$ and $\mathbf{J}$. Subsequently, we form the query-support pairs (*e.g.*, we concatenate their representations) and we pass episodes to the similarity network. We use the Mean Square Error (MSE) loss to train our network:

$$L = \frac{1}{NW} \sum_{n=1}^{N} \sum_{w=1}^{W} (r(\mathbf{R}_{s_{nw}}, \mathbf{R}_q) - \delta(l_{s_{nw}} - l_q))^2, \tag{6}$$

where s_{nw} chooses support images from $\mathcal{I} = \mathcal{I}^* + \mathcal{I}'$, $\mathcal{I}^*$ and $\mathcal{I}'$ are original and hallucinated images, q chooses the query image, r is the similarity network, l is the label of an image, N is the number of classes in an episode, W is the shot number per support class, $\delta(0) = 1$ (0 elsewhere). Note that Eq. (6) does not form foreground-background hallucinated pairs per se. We describe this process in Section 3.3.

3.2. Saliency Map Generation

For brevity, we consider three approaches: deep supervised saliency approaches [46, 37] and an unsupervised shallow method [47]. In this paper, we use saliency maps as a prior to generate foreground and background hypotheses.

In our main experiemnts, we use the deep weakly-supervised slaiency detector MNL [46] due to its superior performance. Moreover, we investigate the deep supervised RFCN approach [37] pre-trained on THUS10K dataset [6], which has no intersection with our few-shot learning datasets. We also investigate the cheap RBD model [47] which performed best among unsupervised models [3].

Figure 3 shows saliency maps generated by the above methods. In the top row, the foreground and background have distinct textures. Thus, both conventional and deep models isolate the foreground well. However, for the scenes whose foreground/background share color and texture composition (bottom row), the unsupervised method fails to detect the correct foreground. As our dataset contains both
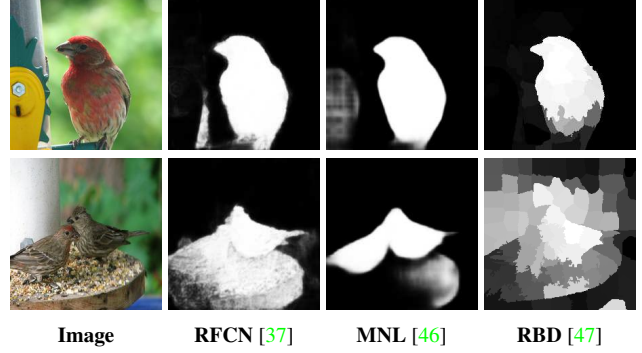

| Image | RFCN [37] | MNL [46] | RBD [47] |

Figure 3: Saliency maps generated by different methods. For a simple scene (top row), the all three methods are able to detect the foreground. However, for a complex scene, the unsupervised method fails to detect the salient object.

simple and complex scenes, the performance of our method is somewhat dependent on the saliency detector *e.g.*, results based on RBD [47] are expected to be worse in comparison to RFCN [37] and MNL [46]. The performance of few-shot learning combined with different saliency detectors will be presented in Section 4.3. Firstly, we detail our strategies for hallucinating additional training data for few-shot learning.

3.3. Data Hallucination

The additional datapoints are hallucinated by the summation of foreground and background feature vector pairs obtained from the feature encoder $\mathbf{f}$ and refined by the encoder $\mathbf{g}$. Taking the N-way W-shot problem as example (see Relation Net [34] or SoSN [45] for the detailed definition of such a protocol), we will randomly sample W images from each of N training classes. Let s_{nw} be the index selecting the w-th image from the n-th class of an episode 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 hallucination of auxiliary datapoints can be formulated.

Strategy I: Intra-class hallucination. For this strategy, given an image index s_{nw}, a corresponding foreground is only mixed with backgrounds of images from the same class n. Thus, we can generate $W - 1$ datapoints for every image. Figure 5 shows that *the intra-class hallucination* produces plausible new datapoints. Note that the image class n 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.

