Progressive Attention Guided Recurrent Network for Salient Object Detection

Xiaoning Zhang[1,*], Tiantian Wang[1,*], Jinqing Qi[1], Huchuan Lu[1], Gang Wang[2]
[1]Dalian University of Technology, China
[2] Alibaba AILabs, China
xiaoningzhang42@gmail.com, tiantianwang.ice@gmail.com,
{jinqing, lhchuan}@dlut.edu.cn, wg134231@alibaba-inc.com

Abstract

Effective convolutional features play an important role in saliency estimation but how to learn powerful features for saliency is still a challenging task. FCN-based methods directly apply multi-level convolutional features without distinction, which leads to sub-optimal results due to the distraction from redundant details. In this paper, we propose a novel attention guided network which selectively integrates multi-level contextual information in a progressive manner. Attentive features generated by our network can alleviate distraction of background thus achieve better performance. On the other hand, it is observed that most of existing algorithms conduct salient object detection by exploiting side-output features of the backbone feature extraction network. However, shallower layers of backbone network lack the ability to obtain global semantic information, which limits the effective feature learning. To address the problem, we introduce multi-path recurrent feedback to enhance our proposed progressive attention driven framework. Through multi-path recurrent connections, global semantic information from the top convolutional layer is transferred to shallower layers, which intrinsically refines the entire network. Experimental results on six benchmark datasets demonstrate that our algorithm performs favorably against the state-of-the-art approaches.

1. Introduction

Salient object detection, which simulates the human vision system to judge the importance of image regions, has received increasing attention in recent years. During the past two decades, many salient object detection methods have been proposed. Conventional saliency methods usually utilize hand-crafted low-level features such as color, intensity, contrast to predict saliency. However, it is of great difficulty for these low-level features based approaches to detect salient objects in complex scenarios.

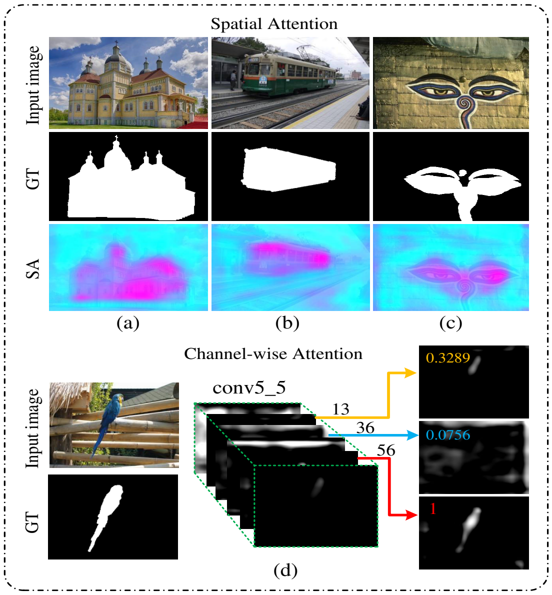


Figure 1. Illustration of spatial and channel-wise attention.

Recently, Convolutional Neural Networks (CNNs), which intelligently extract high-level and multi-scale complex representations from raw images directly, have achieved superior performance in many vision tasks. Due to the semantic information obtained from high-level features, CNN based saliency detection approaches have successfully broken the bottleneck of hand-crafted features. How to design a reasonable network which is able to learn effective features and how to process these features for saliency estimation become the key issues to be addressed.

Many state-of-the-art methods design saliency models by integrating multi-level convolutional features together. However, not all features are of equal importance to saliency detection and some even cause interference. Attention mechanisms, which add weights on image features, provide a feasible solution. To the best of my knowledge, there are not many works that utilize attention mechanisms to process features for saliency estimation. In this paper, we ap-

[*] denotes equal contributions.

ply multiple attention mechanisms to improve features for saliency detection. Figure 1 illustrates the motivation of introducing attention mechanisms.

In an image, not all spatial positions are contributing to saliency prediction in the same way and there sometimes exist background regions that generate distractions. In Figure 1(a)-(c), we can see that spatial attention (SA) can highlight the foreground regions and avoid distractions of some non-salient regions. Similarly, different feature channels have different response to foreground or background. Some channels have high response to foreground while some show obvious response to cluttered background. We visualize the feature maps of conv5_5 of our network in Figure 1(d). It can be seen that feature map of channel-56 (with higher response in foreground) is assigned a larger weight and feature map of channel-36 (with higher response in background) is assigned a small weight by our channel-wise attention mechanism. Based on channel-wise and spatial attention mechanisms, we propose a *progressive attention driven framework*, which selectively integrates multi-level contextual information. Benefiting from this framework, our proposed method outputs more effective features, which can alleviate distractions from background.

Moreover, in most of the state-of-the-art CNN based methods, saliency values are estimated by dealing with multi-scale side-output convolutional features. However, shallower layers of backbone feature extraction network such as VGG and ResNet lack the ability to obtain global semantic information thus generate messy results. It is necessary to propose a method which can refine the network intrinsically. Recurrent based methods such as RFCN [27] build a connection between the output and the input of the network. Saliency map of the previous stage is transmitted to the next stage for refinement. But effects of saliency prior is greatly weakened due to the concatenation with raw images. The network can be improved to some extent, but still can not produce enough effective features. In this paper, *multi-path recurrent feedback* is exploited to iteratively refine our progressive attention guided network. By introducing multi-path recurrent connections, global semantic information from the top convolutional layer is transferred to shallower layers, which intrinsically improves the feature learning ability of our network. The main framework is shown in Figure 2.

We summarize our contributions as follows:

1. Attention mechanisms are introduced into our model to generate powerful attentive features.
2. We propose a novel progressive attention guided module which selectively integrates multiple contextual information of multi-level features.
3. Multi-path recurrent feedback, which transfers global semantic information from the top layer to shallower layers, is exploited to refine the whole network.

