Weakly Supervised Relative Spatial Reasoning for Visual Question Answering

Pratyay Banerjee Tejas Gokhale Yezhou Yang Chitta Baral
Arizona State University
{pbanerj6, tgokhale, yz.yang, chitta}@asu.edu

Abstract

Vision-and-language (V&L) reasoning necessitates perception of visual concepts such as objects and actions, understanding semantics and language grounding, and reasoning about the interplay between the two modalities. One crucial aspect of visual reasoning is spatial understanding, which involves understanding relative locations of objects, i.e. implicitly learning the geometry of the scene. In this work, we evaluate the faithfulness of V&L models to such geometric understanding, by formulating the prediction of pair-wise relative locations of objects as a classification as well as a regression task. Our findings suggest that state-of-the-art transformer-based V&L models lack sufficient abilities to excel at this task. Motivated by this, we design two objectives as proxies for 3D spatial reasoning (SR) – object centroid estimation, and relative position estimation, and train V&L with weak supervision from off-the-shelf depth estimators. This leads to considerable improvements in accuracy for the "GQA" visual question answering challenge (in fully supervised, few-shot, and O.O.D settings) as well as improvements in relative spatial reasoning. Code and data will be released here.

Figure 1. GQA [21] requires a compositional understanding of objects, their properties, and spatial locations (underlined).

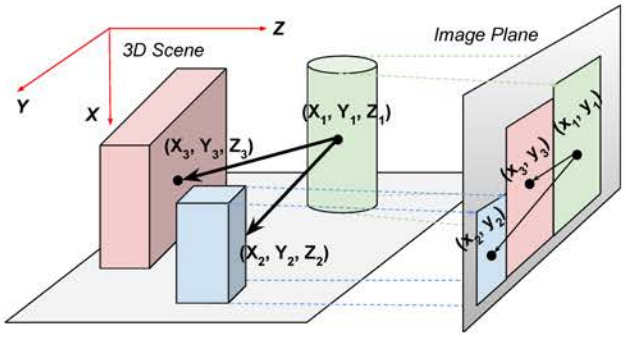

Figure 2. When a camera captures an image, points in the 3D scene are projected onto a 2D image plane. In VQA, although this projected image is given as input, the questions that require spatial reasoning are inherently about the 3D scene.

1. Introduction

"Visual reasoning" is an umbrella term that is used for visual abilities beyond the perception of appearances (objects and their sizes, shapes, colors, and textures). In the V&L domain, tasks such as image-text matching [49, 50, 54], visual grounding [24, 61], visual question answering (VQA) [17, 21], and commonsense reasoning [63] fall under this category. One such ability is spatial reasoning – understanding the geometry of the scene and spatial locations of objects in an image. Visual question answering (such as the GQA challenge shown in Figure 1) is a task that can evaluate this ability via questions that either refer to spatial locations of objects in the image, or questions that require a compositional understanding of spatial relations between objects.

Transformer-based models [51, 33, 10, 13] have led to significant improvements in multiple V&L tasks. However, the underlying training protocol for these models relies on learning correspondences between visual and textual inputs, via pre-training tasks such as image-text matching and cross-modal masked object prediction or feature regression, and then finetuned on specific datasets such as GQA. As such, these models are not trained to reason about the 3D geometry of the scene, even though the downstream task evaluates spatial understanding. As a result, V&L models remain oblivious to the mechanisms of image formation.

Real-world scenes are 3-dimensional, as illustrated by Figure 2, which shows blocks in a scene. When a camera captures an image of this scene, points on the objects are projected onto the same image plane, i.e., all points get mapped to a single depth value, and the z dimension (depth) is lost. This mapping depends on lens equations

Figure 3. Common optical illusions occur because objects closer to the camera are magnified. This illustrates the need to understand 3D scene geometry to perform spatial reasoning on 2D images.

and camera parameters and leads to optical illusions such as Figure 3, due to the fact that the magnification of objects is inversely proportional to the depth and depends on focal lengths [58, 57]. Since the 3D scene is projected to a 2D image, the faraway person appears smaller, and on top of the woman's palm in the left image, and below the woman's shoe in the right image. Such relationships between object sizes, depths, camera calibration, and scene geometries make spatial reasoning from images challenging.

