Hallucinating Saliency Maps for Fine-Grained Image Classification for Limited Data Domains

Carola Figueroa-Flores∗†, Bogdan Raducanu∗, David Berga∗, Joost van de Weijer∗
∗Computer Vision Center
Edifici “O” - Campus UAB
8193 Bellaterra (Barcelona), Spain
†Department of Computer Science and Information Technology
Universidad del Bío Bío, Chile

Abstract—Most of the saliency methods are evaluated on their ability to generate saliency maps, and not on their functionality in a complete vision pipeline, like for instance, image classification. In the current paper, we propose an approach which does not require explicit saliency maps to improve image classification, but they are learned implicitly, during the training of an end-to-end image classification task. We show that our approach obtains similar results as the case when the saliency maps are provided explicitly. Combining RGB data with saliency maps represents a significant advantage for object recognition, especially for the case when training data is limited. We validate our method on several datasets for fine-grained classification tasks (Flowers, Birds and Cars). In addition, we show that our saliency estimation method, which is trained without any saliency groundtruth data, obtains competitive results on real image saliency benchmark (Toronto), and outperforms deep saliency models with synthetic images (SID4VAM).

Index Terms—Fine-grained image classification, Saliency detection, Convolutional neuronal networks

I. INTRODUCTION

Fine-grained image recognition has as objective to recognize many subcategories of a super-category. Examples of well-known fine-grained datasets are Flowers [1], Cars [2] and Birds [3]. The challenge of fine-grained image recognition is that the differences between classes are often very subtle, and only the detection of small highly localized features will correctly lead to the recognition of the specific bird or flower species. An additional challenge of fine-grained image recognition is the difficulty of data collection. The labelling of these datasets requires experts and subcategories can be very rare which further complicates the collection of data. Therefore, the ability to train high-quality image classification systems from few data is an important research topic in fine-grained object recognition.

Most of the state-of-the-art general object classification approaches [4], [5] have difficulties in the fine-grained recognition task, which is more challenging due to the fact that basic-level categories (e.g. different bird species or flowers) share similar shape and visual appearance. Early works have focused on localization and classification of semantic parts using either explicit annotation [6]–[8] or weakly labeling [9], [10]. The main disadvantage of these approach was that they required two different ‘pipelines’, for detection and classification, which made more complicated the joint optimization of the two subsystems. Therefore, more recent approaches are proposing end-to-end strategies with the focus on improving the feature representation from intermediate layers in a CNN through higher order statistics modeling [11], [12].

One recent approach which obtained good fine-grained recognition results, especially with only few labelled data is proposed by Figueroa et al. [13]. The main idea is that a saliency image can be used to modulate the recognition branch of a fine-grained recognition network. We will refer to this technique as saliency-modulated image classification (SMIC). This is especially beneficial when only few labelled data is available. The gradients which are backpropagated are concentrated on the regions which have high saliency. This prevents backpropagation of gradients of uninformative background parts of the image which could lead to overfitting to irrelevant details. A major drawback of this approach is that it requires an explicit saliency algorithm which needs to be trained on a saliency dataset.

In order to overcome the lack of sufficient data for a given modality, a common strategy is to introduce a ‘hallucination’ mechanism which emulates the effect of genuine data. For instance, in [14], they use this ‘hallucination’ strategy for RGB-D object detection. A hallucination network is trained to learn complementary RGB image representation which is taught to mimic convolutional mid-level features from a depth network. At test time, images are processed jointly through the RGB and hallucination networks, demonstrating an improvement in detection performance. This strategy has been adopted also for the case of few-shot learning [15]–[17]. In this case, the hallucination network has been used to produce additional training sample used to train jointly with the original network (also called a neta-learner).

In this paper, we address the major drawback of SMIC, by implementing a hallucination mechanism in order to remove the requirement for providing saliency images for training obtained using one of the existing algorithms [18]. In other words, we show that the explicit saliency branch which requires training on a saliency image dataset, can be replaced with a branch which is trained end-to-end for the task of image classification (for which no saliency dataset is required). We replace the saliency image with the input RGB image (see Fig. [1]). We then pre-train this network for the task of image
classification using a subset from ImageNet validation dataset. During this process, the saliency branch will learn to identify which regions are more discriminative. In a second phase, we initialize the weights of the saliency branch with these pre-trained weights. We then train the system end-to-end on the fine-grained dataset using only the RGB images. Results show that the saliency branch improves recognition significantly. In addition we show, through qualitative and quantitative results, that the saliency branch is actually generating saliency maps, which obtain competitive results on saliency datasets.