2. Related Works

In this section, we briefly introduce the related works in three aspects. At the beginning, several representative salient object detection methods are reviewed. Then we describe the application of attention mechanisms in various vision tasks. Finally, we compare our multi-path recurrent network with other recurrent based works.

2.1. Salient Object Detection

Salient Object Detection methods can be categorized as conventional low-level hand-crafted features based [20, 4, 33, 15, 9, 38, 23, 21] and Convolutional Neural Networks driven [25, 13, 37, 14, 12, 16, 27, 18, 7, 35, 36, 28] approaches. Most of traditional saliency methods are based on low-level manually designed features, such as color, region contrast, etc. Detailed introductions of these methods can be found in recent survey paper [1]. In this paper, we put more emphasis on CNNs based approaches.

Recently, deep convolutional neural networks have set new state-of-the-art on saliency detection. Wang *et al.* [25] adopt two different deep CNNs to learn local information and global contrast respectively. In [13], multi-scale features are extracted from CNNs to estimate saliency of all super-pixels in the image. Zhao *et al.* [37] take both local and global context into account and integrate them into a multi-context deep CNNs for saliency detection. These methods illustrate that deep learning based methods are superior to traditional approaches. However, all these methods take image patches as training and testing samples, which finally leads to high cost for computation. Following the success of end-to-end deep networks in semantic segmentation [19], more works are trained end-to-end to predict pixel-wise saliency maps.

Liu *et al.* [18] propose a two-stage deep hierarchical saliency network that refines the coarse prediction map by recovering details from shallower layers. In [28], Wang *et al.* propose a refinement model by adding low-level detailed features to the saliency map generated in a stagewise manner. In [35], Zhang *et al.* integrate multi-level convolutional features at five resolutions respectively to generate saliency predictions. Motivated by the above mentioned methods, we find that effective features are of great importance to saliency detection. Therefore, we propose a *Progressive Attention Guided Recurrent Network* (PAGRN), which selectively integrates multi-level contextual information, to generate powerful features that possess both high-level information and necessary details for accurate detection.

2.2. Attention Mechanisms

Attention mechanisms have shown its efficiency in various vision tasks such as image captioning [31, 2], visual question answering [34, 30], pose estimation [5] and image classification [24]. In [31], Xu *et al.* first introduce visual

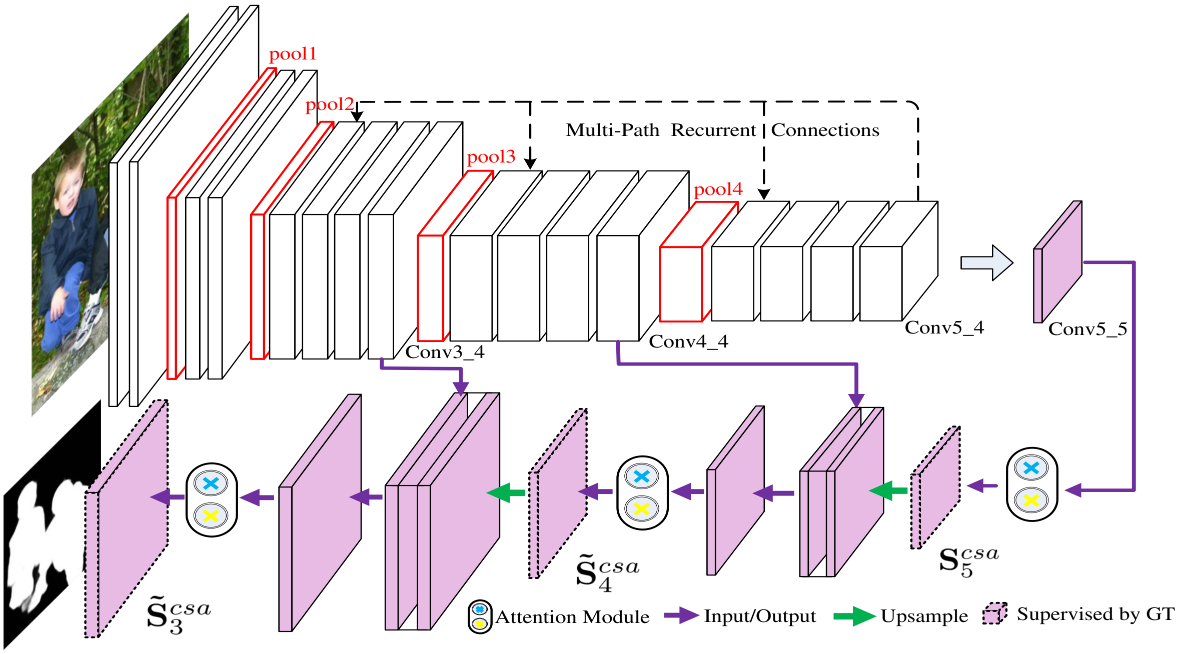


Figure 2. Main framework of our Progressive Attention Guided Recurrent Network. Attention module processes features with channel-wise and spatial attention sequentially. Then multiple layer-wise attention are integrated in a step by step form. Attentive features from high-level stage guide the low-level stage to generate new attentive features. Finally, the whole network is optimized by multi-path recurrent feedback.

attention into image captioning and both ”soft” and ”hard” attention mechanisms are exploited. Multi-context attention mechanisms are proposed by Chu *et al.* [5] for human pose estimation. In [24], Wang *et al.* propose an attention residual learning mechanism to train deep residual networks for image classification. Recently, Chen *et al.* [2] propose a SCA-CNN network that incorporates spatial and channel-wise attention in CNN for image captioning.