If the 3D coordinates of objects (X_i, Y_i, Z_i) are known, it would be trivial to reason about their relative locations, such as the questions in Figure 1. However, images in V&L datasets [21, 17] are crowd-sourced and taken from different monocular cameras with unknown and varying camera parameters such as focal length and aperture size, making it difficult to resolve the 3D coordinates (especially the depth) from the image coordinates. This leads to ambiguities in resolving scene geometry and makes answering questions that require spatial reasoning, a severely ill-posed problem.

In this paper, we consider the task of visual question answering with emphasis on spatial reasoning (SR). We investigate if VQA models can resolve spatial relationships between objects in images from the GQA challenge. Our findings suggest that although models answer some ($\sim$60%) of these questions correctly, they cannot faithfully resolve spatial relationships such as relative locations (left, right, front, behind, above, below), or distances between objects. This opens up a question:

> *Do VQA models actually understand scene geometry, or do they answer spatial questions based on spurious correlations learned from data?*

Towards this end, we design two additional tasks that take 3D geometry into consideration, *object centroid estimation* and *relative position estimation*. These tasks are weakly supervised by inferred depth-maps estimated by an off-the-shelf monocular depth-estimation technique [6] and bounding box annotations for objects. For object centroid estimation, the model is trained to predict the centroids of the detected input objects in a unit-normalized 3D vector space. On the other hand, for relative position estimation, the model is required to predict the distance vectors between the detected input objects in the same vector space.

Our work can be summarized as follows:

1. Our approach combined existing training protocosl for transformer-based language models with novel weakly-supervised SR tasks based on the 3D geometry of the scene – namely, object centroid estimation (OCE) and relative position estimation (RPE).
2. This approach, significantly improves the correlation between GQA performance and SR tasks.
3. We show that our approach leads to an improvement of 2.21% on open-ended questions and 1.77% overall, over existing baselines on the GQA challenge.
4. Our approach also improves the generalization ability to out-of-distribution samples (GQA-OOD [26]) and is significantly better than baselines in the few-shot setting achieving state-of-the-art performance with just 10% of labeled GQA samples.

2. Related Work

Visual Question Answering is a task at the intersection of vision and language in which systems are expected to answer questions about an image as shown in Figure 1. VQA is an active area of research with multiple datasets [7, 3, 17, 21] that encompass a wide variety of questions, such as questions about the existence of objects and their properties, object counting, questions that require commonsense knowledge [63], external facts or knowledge [55, 35] and spatial reasoning (described below).

Spatial Reasoning in VQA has been specifically addressed for synthetic blocks-world images and questions in CLEVR [23] and real-world scenes and human-authored questions in GQA [21]. Both datasets feature questions that require a compositional understanding of spatial relations of objects and their properties. However, the synthetic nature and limited complexity of questions and images in CLEVR make it an easier task; models for CLEVR have reached very high (99.80%) test accuracies [60]. On the other hand, GQA poses significant challenges owing to the diversity of objects and contexts in real-world scenes and visual ambiguities. GQA also brings about linguistic difficulties since questions are crowd-sourced via human annotators. For the GQA task, neuro-symbolic methods have been proposed, such as NSM [20, 22] and TRRNet [59] which try to model question-answering as instruction-following by converting questions into symbolic programs.

3D scene reconstruction is a fundamental to computer vision and has a long history. Depth estimation from multiple observations such as stereo images [44], multiple frames or video [46, 39], coded apertures [67], variable lighting [5], and defocus [56, 52] has seen significant progress. However

monocular (single-image) depth estimation remains a challenging problem, with learning-based methods pushing the envelope [43, 11, 31]. In this work, we utilize AdaBins [6] which uses a transformer-based architecture that adaptively divides depth ranges into variable-sized bins and estimates depth as a linear combination of these depth bins. AdaBins is a state-of-the-art monocular depth estimation model for both outdoor and indoor scenes, and we use it as weak supervision to guide VQA models for spatial reasoning tasks.

