SegEQA: Video Segmentation Based Visual Attention for Embodied Question Answering

Haonan Luo[1,2,*], Guosheng Lin[2], Zichuan Liu[2], Fayao Liu[3], Zhenmin Tang[1], Yazhou Yao[1,4]
[1]Nanjing University of Science and Technology, [2]Nanyang Technological University,
[3]Institute for Infocomm Research A*STAR, [4]IIAI
lhn@njust.edu.cn, gslin@ntu.edu.sg, zliu016@e.ntu.edu.sg,
fayaoliu@gmail.com, tzm.cs@njust.edu.cn, Yazhou.yao@outlook.com

Abstract

Embodied Question Answering (EQA) is a newly defined research area where an agent is required to answer the user's questions by exploring the real world environment. It has attracted increasing research interests due to its broad applications in automatic driving system, in-home robots, and personal assistants. Most of the existing methods perform poorly in terms of answering and navigation accuracy due to the absence of local details and vulnerability to the ambiguity caused by complicated vision conditions. To tackle these problems, we propose a segmentation based visual attention mechanism for Embodied Question Answering. Firstly, We extract the local semantic features by introducing a novel high-speed video segmentation framework. Then by the guide of extracted semantic features, a bottom-up visual attention mechanism is proposed for the Visual Question Answering (VQA) sub-task. Further, a feature fusion strategy is proposed to guide the training of the navigator without much additional computational cost. The ablation experiments show that our method boosts the performance of VQA module by 4.2% (68.99% vs 64.73%) and leads to 3.6% (48.59% vs 44.98%) overall improvement in EQA accuracy.

1. Introduction

Due to the development of deep learning techniques, many relatively low-level visual tasks like classification [12], detection [28], and segmentation [6, 19, 21] have become more and more successful. With the assistance of these mature visual tasks, researchers now pay more attention to various high-level visual reasoning tasks like scene graph generation [37], image captioning [33], visual question answering (VQA) [1], and navigation [34]. Embodied Question Answering (EQA), as one combination of VQA and navigation, also attracts researchers' attention due to its wide potential application areas such as self-driven mobile, in-home robots, and personal assistants. An agent in the EQA task is aimed to answer the user's questions (*e.g.* 'What color is the television?') by interacting with the real-world environment. The agent is arbitrarily initialized at a location in an open environment and resolve user's questions by continuously exploring environment and collecting visual information.

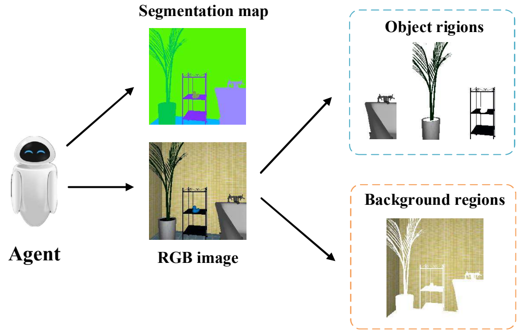

Figure 1. An illustration of the extracting process of our segmentation based visual attention. An agent captures RGB images via its monocular camera and obtains segmentation maps of the current environment through a segmentation network. Then the current scene is split into several regions by the guide of the segmentation masks. We then take these regions as bottom-up attention for subsequent operations.

An EQA system generally consists of two modules: VQA module and navigation module. Given an initial image captured from the environment, the navigation module progressively explores the environment by movements and continuous visual perceptions. The collected visual information and the original question will be further analyzed by the VQA module and convert to a human-interpretable

[*]This work was done when H. Luo was a visiting student in Nanyang Technological University. Corresponding authors: G. Lin & Z. Tang.

answer.

EQA is a challenging task which requires cross-domain knowledge including natural language processing, computer vision and decision process to mine the helpful information from the complex environment. Since the localization, map of the environment and visually semantic information are not directly obtainable, the agent has some difficulties in producing decisions at test phase. Existing EQA frameworks directly apply Convolutional Neural Networks (CNN) [12, 15, 30, 18, 13] to original RGB information of the environment. Without the help of semantic information of the video, existing approaches can not explicitly distinguish the interested object from the background and thus may introduce ambiguities caused by complicated environments. Besides, directly projecting an image to vector representation could reduce its discriminability due to the excessive loss of local details.