Although our intra-class hallucination presents a promising direction, our results will show that sometimes the performance may lie below baseline few-shot learning due to a very simple mixing foreground-background strategy which includes the foreground-background feature vector summa-
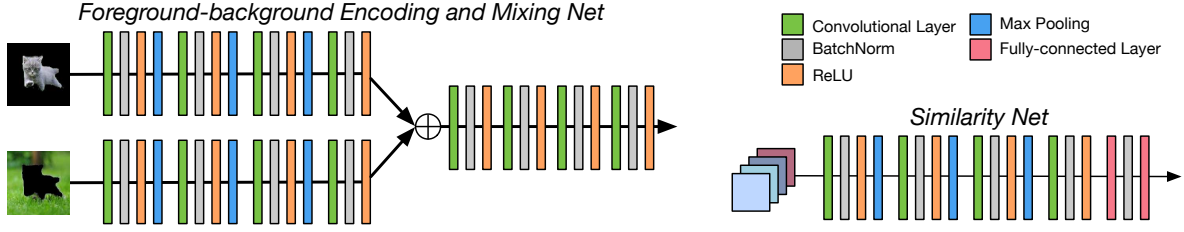
Figure 4: The detailed architecture of Foreground-Background Encoding and Mixing Net and the Similarity Net. Best viewed in color.
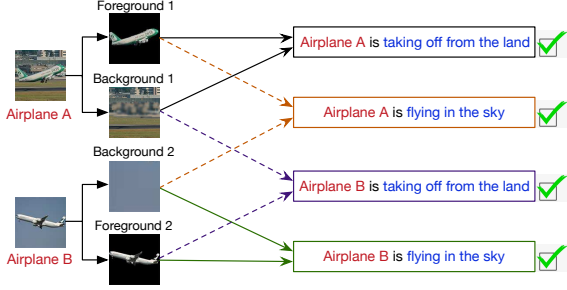

Figure 5: The intra-class datapoint hallucination strategy: the majority of datapoints generated in this way are statistically plausible.
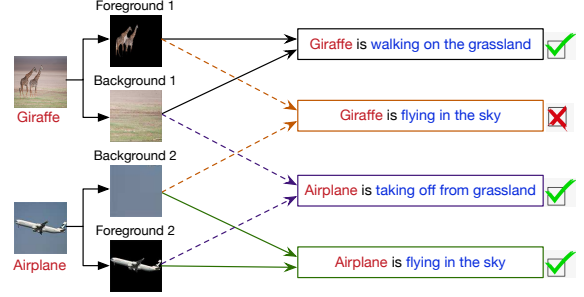

Figure 6: The inter-class datapoint hallucination may generate impossible concepts *e.g.*, 'giraffe in the sky' is an unlikely concept (except for giraffe falling off a helicopter during transportation?).

tion followed by the refining encoder $\mathbf{g}$. Such a strategy incurs possible noises from (i) the substandard saliency maps and/or (ii) mixing incopatible foreground-background pairs.

Therefore, in order to further refine the hallucinated datapoints, we propose to exploit foreground-background mixed pairs $\mathbf{F}_{s_{nw}}$ and $\mathbf{B}_{s_{nw}}$ which come from the same image (*e.g.*, their mixing should produce the original image) and enforce their feature vectors to be close in the ℓ_2-norm sense to some baseline teacher network which does not perform hallucination. Specifically, we take $\boldsymbol{\Phi} = \mathbf{g}(\mathbf{F}_{s_{nw}}, \mathbf{B}_{s_{nw}})$ and encourage its proximity to some teacher representation $\boldsymbol{\Phi}^* = \mathbf{g}^*(\{\mathbf{F}_{s_{nw}}, \mathbf{B}_{s_{nw}}\})$ where $\mathbf{F}_{s_{nw}} + \mathbf{B}_{s_{nw}} = \mathbf{I}_{s_{nw}} \in \mathcal{I}^*$:

$$\Omega = \tfrac{1}{NW} \sum_{n=1}^{N} \sum_{w=1}^{W} \left\| \mathbf{g}(\mathbf{f}(\mathbf{F}_{s_{nw}}) + \mathbf{f}(\mathbf{B}_{s_{nw}})) - \mathbf{g}^*(\{\mathbf{F}_{s_{nw}}, \mathbf{B}_{s_{nw}}\}) \right\|_2^2,$$
$$\text{s.t. } \mathbf{F}_{s_{nw}} + \mathbf{B}_{s_{nw}} = \mathbf{I}_{s_{nw}} \in \mathcal{I}^* \qquad (7)$$
$$L' = L + \beta \Omega,$$

where $\mathcal{I}^*$ is a set of orig. train. images, β adjusts the impact of Ω, L' is the combined loss, and net. $\mathbf{g}^*$ is already trained.