Stimulated by the success of attention in these vision tasks, we propose a progressive attention guided network which generates attentive features by channel-wise and spatial attention mechanisms sequentially. And on this basis, multiple layer-wise attentions are generated stage by stage, where attentive features act as the guidance of the next stage to produce new attentions and attentive features.

2.3. Recurrent Networks

Recently, recurrent neural networks have been used in saliency detection. In [27], Wang *et al.* propose a saliency method based on recurrent fully convolutional networks. At each time step, both the input RGB image and a saliency prior map feed forward through the RFCN to obtain the predicted saliency map and this saliency map will be treated as a saliency prior for the next time step. This kind of recurrent architecture can refine saliency map to a certain extent, however, it can not produce enough effective features for saliency detection because the effects of saliency prior are weakened by raw input images.

Kuen *et al.* [11] also propose a recurrent based network, that is, recurrent attentional convolutional-deconvolution network (RACDNN). It is worthy to mention that the usage of attention in their work is not the same as the one used by us. In RACDNN, a sub-region of input image is selected in each time step by a spatial transformer which achieves spatial attention. Then the attended sub-region will be the input of the next time step and a local saliency map of this sub-region is generated to refine the corresponding region of the whole prediction map. Although attention can drive the model to focus on a specific sub-region at each time step, this will cause redundancy in saliency prediction because of the overlap between these sub-regions.

Different from aforementioned recurrent base methods, we propose an attention guided recurrent network which transfers the high-level semantic information from the top convolutional layer to shallower layers by multi-path recurrent connections. Through multi-path feedback, the learning ability of shallower layers is enhanced. After feed-forward process, the entire network is refined in essence.

3. Attention Guided Recurrent Network

In this paper, we propose a novel progressive attention driven framework which intelligently selects features for integration. Powerful attentive features are generated to conduct saliency prediction. In order to refine the entire network essentially, multi-path recurrent feedback is incorporated to transfer high-level semantic information from the top convolutional layer to shallower layers.

In Section 3.1, we describe the channel-wise and spatial attention mechanisms used in our framework. Then our progressive attention guidance module is introduced in Section 3.2 which is followed by the detailed explanation of multi-path recurrent feedback module in Section 3.3.

3.1. Spatial and Channel-wise Attention

Spatial Attention Mechanism. In general, salient objects only correspond to partial regions of the input image. And there exist some background regions which can distract human attention. Therefore, directly exploiting convolutional features to predict saliency can lead to sub-optimal results because of the distraction of non-salient regions. Instead of considering all spatial positions equally, spatial attention is able to focus more on the foreground regions, which helps to generate effective features for saliency prediction. In Figure 1(a)-(c), we show some examples that spatial attention can highlight the salient object and avoid distractions in the background regions.

We represent convolutional features as $\mathbf{f} \in \mathbb{R}^{W\times H\times C}$. The set of spatial locations is denoted by $\mathbb{L} = \{(x, y)|x = 1, ..., W; y = 1, ..., H\}$, where (x, y) is the spatial coordinate. Spatial attention map is generated through the following steps: At first, a summarized feature map is generated as follows:

$$\mathbf{m} = \mathbf{W_s} * \mathbf{f} + \mathbf{b_s}, \tag{1}$$

where $*$ denotes convolution operation, $\mathbf{W_s}$ represents convolution filters, and $\mathbf{b_s}$ is the bias parameter. $\mathbf{m} \in \mathbb{R}^{W\times H}$ integrates the information of all channels in $\mathbf{f}$.

Then attention weight of feature vector at location l is obtained by applying Softmax operation to $\mathbf{m}$ spatially:

$$\mathbf{a_s}(l) = \frac{e^{\mathbf{m}(l)}}{\sum_{l'\in\mathbb{L}} e^{\mathbf{m}(l')}}, \tag{2}$$

where $\mathbf{m}(l)$ denotes the feature at location l. $\mathbf{a_s}$ is the spatial attention map, where $\sum_{l\in\mathbb{L}} \mathbf{a_s}(l) = 1$.

Channel-wise Attention Mechanism. Spatial attention assigns weights to features from the perspective of space, which relieves the problem of distraction caused by background regions. In fact, channel-wise features suffer from the similar problem. When dealing with convolutional features, most of existing methods treat all channels without distinction. However, different channels of feature in C-NNs generate response to different semantics. For example, in Figure 1 (d), channel-13 and channel-56 have higher response to the bird while channel-36 focuses more on the cluttered background. And our channel-wise attention mechanism assigns larger weights to channels which show higher response to salient objects. To alleviate the interference of the background, it is necessary to introduce channel-wise attention too.

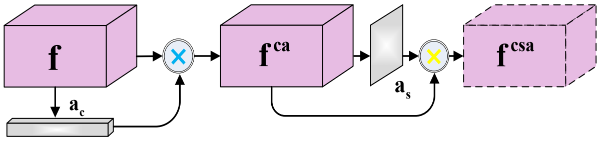


Figure 3. Layer-wise attention and attentive feature. Dotted line represents the saliency map generated by the feature is under the supervision of ground truth.

Next, we will describe the details of channel-wise attention. For channel-wise attention, we unfold $\mathbf{f}$ as $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_C]$, where $\mathbf{f}_i \in \mathbb{R}^{W\times H}$ is the i-th slice of $\mathbf{f}$ and C is the total channel number. We first apply average pooling to each $\mathbf{f}_i$ to obtain a channel feature vector $\mathbf{v} \in \mathbb{R}^C$. Then a convolutional layer is exploited to learn the aggregate feature of each channel:

$$\mathbf{u} = \mathbf{W}_c * \mathbf{v} + \mathbf{b}_c, \tag{3}$$

where $*$ denotes convolution operation, $\mathbf{W_c}$ represents convolution filters, and $\mathbf{b_c}$ is the bias parameter. Following the definition of spatial attention, a Softmax operation is applied to $\mathbf{u}$ to generate attention of each channel i:

$$\mathbf{a_c}(i) = \frac{e^{\mathbf{u}(i)}}{\sum_{i=1}^{C} e^{\mathbf{u}(i)}}, \tag{4}$$

where $\mathbf{u}(i)$ is the feature of channel i and $\mathbf{a_c} \in \mathbb{R}^C$ is the channel-wise attention vector, where $\sum_{i=1}^{C} \mathbf{a_c}(i) = 1$.