To enhance the performance of existing EQA systems, we propose a segmentation based visual attention mechanism which improves the discriminability of visual representations extracted by CNNs. Our segmentation based visual attention injects high-level semantic information of the video into the framework. Therefore our approach has the ability of characterizing local details, while being less susceptible to ambiguities caused by complicated environments.

The extracting process of the segmentation based visual attention is shown in Figure 1. In the VQA phase, instead of directly applying CNN to extract a single visual representation of an image, we first decompose an image into sub-images by the guide of semantic segmentation masks. These sub-images are then encoded into visual representations by a CNN structure, and we use these extracted region features to construct our segmentation based visual attention. As the features of these sub-images contain rich semantic information, using segmentation attention in the VQA module provides detailed object-level information for generating better answers. In the navigation phase, we adopt the semantic information provided by the segmentation masks to guide the training of the navigator, and a behavioral cloning simulation method is applied to jointly optimize the entire system in an end-to-end manner. The experiment demonstrates that our proposed segmentation based visual attention mechanism brings significant performance improvement. Compared with existing methods, our approach boosts the VQA accuracy by 4.2% (68.99% vs 64.73%) and improve the overall EQA accuracy by 3.6% (48.59% vs 44.98%). The contributions of this paper are summarized as follows:

- We develop a segmentation based visual attention mechanism for the VQA module, which significantly improves the performance of the VQA task.

- A segmentation-assisted path-finding algorithm is developed for the navigation module to improve the navigation performance.

- We develop a high-speed video segmentation framework to extract semantic information from videos. The extracted semantic information is used to improve the EQA performance.

- Our approach using segmentation based semantic information is able to improve VQA accuracy by 4.2% (68.99% vs 64.73%), and improve the overall EQA accuracy by 3.6% (48.59% vs 44.98%).

2. Related Work

EQA: EQA task has attracted considerable attention recently. This topic was first introduced by Das et al. in [8]. They built a virtual environment called House3D for EQA tasks. Following the House3D, Gordon et al. [10] presented an interactive question answering system, which automatically search for paths and interact with specific objects. Anderson et al. [2] address the problem of how to use complex human language commands to guide the agent to perform corresponding actions in a photo-realistic environment. Recently, Das et al. [9] propose an EQA approach using hierarchical semantic control.

VQA: A Visual Question Answering (VQA) system aims to answer a question related to a given image. Zhou et al. [40] propose a baseline model called iBOWIMG, which simply concatenates features of the image and question to predict the answer. Ma et al. [23] propose a CNN-only model which not only learns the representation of the image and question, but also learns their inter-modal interactions. Ren et al. [27] present a LSTM based model, which uses the final layer of VGG network to obtain the image encoding. Another representative work is from Kafle et al. [17] who propose a Bayesian framework which has the ability to predict the form of the answer from the question.

Bottom-up attention: One relevant work is from Zhu et al. in [42] where they use object bounding boxes to obtain object-level semantic information to improve the performance of the VQA task. Another relevant work is the method proposed by Anderson et al. [1], in which they use fast R-CNN method to automatically extract bounding boxes of all potential objects in the image as the bottom-up attention. Teney et al. describes the detailed implementation of [1] in [32].

3. Proposed Methods

In this section, we present our segmentation based visual attention mechanism for the Embodied Question Answering(EQA) tasks. It exploits the potential semantic information and refines the scope of the system to object-level.