We investigate $\mathbf{g}^*$ that encodes (i) the original images only *i.e.*, $\mathbf{g}^*(\mathbf{f}(\mathbf{I}_{nw}))$ or (ii) foreground-background pairs from original images *i.e.*, $\mathbf{g}^*(\mathbf{f}(\mathbf{F}_{s_{nw}}) + \mathbf{f}(\mathbf{B}_{s_{nw}}))$. We call Ω as *Real Representation Regularization* (*TriR*). Our experiments will demonstrate that TriR improves the final results.

Strategy II: Inter-class hallucination. For this strategy, we allow mixing the foregrounds of support images with all available backgrounds (between-class mixing is allowed) in the support set. Compared to the intra-class generator, *the inter-class hallucination* can generate $W-1+W(N-1)$ new datapoints. However, many foreground-background pairs

will be statistically implausible, as shown in Figure 6, which would cause the degradation of the classification accuracy.

To eliminate the implausible foreground-background pairs from the inter-class hallucination process, we design a similarity prior which assigns probabilities to backgrounds in regard of their compatibility with a given foreground.

Numerous similarity priors can be proposed *e.g.*, one can use the label information to specify some similarity between two given classes. Intuitively, backgrounds between images containing dogs and cats should be more correlated than backgrounds of images of dogs and radios. However, modeling such relations explicitly may be cumbersome and it has its shortcomings *e.g.*, backgrounds of images containing cars may also be suitable for rendering animals on the road or sidewalk, despite of an apparent lack of correlation between say cat and car classes. Thus, we ignore 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 s_{nw} and all other backgrounds to assign a probability score of how similar two backgrounds are, thus:

$$d(\mathbf{B}_{s_{nw}}, \mathbf{B}_{s_{n'w'}}) = \left\| \mathbf{f}(\mathbf{B}_{s_{nw}}) - \mathbf{f}(\mathbf{B}_{s_{n'w'}}) \right\|_2^2, \quad (8)$$

$$p(\mathbf{B}_{s_{n'w'}} | \mathbf{B}_{s_{nw}}) = \frac{2e^{-\alpha d(\mathbf{B}_{s_{nw}}, \mathbf{B}_{s_{n'w'}})}}{1 + e^{-\alpha d(\mathbf{B}_{s_{nw}}, \mathbf{B}_{s_{n'w'}})}}, \quad (9)$$

where α is a hyper-parameter to control our probability profile function $p(d)$ shown in Figure 7: a Sigmoid reflected along its y axis. We apply the profile p to hallucinated outputs of $\mathbf{g}$ to obtain $\mathbf{g}'$. We show this strategy in Figure 8 and
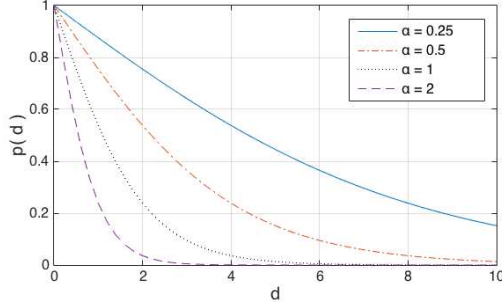
Figure 7: The probability profile p w.r.t. the dist. d and various α.
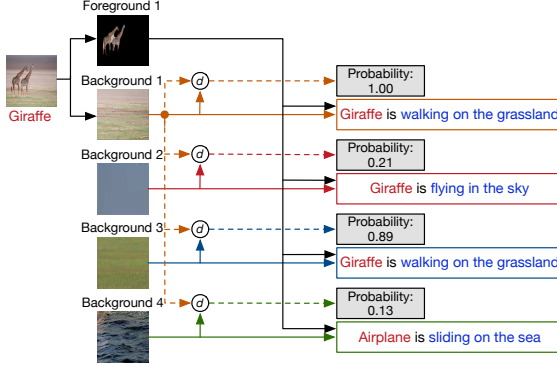

Figure 8: The inter-class hallucination strategy with the similarity prior. We assign likelihoods to generated datapoints based on the similarity of a background of a given image to other backgrounds.

we call it as *Soft Similarity Prior* (*SSP*):

$$\mathbf{g}'(\mathbf{F}_{s_{nw}}, \mathbf{B}_{s_{n'w'}}) = p(\mathbf{B}_{s_{n'w'}} | \mathbf{B}_{s_{nw}}) \, \mathbf{g}(\mathbf{f}(\mathbf{F}_{s_{nw}}), \mathbf{f}(\mathbf{B}_{s_{n'w'}})). \tag{10}$$