We briefly summarize below our main contributions:

- we propose an approach which hallucinates saliency maps that are fused together with the RGB modality via a modulation process,
- our saliency does not require any saliency maps for training (like previous work [13], [19]) but instead is trained indirectly in an end-to-end fashion by training the network for image classification,
- our method improves the classification accuracy on three fine-grained datasets when compared to the baseline. We obtain similar results as previous work [13] without the need of a saliency network trained on groundtruth saliency data,
- the saliency maps which we obtain without using any saliency groundtruth data, obtain competitive results on real images, and outperforms deep saliency estimation methods in synthetic images.

The paper is organized as follows. Section II is devoted to review the related work in fine-grained image classification and saliency estimation. Section III presents our approach. We report our experimental results in Section IV. Finally, Section V contains our conclusions.

II. RELATED WORK

A. Fine-grained image classification

A first group of approaches on fine-grained recognition operate on a two-stage pipeline: first detecting some object parts and then categorizing the objects using this information.

The work of Huang et al. [20] first localizes a set of part keypoints, and then simultaneously processes part and object information to obtain highly descriptive representations. Mask-CNN [21] also aggregates descriptors for parts and objects simultaneously, but using pixel-level masks instead of keypoints. The main drawback of these models is the need of human annotation for the semantic parts in terms of keypoints or bounding boxes. To partially alleviate this tedious task of annotation, Xiao et al. [22] propose an weakly-supervised approach based on the combination of three types of attention in order to guide the search for object parts in terms of ‘what’ and ‘where’. A further improvement has been reported in Zhang et al. [23], where the authors propose and approach free of any object / part annotation. Their method explores a unified framework based on two steps of deep filter response picking.

A second group of approaches merges these two stages into an end-to-end learning framework which optimize simultaneously both part localization and fine-grained classification. This is achieved by first finding the corresponding parts and then comparing their appearance [12]. In [24], their framework first performs unsupervised part candidates discovery and global object discovery which are subsequently fed into a two-stream CNN in order to model jointly both the local and global features. In [9], they propose a novel part learning approach by a multi-attention convolutional neural network (MA-CNN) without bounding box/part annotations. MA-CNN jointly learns part proposals (defined as multiple attention areas with strong discrimination ability) and the feature representations on each part. Some other approaches which are based on attention mechanisms are presented in [25] and [26].

In another direction, some end-to-end frameworks aim to enhance the intermediate representation learning capability of a CNN by encoding higher-order statistics. For instance in [27] they capture the second-order information by taking the outer-product over the network output and itself. Other approaches focuses on reducing the high feature dimensionality [28] or extracting higher order information with kernelized modules [11]. In [12], they learn a bank of convolutional filters that capture class-specific discriminative patches without extra part or bounding box annotations. The advantage of this approach is that the network focuses on classification only and avoids the trade-off between recognition and localization.

Regardless, most fine-grained approaches use the object ground-truth bounding box at test time, achieving a significantly lower performance when this information is not available. Moreover, automatically discovering discriminative parts might require large amounts of training images. Our approach is more general, as it only requires image level annotations at training time and could easily generalize to other recognition tasks.

B. Saliency estimation

Initial efforts in modelling saliency involved multi-scale representations of color, orientation and intensity contrast. These were often biologically inspired such as the well-known work by Itti et al. [29]. From that model, a myriad of models were based on handcrafting these features in order to obtain an accurate saliency map [30], [31], either maximizing [32] or learning statistics of natural images [33], [34]. Saliency research was propelled further by the availability of large data sets which enabled the use machine learning algorithms [35], mainly pretrained on existing human fixation data.