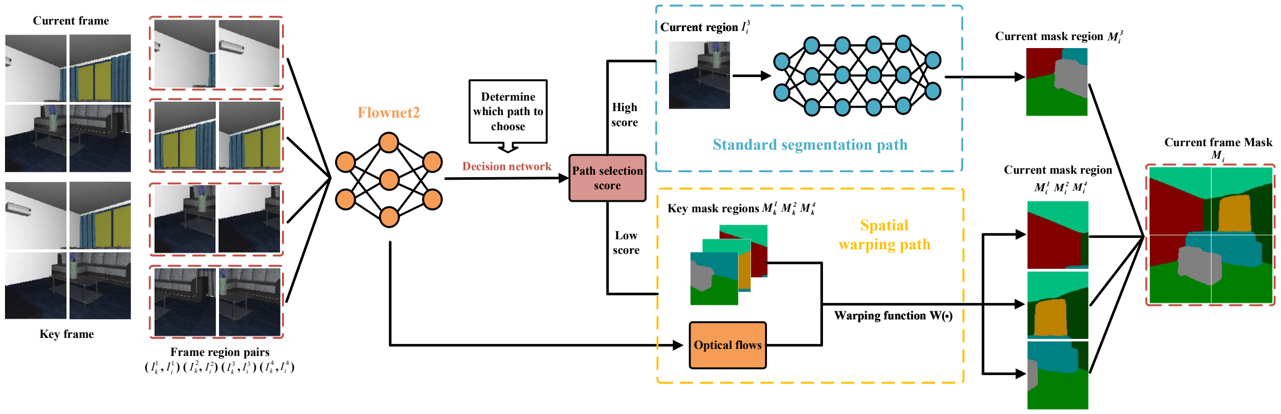

Figure 2. The framework of our high-speed video semantic segmentation method. Firstly, the input frames are divided into several separate regions. Secondly, all of the region pairs are fed into a shallow network to distill the difference of region pairs between I_k and I_i. A decision network (DN) analyzes these distilled differences and evaluates the path-selection scores for every region separately. Finally, current frame regions are forwarded to different paths to generate their regional segmentation masks based on the path-selection score produced by DN.

Applying our proposed algorithm to the EQA task, the accuracy of both VQA and navigation modules are both significantly improved, leading to the overall performance improvement for the whole EQA system. Section 3.1 presents a high-speed video segmentation framework for real-time segmentation masks extraction. By taking the extracted segmentation masks as semantic clues, section 3.2 describes the bottom-up attention mechanism which is used in the VQA module. In section 3.3, a segmentation-assisted navigation module is given in detail. All algorithms described in this article are deployed in the House3D [36] virtual environment and EQA v1 [8] question dataset.

3.1. A High-Speed Video Segmentation Framework

In order to explore the potential semantic information in videos and use them to assist the EQA task, we choose segmentation masks as semantic clues to generate visual attention. In reality, agents cannot obtain semantic segmentation maps directly from the camera. Thus a predicting network is needed to produce the semantic segmentation maps of the current environment. Since the agent keeps moving, the EQA system requires higher level of time efficiency for ensuring the consistency of subsequent operations. Directly applying general image segmentation methods to each frame of the video usually leads to massive time consumption. Therefore, it is necessary to develop an efficient framework perform high-speed video segmentation for EQA tasks. Inspired by recent achievements of semantic segmentation, we develop an efficient framework for the extraction of segmentation masks of videos. The overview of our proposed framework is illustrated in Figure 2.

Firstly, the input frames are divided into several separate regions, for example, four regions. We adopt keyframe scheduling policy used by DFF [41] to capture one keyframe in every l $(l = 10)$ consecutive frames. In Figure 2, I_i represents the current frame, I_k represents its neighboring keyframe, and M_i represents the segmentation mask of the current frame. Then, all of the region pairs are fed into a shallow network to distill the difference of region pairs of I_k and I_i. Here, we use Flownet2 [14] to perform this task.

Then the decision network (DN) analyzes these differences and output a path-selection score for each region separately. We compare the resulting path-selection score of each region against a predetermined threshold. If the score is lower than the threshold, the corresponding region will be input to a pre-trained standard image segmentation network (*e.g.*, RefineNet-152). Otherwise, it will be fed to a spatial warping unit. DN is a simple network consisting of only a single convolutional layer and three fully connected layers. The role of DN is to evaluate the similarity of the region pair of I_k and I_i. If the regions are sufficiently similar, spatial warping will be performed to generate satisfactory segmentation for the regions.

Finally, current frame regions are forwarded to different paths for generating their regional segmentation masks based on the path-selection score produced by DN. For the path of the spatial warping unit, a particular warping function W(·) [41] is employed to process the optical flow F^r and the segmentation map of keyframe region M_k^r to generate a new segmentation mask M_i^r for the current frame region. The formulation of the spatial warping function is defined as:

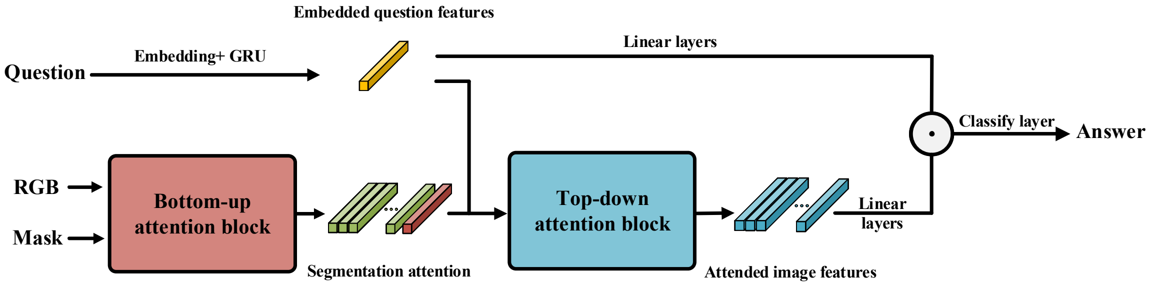

Figure 3. An overall illustration of our SegVQA module. This module takes the question and the environment state (RGB images, segmentation maps) as inputs, and predict an answer to the question according to the state of current environment. This module contains two main sub-modules: the bottom-up attention block and the top-down attention block. Details of these two modules are described in Figure 4 and Figure 5, respectively.

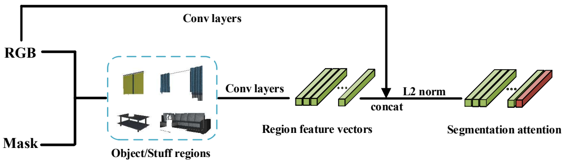

Figure 4. Details of the bottom-up attention block. This submodule takes RGB images and extracted segmentation masks as input to produce segmentation attention.

$$C_i^r(p) = \sum_q G(q, p + \triangle p) C_k^r(q). \quad (1)$$

Here C_k^r is an intermediate feature map of the keyframe region, and C_i^r is the corresponding intermediate feature map of the current frame region. The location p in the current frame region corresponds to the location $p + \triangle p$ in the keyframe region, q numerates all spatial location in the feature map. $G(\cdot, \cdot)$ represents the bilinear interpolation kernel. Note that G operates on two dimensions and it can be decomposed into two one-dimensional kernels as:

$$G(q, p + \triangle p) = g(q_x, p_x + \triangle p_x) g(q_y, p_y + \triangle p_y), \quad (2)$$

where $g(a, b) = \max(0, 1 - |a - b|)$.

3.2. SegVQA: VQA with Segmentation Attention

For a virtual environment such as House3D that contains both indoor and outdoor scenes, the environmental images obtained by the agent are commonly diverse and changeable. When analyzing these images, the system may be disturbed by unclear object boundaries and background textures, which may decline the accuracy of answer prediction.

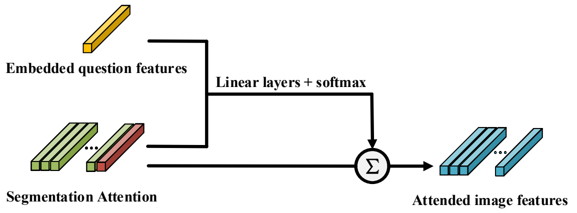

Figure 5. Details of the top-down attention block. This sub-module takes features of question and segmentation attention as input, and produces attended image features for subsequent operations.

Therefore, with the extracted semantic segmentation maps of the environment, we constructs a segmentation based bottom-up visual attention mechanism in the VQA module to help EQA tasks.

When the agent arrives at the target position, the RGB image I of the current environment will be obtained by the monocular camera carried by the agent, and at the same time, the corresponding semantic segmentation map S will be generated by the proposed fast video segmentation framework described in the previous section. As is shown in Figure 3, taking I and S as input, a bottom-up attention block is proposed to extract segmentation based bottom-up attention. The output of this block is denoted as segmentation attention. On the other side, user's question is turned into a vector representation using a look-up table, which is initialized with the pretrained Global Vectors word embedding [26]. The resulting sequence of word embedding is then passed through a Recurrent Gated Unit (GRU [7]) for producing embedded question features. By taking the segmentation attention and the embedded question features as input, a top-down attention block is used to weight the features of bottom-up attention by the values of top-down attention weights. The output of this block is referred as attended image features. Finally, the representations of the question and the attended image features are passes to convolution layers. They are then merged in an element-wise manner to produce the final answer. The detail of the bottom-up attention block and top-down attention block are described in Figure 4 and Figure 5, respectively.