Also, 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 certain τ:

$$\mathbf{g}'(\mathbf{F}_{s_{nw}}, \mathbf{B}_{s_{n'w'}}) = \begin{cases} \mathbb{0}, & \text{if } p(\mathbf{B}_{s_{n'w'}} | \mathbf{B}_{s_{nw}}) \leq \tau, \\ \mathbf{g}(\mathbf{f}(\mathbf{F}_{s_{nw}}), \mathbf{f}(\mathbf{B}_{s_{n'w'}})), & \text{otherwise.} \end{cases} \tag{11}$$

We will show in our experiments that the use of priors significantly enhances the performance of the inter-class hallucination, especially for the 1-shot protocol, to which the intra-class hallucination is not applicable. We will show in Section 4 that both HSP and SSP improve the performance of few-shot learning; SSP being a consistent performer on all protocols. Firstly, we detail datasets and then we show the usefulness of our approach experimentally.

4. Experiments

Our network is evaluated in the few-shot learning scenario on the *mini*Imagenet [36] dataset and a recently proposed Open MIC dataset [16] which was used for few-shot learning by the SoSN approach [45]. Our implementation is based on PyTorch and models are trained on a Titan Xp

Table 1: Evaluations on the *mini*Imagenet dataset. See [34, 45] 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 the same. The astersik (*) denotes the 'sanity check' results on our proposed pipeline given disabled both saliency segmentation and hallucination (see the supp. material for details).

Model	Fine Tune	5-way Acc. 1-shot	5-shot
Matching Nets [36]	N	43.56 ± 0.84	55.31 ± 0.73
Meta Nets [27]	N	49.21 ± 0.96	-
Meta-Learn Nets [29]	N	43.44 ± 0.77	60.60 ± 0.71
Prototypical Net [33]	N	49.42 ± 0.78	68.20 ± 0.66
MAML [10]	Y	48.70 ± 1.84	63.11 ± 0.92
Relation Net [34]	N	51.36 ± 0.86	65.63 ± 0.72
SoSN [45]	N	52.96 ± 0.83	68.63 ± 0.68
SalNet w/o Sal. Seg. (*)	N	53.15 ± 0.87	68.87 ± 0.67
SalNet w/o Hal.	N	55.57 ± 0.86	70.35 ± 0.66
SalNet Intra. Hal.	N		71.78 ± 0.69
SalNet Inter. Hal.	N	57.45 ± 0.88	72.01 ± 0.67

GPU via the Adam solver. The architecture of our saliency-guided hallucination network is shown in Fig. 2 and 4. The results are compared with several state-of-the-art methods for one- and few-shot learning.

4.1. Datasets

Below, we describe our setup, datasets and evaluations.

***mini*Imagenet** [36] consists of 60000 RGB images from 100 classes. We follow the standard protocol [36] 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 with other methods. We also investigate larger sizes, *e.g.* 224×224, as our SalNet model can use richer spatial information from larger images to obtain high-rank auto-correlation matrices without a need to modify the similarity network to larger feature maps.

Open MIC is a recently proposed Open Museum Identification Challenge (Open MIC) dataset [16] which 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 was captured with wearable cameras. This makes Open MIC a perfect candidate for testing one-shot learning algorithms. Following the setup in SoSN [45], we combine (*shn+hon+clv*), (*clk+gls+scl*), (*sci+nat*) and (*shx+rlc*) splits into subproblems *p1, ..., p4*. We randomly select 4 out of 12 possible pairs in which subproblem x is used for training and y for testing (x→y).

Relation Net [34] and SoSN [45] are employed as baselines against which we compare our SalNet approach.

Table 2: Evaluations on the Open MIC dataset. 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.