Weak Supervision in V&L. Weak supervision is an active area of research in vision tasks such as action/object localization [48, 66] and semantic segmentation [27, 64]. While weak supervision *from* V&L datasets has been used to aid image classification [14, 42], the use of weak supervision *for* V&L and especially for VQA, remains underexplored. While existing methodologies have focused on learning cross-modal features from large-scale data, annotations other than objects, questions, and answers have not been extensively used in VQA. Kervadec *et al*. [25] use weak supervision in the form of object-word alignment as a pre-training task, Trott *et al*. [53] use the counts of objects in an image as weak supervision to guide VQA for counting-based questions, Gokhale *et al*. [16] use rules about logical connectives to augment training datasets for yes-no questions, and Zhao *et al*. [65] use word-embeddings [36] to design an additional weak-supervision objective. Weak supervision from captions has also been recently used for visual grounding tasks [19, 37, 12, 4].

3. Relative Spatial Reasoning

In V&L understanding tasks such as image-based VQA, captioning, and visual dialog, systems need to reason about objects present in an image. Current V&L systems, such as [2, 51, 9, 33] extract FasterRCNN [40] object features to represent the image. These systems incorporate positional information by projecting 2D object bounding-box co-ordinates and adding them to the extracted object features. While V&L models are pre-trained with tasks such as image-caption matching, masked object prediction, and masked-language modeling, to capture object–word contextual knowledge, none of these tasks explicitly train the system to learn spatial relationships between objects.

In the VQA domain, spatial understanding is evaluated indirectly, by posing questions as shown in Figure 1. However, this does not objectively capture if the model can infer locations of objects, spatial relations, and distances. Previous work [1] has shown that VQA models learn to answer questions by defaulting to spurious linguistic priors between question-answer pairs from the training dataset, which doesn't generalize when the test set undergoes a change in these linguistic priors. In a similar vein, our work seeks to disentangle spatial reasoning (SR) from the linguistic priors of the dataset, by introducing two new geometry-based training objectives – object centroid estimation (OCE) and relative position estimation (RPE). In this section, we describe these SR tasks.

3.1. Pre-Processing

Pixel Coordinate Normalization. We normalize pixel coordinates to the range $[0, 1]$ for both dimensions. For example, for an image of size $H \times W$, coordinates of a pixel (x, y) are normalized to $(\frac{x}{H}, \frac{y}{W})$.

Depth Extraction. Although object bounding boxes are available with images in VQA datasets, they lack depth annotations. To extract depth-maps from images, we utilize an open-source monocular depth estimation method, AdaBins [6], which is the state-of-the-art on both outdoor [15] and indoor scene datasets [47]. AdaBins utilizes a transformer that divides an image's depth range into bins whose center value is estimated adaptively per image. The final depth values are linear combinations of the bin centers. As depth values for images lie on vastly different scales for indoor and outdoor images, we normalize depth to the $[0, 1]$ range, using the maximum depth value across all indoor and outdoor images. We thus obtain depth-values $d(i, j)$ for each pixel $(i, j), i \in \{1, H\}, j \in \{1, W\}$ in the image.

Representing Objects using Centroids. Given the bounding boxes for each object in the image, $[(x_1, y_1), (x_2, y_2)]$ we can compute (x_c, y_c, z_c) coordinates of the object's centroid. x_c and y_c are calculated as the mean of the top-left corner (x_1, y_1) and bottom-right corner (x_2, y_2) of the bounding box, and z_c is calculated as the mean depth of all points in the bounding box:

$$x_c = \frac{x_1 + x_2}{2}, \quad y_c = \frac{y_1 + y_2}{2} \tag{1}$$

$$z_c = \sum_{i \in [x_1, x_2], j \in [y_1, y_2].} d(i, j). \tag{2}$$

Thus every object V_k in object features can be represented with 3D coordinates of its centroid. These coordinates act as weak supervision for our spatial reasoning tasks below.

3.2. Object Centroid Estimation (OCE)

Our first spatial reasoning task trains models to predict centroids of each object in the image.

In **2D OCE**, we model the task as prediction of the 2D centroid co-ordinates (x_c, y_c) of the input objects. Let V denote the features of the input image and let Q be the textual input. Then the 2D estimation task requires the system to predict the centroid co-ordinates, (x_{c_k}, y_{c_k}), for all objects $k \in \{1 \dots N\}$ present in object-features V.

In **3D OCE**, we also predict the depth co-ordinate of the object. Hence the task requires the system to predict the

3D centroid co-ordinates, $(x_{c_k}, y_{c_k}, z_{c_k})$, for all objects $k \in \{1 \dots N\}$ present in object-features V.