In Figure 4, according to the category labels in S, I is disassembled into several sub-images, and each of them contains different object or background categories. These sub-images, re-scaled into the same resolution, are merged into a matrix with b channels, and then the matrix is input to a CNN network, resulting in b set of region feature vectors. This CNN network follows the CNN structure described in [8] for generating image features, and it is pretrained in ImageNet. We also feed the whole RGB image of the current environment into the same CNN network to extract its feature vector. The whole-image feature vector and the region feature vectors are then merged together to generate the the bottom-up attention feature vectors (segmentation attention) v_i (i=1,...,$b+1$).

In Figure 5, the top-down attention block is similar to many modern VQA models (e.g. [42, 38, 5, 16, 39, 3, 4]). Specifically, the feature vectors of segmentation attention are concatenated with the question embedding q. q is repeated to match the number of segmentation attention vectors before concatenation. Then they are passed through a nonlinear layer f_a (Relu layer) and a linear layer to obtain a scalar attention weight a_i which corresponds to a particular object/stuff region. The segmentation attention feature vectors are then multiplied with normalized a and summed to obtain a single 3200-sized vector for representing the attended image features. Formally,

$$a_i = w_a^{\mathsf{T}} f_a([v_i, q]), \tag{3}$$

$$\hat{a}_i = softmax(a_i), \tag{4}$$

$$\hat{v} = \sum_{i=1}^{b} \hat{a}_i v_i, \tag{5}$$

where v_i is the bottom-up attention vectors, w_a is a learned parameter vector, $\hat{v}$ denotes the attended image features.

3.3. Segmentation-Assisted Navigation Module

The training of our navigation module can be divided into two steps. Firstly, the navigator is pretrained by segmentation-assisted behavioral cloning algorithm under the guidance of the expert path. Secondly, the entire architecture which includes navigation and answering modules are jointly fine-tuned using reinforce policy gradients.

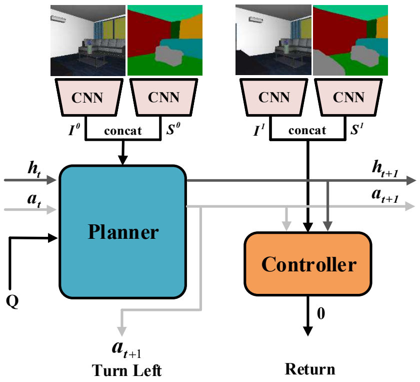

Figure 6. An overview of our SegNavigation module. It takes the question and the current environment state (RGB images, segmentation maps) perceived by the agent as input, and trains a navigator which generates the next action based on the present environment.

3.3.1 Segmentation-Assisted Imitation Learning

We employ the approach of intermediate feature fusion to improve the navigator. The proposed algorithm is denoted as SegNavigation. Specifically, the agent obtains RGB image I and extracts the corresponding segmentation map S using our fast video segmentation framework presented in section 3.1. I and S are sent to a standard CNN network to extract image features, and then these image features are merged into a single vector which will be passed to subsequent navigation operations latter.

Same as in EQA [8], we build the navigator using the Adaptive Computation Time (ACT) [11] algorithm. It decomposes the navigation process into a ‘planner’ which selects actions, and a ‘controller’ which executes these primitive actions for a varying number of times before returning control to the planner. We instantiate the planner as an LSTM and instantiate the controller as a multi-layer perceptron with one hidden layer. After obtaining the environmental information from the system, the controller determines whether to execute the current policy or not. If the decision is yes, the agent repeats the action according to the strategy generated by the planner. Otherwise, control is returned to the planner and a new action strategy will be chosen.

Correct navigation is not unique for most questions. Therefore, we use the shortest path from the agents spawn location to the target as an expert guidance for the imitation learning algorithm. Given the history encoding, question encoding, current frame and segmentation frame, the model is trained to predict the action that would keep it on the shortest path. In the training process, we use a cross-entropy loss and train the model for 25 epochs with a batch size of 20. Figure 6 provides an overview of our SegNavi-

datasets	pascalcontext [24]	sunrgbd [31]	nyud v2 [29]
mIoU	67.88%	70.12%	72.59%

Table 1. Comparison of three candidate pre-training datasets. We randomly chose 500 scene images in House3D to construct the test set for the calculation of mIoU.