3.2. Progressive Attention Guidance Module

Due to repeated down-sampling operations such as pooling and convolution, the resolution of prediction map is greatly reduced, which leads to blurred object boundaries. Features from deep layers learn more about high-level semantic information, while features of shallow layers keep rich spatial details. To accurately locate salient objects and obtain sharper boundaries simultaneously, it is necessary to combine multi-level features together. However, FCN-based methods, which directly integrate multi-level features indiscriminately, are defective due to the redundant details and distractions from background.

To address the problem, we propose a novel attention driven network, which progressively encodes multi-level contextual information to produce more effective features for saliency estimation. Based on the discussions in Section 3.1, we can see that features should be assigned different weights both from spatial and channel-wise aspects. Therefore, both of the two attention mechanisms are utilized in our network to generate layer-wise attentive features.

Layer-wise Attentive Features. As displayed in Figure 3, layer-wise attentions and attentive features are generated by applying channel-wise attention and spatial attention sequentially. Given convolutional features $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_C]$, channel-wise attention vector $\mathbf{a_c}$ is generated by (4). Then $\mathbf{a_c}$ is applied to each slice of $\mathbf{f}$ as follows:

$$\mathbf{f}_i^{\mathbf{ca}} = \mathbf{a_c}(i) \times \mathbf{f}_i, \tag{5}$$

where $i \in \{1, ..., C\}$ and $\mathbf{f^{ca}}$ is the channel-wise attentive feature. Based on $\mathbf{f^{ca}}$, the spatial attention $\mathbf{a_s}$ is obtained through (2). We apply $\mathbf{a_s}$ to feature $\mathbf{f^{ca}}$ as follows:

$$\mathbf{f^{csa}} = \mathbf{a_s} \star \mathbf{f^{ca}}, \tag{6}$$

where $\star$ denotes channel-wise Hadamard matrix product operation. $\mathbf{f^{csa}}$ is the layer-wise attentive feature generated by our attention module. Then we use the attentive feature to predict saliency and the saliency map generated by each layer-wise attentive feature is supervised by the ground truth.

Progressive Attention Guidance Mechanism. In our model, layer-wise attentive information serves as a guidance for the next stage to adaptively generate new attentions as shown in Figure 2. Multiple layer-wise attentions and attentive features are generated stage by stage, which selectively introduce contextual information from multi-level features to refine features in a coarse-to-fine manner. Consider a CNN which is composed of L convolutional blocks. Denote $\mathbf{S}_\ell$ as the side-output convolutional feature of the ℓ-th block. Beginning from the L-th block, attentive feature of $\mathbf{S}_L$, expressed as $\mathbf{S}_L^{\mathbf{csa}}$, is constructed through (5) and (6). Then it acts as a guidance for the side-output feature of the (L-1)-th block to produce attention and attentive feature. Under the guidance of $\mathbf{S}_L$, $\mathbf{S}_{L-1}$ is transformed into:

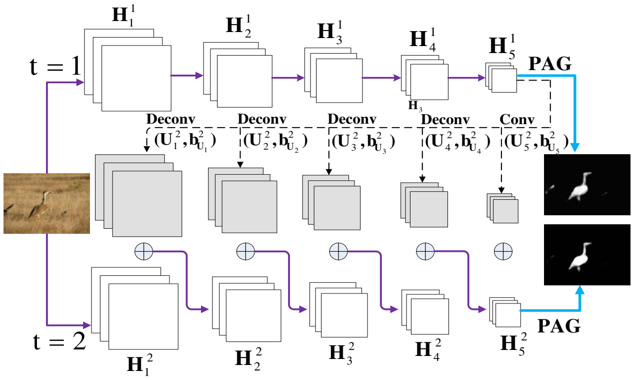
$$\tilde{\mathbf{S}}_{L-1} = UP(\mathbf{S}_L^{\mathbf{csa}})_2 \oplus \mathbf{S}_{L-1} \tag{7}$$

where $UP(\cdot)_2$ denotes upsampling feature maps by a factor of 2, and $\oplus$ represents element-wise addition operation. Next, layer-wise attentive feature $\tilde{\mathbf{S}}_{L-1}^{\mathbf{csa}}$ will be generated and it will serve as the guidance for feature of the (L-2)-th block. The rest layer-wise attentions and attentive features can be obtained in the same manner. In our proposed network, $L = 5$ and attentive features of $\mathbf{S}_{\ell \in \{3,4,5\}}$ are produced stage by stage. Attentive features of the final stage (i.e., $\tilde{\mathbf{S}}_3^{csa}$) is exploited to predict the final saliency map.

In Figure 7, we compare saliency maps of our progressive attention guidance mechanism (CA and CSA) with FCN-based method (FCN). It demonstrates that our proposed guidance mechanism can better integrate multi-level contextual information and avoid distractions of background. Comparing the results of CA and CSA, we can see that it is reasonable to apply this two kinds of attentions together. More results are displayed in Section 4.