Model	N-way	W-shot	$p1 \to p2$	$p2 \to p3$	$p3 \to p4$	$p4 \to p1$
Relation Net[34]	5	1	70.1	49.7	66.9	46.9
SoSN [45]	5	1	78.0	60.1	75.5	57.8
Intra.-Hal.	5	1	78.2	60.3	75.9	58.1
Inter.-Hal.	5	1	79.3	61.4	76.6	59.2
Relation Net[34]	5	2	75.6	55.2	72.3	56.0
SoSN [45]	5	2	84.6	68.1	82.7	66.8
Intra.-Hal.	5	2	85.7	69.2	84.1	67.5
Inter.-Hal.	5	2	86.4	70.0	84.3	67.8
Relation Net[34]	5	3	80.9	61.9	78.5	58.9
SoSN [45]	5	3	87.1	72.6	85.9	72.8
Intra.-Hal.	5	3	87.5	73.9	86.5	73.6
Inter.-Hal.	5	3	88.1	74.2	87.1	73.9
Relation Net[34]	10	1	54.4	35.3	53.1	35.5
SoSN [45]	10	1	67.2	46.2	63.9	46.6
Intra.-Hal.	10	1	67.6	46.7	64.3	47.0
Inter.-Hal.	10	1	68.3	47.5	65.4	48.4
Relation Net[34]	10	2	65.5	40.9	62.6	41.5
SoSN [45]	10	2	74.4	54.6	73.0	54.2
Intra.-Hal.	10	2	75.8	56.3	73.8	55.3
Inter.-Hal.	10	2	75.6	56.4	74.2	55.6
Relation Net[34]	10	3	69.0	45.7	67.5	46.3
SoSN [45]	10	3	78.0	56.3	77.5	58.6
Intra.-Hal.	10	3	79.2	58.3	78.3	59.1
Inter.-Hal.	10	3	79.3	58.5	78.6	59.9

Figure 9: The accuracy as a function of (*left*) W-shot (5-way) and (*right*) N-way (5-shot) numbers on *mini*Imagenet given different methods. Our models improve results over all baselines.

Figure 10: The accuracy on *mini*Imagenet as a function of (a) β of TriR from Eq. (7) (5-shot 5-way) and (b) α of SSP from Eq. (10) (1-shot 5-way).

4.2. Experimental setup

For the *mini*Imagenet dataset, we perform 1- to 10-shot experiments in 5-way scenario to demonstrate the improvements obtained with our SalNet on different number of W-shot images. For every training and testing episode, we randomly select 5 and 3 query samples per class. We average the final results over 600 episodes. The initial learning rate is set to $1e-3$. We train the model with 200000 episodes.

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 $p4 \to p1$. Firstly, 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 less than 3 images, we apply 5-way 1-shot to 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 over 1000 episodes for the final score. The initial learning rate is set to $1e-4$. The models are trained with 50000 episodes.

4.3. Results

For *mini*Imagenet dataset, Table 1 shows that our proposed SalNet outperforms all other state-of-the-art methods on standard 5-way 1- and 5-shot protocols. Compared with current state-of-the-art methods, our *SalNet Inter-class Hal.* model achieves $\sim$4.4% and $\sim$3.3% higher top-1 accuracy than SoSN on 1- and 5-shot protocols, respectively, while our *SalNet Intra-class Hal.* yields improvements of $\sim$2.5% and $\sim$3.1% accuracy over SoSN.

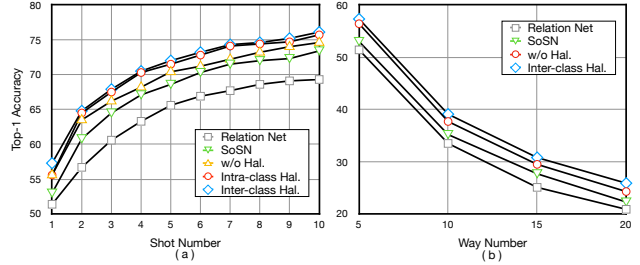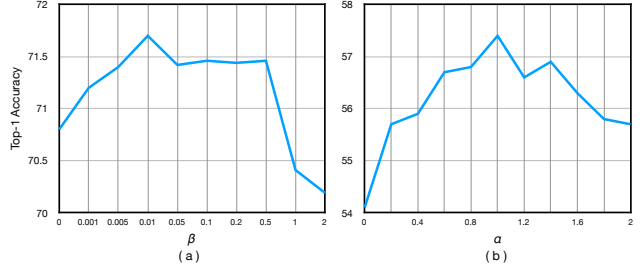Table 2 presents results on Open MIC. The improvements of *SalNet Inter-class Hal.* and *SalNet Intra-class Hal.* on this dataset are consistent with *mini*Imagenet. However, the improvements on some splits are small (*i.e.*, $\approx$ 1.1%) due to the difficulty of these splits *e.g.*, jewellery, fossils, complex non-local engine installations or semi-transparent exhibits captured with wearable cameras cannot be easily segmented out by saliency detectors.