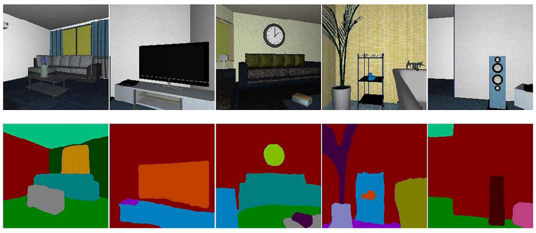

Figure 7. Visualization results of our high-speed video segmentation framework. The first row shows the original RGB image, and the second row shows our segmentation results.

gation module.

3.3.2 Target-aware Fine Tuning

In the EQA task, SegVQA and SegNavigation modules mentioned above need to work together to complete the task. Similar to the work of [8], to avoid the confusion of noisy or invalid information to the answer prediction, we freeze the visual question answering module while fine-tuning the navigator. Two types of reward signals to the navigator are provided: the question answering accuracy achieved at the end of the navigation and a reward shaping [25] term that gives intermediate rewards for getting closer to the target. We train the agent with reinforce policy gradients [35] with the average of two answer rewards. For the imitation learning setting, we follow a process of increasing distance between spawn and target locations from 10 to 50.

4. Experiment

The goal of intelligent agents in the EQA task is to answer the questions correctly. The performance of the modules which constitutes the entire system is crucial since they directly determine the final accuracy of question answering. In this section, a quantitative evaluation of our segmentation module, SegVQA module, SegNavigation module, and the overall system will be given. Comprehensive ablation analysis will be performed. Furthermore, the time complexity is also analyzed in the end.

4.1. Video Segmentation

In the EQA task, it is expected that rich category information is captured in segmentation masks, and the whole scene is fully segmented. Therefore, we choose three full-scene segmentation datasets pascalcontext [24], sunrgbd [31], and nyud v2 [29] as candidates for the pre-training of our video segmentation framework. After the pre-training process, 2000 randomly selected scene images in House3D are used to fine tune the model. As shown in Table 1, among the three candidate pre-training databases, nyud v2 [29] performs the best, because it contains abundant helpful information of indoor scenes which is similar to the scenes of House3D environment.

Some visualized results of our high-speed video segmentation are given in Figure 7, which shows the predicted segmentation masks basically retain outlines and main areas of potential targets. Therefore they can provide necessary semantic information for the subsequent operations in the EQA system.

	Accuracy	MR
VQA [8](baseline)	64.73%	2.01
SMem [38]	61.86%	2.26
Visual7W [42]	61.99%	2.24
Co-Attention [22]	63.77%	2.05
SAN [39]	64.39%	2.01
Up-Down [1]	66.04%	1.96
SegVQA(ours)	**68.99%**	**1.89**

Table 2. Quantitative evaluations of EmbodiedQA agents on question answering metrics for the EQA v1 test set. The performance of our SegVQA is significantly better than other VQA methods in terms of accuracy and MR.

4.2. Evaluation of SegVQA Module

In the VQA module, we term our segmentation based visual attention question answering algorithm as SegVQA. Accuracy and MR (Mean rank) are used as evaluation measures. We compare our SegVQA with the baseline VQA algorithm used in [8], SMem [38], Visual7W [42], Co-Attention [22], SAN [39], and Up-Down proposed in [1]. In Table 2, all experimental ablations are performed based on the averaged results of multiple Q&A processes. The comparison results show that the performance of SegVQA is significantly better than other existing VQA methods. Comparing to our baseline, the accuracy enhancement is above 4.2% (68.99% vs 64.73%), and the MR decrease is more than 0.1.

4.3. Ablation Analysis of SegVQA

The proposed segmentation-based visual attention mechanism is not the only way of utilizing video's semantic information to help the EQA task. There are some other relatively simple feature fusion approaches which can be applied to exploit the information encoded in segmentation masks and RGB images. These simple feature fusion ap-

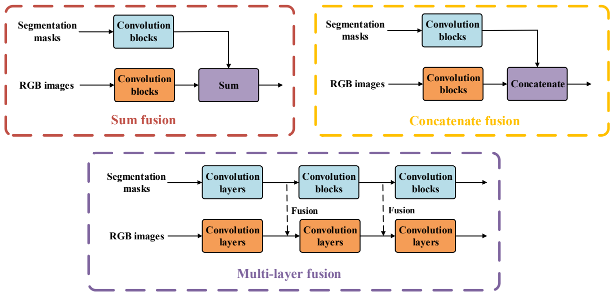

Figure 8. Illustration of different feature fusion approaches. All three approaches take RGB images and segmentation masks as input. These inputs are then forwarded into separate convolution blocks. The difference of these approaches is the way of intermediate feature fusion.

