Exploring Duality in Visual Question-Driven Top-Down Saliency

Shengfeng He, *Member, IEEE*, Chu Han, Guoqiang Han, and Jing Qin, *Member, IEEE*

***Abstract*— Top-down, goal-driven visual saliency exerts a huge influence on the human visual system for performing visual tasks. Text generations, like visual question answering (VQA) and visual question generation (VQG), have intrinsic connections with top-down saliency, which is usually involved in both VQA and VQG processes in an unsupervised manner. However, it is shown that the regions that humans choose to look at to answer questions are very different from the unsupervised attention models. In this brief, we aim to explore the intrinsic relationship between top-down saliency and text generations, and to figure out whether an accurate saliency response benefits text generation. To this end, we propose a dual supervised network with dynamic parameter prediction. Dual-supervision explicitly exploits the probabilistic correlation between the primal task top-down saliency detection and the dual task text generation, while dynamic parameter prediction encodes the given text (i.e., question or answer) into the fully convolutional network. Extensive experiments show the proposed top-down saliency method achieves the best correlation with human attention among various baselines. In addition, the proposed model can be guided by either questions or answers, and output the counterpart. Furthermore, we show that combining human-like visual question-saliency improves the performance of both answer and question generations.**

***Index Terms*— Dual learning, saliency, visual question answering (VQA), visual question generation (VQG).**

I. Introduction

Given a specific goal, humans have an excellent ability in rapidly locating relevant regions of the image, known as top-down visual attention. Various applications in computer vision, like image captioning [1], [2] and visual question answering (VQA) [3]–[5], exploit the mechanism of top-down saliency to focus on selective regions while generating or understanding a description. The obtained saliency maps usually explain the internal representations of the learned convolutional neural networks (CNNs). Unlike image captioning, the attention maps of VQA vary according to the given questions in order to obtain correct answers. However, a recent study [6] shows that the learned attentions look at different regions from what humans do (see Fig. 1). In other words, existing unsupervised models cannot simulate the process of question-driven top-down visual attention. Traditional classification-based saliency methods detect distinct object regions [7]–[9], but cannot be applied to textual information. On the other hand, whether an accurate saliency response benefits VQA performance remains a question.

In this article, we aim to explore the intrinsic relationship between top-down saliency and visual questions and find out how a human-like saliency influences text generations. To this end, we introduce visual question generation (VQG) [12], [13] into the saliency prediction network. VQG aims to generate questions based on the image content. This process requires additional information, typically an answer [14], to confine the output space. Interestingly, we find that VQG and text-driven top-down saliency can be modeled in a dual form, i.e., the input and output of text-driven top-down saliency (a question and a saliency map) can be the output and input of VQG. Dual tasks can intrinsically complement each other. Therefore, learning one task may help the other task and vice versa.

Therefore, we formulate text-driven top-down saliency as the primal task and text generation as the dual task. These two tasks are trained simultaneously in a dual supervised manner. Unlike multi-task learning that shares the same representation of two tasks, we govern the training process of two different networks by adding a regularization term. This learning scheme exploits the structural relationship between two tasks for effective learning. In order to encode visual questions into a fully convolutional network, we involve a dynamic parameter layer, where the weights of this layer are adaptively based on the input text information. Extensive experiments demonstrate that the proposed text-driven top-down saliency network outperforms the state-of-the-art attention models and various baselines. In addition, the proposed dual network is flexible, and we show that top-down saliency can also be guided by visual question answers, leading to more precise answers for the VQA network. Finally, we investigate the saliency-driven VQG, and we demonstrate that saliency maps may serve as additional information for the better question generation.

In summary, the contributions of this work are threefold.

1) We propose a dual supervised network with dynamic parameter prediction, which demonstrates the effectiveness for top-down saliency and text generation.
2) We delve into the learning interaction between text generations and top-down saliency. In particular, we demonstrate a better saliency map benefit both the VQA and VQG tasks.
3) We achieve state-of-the-art performances on three different tasks, i.e., text-driven top-down saliency, VQA, and VQG.

II. Related Work

Top-Down Saliency can be classified into weakly supervised and supervised methods. Weakly supervised methods aim to discover the saliency response during the prediction of CNNs [15], [16]. Besides locating relevant objects in the task of image-level classification, it interprets the internal relationship of the learned CNNs. Given a target class, supervised methods simulate the process of visual search with the supervision of object segments [17], [18]. Some methods explored top-down saliency that is driven by different factors, e.g., exemplars [19] or keywords [20]. In this work, we extend the idea to explore the understanding between top-down saliency and visual questions.