3.3. Multi-Path Recurrent Guidance Module

Most of saliency works predict saliency through side-output features of certain convolutional blocks. This kinds of methods depend a lot on the backbone feature extraction network. However, due to the lack of ability to obtain high-level semantic information, features learned by shallow layers are messy. Side-output features from these layers usually contain redundant details. Designing a method to refine these side-output features can alleviate the problem but not enough. The critical point lies in how to improve the feature learning procedure of the backbone network. To remedy for the problem, we propose a multi-path recurrent guidance module that transfers global semantic information from top layer to shallower layers. Having access to high-level information, shallow layers can learn more powerful features, which refines the backbone network essentially.



t = 1
$\mathbf{H}_1^1$
$\mathbf{H}_2^1$
$\mathbf{H}_3^1$
$\mathbf{H}_4^1$
$\mathbf{H}_5^1$
PAG
$\mathbf{H}_3$
Deconv
$(\mathbf{U}_1^2, \mathbf{b}_{\mathbf{U}_1}^2)$
Deconv
$(\mathbf{U}_2^2, \mathbf{b}_{\mathbf{U}_2}^2)$
Deconv
$(\mathbf{U}_3^2, \mathbf{b}_{\mathbf{U}_3}^2)$
Deconv
$(\mathbf{U}_4^2, \mathbf{b}_{\mathbf{U}_4}^2)$
Conv
$(\mathbf{U}_5^2, \mathbf{b}_{\mathbf{U}_5}^2)$
t = 2
$\mathbf{H}_1^2$
$\mathbf{H}_2^2$
$\mathbf{H}_3^2$
$\mathbf{H}_4^2$
$\mathbf{H}_5^2$
PAG

Figure 4. Illustration of multi-path recurrent connections. PAG denotes progressive attention guidance module.

Multi-Path Recurrent Connections. In [10], Jin *et al.* propose a Multi-Path Feedback Recurrent Neural Network (MPF-RNN). Inspired by their work, we apply multi-path recurrent connections to our saliency model where high-level information is adaptively transferred back to different shallower layers. Figure 4 illustrates the unfolded recurrent network with multi-path recurrent connections. For simplicity, we consider a network only have L (L=5, in Figure 4) convolutional layers and unfold our network with $T = 2$. Denote $\mathbf{H}_\ell$ as the feature of the ℓ-th convolutional layer. In feed-forward process, it can be expressed as:

$$\mathbf{H}_\ell = f_\ell(\mathbf{W}_\ell * \mathbf{H}_{\ell-1} + \mathbf{b}_{\mathbf{W}_\ell}), \tag{8}$$

where $\mathbf{W}_\ell$ and $\mathbf{b}_{\mathbf{W}_\ell}$ denote kernel and bias parameters, and $f_\ell(\cdot)$ is a composite of multiple functions including activation function, pooling, etc. When multi-path recurrent connections from the top layer are introduced to hidden layers, (8) can be rewritten as:

$$\mathbf{H}_\ell^t = \begin{cases} f_\ell(\mathcal{N}(\mathbf{W}_\ell * \mathbf{H}_{\ell-1}^t + \mathbf{b}_{\mathbf{W}_\ell}) \\ \quad + \mathcal{N}(\mathbf{U}_\ell^t * \mathbf{H}_L^{t-1} + \mathbf{b}_{\mathbf{U}_\ell}^t)), & \ell \in \mathbf{R} \\ f_\ell(\mathbf{W}_\ell * \mathbf{H}_{\ell-1}^t + \mathbf{b}_{\mathbf{W}_\ell}), & otherwise \end{cases} \tag{9}$$

where $\mathcal{N}(\cdot)$ represents normalize operation by l_2-$norm$ and $\mathbf{H}_L^{t-1}$ denotes global convolutional features of top layer at time $t-1$. $\mathbf{U}_\ell^t$ and $\mathbf{b}_{\mathbf{U}_\ell}^t$, which are time variable, are adaptively learned feedback parameters for the ℓ-th layer. And $\mathbf{R} = \{r_m | m = 1, ...M\}$ is the set of layers with recurrent connections, where $r_m \in \{1, ..., L\}$ indexes the layers.

Recurrent Guidance. CNNs with fewer convolutional layers usually can not abstract global contextual information well. However, training a very deep convolutional network is costly and time consuming. In our paper, through multi-path recurrent connections, the network is enhanced in modeling long-range contextual information. Global semantic information is adaptively applied to guide the shallow layers to generate more effective features. With recurrent guidance, our network can obtain the learning ability of deeper networks. The output saliency maps at the final time step achieve state-of-the-art performance.

4. Experiments and Results

4.1. Experimental Setup

Datasets: To evaluate the performance of our algorithm, we conduct experiments on six benchmark datasets: ECSSD [32], HKU-IS [13], THUR15K [3], PASCAL-S [17], DUT-OMRON [33] and DUTS (the testing dataset which contains 5019 images) [26].

Implementation Details: The proposed algorithm is based on Caffe [8]. We use a fixed learning rate 1e-10 with a weight decay of 0.0005. The parameter of backbone feature extraction layer is initialized by the pre-trained VGG-19 model [22]. The training dataset of DUTS [26], which contains 10,553 images, is utilized to train our network for salient object detection. In our experiments, all input images are resized to 353×353.

Evaluation Metrics: We adopt precision-recall (PR) curves, F-measure, mean absolute error (MAE) and recently proposed S-measure [6] as our evaluation metrics. The F-measure, which is an overall performance measurement, is defined as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{10}$$

where $\beta^2 = 0.3$ is employed to emphasize the precision. And MAE is defined as the average pixel-wise absolute difference between the binary ground truth and the saliency map:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)|, \tag{11}$$

where W and H denote width and height of saliency map S.

4.2. Comparison with the State-of-the-Art

Our algorithm is compared with thirteen state-of-the-art salient object detection methods, including DRFI [9], BL [23], KSR [29], LEGS [25], ELD [12], MDF [13], DS [16], MCDL [37], DCL [14], RFCN [27], DHS [18], UCF [36] and Amulet [35]. For fair comparison, we use the recommended parameter settings to implement these methods or utilize the saliency maps provided by the authors.