The question of whether saliency is important for object recognition and object tracking has been raised in [36]. Latest methods [35] take advantage of end-to-end convolutional architectures by finetuning over fixation prediction [37]–[39]. But the main goal of these works was to estimate a saliency map, not how saliency could contribute to object recognition. In this paper instead, we propose an approach which does not require explicit saliency maps to improve image classification,
Step I: Training on Imagenet

Step II: Finetuning on a fine-grained dataset

**Finetuning:**
- Finetuned layers
- Frozen layers

**Initialize weights:**
- Random initialization
- Pretrained network

**Approach A**

**Approach B**

**Fig. 1:** Overview of our method. We process an RGB input image through two branches: one branch extracts the RGB features and the other one is used to learn saliency maps. The resulting features are merged via a modulation layer, which continues with a few more convolutional layers and a classification layer. The network is trained in two steps.

but they are learned implicitly, during the training of an end-to-end image classification task. We show that our approach obtains similar results as the case when the saliency maps are provided explicitly.

### III. Proposed Method

Several works have shown that having the saliency map of an image can be helpful for object recognition and fine-grained recognition in particular [13], [19]. The idea is twofold, the saliency map can help focus the attention on the relevant parts of the image to improve the recognition. Also, it can help guide the training, by focusing the backpropagation to the relevant image regions. In earlier work, we have shown that saliency-modulated image classification (SMIC) is especially efficient for training on datasets with few labeled data [13]. The main drawback of these methods is that they require a trained saliency method. Here we show that this restriction can be removed and that we can hallucinate the saliency image from the RGB image. By training the network for image classification on the imageNet dataset we can obtain the saliency branch without human groundtruth images.

**A. Overview of the Method**

The overview of our proposed network architecture is illustrated in Figure 1. Our network consists of two branches: one to extract the features from an RGB image, and the other one (saliency branch) to generate the saliency map from the same RGB image. Both branches are combined using a modulation layer (represented by the \( \times \) symbol) and are then processed by several shared layers of the joint branch which finally ends up with a classification layer. The RGB branch followed by the joint branch resembles a standard image classification network. The novelty of our architecture is the introduction of the saliency branch, which transforms the generated saliency image into a modulation image. This modulation image is used to modulate the characteristics of the RGB branch, putting more emphasis on those characteristics that are considered important for the fine-grained recognition task. In the following sections we provide the details of the network architecture, the operation of the modulation layer, and finally, how our saliency map is generated. We explain our model using AlexNet [5] as the base classification network, but the theory could be extended to other convolutional neural network architectures. For instance, in the experimental results section, we also consider the ResNet-152 architecture [40].

**B. Hallucination of saliency maps from RGB images**

Our experiments aim to enforce top-down prominence driven by a specific ranking task, rather than bottom-up prominence. In other words, our visual attention maps focus on the location of the characteristics necessary to identify the target classes, ignoring anything else that may be irrelevant to the classification task. Therefore, given an input RGB image, our saliency branch should be able to produce a map of the most salient image locations useful for classification purposes.

To achieve that, we apply a CNN-based saliency detector consisting of four convolutional layers (based on the AlexNet architecture). The output from the last convolutional layer, i.e. one with 384 dimensional feature maps with a spatial resolution of 13 13 (for a 227 227 RGB input image), is further processed using a 1 1 convolution and then a function of activation ReLU. This is to calculate the saliency score for
each "pixel" in the feature maps of the previous layer, and to produce a single channel map. Finally, to generate the input for the subsequent classification network, the 13 13 saliency maps are upsampled to 27 27 (which is the default input size of the next classification module) through bilinear interpolation. We justify the size of the output maps by claiming that saliency is a primitive mechanism, used by humans to direct attention to objects of interest, which is evoked by coarse visual stimuli. Therefore, our experiments (see section IV) show that 13 13 feature maps can encode the information needed to detect salient areas and drive a classifier with them.

C. Fusion of RGB and Saliency Branches

Consider an input image \( I(x, y, z) \), where \( z = \{1, 2, 3\} \) indicate the three color channels of the image. Also consider a saliency map \( s(x, y) \). In previous work, a network \( h(I, s) \) was trained which performed image classification based on the input image \( I \) and the saliency map \( s \). Here, we replace the saliency map (which was generated by a saliency algorithm) by a hallucinated saliency map \( h(I, \hat{s}(I)) \). The hallucinated saliency map \( \hat{s} \) is trained end-to-end and estimated from the same input image \( I \).