3.3. Relative Position Estimation (RPE)

The model is trained to predict the distance vector between each pair of distinct objects in the projected unit-normalized vector space. These distance vectors real-valued vectors $\in \mathbb{R}^3_{[-1,1]}$. Therefore, for a pair of centroids (x_1, y_1, z_1) and (x_2, y_2, z_2) for two distinct objects, given V and Q, the model is trained to predict the vector $[x_1 - x_2, y_1 - y_2, z_1 - z_2]$. RPE is not symmetric and for any two distinct points A, B, $\texttt{dist}(A, B) = -\texttt{dist}(B, A)$.

Regression vs. Bin Classification. In both tasks above, predictions are real-valued vectors. Hence, we evaluate two variants of these tasks: (1) a regression task, where models predict real-valued vectors in $\mathbb{R}^3_{[-1,1]}$, and (2) bin classification, for which we divide the range of real values across all three dimensions into C log-scale bins. Bin-width for the c^{th} bin is given by (with hyper-parameter $\lambda = 1.5$):

$$b_c = \frac{1}{\lambda^{C - |c - \frac{C}{2}| + 1}} - \frac{1}{\lambda^{C - |c - \frac{C}{2}| + 2}} \quad \forall c \in \{0..C\text{-}1\}. \quad (3)$$

Log-scale bins lead to a higher resolution (more bins) for closer distances and lower resolution (fewer bins) for farther distances, giving us fine-grained classes for close objects. Models are trained to predict the bin classes as outputs for all 3 dimensions, given a pair of objects. We evaluate different values for the number of bins: $C \in \{3, 7, 15, 30\}$, to study the extent of V&L model's ability to differentiate at a higher resolution of spatial distances. For example, the simplest form of bin classification is a three-class classification task with bin-intervals $[-1, 0)$, $[0]$, $(0, 1]$.

4. Method

We adopt LXMERT [51], a state-of-the-art vision and language model, as the backbone for our experiments. LXMERT and other popular transformer-based V&L models methods [33, 9], are pre-trained on a combination of multiple VQA and image captioning datasets such as Conceptual Captions [45], SBU Captions [38], Visual Genome [30], and MSCOCO [32]. These models use object features of the top 36 objects extracted by the FasterRCNN object detector [40] as visual representations for input images. A transformer encoder takes these object features along with textual features as inputs, and outputs cross-modal `[CLS]` tokens. The model is pre-trained by optimizing for masked language modeling, image-text matching, masked-object prediction and image-question answering.

4.1. Weak Supervision for SR

Let $v \in \mathbb{R}^{36 \times H}$ be the visual features, $x \in \mathbb{R}^{1 \times H}$ be the cross-modal features, and $t \in \mathbb{R}^{L \times H}$ be the text features, produced by the cross-modality attention layer of the LXMERT encoder. Here H is the hidden dimension, and L is the number of tokens. These outputs are used for fine-tuning the model for two tasks: VQA using x as input, and the spatial reasoning tasks using v as input. Let D be the number of coordinate dimensions (2 or 3) that we use in spatial reasoning. For the SR-regression task, we use a two-layer feed-forward network f_{reg} to project v to a real-valued vector with dimensions $36 \times D$, and compute the loss using mean-squared error (MSE) with respect to the ground-truth object coordinates y_{reg} .

$$\mathcal{L}_{SR\text{-}reg} = \mathcal{L}_{MSE}(f_{reg}(v), y_{reg}). \quad (4)$$

For the bin-classification task, we train a two-layer feed-forward network f_{bin} to predict $36 \times C \times D$ bin classes for each object along each dimension, where C is the number of classes, trained using cross-entropy loss:

$$\mathcal{L}_{SR\text{-}bin} = \mathcal{L}_{CE}(f_{bin}(V), y_{bin}), \quad (5)$$

where y_{bin} are the ground-truth object location bins.
The total loss is given by:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{VQA} + \beta \cdot \mathcal{L}_{SR}, \quad where\ \alpha, \beta \in (0, 1]. \quad (6)$$

y_{reg} and y_{bin} are obtained from the object centroids computed during preprocessing (Sec. 3.1) from depth estimation networks and object bounding boxes. Since the real 3D coordinates of objects in the scene are unknown, these y_{reg} and y_{bin} act as proxies and therefore can weakly supervise our spatial reasoning tasks.