Visual Question Answering and Generation are cross-discipline tasks that require an understanding of texts and images. Image

Manuscript received January 29, 2019; revised April 24, 2019; accepted August 2, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61472145 and Grant 61702194, in part by the Innovation and Technology Fund of Hong Kong under Project ITS/319/17, in part by the Special Fund of Science and Technology Research and Development on Application From Guangdong Province (SF-STRDA-GD) under Grant 2016B010127003, in part by the Guangzhou Key Industrial Technology Research fund under Grant 201802010036, and in part by the Guangdong Natural Science Foundation under Grant 2017A030312008. *(Corresponding author: Chu Han.)*

S. He and G. Han are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: hesfe@scut.edu.cn; csgqhan@scut.edu.cn).

C. Han is with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong (e-mail: chan@cse.cuhk.edu.hk).

J. Qin is with the Department of Nursing, The Hong Kong Polytechnic University, Hong Kong (e-mail: harry.qin@polyu.edu.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2933439

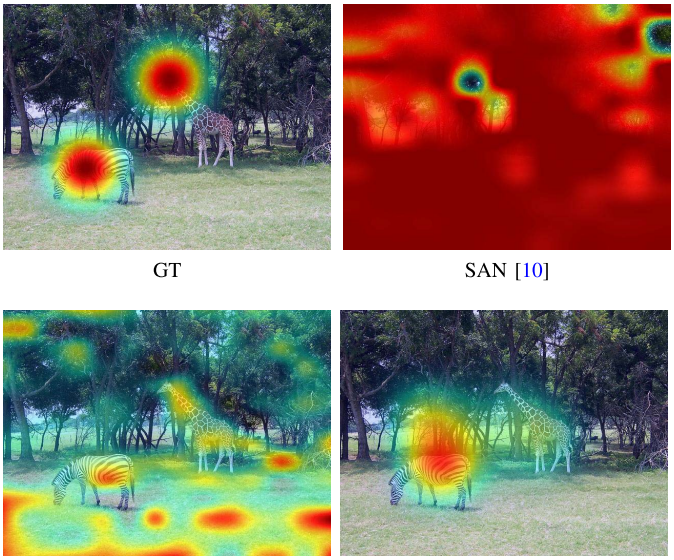

Fig. 1. Humans look at different regions compared with unsupervised attention (second and third examples) to answer questions. We propose to predict top-down saliency guided by visual questions.

features are extracted by CNNs, while texts are handled by recurrent networks. These two features are then concatenated in different forms. An alternative [21] uses dynamic parameter prediction to embed questions into the fully connected layer of the classification network. We adopt this idea and integrate it into the fully convolutional network for image processing tasks. In order to extract the underlying relationship between questions and images, VQA usually involves the top-down attention mechanism to highlight relevant regions of the questions [5], [10], [11], [22]–[25]. These methods, however, are unsupervised attention models, and the obtained saliency maps are largely different from human attention to answer the questions [6]. VQG, on the other hand, attracts research attention since the first data set is proposed by Mostafazadeh *et al.* [12]. Zhang *et al.* [13] propose to use region captions as guidance for generating questions. Jain *et al.* [26] present a diverse VQG model by combining the variational autoencoder and the long short term memory. Other than visual question processing, Cornia *et al.* [27], [28] propose to explicit predicting caption-driven saliency for image captioning, and they show human-like saliency benefits the resulted image captions. Different from existing works, our focus is on detecting visual question saliency and exploring the duality between saliency and questions.

Multi-Task and Dual Learning both can be used to train multiple tasks jointly. Multi-task learning introduces multiple losses for supervising all tasks. Although different strategies [29]–[31] are proposed to integrate multiple losses, the target tasks are required to share the same input space and representations. Differently, dual learning aims to optimize the training process by levering the cycle consistency. It can be used in two separated networks with invertible input–output. He *et al.* [32] propose dual learning to machine translation, where A-to-B and B-to-A translations can be intuitively modeled in a closed loop. Li *et al.* [14] model VQG and VQA as dual tasks, and they leverage Q-A dependences to regularize the training process. Different from existing works, we make the first attempt to explore the duality between top-down saliency and text generation.

III. Approach

The pipeline of our proposed approach is shown in Fig. 2. Our method consists of two components: a text-driven top-down saliency network and a saliency-driven VQG network. The first component takes an image and a question as input and yields a saliency map that correlates with the question. The second component takes an image and a saliency map as the input, and outputs an appropriate question according to the salient regions. These two networks do not share parameters, and they are trained simultaneously as dual tasks. We will describe the two networks in Sections III-A and III-B, respectively. Our dual learning strategy is presented in Section III-C

A. Text-Driven Top-Down Saliency Network

Given an input image I and a question q, our text-driven top-down saliency network predicts the salient regions m that draw human attention to answer question q. This network consists of two subnetworks [Fig. 2 (left)]. The first subnetwork predicts a saliency map based on the conv–deconv pipeline (i.e., fully convolutional network [33]). The convolution part is based on ResNet-152 [34] except the last classification layer, while the deconvolution part is set as the mirror architecture of the first part.

The second subnetwork encodes the input question into textual features using gated recurrent units (GRUs) [35]. Traditional VQA and image captioning methods combine textual and visual features by simply concatenating them, but it may not fully explore the correlation between two sources of knowledge. We combine both the types of information using a parameter-prediction layer.

1) Dynamic Parameter Prediction: The text-driven top-down saliency problem is formulated as a binary classification problem, and each predicted value of pixel m_i in m is defined as

$$m_i = p(c_1|I, q; \theta) \tag{1}$$

where c_1 indicates the class of salient regions. I and q denote the input image and the question. θ denotes the parameters of the network. Note that θ represents the parameters of both the saliency network and the GRU cells. Once properly trained, θ is fixed for any image and question. This prevents the trained network from generalizing to different input images and questions. We introduce the parameter prediction mechanism to (1)

$$m_i = p(c_1|I, q; \theta, \theta_d(q)) \tag{2}$$

where $\theta_d(q)$ denotes the parameters of the added parameter prediction layer. $\theta_d(q)$ is dynamically predicted according to the given question q, and thus enables question-dependent textual knowledge embedding.