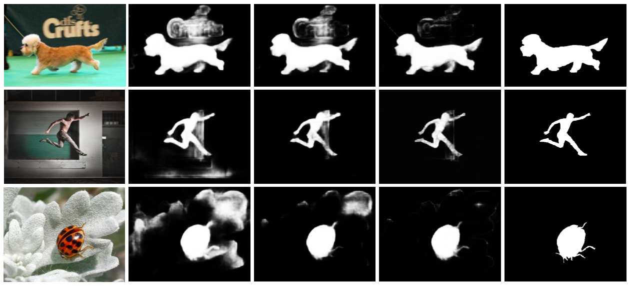

(a) Input (b) FCN (c) CA (d) CSA (e) GT

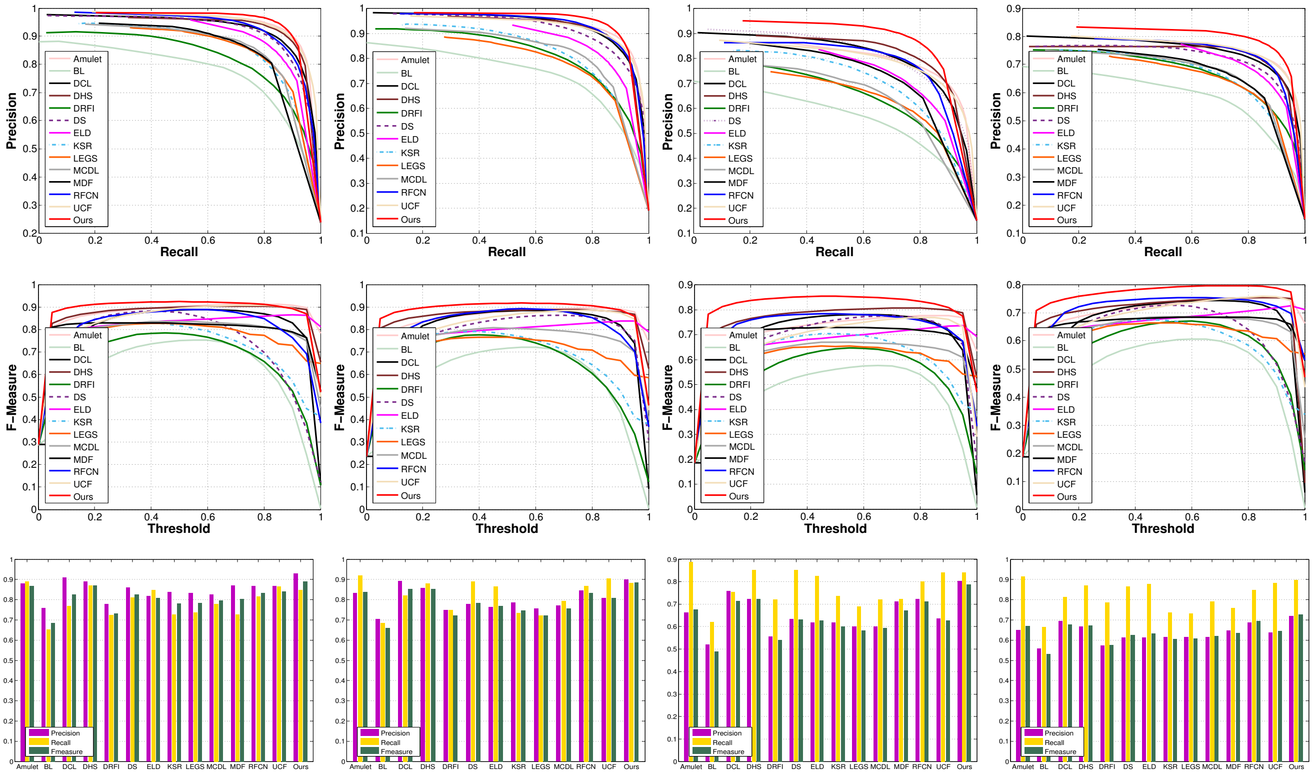
Figure 7. Saliency maps of our attention guided model and FCN-based method. CA denotes that only channel-wise attention is exploited while CSA incorporates both channel-wise and spatial attention into our model.

For quantitative evaluation, P-R curves, F-measure curves and F-measure scores are displayed in Figure 5. We can see that our proposed method performs favorably against other methods on all datasets and evaluation metrics. Furthermore, we compare our algorithm with other state-of-the-art in the form of F-measure and MAE. Table 1 illustrates that our method ranks first on almost all datasets. Comparing F-measure scores, our PAGRN outperforms the second best method by 2.3%, 3.9%, 4.9%, 3.9%, 4.0%, 8.8% over ECSSD, HKU-IS, THUR15K, PASCAL-S, DUT-OMRON, DUTS respectively. As for MAE, our model lower the value by 7.7%, 14.6%, 3.2%, 19.1%, 16.4% on HKU-IS, THUR15K, PASCAL-S, DUT-OMRON, DUTS respectively.

On the other hand, we also show the qualitative evaluation results in Figure 6. It can be seen that our method uniformly highlights the foreground regions even in very challenging scenes. In the third row of Figure 6, almost all methods wrongly assign foreground label to the shadow of the dogs except for our method. Beneficial from attention guidance, our method can effectively suppress the distractions from the background.

4.3. Ablation Analysis

Comparison with FCN-based structure. To verify the importance of progressive attention guided module, we compare our model with a similar FCN-based structure which is conducted without attention guidance. Figure 7 shows the output saliency maps of our attention guided models (i.e., CA and CSA) and FCN-based method. We can see that FCN-based method suffers from background interference, which mainly due to the unselective combination of multi-level features. Our proposed progressive attention guided



(a) ECSSD dataset (b) HKU-IS dataset (c) DUTS dataset (d) THUR15K dataset

Figure 5. The first row shows the P-R curve of the proposed method with other state-of-the-art methods. The second shows F-measure curves. The last show the average precision, recall, and F-measure scores across four datasets. The proposed method performs best among all datasets in terms of all metrics.