4.2. Spatial Pyramid Patches

As LXMERT only takes as input the distinct object and the 2D bounding box features, it inherently lacks the depth information required for 3D spatial reasoning task. This is confirmed by our evaluation on the 2D and 3D spatial reasoning tasks, where the model has strong performance in 2D tasks, but lacks on 3D tasks, as shown in Table 1. In order to incorporate spatial features from the original image to capture relative object locations as well as depth information, we propose the use of *spatial pyramid patch features* [4] to represent the given image into a sequence of features at different scales. The image I is divided into a set of patches: $p_n = \{I_{i_1}, \dots, I_{i_n}\}$, each I_{i_j} being a $i_j \times i_j$ grid of patches, and ResNet features are extracted for each patch. Larger patches encode global object relationships, while smaller patches contain local relationships.

4.3. Fusion Transformer

In order to combine the spatial pyramid patch features and features extracted from LXMERT, we propose a fusion transformer with e-layers of transformer encoders, containing self-attention, a residual connection and layer normalization after each sub-layer. We concatenate the p_n patch

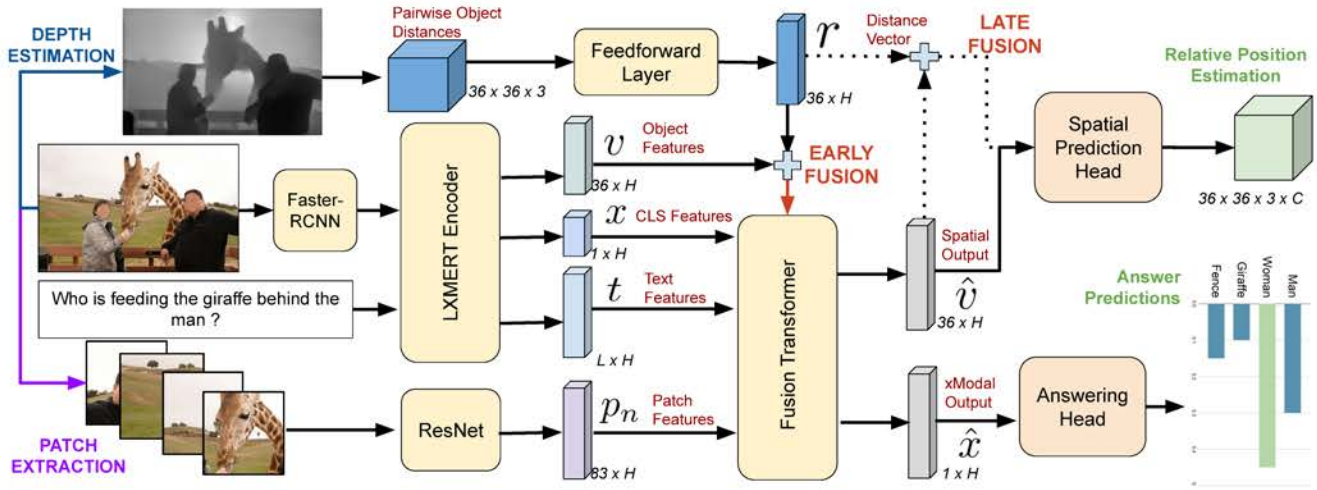

Figure 4. Overall architecture for our approach shows conventional modules for object feature extraction, cross-modal encoding, and answering head, with our novel weak supervision from depthmaps, patch extraction, fusion mechanisms, and spatial prediction head.

features with v visual, x cross-modal and t textual hidden vector output representations from LXMERT, to create the fused vector h, which is fed into the fusion transformer. Let M be the length of the sequence after concatenating all hidden vectors, then for any hidden vector m in the sequence:

$$h^0 = [X, V, T, P_n].$$
$$\hat{h}^e_m = \texttt{Self-Att}(h^{e-1}_m, [h^{e-1}_1, \ldots, h^{e-1}_M]);\ \forall e. \quad (7)$$

The output of fusion transformer $\hat{h}^e = [\hat{x}, \hat{v}, \hat{t}, \hat{p}_n]$ is then separated into its components, of which, $\hat{x}$, $\hat{v}$ are used as inputs for VQA and SR task, on the same lines as Section 4.1.