Specifically, the parameter prediction layer is implemented as a convolution layer and is inserted in the middle of the conv–deconv pipeline. The output of the parameter-prediction layer with an input feature map f_s is

$$f_o = W_d(q) * f_s + b \tag{3}$$

where $W_d(q)$ denotes the predicted matrix by the parameter prediction network, $*$ is the convolution operator, and b is the bias. By doing so, the saliency network is parameterized by the input question q.

Assume that the output embedding vector of GRU cells is $V(q)$. In order to predict $W_d(q)$, we apply a fully connected layer to the embedding vector

$$\hat{W_p}(q) = W_g V(q) \tag{4}$$

where W_g denotes the parameters of this fully connected layer. $\hat{W_p}(q)$ is a 1-D vector, which cannot be directly used in the parameter-prediction layer. We first set the output length of the fully connected

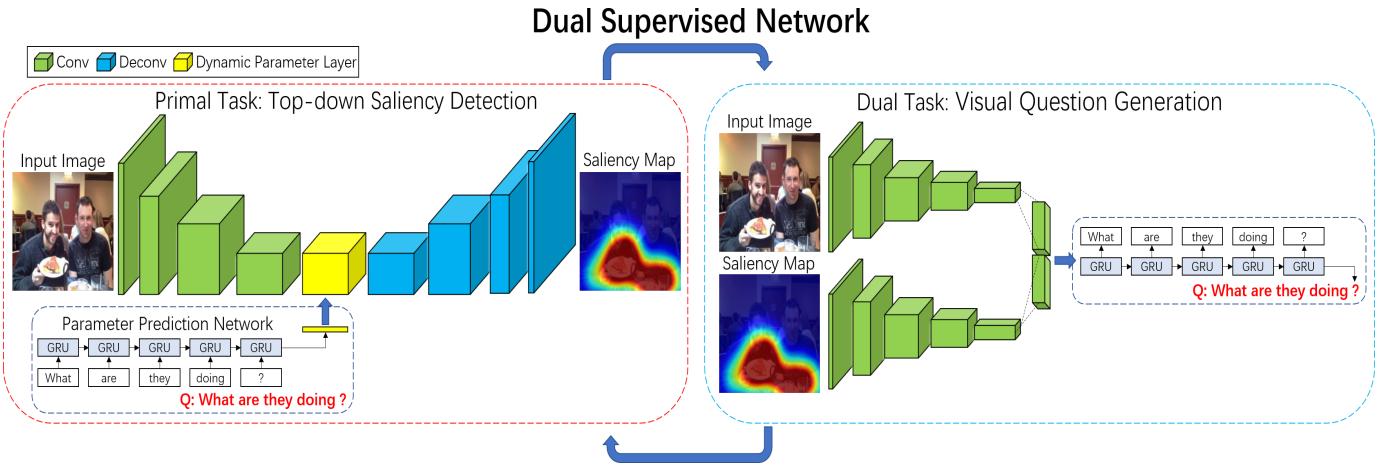

Fig. 2. Pipeline of our proposed method, which involves two tasks, top-down saliency detection, and VQG. The input–output of one task is the output/input of the other; therefore, these two tasks are simultaneously trained using dual learning. Dynamic parameter prediction is used to embed visual questions into the fully convolutional network.

layer to be the same as the depth of $W_d(q)$. It is then repeated spatially to match the size of $W_d(q)$, and the final predicted matrix is denoted as $W_p(q)$. Equation (3) can be rewritten as

$$f_o = W_p(q) * f_s + b. \tag{5}$$

2) Backpropagation: The introduced parameter prediction layer can be trained in an end-to-end manner using standard backpropagation. The derivative of the predicated parameters $g(q)$ can be computed as

$$\frac{\partial \ell_1}{\partial g(q)} = f_s \frac{\partial \ell_1}{\partial f_o} \tag{6}$$

where ℓ_1 denotes the loss function of the top-down saliency task.

B. Saliency-Driven VQG Network

Given an image and a saliency map, the saliency-driven VQG network predicts a question that corresponds to the salient regions. It is set as a dual task in our framework, where the network design is independent of the primal task. We use a simple architecture for VQG [Fig. 2 (right)]. First, two images are fed to the ResNet-based Siamese network [36] for extracting feature representations. Then, two image features are concatenated together as the input for GRU cells, and each unit outputs a word for predicting the question. Each word of the question is generated as follows:

$$w_t = \arg\max_{\hat{w} \in \mathbb{W}} p(\hat{w}|I, m, w_0, \ldots, w_{t-1}) \tag{7}$$

where w_t denotes the word generated at step t and $\mathbb{W}$ denotes the word vocabulary.