	Accuracy	MR
VQA [8](baseline)	64.73%	2.01
Sum fusion(baseline)	64.94%	2.01
Concatenate fusion(baseline)	65.69%	1.98
Multi-layer fusion(baseline)	65.81%	1.97
SegVQA(ours)	**68.99%**	**1.89**

Table 3. Analysis results of different integration approaches of segmentation masks for the VQA module. All segmentation mask integration methods obtain certain performance improvements, while our algorithm achieves the largest performance gain.

proaches include sum fusion, concatenate fusion and multi-layer fusion. The details of these three fusion approaches can be found in Figure 4.3. To further prove the validness of our approach, we conduct an analysis experiment to compare these feature fusion approaches with our SegVQA. As is shown in Table 3, all approaches that utilize segmentation masks gain some performance improvement than the original VQA (baseline) [8], which indicates that segmentation masks are helpful for the VQA module. Due to the simplicity of these fusion approaches, their performance improvements are not significant. In contrast, our segmentation based visual attention SegVQA shows large improvement over the baseline. This clearly verify the superiority of our proposed SegVQA in exploring pixel-level semantic information for the EQA task.

4.4. Evaluation of SegNavigation Module

In the navigation module, our segmentation-assisted navigation algorithm is termed as SegNavigation. We evaluate the navigation performance in the EQA task by reporting the changes in distance to target from the initial to the final position ($d_\triangle$) and the distance to the target object at navigation termination (d_T). To overcome the difficulty of the task at test time, we spawn the agent 10, 30, or 50 actions away from the target and report each metric for the 10, 30, 50 settings. Table 4 lists the comparison of our SegNavigation against the baseline navigation algorithm used in [8] and IQA proposed in [10]. For the comparison with IQA, we reconstructed partial components of IQA algorithm for running in House3D environment. Two oracles are also used for comparison: HumanNav denotes goal-driven navigations by AMT workers remotely operating the agent, and ShortestPaths+VQA denotes QA performance when the shortest paths are available at test time. In the navigation process, one forward step corresponds to at most 0.25 meters, and it takes 40 turns to turn 360. Backward and strafe motions are not allowed. All the experimental ablations are calculated based on the averaged results of multiple navigation processes, and all results are measured in meters along the shortest path to the target. It can be seen from Table 4 that our SegNavigation algorithm is leading at $d_\triangle$ and d_T in all settings.

4.5. Ablation Analysis of Overall System

After the training of VQA and navigation modules, a reinforcement learning algorithm is used to fine tune the entire system to improve the answering performance. In the case the agent is spawned at 50 actions away from the target, we evaluate different combinations of the algorithm proposed in this paper and the baseline algorithms used in [8]. Table 5 shows the performance of different algorithm combinations in terms of accuracy and MR. It shows that all combinations with our proposed modules bring performance improvement, and the combination of our proposed two modules achieves the best performance.

4.6. Analysis of Video Segmentation

High-speed video segmentation is an essential component of the whole system. Large time consumption in video segmentation will make it difficult for agents to coherently explore the surroundings in practical applications. Our method is able to perform high-speed video segmentation. We use a desktop machine equipped with a GPU 1080Ti and a CPU i7-8700K to measure the time consumption. For images in House3D environment, we achieve a frame rate of 45 FPS for optical flow extraction and 57 FPS for the remaining segmentation components, which is much faster than the per-frame segmentation strategy.

The segmentation performance of our approach is similar to the the per-frame segmentation method which performs standard image segmentation on every single frame of the video. Here we use RefineNet [20] as the image segmentation method and measure the performance on our collected EQA images. The mIoU of our high-speed video segmentation method only shows a slight decline of 0.7% than the standard image segmentation method with a per-frame segmentation strategy. In our VQA task, using segmentation masks generated by our fast video segmentation ap-