The combination of the hallucinated saliency map \( \hat{s} \), which is the output of the saliency branch, and the RGB branch is done with modulation. Consider the output of the \( i^{th} \) layer of the network, \( \hat{l}_i \), with dimension \( w_i \times h_i \times z_i \). Then we define the modulation as
\[
\hat{l}_i(x, y, z) = \hat{l}_i(x, y, z) \cdot \hat{s}(x, y),
\]
resulting in the saliency-modulated layer \( \hat{l}_i \). Note that a single hallucinated saliency map is used to modulate all \( i \) feature maps of \( \hat{l} \).

In addition to the formula in Eq. (1) we also introduce a skip connection from the RGB branch to the beginning of the joint branch, defined as
\[
\hat{l}_i(x, y, z) = l_i(x, y, z) \cdot (\hat{s}(x, y) + 1).
\]
This skip connection is depicted in Fig. 1 (+ symbol). It prevents the modulation layer from completely ignoring the features from the RGB branch. This is inspired by our previous work [13] that found this approach beneficial when using attention for network compression.

We train our architecture in an end-to-end manner. The backpropagated gradient for the modulation layer into the image classification branch is equal defined as:
\[
\frac{\partial L}{\partial \hat{l}} = \frac{\partial L}{\partial \hat{l}} \cdot (\hat{s}(x, y) + 1),
\]
where \( L \) is the loss function of the network. We can see that the saliency map modulates both the forward pass (see Eq. (2)) as well as the backward pass in the same manner; in both cases putting more weight on the features that are on locations with high saliency, and putting less weight on the irrelevant features. We show in the experiments that this helps the network train more efficiently, also on datasets with only few labeled samples. The modulation prevents the network from overfitting to the background.

D. Training on Imagenet and fine-tuning on a target dataset

As can be seen in Fig. 1 the training of our approach is divided into two steps: first, training on Imagenet and second, fine-tuning on a target dataset.

Step 1: Training of saliency branch on Imagenet.
As explained above, the aim of the saliency branch is to hallucinate (generate) a saliency map directly from an RGB input image. This way, we the do not need any pre-computed saliency maps, as was the case in [13].

This network is constructed by initializing the RGB branch with pretrained weights from Imagenet. The weights of the saliency branch are initialized randomly using the Xavier method (see Fig. 1 left image). The network is then trained selectively, using the ImageNet validation set: we allow to train only the layers corresponding to the saliency branch (depicted by the surrounding dotted line) and freeze all the remaining layers (depicted through the continuous line boxes).

During the training the saliency branch learns to focus on those regions of the image which are important for the classification of 1000 classes. As such it is learning to estimate the salient regions in the images as can be seen from Fig. 2 and Fig. 3.

Step 2: Fine-tuning on a target dataset. In this step, we initialize the RGB branch with the weights pre-trained from Imagenet and the saliency branch with the corresponding pre-trained weights from Step 1. The weights of the top classification layer are initialized randomly, using the Xavier method. Then, this network is then further fine-tuned on a target dataset, selectively. We distinguish two cases:

- **Approach A**: We freeze the layers of the saliency branch and we allow all the other layers in the network to be trained. This process is depicted by the continuous line surrounding the saliency branch and the dotted line for the rest (see the Fig. 1 middle image).
- **Approach B**: We allow all layers to be trained. Since we consider training on datasets with only few labels this could results in overfitting, since it requires all the weights of the saliency branch to be learned (see the Fig. 1 right image).

In the experiments we evaluate both approaches to training the network.

IV. EXPERIMENTS

A. Experimental Setup

**Datasets.** To evaluate our approach, we used three standard datasets used for fine-grained image classification:

- **Flowers**: Oxford Flower 102 dataset [1] has 8,189 images divided in 102 classes.
- **Birds**: CUB200 has 11,788 images of 200 different bird species [3].
- **Cars**: the CARS-196 dataset in [2] contains 16,185 images of 196 car classes.