C. Dual Learning

The above two networks have interchangeable input–output. Inspired by [14], [37], we formulate these two tasks in a dual form, where the primal task learns to map a question q to a saliency map m, and the dual task learns to map a saliency map m to a question q. These mapping functions are equal to learning the joint probability of $p(m|q; \theta_{qm})$ (primal task) and $p(q|m; \theta_{mq})$ (dual task), where θ_{qm} and θ_{mq} are the parameters of two networks, respectively. In standard supervised learning, the parameters of these two networks are learned by optimizing the following equations:

$$\min_{\theta_{qm}} \left(\frac{1}{n}\right) \Sigma_{i=1}^{n} \ell_1(m_i, m_i^{gt}) \tag{8}$$

$$\min_{\theta_{mq}} \left(\frac{1}{n}\right) \Sigma_{i=1}^{n} \ell_2(q_i, q_i^{gt}) \tag{9}$$

where n is the total number of training samples, m^{gt} and q^{gt} are the ground truth, and ℓ_1 and ℓ_2 are the loss functions of two tasks. Traditionally, these two tasks are trained independently and separately, and thus cannot explore the bidirectional relationship between two tasks. Here, the joint probability of two tasks $p(q, m)$ is used to define the duality, and ideally, two learned networks should satisfy the following condition:

$$p(q, m) = p(q)p(m|q; \theta_{qm}) = p(m)p(q|m; \theta_{mq}) \tag{10}$$

where $p(q)$ and $p(m)$ denote the marginal distributions. This is a necessary condition for the networks that are trained dually. It is obvious that traditional learning strategies cannot guarantee the above condition. To make (10) hold, we solve the following multi-objective optimization problem:

$$\begin{aligned} &\text{obj. 1: } \min_{\theta_{qm}} \left(\frac{1}{n}\right) \Sigma_{i=1}^{n} \ell_1(m_i, m_i^{gt}) \\ &\text{obj. 2: } \min_{\theta_{mq}} \left(\frac{1}{n}\right) \Sigma_{i=1}^{n} \ell_2(q_i, q_i^{gt}) \\ &\text{s.t. } p(q)p(m|q; \theta_{qm}) = p(m)p(q|m; \theta_{mq}). \end{aligned} \tag{11}$$

Equation (11) can be solved by introducing the Lagrange multipliers [38]. First, the duality constraint in (11) can be converted into a regularization term as follows:

$$\ell_d = (\log p(q) + \log p(m|q, \theta_{qm}) - \log p(m) - \log p(q|m; \theta_{mq}))^2. \tag{12}$$

This regularization term is involved in the training process of dual tasks with a weighted combination, and the two tasks can be trained by minimizing the following objective functions:

$$\min_{\theta_{qm}} \left(\frac{1}{k}\right) \Sigma_{j=1}^{k} \left[\ell_1(m_j, m_j^{gt}) + \lambda_{qm} \ell_d(q_j, m_j; \theta_{qm}, \theta_{mq})\right] \tag{13}$$

$$\min_{\theta_{mq}} \left(\frac{1}{k}\right) \Sigma_{j=1}^{k} \left[\ell_2(q_j, q_j^{gt}) + \lambda_{mq} \ell_d(q_j, m_j; \theta_{qm}, \theta_{mq})\right] \tag{14}$$

where k is the total number of training sample pairs of two tasks and λ_* indicates the regularization weight. In our experiment, we train (15) by using an Adam optimizer [39].

In practice, it is usually impossible to obtain the ground-truth marginal distributions of $p(q)$ and $p(m)$. We use empirical marginal distributions instead. We use our pretrained GRU-based language model [35] to generate the marginal distributions. An input question q is modeled by GRU cells, and the marginal distribution of the ith word q_i in q is defined as

$$\Pi_{i=1}^{E_q} p(q_i|q_1, \ldots, q_{i-1}) \tag{15}$$

where E_q denotes the total number of words in q. We use our pretrained text-driven saliency network for modeling the saliency distribution. For a saliency map m with T pixels, the saliency distribution is defined as

$$\Pi_{i=1}^{T} p(m_i|m_1, \ldots, m_{i-1}) \tag{16}$$

where m is reshaped to a 1-D vector and m_i is the ith saliency value.

D. Training the Networks

Our framework contains two base models, ResNet-152 [34] for images and GRU [35] for questions. The primal and dual networks are pretrained independently. For the purpose of visual feature extraction in VQG, we directly used the pretrained model of ResNet-152 based on ImageNet, and this part is fixed during the training of VQG. The VQG subnetwork is pretrained on the VQG benchmark [12], while the top-down saliency model is pretrained on the traditional saliency benchmark [40]. Given different input questions, the distributions in the parameter-prediction layer vary significantly, which may prevent the saliency model from obtaining optimal results. We apply batch normalization [41] in the parameter-prediction layer as well as all the conv and deconv layers for regularizing the training process.

To enhance the textual embedding ability of GRU, we pretrain our GRU on a book-collection corpus [42]. It contains more than 74M sentences. This model is pretrained in an unsupervised fashion to predict surrounding sentences according to the embedding sentences. By learning from a large number of sentences, generic textual knowledge is embedded in the GRU model with richer representations.

IV. Experiments

In this section, we compare the proposed method with various baselines, explore whether answers can be used to generate saliency maps, and evaluate our saliency-driven VQG model. The proposed method is implemented using Pytorch and tested on a PC with an i7 3.4-GHz CPU, an Nvidia Titan Pascal GPU, and a 32-GB RAM. In our experiments, all the compared networks are trained with the same amount of data (i.e., the same number of epochs). The entire training procedure (including pretraining) takes about one week before convergence.