Ablation study. The network proposed in our paper builds on the baseline framework [34]. However, we have added several non-trivial units/sub-networks to accomplish our goal of the datapoint hallucination in the feature space. Thus, we perform additional experiments to show that the achieved accuracy gains stem from our contributions. We also break down the accuracy w.r.t. various components.

Firstly, Table 1 shows that if the saliency segmentation and data hallucination are disabled in our pipeline (*SalNet w/o Sal. Seg.*), the performance on all protocols drops down to the baseline level of SoSN.

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

Figure 9 (a) shows the accuracy of our (*SalNet Intra-class Hal.*) model on *mini*Imagenet for 5-shot 5-way case
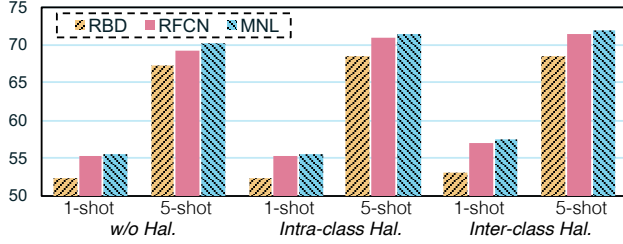
Figure 11: The results on RBD [47], RFCN [37] and MNL [46] saliency methods for *mini*Imagenet.

as a function of the parameter β of our regularization loss TriR. We observe that for $\beta = 0.01$ we gain $\sim 1\%$ accuracy over $\beta = 0$ (TriR disabled). Importantly, the gain remains stable over a large range $0.005 \leq \beta \leq 0.5$. Table 3 verifies further the usefulness of our TriR regularization in combination with the intra- and inter-class hallucination SalNet (*Intra.-Hal.+TriR*) and (*Inter.-Hal.+TriR*) with gains up to 1.6% and 1.5% accuracy on *mini*Imagenet. We conclude that TriR helps our end-to-end training by forcing encoder **g** to mimic teacher $\mathbf{g}^*$ for real foreground-background pairs ($\mathbf{g}^*$ is trained on such pairs only to act as a reliable superv.).

Figure 9 (b) shows the accuracy of our (*SalNet Inter-class Hal.*) model on *mini*Imagenet for 1-shot 5-way as a function of the Soft Similarity Prior (SSP). The maximum observed gain in accuracy is $\sim 3.3\%$. Table 3 further compares the hard and soft priors (*SalNet Inter-class Hal.+HSP*) and (*SalNet Inter-class Hal.+SSP*) with SSP outperforming HSP by up to $\sim 2.2\%$.

Lastly, Figure 11 compares several saliency methods in terms of few-shot learning accuracy. The complex saliency methods perform equally well. However, the use of the RBD approach [47] results in a significant performance loss due to its numerous failures *e.g.*, see Figure 3.

Saliency Map Dilation. As backgrounds extracted via a saliency detector contain 'cut out' silhouettes, they unintentionally carry some foreground information. Figure 12 suggests that if we apply the Gaussian blur and a threshold over the masks to eliminate the silhouette shapes, we can prevent mixing the primary foreground with a foreground corresponding to silhouettes. Table 4 shows that pairing each foreground with background images whose silhouettes
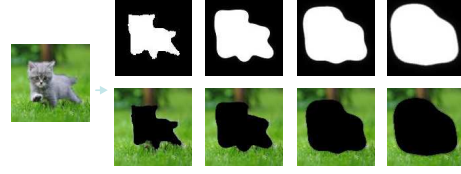

Figure 12: Gradual dilation of the foreground mask.

Model	5-way 1-shot	5-way 5-shot
Intra-class Hal.	55.57 ± 0.86	71.78 ± 0.69
Intra-class Hal.+Dilation	$\mathbf{56.67} \pm 0.85$	$\mathbf{72.15} \pm 0.68$

Table 4: Results for dilating contours of silhouettes.

were removed by dilating according to two different radii (*Dilation*) leads to further improvements due to doubling of possible within-class combinations for (*Intra-class Hal.*).