Additionally, we also evaluate our saliency estimation results on two datasets:
| #train images | 1  | 2  | 3  | 5  | 10 | 15 | 20 | 25 | 30 | \( K \) | AVG |
|---------------|----|----|----|----|----|----|----|----|----|-------|-----|
| **Flowers**   |    |    |    |    |    |    |    |    |    |       |     |
| Baseline-RGB  | 31.8 | 45.8 | 53.1 | 63.6 | 72.4 | 76.9 | 81.2 | 85.1 | 87.2 | 87.8 | 68.3 |
| Baseline-RGB + scratch SAL | 34.3 | 48.9 | 54.3 | 65.9 | 73.1 | 77.4 | 82.3 | 85.9 | 88.9 | 91.1 | 70.0 |
| SMIC [13] *  | 37.6 | **51.9** | 57.1 | 68.5 | 75.2 | **79.7** | **84.9** | 88.2 | 91.2 | 92.3 | **72.7** |
| Approach A    | 36.9 | 51.3 | 56.9 | 67.8 | 74.9 | 78.4 | 82.9 | 88.1 | 90.9 | 92.0 | 72.0 |
| Approach B    | 37.3 | 51.7 | **57.2** | **68.7** | **75.6** | 78.7 | 83.8 | **88.4** | **91.7** | **92.5** | 72.6 |
| **Cars**      |    |    |    |    |    |    |    |    |    |       |     |
| Baseline-RGB  | 4.1 | 7.8 | 11.7 | 17.3 | 25.5 | 31.1 | 38.5 | 42.2 | 47.2 | 60.0 | 28.5 |
| Baseline-RGB + scratch SAL | 5.9 | 10.7 | 14.4 | 19.1 | 27.4 | 32.9 | 38.5 | 44.0 | 48.7 | 61.5 | 30.3 |
| SMIC [13] *  | 9.3 | 14.0 | 18.0 | 22.8 | **30.0** | **34.7** | **40.4** | **46.0** | **50.0** | **61.4** | 32.7 |
| Approach A    | 9.3 | 14.3 | 17.4 | 22.3 | 28.4 | 35.3 | 39.7 | 45.7 | 50.1 | 61.9 | 32.4 |
| Approach B    | **9.8** | **15.1** | **18.4** | **22.9** | 28.8 | **35.1** | 39.9 | 45.8 | 49.7 | **62.9** | **32.8** |
| **Birds**     |    |    |    |    |    |    |    |    |    |       |     |
| Baseline-RGB  | 9.1 | 13.6 | 19.4 | 27.7 | 37.8 | 44.3 | 48.0 | 50.0 | 54.2 | 57.0 | 34.8 |
| Baseline-RGB + scratch SAL | 10.4 | 14.9 | 20.3 | 28.3 | 38.6 | 43.9 | 46.9 | 48.4 | 50.7 | 55.7 | 35.8 |
| SMIC [13] *  | **13.1** | **18.9** | **22.2** | **30.2** | 38.7 | 44.3 | 48.0 | **50.0** | **54.2** | **57.0** | **37.7** |
| Approach A    | 11.8 | 18.3 | 22.1 | 29.3 | 39.1 | **44.4** | 47.8 | 49.7 | 53.1 | 56.5 | 37.2 |
| Approach B    | 12.9 | 18.7 | **22.7** | **29.7** | **39.7** | **44.1** | **48.2** | **49.9** | **53.9** | **57.7** | **37.7** |

**TABLE I:** Classification accuracy for Flowers, Cars, and Birds dataset. Results are provided for varying number of training images, from 1 until 30; \( K \) refers to using the number of training images used in the official dataset split. The rightmost column shows the average. The * indicates that the method requires an explicit saliency method. Our method (Approach B) obtains similar results as SMIC but without the need of a pretrained saliency network trained on a saliency dataset.

- **Toronto:** Human fixation on real images [32], containing a total of 120 images.
- **SID4VAM:** Human fixations on synthetic images [41] with available fixation data and clear pop-out (salient) objects on a total of 230 images.