A. Data Set and Evaluation Metrics

We evaluate the proposed top-down saliency method on the VQT-HAT data set [6], which is the only data set containing both visual questions and the corresponding saliency maps. In total, there are 58 475 training and 1374 validation question-image pairs annotated by 800 unique workers. For question generation and answering, we evaluate our models on the VQA [3] and GNQ [12] benchmarks.

1) Saliency Metrics: We follow the VQT-HAT data set [6] and evaluate the resulted saliency maps using rank correlation. Normalized scanpath saliency (NSS) [43] metric is further used to evaluate the distributions of saliency maps. For the rank-correlation metric, saliency maps are first downsampled to 14×14. Each pixel is ranked based on its spatial attention, which results in a ranked list for each saliency map. It is then compared with the ground-truth saliency map ranked list by computing their correlation. This metric shows how the machine-generated saliency maps correlate with human attention maps.

TABLE I

Comparisons With Respect to Mean Rank-Correlation and NSS (Higher Is Better). Error Bar in Rank-Correlation Indicates the Standard Error of Means. The Proposed Method "Ours-Q" Achieves Comparable Performance With Human Attention and Outperforms All Unsupervised Methods and Traditional Bottom-Up Approach in Both the Metrics

Method	Mean Rank-correlation	NSS
SAN [10]	0.249 ± 0.004	1.16
HCA-W [11]	0.246 ± 0.004	1.17
HCA-P [11]	0.256 ± 0.004	1.18
HCA-Q [11]	0.264 ± 0.004	1.21
MFB [23]	0.307 ± 0.004	1.42
MUTAN [24]	0.289 ± 0.004	1.30
Judd *et al.*[44]	0.497 ± 0.004	1.79
B1: Without Guidance	0.518 ± 0.004	1.87
B2: Concatenated features	0.571 ± 0.003	2.06
B3: Independent learning	0.598 ± 0.003	2.15
Ours-A	**0.593 ± 0.003**	**2.12**
Ours-Q	**0.620 ± 0.003**	**2.33**
Human	0.623 ± 0.003	2.36

2) VQG Metrics: Similar to [26], we use the corpus-level bilingual evaluation understudy (BLEU) and Metric for Evaluation of Translation with Explicit ORdering (METEOR) scores to measure the generated questions. BLEU is designed to measure the performance of the task of machine translation, and it is a traditional metric that achieved good correlation with human judgment. The METEOR score is another popular machine translation metric that uses F-measure to measure word matches.

3) VQA Metrics: The proposed method is flexible, and we can replace questions by answers in our framework to see whether the answers can be correctly predicted only by saliency maps. We use top-1 accuracy (Acc1) and top-5 accuracy (Acc5) to measure the predicted answers.

B. Text-Driven Top-Down Saliency Evaluations

1) Baselines: We evaluate the primal task by comparing with four state-of-the-art attention-based VQA methods: SAN [10], HCA [11], MFB [23], and MUTAN [24], and one classic saliency prediction model [44]. HCA contains three levels of saliency maps (i.e., word, phrase, and question), and we compare all of them. In addition, three baselines are compared.

1) *B1: Without Guidance:* This baseline directly learns the image-saliency mapping without question guidance using the same saliency-detection network.
2) *B2: Concatenated Features:* This baseline concatenates the embedded question features directly to the middle layer of the conv–deconv pipeline.
3) *B3: Independent Learning:* This baseline trains the text-driven top-down saliency network independently without involving the VQG process.

All the three baselines are trained without dual learning, and they are served for the ablation study of the proposed method.

2) Comparisons: Table I shows the mean rank-correlation and NSS of the validation set on the VQA-HAT data set. The last row "Human" indicates the interhuman agreement on the validation set for reference, which are computed by the average rank-correlation and