	ECSSD		HKU-IS		THUR15K		PASCAL-S		DUT-OMRON		DUTS	
	MAE	F-measure	MAE	F-measure	MAE	F-measure	MAE	F-measure	MAE	F-measure	MAE	F-measure
DRFI	0.166	0.733	0.145	0.722	0.150	0.576	0.207	0.618	0.138	0.550	0.175	0.541
BL	0.217	0.684	0.207	0.660	0.219	0.530	0.249	0.574	0.239	0.499	0.238	0.490
KSR	0.135	0.782	0.120	0.747	0.123	0.604	0.157	0.704	0.131	0.591	0.121	0.602
LEGS	0.119	0.785	0.119	0.723	0.125	0.607	0.155	0.697	0.133	0.592	0.138	0.585
ELD	0.082	0.810	0.074	0.769	0.098	0.634	0.123	0.718	0.092	0.611	0.093	0.628
MDF	0.108	0.805	-	-	0.109	0.636	0.146	0.709	0.092	0.644	0.100	0.673
DS	0.124	0.826	0.078	0.785	0.116	0.626	0.176	0.659	0.120	0.603	0.091	0.632
MCDL	0.102	0.796	0.092	0.757	0.103	0.620	0.145	0.691	0.089	0.625	0.105	0.594
DCL	0.151	0.827	0.136	0.853	0.161	0.676	0.181	0.714	0.157	0.684	0.149	0.714
RFCN	0.109	0.834	0.089	0.835	0.100	0.695	0.133	0.751	0.111	0.627	0.090	0.712
DHS	0.063	0.871	0.054	0.852	0.082	0.673	0.095	0.773	-	-	0.067	0.724
UCF	0.080	0.841	0.074	0.808	0.112	0.645	0.127	0.701	0.132	0.613	0.117	0.629
Amulet	0.061	0.869	0.052	0.839	0.094	0.670	0.100	0.763	0.098	0.647	0.085	0.678
Ours	0.064	0.891	0.048	0.886	0.070	0.729	0.092	0.803	0.072	0.711	0.055	0.788

Table 1. MAE (lower is better) and F-measure (higher is better) comparisons with 13 methods on 6 benchmark datasets. The best three results are shown in red, green, and blue fonts respectively. Our algorithm ranks first on almost all datasets.

network addresses the problem effectively. To make quantitative evaluations, histograms of F-measure and MAE on HKU-IS and THUR datasets are shown in Figure 8. Our attention guided module can selectively integrate multi-level information thus achieve better performance.

Selection of attention mechanisms. As mentioned in Section 3.2, attention mechanisms assign weights to features from both spatial and channel-wise perspective. We separately utilize channel-wise attention mechanism and spatial attention mechanism in a progressive manner (Denoted as CA and SA). Comparing the F-measure and MAE results in Figure 8, we can see that both of these two attention mechanisms are helpful and combining them together can have a better performance.

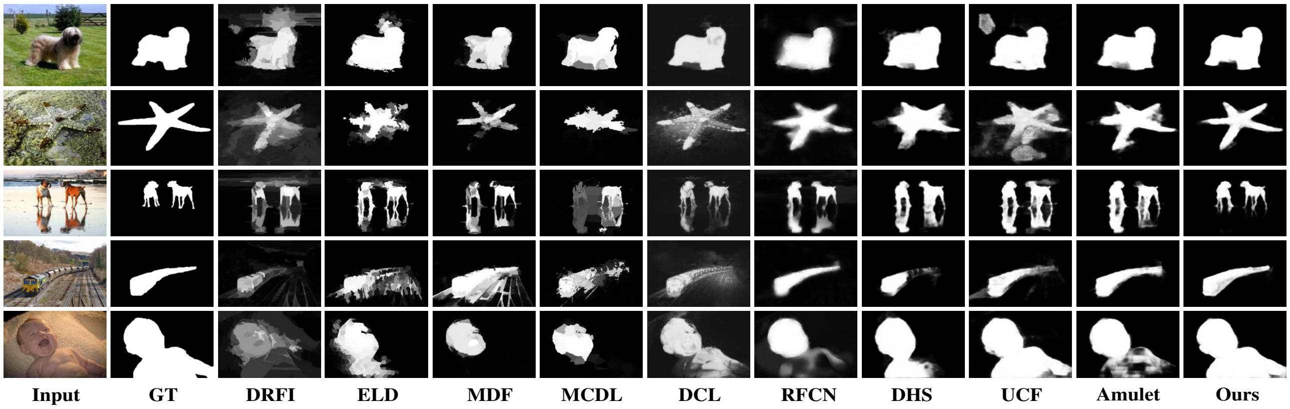
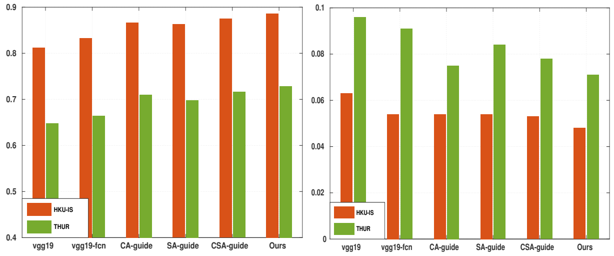


Figure 6. Visual comparison between our results and state-of-the-art methods.



(a) F-measure (b) MAE

Figure 8. Histograms of F-measure and MAE on HKU-IS and THUR datasets. Our attention guided network outperforms FCN-based method in a similar framework.