**Networks architectures.** We evaluate our approach using two network architectures: Alexnet [5] and Resnet-152 [40]. In both cases, the weights were pretrained on Imagenet and then finetuned on each of the datasets mentioned above. The networks were trained for 70 epochs with a learning rate of 0.0001 and a weight decay of 0.005. The top classification layer was initialized from scratch using Xavier method [42]. The saliency branch consists of four convolutional layers. The fusion took place after the second convolution layer in the RGB branch for Alexnet and after the forth residual block for Resnet-152. We demonstrated in [13] that using these settings we achieve the optimal performance.

**Evaluation protocol.** To validate our approach, we follow the same protocol as in our previous work [13]. For the image classification task, we train each model with subsets of \( k \) training images for \( k \in \{1, 2, 3, 5, 10, 15, 20, 25, 30, K\} \), where \( k \) is the total number of training images for the class. We keep 5 images per class for validation and 5 images per class for test. We report the performance in terms of accuracy, i.e. percentage of correctly classified samples. We show the results as an average over several runs.

**B. Fine-grained Image Classification Results**

**Evaluation on scarce data domain:** As described in section III, we consider two alternative ways to train the saliency branch on the target dataset: keeping the saliency branch fixed (Approach A) or allowing it to finetune (Approach B). In this section, we compare these two approaches with respect to the Baseline-RGB and Baseline-RGB + scratch SAL (where Saliency branch is initialized from scratch without pretraining on Imagenet) and our previous work, called SMIC [13]. We do not compare to other fine-grained methods here, because they do not report results when only considering few labeled images. The experiments are performed on **Flowers, Cars** and **Birds** datasets and can be seen in Table II. The average improvement of accuracy of our **Approach A** and **B** with respect the Baseline-RGB is 3.7% and 4.3%, respectively for the **Flowers** dataset; 3.9% and 4.3%, respectively for the **Cars** dataset; and 2.4% and 2.9%, respectively for the **Birds** dataset. Our **Approach B** is especially advantageous, if we compare it with our previous SMIC approach [13], where we needed an additional algorithm to generate the salience map. It is therefore advantageous to also finetune the saliency branch on the target data even when we only have a few labeled images per class.

**Comparison with other state-of-the-art approaches:** In the past experiments, we used a custom data split consisting of a
fixed subset of $k$ training images. To compare our approach with other state-of-the-art methods, we followed the standard data split for training and evaluation of each dataset. Note that our main purpose is to evaluate on domains with little labeled data but we have included this results for comparison. This results are presented in Table [I]. For the current comparison, we use our both approaches with ResNet-152 as base network, which is equivalent to the network architecture used by the most of the recent works. It can be appreciated that both our methods show similar performance with other fine-grained specialized approaches which often use more complex architectures including part-localization modules.

C. Evaluation benchmark of saliency hallucination

Here we compare the saliency estimation which is obtained after only performing Step I in Fig. [I] with existing saliency methods. This saliency estimation is trained without access to any groundtruth saliency data, and is obtained while training the image classification task on Imagenet.

Saliency prediction metrics assign a score depending on how well the predicted saliency map is able to match with locations of human fixations (see definitions in Borji et al. [53] and Bylinskii et al. [54]). We selected the Area Under ROC (AUC), Kullback-Leibler divergence (KL), similarity (SIM), shuffled AUC (sAUC) and Information Gain (IG) metrics considering its consistency of predictions of human fixation maps as well as towards to the center bias. We compare scores with classical saliency models, both with handcrafted low-level features (i.e. IKN [29], AIM [32], SDLF [33] and GBVS [34]) and state-of-the-art deep saliency models (i.e. DeepGazeII [37], SAM-ResNet [39], SalGAN [38]) mainly pretrained on human fixations. The results are surprising, our method which has not been trained on any groundtruth data, obtains competitive results. For the case of Toronto (Table [III]) Left the best models are GBVS and OpenSALICON, followed by our model that scores in the top-3 of KL and SAM-ResNet that scores slightly higher in InfoGain metric. For the case of SID4VAM (Table [III]) Right our approach gets best scores for most metrics compared to other deep saliency models, being mainly the top-2 acquiring similar scores to GBVS in most metrics (outperforming it in AUC measures).