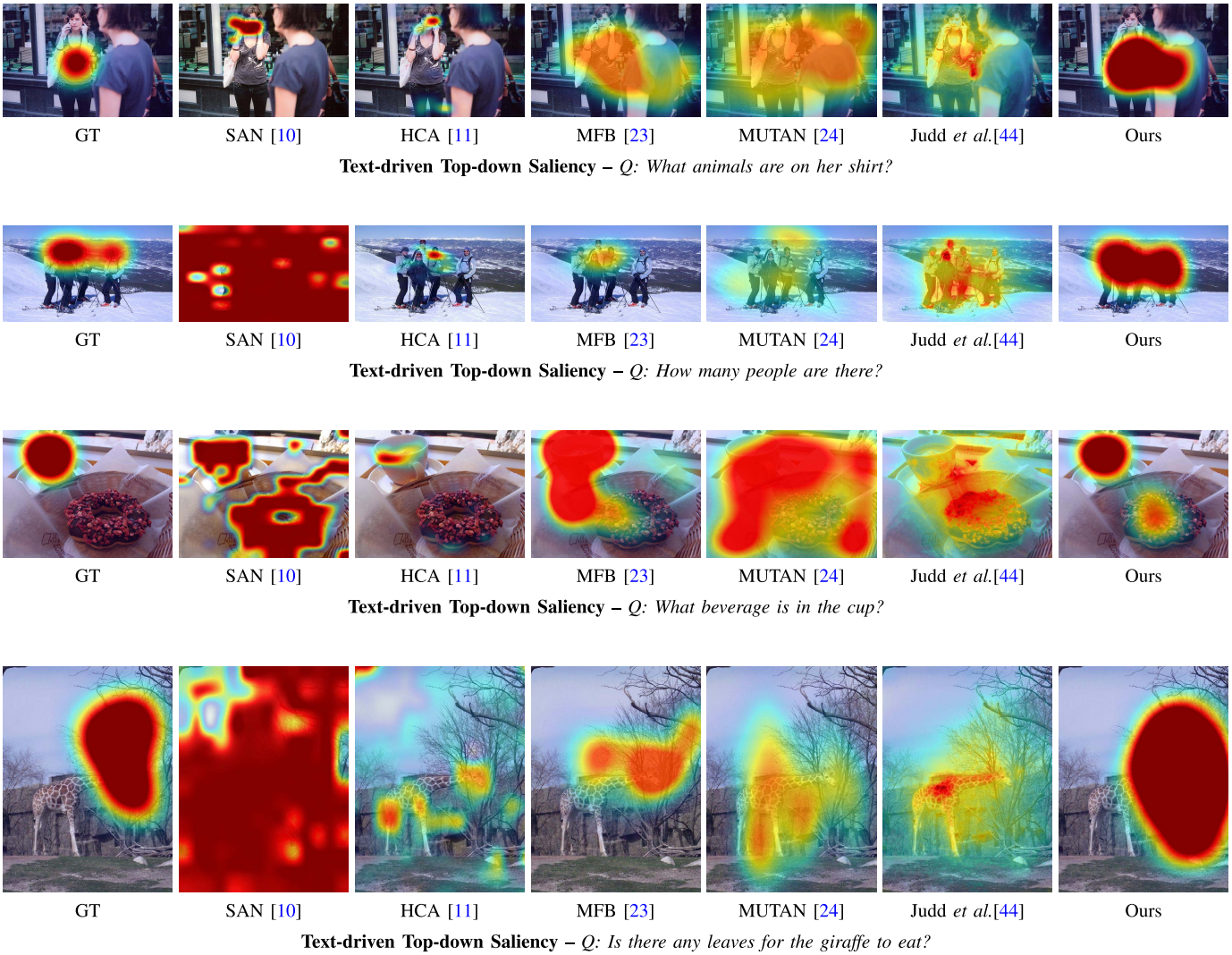

Fig. 3. Qualitative evaluations of the proposed methods on text-driven top-down saliency detection. Results show the proposed method predicts saliency better than attention-based VQA methods as well as the traditional bottom-up approach.

NSS of three saliency maps generated by three users. All reported results are the average values of these three saliency maps. Uniformly weighted regions may lead to ranking variations, and we add random noise (order of 10^{-14}) to alleviate this problem.

As can be seen, four state-of-the-art attention-based methods show minor correlations with human attention maps. In particular, MFB [23] achieves the best correlations and NSS scores due to the accurate localization of objects. MUTAN [24], on the other hand, produces widely distributed saliency maps (see Fig. 3) that show less correlations with human attention maps. Traditional saliency-detection method [44] achieves higher correlation and NSS than attention-based VQA models. The main reason is that it includes center bias in the prediction model. Our first baseline B1 learns to directly detect saliency without guidance, and it performs only slightly better than [44]. This implies that unlike learning bottom-up saliency, top-down saliency is ambiguous without the given task. Our baseline B2 adds questions as guidance by simply concatenating image and text features. The performance improves significantly due to the introduced guidance. In B3, we replace the concatenated features by the proposed parameter-prediction layer, and it leads to about 5% improvement. This indicates that the proposed parameter-prediction layer provides richer question representations than simply concatenating text features. Finally, "Ours-Q" shows the final results with dual learning. It shows mutually optimizing two tasks helps detecting the salient object, and it is able to achieve comparable result with human attention.

TABLE II

QUESTION-GENERATION PERFORMANCE ON THE GNQ BENCHMARK [12]. GNQ [12] GENERATES QUESTIONS SOLELY BASED ON IMAGES, WHILE GDQ [26] INVOLVES ANSWERS INTO THE PROCESS. OUR MODEL CAN BE GUIDED BY SALIENCY MAPS AND ANSWERS. BOTH THE SALIENCY GUIDANCE AND DUAL LEARNING BOOST THE QUESTION-GENERATION PERFORMANCE

Method	BLEU	METEOR
Image only		
GNQ [12]	0.195	0.202
Image + answer		
GDQ [26]	0.350	0.201
Image + saliency		
Ours (independent learning)	0.227	0.199
Ours (multi-task learning)	0.215	0.192
Ours	0.248	0.203
Image + answer + saliency		
Ours	**0.362**	**0.210**

GT SAN [10] HCA [11] MFB [23] MUTAN [24] Judd *et al.*[44] Ours

Answers-driven Top-down Saliency – *A: Grocery.*

GT SAN [10] HCA [11] MFB [23] MUTAN [24] Judd *et al.*[44] Ours

Answers-driven Top-down Saliency – *A: Laying down.*

Input Saliency Map Input Saliency Map

Saliency-driven Answer Prediction – *A: Giraffe.* **Saliency-driven Answer Prediction –** *A: Kite.*

Input Saliency Map Input Saliency Map

Saliency-driven VQG – *A: What kind of fruit is here?* **Saliency-driven VQG –** *Q: Which player is wearing red?*

Fig. 4. We show different applications of the proposed network: answer-driven top-down saliency detection, and saliency-driven answer prediction and question generation. These demonstrate that the proposed dual model is flexible and can be used to model different invertible applications.