5. Conclusions

In this paper, we have presented two novel light-weight data hallucination strategies for few-shot learning. in contrast to other costly hallucination methods based on GANs, we have leveraged the readily available saliency network to obtain foreground-background pairs on which we trained our SalNet network in end-to-end manner. To cope with noises of saliency maps, we have proposed a Real Representation Regularization (TriR) which regularizes our network with viable solutions for real foreground-background pairs. To alleviate performance loss caused by implausible foreground-background hypotheses, we have proposed a similarity-based priors effectively reduced the influence of incorrect hypotheses. For future work, we will investigate a self-supervised attention module for similarity perception and study relaxations of saliency segmentation methods.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based clas-

Table 3: 5-way evaluations on the *mini*Imagenet dataset for different N-shot numbers. Refer to [34, 45] for details of baselines.

	5-way Accuracy									
W-shot	1	2	3	4	5	6	7	8	9	10
Relation Net [34]	51.4 ± 0.7	56.7 ± 0.8	60.6 ± 0.8	63.3 ± 0.7	65.6 ± 0.7	66.9 ± 0.7	67.7 ± 0.7	68.6 ± 0.6	69.1 ± 0.6	69.3 ± 0.6
SoSN [45]	53.0 ± 0.8	60.8 ± 0.8	64.5 ± 0.8	67.1 ± 0.7	68.6 ± 0.7	70.3 ± 0.7	71.5 ± 0.6	72.0 ± 0.6	72.3 ± 0.6	73.4 ± 0.6
w/o Sal. Seg.	53.1 ± 0.9	60.9 ± 0.8	64.7 ± 0.8	67.3 ± 0.7	68.9 ± 0.7	70.6 ± 0.7	71.7 ± 0.6	72.1 ± 0.6	72.6 ± 0.6	73.6 ± 0.6
w/o Hal.	55.6 ± 0.9	63.5 ± 0.8	66.2 ± 0.8	68.2 ± 0.7	70.4 ± 0.7	71.2 ± 0.7	72.2 ± 0.7	73.2 ± 0.6	74.0 ± 0.6	74.6 ± 0.6
Intra.-Hal.	55.6 ± 0.9	63.1 ± 0.8	65.9 ± 0.7	68.7 ± 0.7	70.8 ± 0.7	71.8 ± 0.7	73.6 ± 0.6	73.8 ± 0.6	74.1 ± 0.6	75.2 ± 0.6
Intra.-Hal.+TriR	55.6 ± 0.9	64.5 ± 0.8	67.5 ± 0.7	70.3 ± 0.7	71.8 ± 0.7	72.8 ± 0.7	74.1 ± 0.6	74.4 ± 0.6	74.7 ± 0.6	75.7 ± 0.6
Inter.-Hal.	53.7 ± 0.9	58.9 ± 0.8	62.4 ± 0.8	65.2 ± 0.7	67.7 ± 0.7	68.5 ± 0.7	69.6 ± 0.7	69.9 ± 0.6	70.6 ± 0.6	71.1 ± 0.6
Inter.-Hal.+TriR	54.1 ± 0.9	60.1 ± 0.8	63.4 ± 0.7	65.8 ± 0.7	67.9 ± 0.7	69.6 ± 0.7	70.5 ± 0.6	71.0 ± 0.7	72.1 ± 0.6	72.5 ± 0.7
Inter.-Hal.+TriR+HSP	56.4 ± 0.9	63.0 ± 0.8	67.3 ± 0.8	69.2 ± 0.7	71.0 ± 0.6	71.8 ± 0.7	72.1 ± 0.6	73.0 ± 0.6	74.2 ± 0.6	75.4 ± 0.6
Inter.-Hal.+TriR+SSP	57.5 ± 0.9	64.8 ± 0.8	67.9 ± 0.8	70.5 ± 0.7	72.0 ± 0.7	73.2 ± 0.7	74.3 ± 0.6	74.6 ± 0.6	75.2 ± 0.6	76.1 ± 0.6

sification. In *CVPR*, pages 819–826, 2013. 2

[2] Evgeniy Bart and Shimon Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, pages 672–679, 2005. 2

[3] A. Borji, M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *TIP*, 24(12):5706–5722, 2015. 3, 4

[4] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A Theoretical Analysis of Feature Pooling in Vision Algorithms. In *ICML*, 2010. 3

[5] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic Segmentation with Second-Order Pooling. In *ECCV*, 2012. 3

[6] M. Cheng, G. Zhang, N.J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011. 4