These saliency prediction results show that our model has robust metric scores on both real images and synthetic images for saliency prediction. Again, we would like to stress that our model is not trained on fixation prediction datasets and does not add a center Gaussian to leverage some metrics due to the center bias. Our model performs best on detecting pop-out effects (from visual attention theories [29]), whilst performing similarly for real image datasets (Fig. [3]). Some deep saliency models use several mechanisms to leverage (or/and train) performance for improving saliency metric scores, such as smoothing/thresholding (see Fig. [2] rows 4-5) or a center gaussian (see Fig. [3] row 5). We also consider that some of these models are already finetuned for synthetic images (e.g. SAM-ResNet [39]). Our Approach (that has not been trained in these type of datasets) has shown to be robust on these two distinct scenarios/domains.

V. CONCLUSIONS

In this work, we proposed a method to improve fine-grained image classification by means of saliency maps. Our method does not require explicit saliency maps, but they are learned implicitly during the training of an end-to-end deep convolutional network. We validated our method on several datasets for fine-grained classification tasks (Flowers, Birds and Cars). We showed that our approach obtains similar results as our previous work which required explicit saliency maps. We showed that combining RGB data with saliency maps represents a significant advantage for object recognition, especially for the case when training data is limited. In addition, we showed that our saliency estimation method, which is trained without any saliency groundtruth data, obtains competitive results on a real image saliency benchmark, and obtains similar to state-of-the-art results on a synthetic saliency benchmark.

TABLE III: Comparison our saliency output with on standard benchmark methods over synthetic image datasets (Left: Toronto, Right: SID4VAM) for saliency prediction. (Top) Baseline low-level saliency models. (Bottom) State-of-the-art deep saliency models. Best score for each metric is defined as bold and TOP-3 scores are underlined.

| Method | AUC | KL ↓ | SIM | sAUC | InfoGain |
|--------|-----|------|-----|------|----------|
| IKN [29]| 0.782| 1.249| 0.366| 0.650| -0.024 |
| AIM [32]| 0.716| 1.672| 0.314| 0.663| -0.580 |
| SDLF [33]| 0.703| 1.518| 0.304| 0.664| -0.398 |
| GBVS [34]| 0.803| 1.168| 0.397| 0.632| 0.077 |
| DeepGazeII [37]| 0.838| 1.367| 0.325| 0.676| -0.200 |
| SAM-ResNet [39]| 0.725| 2.420| 0.516| 0.666| -1.555 |
| OpenSALICON [51,52]| 0.771| 1.113| 0.428| 0.716| 0.232 |
| SalGAN [38]| 0.818| 1.272| 0.435| 0.715| 0.392 |

| Our Approach (Step I) | 0.731| 1.513| 0.394| 0.589| -0.418 |
| GroundTruth (Humans) | 0.954| 0.000| 1.000| 0.902| 2.425 |

| Method | AUC | KL ↓ | SIM | sAUC | InfoGain |
|--------|-----|------|-----|------|----------|
| IKN [29]| 0.679| 1.748| 0.380| 0.608| -0.233 |
| AIM [32]| 0.566| 14.472| 0.224| 0.557| -18.181 |
| SDLF [33]| 0.607| 3.954| 0.322| 0.596| -3.344 |
| GBVS [34]| 0.718| 1.363| 0.413| 0.628| 0.331 |
| DeepGazeII [37]| 0.610| 1.345| 0.335| 0.571| -0.964 |
| SAM-ResNet [39]| 0.673| 2.510| 0.388| 0.600| -1.475 |
| OpenSALICON [51,52]| 0.673| 1.349| 0.375| 0.615| 0.052 |
| SalGAN [38]| 0.662| 2.506| 0.373| 0.593| -1.350 |
| Our Approach (Step I) | 0.723| 1.663| 0.409| 0.622| -0.125 |
| GroundTruth (Humans) | 0.882| 0.000| 1.000| 0.860| 2.802 |

REFERENCES

[1] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” in Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008, pp. 722–729.
[2] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3d object representations,” in 4th IEEE Workshop on 3D Representation and Recognition, at ICCV, 2013, pp. 1–8.
[3] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, “Caltech-UCSD Birds 200,” California Institute of Technology, Tech. Rep. CNS-TR-2010-001, 2010.
[4] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.
Fig. 2: Qualitative results for real images (Toronto dataset). Each image is represented in a different column and each model saliency map in each row. The ground truth density map of human fixations is represented in the 2nd row.