Figs. 1 and 3 show examples of the proposed top-down saliency results. We can see that the proposed method is able to correctly locate relevant regions to the questions. On the contrary, unsupervised attention models may focus on wrong regions.

C. Questions Generation

The dual-task, saliency-driven question generation is also evaluated. To have a fair comparison, we evaluate the proposed method on the GNQ benchmark [12]. As a result, we use the entire VQA-HAT data set for training in this comparison. We compare with the publicly available models GNQ [12] and GDQ [26]. GNQ [12] generates questions solely based on images, while GDQ [26] involves answers into the process. We also use our model without dual learning as a baseline. Three types of guidance, saliency, answer, and their combination are used in our model. The corpus-level BLEU and METEOR scores are shown in Table II. We can see that the proposed models with saliency guidance outperform GNQ [12], which indicates that additional guidance (in this case, saliency maps) would be useful for question generation. In addition, we can see dual learning boosts the generation performance and leads to more correct questions. To verify whether the gained performance is due to additional training or the proposed dual learning, we further train our networks separately with multi-task learning. As can be seen, training with additional saliency data cannot lead to better performance. Instead, the VQG performance is slightly decreased, and it may be because two tasks cannot build an inherent connection with simple multi-task learning. GDQ [26] uses answers to generate their corresponding questions, leading to significant improvement. Comparing with the saliency map, using answer is less ambiguous and easy to predict the correct question. Finally, we combine the answer and saliency map together for guiding the question-generation process. Our model outperforms the state-of-the-art VQG models, and using saliency map provides complementary information to answer. Therefore, our model is able to generate precise and appropriate questions with contextual understanding.

TABLE III

COMPARISON ON THE SALIENCY-DRIVEN ANSWERS PREDICTION ON THE VQA VALIDATION SET [3]. FOUR STATE-OF-THE-ART METHODS ARE GUIDED BY QUESTIONS, WHILE THE PROPOSED METHODS ARE GUIDED BY SALIENCY MAPS. SURPRISINGLY, SALIENCY MAPS CONTAIN SUFFICIENT INFORMATION TO PREDICT ANSWERS WITHOUT KNOWING THE INPUT QUESTIONS. COMBINING THREE TYPES OF INFORMATION ACHIEVES SUPERIOR PERFORMANCE

Method	Acc1	Acc 5
Image only	27.8%	45.7%
Question only	47.8%	73.0%
Image + Question		
SAN [10]	55.7%	81.3%
HCA [11]	58.1%	83.9%
MFB [23]	63.2%	87.3%
MUTAN [24]	64.0%	87.6%
Ours	65.6%	88.4%
Image + Saliency		
Ours (independent learning)	42.6%	68.5%
Ours (multi-task learning)	42.1%	67.9%
Ours	47.3%	72.7%
Image + Saliency + Question		
Ours	**67.5%**	**89.2%**

The last row of Fig. 4 shows two generated questions guided by saliency maps. Based on the predicted saliency map, the first example correctly predicts the question. For the second example, the ground-truth question is "*Which player is wearing number 2?*" Given an input saliency, which focuses on the left player of the image, the proposed method is able to predict close question to the ground truth. This ambiguous prediction problem can be addressed by using additional answer information as the input.

D. Driven by Answers

Owing to our flexible structure, we can produce top-down saliency maps driven by answers instead of questions, by simply replacing the input–output questions to answer. In this experiment, we evaluate this model on the validation set of the VQA data set [3]. Again, we train our model on the entire VQA-HAT data set. However, the answers "*yes*" or "*no*" are too uninformative to predict reasonable saliency maps. We filter out images with these two answers for the VQA data set. As the answers are modeled as classification options, we simply convert the answers into a one-hot vector and concatenate to the beginning of the fully convolutional network.

As the primal task, the rank-correlation result of this answer-driven top-down saliency model on the filtered data set is shown in the third last row ("Ours-A") of Table I. Interestingly, although it is not as good as driven by questions, it generates good saliency map compared with others. The first two rows of Fig. 4 show two examples of our answer-driven top-down saliency model. Given the answers "*Grocery*" and "*Laying down*," the proposed method is able to cover the relative regions, which is similar to a visual search task.

The dual task of this model now becomes saliency-driven answer prediction. Note that in this task, answers can be predicted with or without questions. We show the prediction performance in Table III. Again, we compare with four VQA models SAN [10], HCA [11], MFB [23], and MUTAN [24] on the filtered VQA validation set. It is obvious that questions contain richer information than saliency maps, and thus, four VQA models obtain better accuracies than using saliency only. However, it is still surprising that saliency map can be used to predict answers and our method achieves reasonable performance. The proposed dual learning is also effective in this task, and achieves about 5% improvement than independent learning. Similar to VQG, multi-task learning cannot boost performance in the VQA task. We also report the results with image or question only. These indicate that using images only cannot predict correct results, while introducing saliency maps may bring hints of the corresponding questions. The result of our model trained with image and question also demonstrates the benefits of the proposed dual setting.