[7] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009. 2

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611, 2006. 1, 2

[9] Michael Fink. Object classification from a single example utilizing class relevance metrics. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *NIPS*, pages 449–456, 2005. 2

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 2, 6

[11] Mehrtash Harandi, Mathieu Salzmann, and Richard Hartley. Joint dimensionality reduction and metric learning: A geometric take. In *ICML*, page 14041413, 2017. 1

[12] Bharath Hariharan and Ross B Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, pages 3037–3046, 2017. 1, 2

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 3203–3212, July 2017. 2

[14] H. Jégou, M. Douze, and C. Schmid. On the Burstiness of Visual Elements. In *CVPR*, pages 1169–1176. IEEE, 2009. 3

[15] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015. 2

[16] Piotr Koniusz, Yusuf Tas, Hongguang Zhang, Mehrtash Harandi, Fatih Porikli, and Rui Zhang. Museum exhibit identification challenge for the supervised domain adaptation. *CoRR:1802.01093*, 2018. 6

[17] P. Koniusz, F. Yan, P. Gosselin, and K. Mikolajczyk. Higher-order Occurrence Pooling on Mid- and Low-level Features: Visual Concept Detection. *Technical Report*, 2013. 3

[18] Piotr Koniusz, Fei Yan, Philippe-Henri Gosselin, and Krystian Mikolajczyk. Higher-order occurrence pooling for bags-of-words: Visual concept detection. *PAMI*, 39(2):313–326, 2017. 3

[19] P. Koniusz, F. Yan, and K. Mikolajczyk. Comparison of Mid-Level Feature Coding Approaches And Pooling Strategies in Visual Concept Detection. *CVIU*, 2012. 3

[20] Piotr Koniusz, Hongguang Zhang, and Fatih Porikli. A deeper look at power normalizations. In *CVPR*, pages 5774–5783, 2018. 3

[21] Martin Kstinger, Martin Hirzer, Paul Wohlhart, Peter M.

Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012. 1

[22] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. *CogSci*, 2011. 2

[23] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008. 2

[24] Fei Fei Li, Rufin VanRullen, Christof Koch, and Pietro Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14):9596–9601, 2002. 2

[25] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *CVPR*, pages 1–8, 2007. 1

[26] E. G. Miller, N. E. Matsakis, and P. A. Viola. Learning from one example through shared densities on transforms. In *CVPR*, volume 1, pages 464–471, 2000. 2

[27] Tsendsuren Munkhdalai and Hong Yu. Meta networks. *CoRR:1703.00837*, 2017. 6

[28] Jagath Chandana Rajapakse and Lipo Wang. *Neural Information Processing: Research and Development*. Springer-Verlag Berlin and Heidelberg GmbH & Co. KG, 2004. 2

[29] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 6

[30] A. Romero, M. Y. Terán, M. Gouiffès, and L. Lacassagne. Enhanced local binary covariance matrices (ELBCM) for texture analysis and object tracking. *MIRAGE*, pages 10:1–10:8, 2013. 3

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 2

[32] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, pages 4967–4976, 2017. 1

[33] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087, 2017. 1, 2, 6

[34] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. *CoRR:1711.06025*, 2017. 1, 2, 3, 4, 6, 7, 8

[35] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 2006. 3

[36] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016. 1, 2, 6

[37] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, pages 825–841, 2016. 2, 3, 4, 8

[38] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. *CoRR:1801.05401*, 2018. 1, 2

[39] Kilian Q Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *NIPS*, pages 1473–1480, 2006. 1

[40] R. S. Woodworth and E. L. Thorndike. The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review (I)*, 8(3):247–261, 1901. 2

[41] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. *CoRR:1712.00981*, 2017. 2

[42] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[43] Hongguang Zhang and Piotr Koniusz. Model selection for generalized zero-shot learning. In *ECCV*, pages 198–204. Springer, 2018. 2

[44] Hongguang Zhang and Piotr Koniusz. Zero-shot kernel learning. In *CVPR*, pages 7670–7679, 2018. 2

[45] Hongguang Zhang and Piotr Koniusz. Power normalizing second-order similarity network for few-shot learning. In *WACV*, pages 1185–1193. IEEE, 2019. 2, 3, 4, 6, 7, 8

[46] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *CVPR*, June 2018. 1, 3, 4, 8

[47] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *CVPR*, pages 2814–2821, 2014. 2, 3, 4, 8