Fig. 3: Qualitative results for synthetic images (SID4VAM dataset). Each image is represented in a different column and each model saliency map in each row. The ground truth density map of human fixations is represented in the 2nd row.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[6] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, “Part-based r-cnn for
fine-grained category detection,” in European Conference on Computer Vision, 2014, pp. 834–849.

[7] D. Lin, X. Shen, C. Lu, and J. Jia, “Deep lac: deep localization, alignment and classification for fine-grained recognition,” in Proc. of CVPR, 2015, pp. 1666–1774.

[8] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, “Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1143–1152.

[9] H. Zheng, J. Fu, T. Mei, and J. Luo, “Learning multi-attention convolutional neural network for fine-grained image recognition,” in Proc. of ICCV, 2017, pp. 5209–5217.

[10] J. Fu, H. Zheng, and T. Mei, “Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition,” in Proc. of CVPR, 2017, pp. 4438–4446.

[11] C. Cai, W. Zhao, and L. Zhang, “Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization,” in Proc. of ICCV, 2017, pp. 511–520.

[12] Y. Wang, V. I. Morariu, and L. S. Davis, “Learning a discriminative filter bank within a cnn for fine-grained recognition,” in Proc. of CVPR, 2018, pp. 4148–4157.

[13] C. F. Flores, A. Gonzalez-Garcia, J. van de Weijer, and B. Raducanu, “Saliency for fine-grained object recognition in domains with scarce training data,” Pattern Recognition, vol. 94, pp. 62–73, 2019.

[14] J. Hoffmann, S. Gupta, and T. Darrell, “Learning with side information through modality hallucination,” in Proc. of CVPR, 2016, pp. 826–834.

[15] B. Hariharan and R. Girshick, “Low-shot visual recognition by shrinking and hallucinating features,” in Proc. of ICCV, 2017, pp. 3018–3027.

[16] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, “Low-shot learning from imaginary data,” in Proc. of CVPR, 2018, pp. 7278–7286.

[17] H. Zhang, J. Zhang, and P. Kornius, “Few-shot learning via saliency-guided hallucination of samples,” in Proc. of CVPR, 2019, pp. 2770–2779.

[18] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, “Mit saliency benchmark,” http://saliency.mit.edu/.

[19] F. Murabito, C. Spannuntato, S. Palazzo, K. Pogorelov, and M. Rieger, “Top-down saliency detection driven by visual classification,” Computer Vision and Image Understanding, vol. 172, 09 2017.

[20] S. Huang, Z. Xu, D. Tao, and Y. Zhang, “Part-stacked cnn for fine-grained visual categorization,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1173–1182.

[21] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, “Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization,” Pattern Recognition, vol. 76, pp. 704 – 714, 2018.

[22] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in Proc. of CVPR, 2015, pp. 842–850.

[23] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, “Picking deep filter responses for fine-grained image recognition,” in Proc. of CVPR, 2016, pp. 1134–1142.

[24] G.-S. Xie, X.-Y. Zhang, W. Yang, M. Xu, S. Yan, and C.-L. Liu, “Lg-cnn: From local parts to global discrimination for fine-grained recognition,” Pattern Recognition, vol. 71, pp. 118–131, 2017.

[25] M. Sun, Y. Yuan, F. Zhou, and E. Ding, “Multi-attention multi-class constraint for fine-grained image recognition,” in IEEE Conference on ECCV, 2018, pp. 834–850.

[26] W. Luo, X. Yang, X. Mo, Y. Lu, L. S. Davis, J. Li, J. Yang, and S.-N. Lim, “Cross-x learning for fine-grained visual categorization,” in Proc. of ICCV, 2019, pp. 8242–8251.

[27] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” in Proc. of CVPR, 2016, pp. 317–326.

[28] S. Kong and C. Fowlkes, “Low-rank bilinear pooling for fine-grained classification,” in Proc. of CVPR, 2017, pp. 365–374.

[29] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

[30] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 185–207, jan 2013.

[31] Z. Bylinskii, E. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J. Tostos, “Towards the quantitative evaluation of visual attention modeling,” Vision Research, vol. 116, pp. 258–268, nov 2015.