The third row of Fig. 4 shows two examples of the predicted answers. It is not easy to accurately guess the question of this image, and therefore difficult to accurately predict the answer without extra information. In this case, the questions of the first image are: "*Are the deer afraid of the giraffe?*" and "*Which way is the giraffe facing?*" Our system predicts "*Giraffe*" according to the input saliency map, which is not the correct answer for both questions. This demonstrates that questions are still the most important information in VQA. In the second example, the predicted salient region is highly correlated with the question "*What is flying in the sky?*," and our system predicts the correct answer.

Finally, we train an additional VQA model with three types of inputs: image, question, and text-driven saliency map. The input image and the saliency map are concatenated, and the question is encoded by the parameter-prediction layer [similar to Fig. 2 (left)]. As can be seen in Table III, the combined network performs superior to the four state-of-the-art models. Unlike unsupervised models that implicitly learn attention in the process, our model explicitly exploits visual question (VQ)-saliency and it performs better, which implies that learning human-like VQ-saliency benefits predicting correct answers.

V. CONCLUSION

We present the first attempt to explore the relationship between text generation and top-down saliency. It is formed as dual tasks between top-down saliency and text generation. Specifically, we propose a dual learning network that trains the primal task and the dual task simultaneously, and it exploits the bidirection relationship between two tasks. In order to encode textual information to the fully convolutional network, we introduce the parameter-prediction layer, where the parameters of the top-down saliency network are adaptively determined by the input questions. Experiments show that our top-down saliency method achieves similar performance to human. Furthermore, we demonstrate an accurate saliency response that benefits the generation of question and answer.

REFERENCES

[1] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, Jul. 2015, pp. 2048–2057.
[2] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. CVPR*, Jun. 2016, pp. 4651–4659.
[3] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. ICCV*, Dec. 2015, pp. 2425–2433.
[4] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Proc. CVPR*, Jul. 2017, pp. 6904–6913.
[5] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.
[6] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Comput. Vis. Image Understand.*, vol. 163, pp. 90–100, Oct. 2017.
[7] G. Li and Y. Yu, "Contrast-oriented deep neural networks for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6038–6051, Dec. 2018.
[8] Z. Liu, X. Wang, and S. Bu, "Human-centered saliency detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1150–1162, Jun. 2016.

[9] S. Huo, Y. Zhou, W. Xiang, and S. Y. Kung, "Semisupervised learning based on a novel iterative optimization model for saliency detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 225–241, Jan. 2019.

[10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. CVPR*, Jun. 2016, pp. 21–29.

[11] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.

[12] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, "Generating natural questions about an image," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Aug. 2016, pp. 1802–1813.

[13] S. Zhang, L. Qu, S. You, Z. Yang, and J. Zhang, "Automatic generation of grounded visual questions," in *Proc. IJCAI*, 2017, pp. 4235–4243.

[14] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, and X. Wang, "Visual question generation as dual task of visual question answering," Sep. 2017, *arXiv:1709.07192*. [Online]. Available: https://arxiv.org/abs/1709.07192

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, Jun. 2016, pp. 2921–2929.

[16] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 543–559.

[17] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "SUN: Top-down saliency using natural statistics," *Vis. Cognit.*, vol. 17, nos. 6–7, pp. 979–1003, 2009.

[18] J. Yang and M.-H. Yang, "Top-down visual saliency via joint CRF and dictionary learning," in *Proc. CVPR*, Jun. 2012, pp. 2296–2303.

[19] S. He and R. W. Lau, "Exemplar-driven top-down saliency detection via deep association," in *Proc. CVPR*, Jun. 2016, pp. 5723–5732.

[20] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proc. CVPR*, Jul. 2017, pp. 7206–7215.

[21] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. CVPR*, Jun. 2016, pp. 30–38.

[22] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. CVPR*, Jun. 2016, pp. 4995–5004.

[23] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. ICCV*, Oct. 2017, pp. 1839–1848.

[24] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proc. ICCV*, Oct. 2017, pp. 2612–2620.

[25] P. Anderson *et al.*, "Bottom-Up and top-down attention for image captioning and visual question answering," in *Proc. CVPR*, Jun. 2018, pp. 6077–6086.

[26] U. Jain, Z. Zhang, and A. Schwing, "Creativity: Generating diverse questions using variational autoencoders," in *Proc. CVPR*, Jul. 2017, pp. 6485–6494.

[27] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Visual saliency for image captioning in new multimedia services," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 309–314.

[28] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," *ACM Trans. Multimedia Comput. Commun., Appl.*, vol. 14, no. 2, p. 48, May 2018.

[29] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[30] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proc. CVPR*, Jun. 2016, pp. 3994–4003.

[31] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[32] D. He *et al.*, "Dual learning for machine translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 820–828.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[35] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," Dec. 2014, *arXiv:1412.3555*. [Online]. Available: https://arxiv.org/abs/1412.3555

[36] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. 6th Int. Conf. Neural Inf. Process. Syst.*, 1993, pp. 737–744.

[37] Y. Xia, T. Qin, W. Chen, J. Bian, N. Yu, and T.-Y. Liu, "Dual supervised learning," in *Proc. ICML*, Aug. 2017, pp. 3789–3798.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[40] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT Tech. Rep., 2012.

[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, Jul. 2015, pp. 448–456.

[42] R. Kiros *et al.*, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3294–3302.

[43] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 8, pp. 2397–2416, 2005.

[44] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2106–2113.