Settings	THUR			DUTS		
	F-measure	S-measure	MAE	F-measure	S-measure	MAE
w or w/o middle supervision						
CSA	0.716	0.808	0.078	0.761	0.797	0.065
CSA(w/o ms)	0.706	0.807	0.081	0.750	0.804	0.066
layers with recurrent connection(recurrent once, T=2)						
5	0.720	0.825	0.074	0.774	0.825	0.058
45	0.719	0.826	0.075	0.780	0.829	0.056
345	0.729	0.830	0.070	0.788	0.825	0.055
2345	0.721	0.827	0.072	0.785	0.833	0.054
12345	0.727	0.826	0.071	0.784	0.823	0.056
recurrent twice, T=3						
r2-345	0.723	0.829	0.072	0.783	0.827	0.055
merge the results of each time step						
r1-merge	0.723	0.829	0.071	0.775	0.825	0.058
r2-merge	0.716	0.828	0.074	0.772	0.831	0.057

Table 2. Ablation analysis using F-measure, S-measure and MAE metrics. The results of the top two are shown in red and green.

Effectiveness of middle supervision. In our progressive attention module, there are three attentive features generated stage by stage (i.e., $\mathbf{S}_5^{\mathbf{csa}}$, $\tilde{\mathbf{S}}_4^{\mathbf{csa}}$ and $\tilde{\mathbf{S}}_3^{\mathbf{csa}}$). To make attentions at each stage meaningful, we use the three attentive features to estimate saliency respectively and the three output saliency maps are all supervised by ground truth. The supervision on $\mathbf{S}_5^{\mathbf{csa}}$ and $\tilde{\mathbf{S}}_4^{\mathbf{csa}}$ is called middle supervision. In Table 2, we list the F-measure, S-measure and MAE results with or without middle supervision on THUR and DUTS datasets. The results verify the rationality of middle supervision.

Analysis of multi-path recurrent guidance module. We analyze the multi-path recurrent guidance module from three aspects: 1) how to choose the layer set $\mathbf{R}$ with recurrent connections; 2) recurrent times; 3) merging or not merging the results of each time step. The F-measure, S-measure and MAE values are shown in Table 2. For simplicity, we denote ”5” as conv5_1, ”45” as $\{$conv5_1, conv4_1$\}$ and other settings are in the same form. By adding recurrent paths, the model is refined gradually. It is observed that introducing recurrent connections to layer conv2_1 and conv1_1 can not make improvements. The main reason is that these two layers are too shallow to possess ability of effective feature learning. As for recurrent times, we conduct recurrent twice and the results are represented by ”r2-345”. We find that the model is saturated and can not be further improved. Every time step can generate saliency maps. Merging them together reduces the performance which indicates that our model generates best results at the final time step. With the settings of ”345” and ”T=2”, the best result of multi-path recurrent is obtained. By comparing this result with CSA, we can see that multi-path recurrent guidance module intrinsically refines the whole network to achieve state-of-the-art results.

5. Conclusion

In this paper, we propose a novel progressive attention guided recurrent network, which selectively integrates contextual information from multi-level features to generate powerful attentive features. By introducing multi-path recurrent connections, global semantic information is utilized to guide the feature learning procedure of shallower layers, which refines the entire network essentially. Extensive evaluations demonstrate the effectiveness of our network.

Acknowledgements. This work was supported by the Natural Science Foundation of China under Grant 61725202 and 61472060.

References

[1] A. Borji, M. M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *TIP*, 24(12):5706–5722, 2015. 2

[2] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, June 2017. 2, 3

[3] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu. Salientshape: group saliency in image collections. *The Visual Computer*, 30:443–453, Apr 2014. 6

[4] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu. Global contrast based salient region detection. *TPAMI*, 37(3):569–582, 2015. 2

[5] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *CVPR*, June 2017. 2, 3

[6] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017. 6

[7] Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In *CVPR*, June 2017. 2

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM*, pages 675–678, 2014. 6

[9] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, June 2013. 2, 6

[10] X. Jin, Y. Chen, Z. Jie, J. Feng, and S. Yan. Multi-path feedback recurrent neural networks for scene parsing. In *AAAI*, pages 4096–4102, 2017. 5

[11] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *CVPR*, June 2016. 3

[12] G. Lee, Y.-W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, June 2016. 2, 6

[13] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *CVPR*, June 2015. 2, 6

[14] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *CVPR*, June 2016. 2, 6

[15] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, December 2013. 2

[16] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *TIP*, 25(8):3919–3930, 2016. 2, 6

[17] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, June 2014. 6

[18] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, June 2016. 2, 6

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, June 2015. 2

[20] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012. 2

[21] Y. Qin, H. Lu, Y. Xu, and H. Wang. Saliency detection via cellular automata. In *CVPR*, June 2015. 2

[22] M. Simon, E. Rodner, and J. Denzler. Imagenet pre-trained models with batch normalization. 2016. 6

[23] N. Tong, H. Lu, X. Ruan, and M.-H. Yang. Salient object detection via bootstrap learning. In *CVPR*, June 2015. 2, 6

[24] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *CVPR*, June 2017. 2, 3

[25] L. Wang, H. Lu, X. Ruan, and M.-H. Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, June 2015. 2, 6

[26] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, June 2017. 6

[27] 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, 6

[28] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017. 2

[29] T. Wang, L. Zhang, H. Lu, C. Sun, and J. Qi. Kernelized subspace ranking for saliency detection. In *ECCV*, pages 450–466, 2016. 6

[30] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, pages 451–466, 2016. 2

[31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2

[32] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, June 2013. 6

[33] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, June 2013. 2, 6

[34] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, June 2016. 2

[35] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017. 2, 6

[36] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017. 2, 6

[37] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *CVPR*, June 2015. 2, 6

[38] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *CVPR*, June 2014. 2