	$d_{\triangle}$-10	$d_{\triangle}$-30	$d_{\triangle}$-50	d_T-10	d_T-30	d_T-50
Navigation [8](baseline)	-1.35	0.03	1.51	0.46	1.50	2.74
SegNavigation(ours)	**-1.22**	**0.15**	**1.62**	**0.34**	**1.31**	**2.52**
IQA [10]	-1.69	-0.56	-0.03	1.03	1.79	3.45
HumanNav(Oracle)	0.44	1.62	2.85	0.81	0.81	0.81
ShortestPath+VQA(Oracle)	0.85	2.78	4.86	-	-	-

Table 4. Quantitative evaluations of navigation processing on the EQA v1 test set. HumanNav and ShortestPath+VQA act as two upper bounds which exhibit the best performance that oracles can achieve. It shows that our SegNavigation performs significantly better than other navigation methods in terms of $d_{\triangle}$ and d_T in all settings.

	Accuracy	MR
VQA [8](baseline) + Navigation [8](baseline)	44.98%	2.33
SegVQA(ours) + Navigation [8](baseline)	47.25%	2.29
VQA [8](baseline) + SegNavigation(ours)	45.75%	2.32
SegVQA(ours) + SegNavigation(ours)	**48.59%**	**2.24**

Table 5. Ablation comparisons of different algorithm combinations. The combination of the two modules proposed here performs the best, and any combination with our proposed modules brings performance gain.

proach shows a very similar performance to per-frame segmentation – it only shows a decline of 0.09% (68.99% vs 69.08%) in question answering accuracy.

	Accuracy	MR
Bounding boxes attention [1]	66.91%	1.94
SegVQA(ours)	68.99%	1.89
VQA(baseline)	64.73%	2.01
Upper bound(GT segmentation)	69.87%	1.85

Table 6. Comparison for discussion.

4.7. Some Discussion

Recently researchers attempt to use bounding boxes to generate attention regions in VQA tasks. Comparing with bounding boxes, segmentation has several advantages in the EQA task. (1) Segmentation masks are more accurate than bounding boxes to locate objects. (2) It is difficult for bounding boxes to recognize background/stuff regions (such as wall, ground, etc.) which are useful for EQA tasks. In contrast to bounding boxes, segmentation masks are able to accurately locate both things (objects) and stuff (amorphous background regions). The first and second rows in Table 6 show that our segmentation based attention mechanism performs better than bounding boxes based attention [1] in the EQA task.

Our main focus is on how to explore semantic segmentation to help the EQA task. The upper bound performance of our segmentation based attention mechanism can be achieved by using ideal segmentation maps which are rendered by House3D environment. The fourth row of Table 6 lists the upper bound of our VQA performance by directly utilizing ground truth segmentation maps. It shows that using ground truth segmentation maps significantly outperforms the baseline. This verifies the importance of segmentation based scene understanding for the EQA task.

5. Conclusion

We have present a novel video segmentation based visual attention mechanism to improve the performance of EQA systems. Our approach first extracts object-level semantic features of videos by a high-speed video segmentation framework, and then incorporate the semantic features into the VQA and navigation modules to gain improvements. Finally, the two modules are combined and fine-tuned to perform the EQA task. Because of the effectiveness of the proposed components for using pixel-level semantic information, our approach leads to a significant improvement for each module and the overall EQA system.

Acknowledgements. The authors would like to thank the financial support from the program of China Scholarships Council (No.201806840059). This work is partly supported by the National Research Foundation Singapore under its AI Singapore Programme [AISG-RP-2018-003] and the MOE Tier-1 research grant [RG126/17 (S)]. We would like to thank NVIDIA for GPU donation.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, page 6, 2018.

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real

environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018.

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[8] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Neural modular control for embodied question answering. *arXiv preprint arXiv:1810.11181*, 2018.

[10] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. *arXiv preprint arXiv:1712.03316*, 1, 2017.

[11] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2462–2470, 2017.

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[16] Aiwen Jiang, Fang Wang, Fatih Porikli, and Yi Li. Compositional memory for visual question answering. *arXiv preprint arXiv:1511.05676*, 2015.

[17] Kushal Kafle and Christopher Kanan. Answer-type prediction for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4976–4984, 2016.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Guosheng Lin, Fayao Liu, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[20] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[21] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2016.

[22] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[23] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *AAAI*, volume 3, page 16, 2016.

[24] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[25] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[27] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[29] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[32] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017.

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[34] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019.

[35] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[36] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018.

[37] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017.

[38] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[39] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[40] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

[41] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2017.

[42] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.