4.4. Relative Position Vectors as Inputs

The final set of features that we utilize are the pair-wise relative distance vectors between objects as described in section 3.3. In this case, the pairwise distances are used as inputs, in addition to visual, textual, cross-modal and patch features, and the model is trained to reconstruct the pairwise distances. This makes our model an auto-encoder for the regression task. For each input visual object feature v_k, we create a relative position feature r_k using the pair-wise distance vectors projected from the input dimensions of 36×3 to $36 \times H$ using a feed-forward layer, where H is the size of the hidden vector representations. We evaluate two-modes of fusion of these features. In **Early Fusion**, r_k is added to v_k the output of the LXMERT encoder. In **Late Fusion**, r_k is added to $\hat{v}_k$ the output of the fusion transformer. Figure 4 shows the architecture for the final model that utilizes both the patch features and relative positions as input.

5. Experiments

Datasets. We evaluate our methods on two popular benchmarks, GQA [21] and GQA-OOD [26], both of which contain spatial reasoning visual questions requiring compositionality and relations between objects present in natural non-iconic images. Both datasets have a common training set, but differ in the test set: GQA uses an i.i.d. split, while GQA-OOD contains a distribution shift. There are 2000 unique answers in these datasets, and questions can be categorized based on the type of answer: binary (yes/no answers) and open-ended (all other answers).

Evaluation Metrics. For evaluating performance in fully-supervised, few-shot, as well as O.O.D. settings for the GQA task, we use metrics defined in [21]. These include exact match accuracy, accuracy on the most frequent head answer-distribution, infrequent tail answer-distribution, consistency to paraphrased questions, validity, and plausibility of spatial relations[1]. We evaluate SR tasks using mean-squared error (MSE) for SR-Regression and classification accuracy for SR bin-classification.

Model Architectures. LXMERT contains 9 language transformer encoder layers, 5 visual layers, and 5 cross-modal layers. This feature extractor can be replaced by any other transformer-based V&L model. Our Fusion transformer has 5 cross-modal layers with a hidden dimension of $H = 512$. For visual feature extraction, we use ResNet-50 [18] pre-trained on ImageNet [41] to extract image patch features, with 50% overlap, and Faster RCNN pre-trained on Visual Genome [30] to extract the top 36 object features. We use 3×3, 5×5, 7×7 patches, and the entire image as the spatial image patch features. The image is uniformly divided into a set of overlapping patches at multiple scales.

[1]Detailed definitions of these metrics can be found in Section 4.4. of Hudson *et al.* [21] or accessed on the GQA Challenge webpage

Model	GQA-Val↑	2D-Reg↓	2D Bin Classification			GQA-Val↑	3D-Reg↓	3D Bin Classification		
			2D-3w↑	2D-15w↑	2D-30w↑			3D-3w↑	3D-15w↑	3D-30w↑
LXMERT + SR	59.85	0.64	88.20	76.75	55.12	60.05	0.44	55.66	52.80	48.15
+ Late Fusion	59.90	0.47	92.60	81.24	60.42	60.18	0.29	71.20	69.45	52.84
+ Early Fusion	60.10	0.36	96.40	82.48	64.85	61.32	0.24	78.67	74.20	54.73
+ Patches	60.52	0.41	89.60	79.56	59.40	60.64	0.28	73.21	71.74	50.94
+ Late Fusion + Patches	60.80	0.33	95.20	82.10	67.38	61.80	0.21	85.35	79.60	65.45
+ Early Fusion + Patches	**60.95**	**0.29**	**97.40**	**84.60**	**71.46**	**62.32**	**0.17**	**89.58**	**81.47**	**68.20**

Table 1. Results for the LXMERT model trained for the spatial reasoning task (LXMERT + SR), on 2D and 3D Relative Position Estimation (RPE), for regression as well as C-way bin classification tasks. A comparison with the same model weakly supervised with additional features (image patches) and weak supervision with relative position vectors extracted from depth-maps is shown. GQA-Val scores are for the best performing weak-supervision task, which are 2D-15w and 3D-15w respectively. Regression scores are in terms of mean-squared error, and classification scores are percentage accuracy. *15w: 15-way bin-classification.*

Model	GQA-Val↑
LXMERT + SR	59.40
+ 2D OCE (Regression)	57.33
+ 3D OCE (Regression)	58.28
+ 2D RPE (Regression)	59.85
+ 3D RPE (Regression)	59.54
+ 2D OCE (15-bin Classification)	58.64
+ 3D OCE (15-bin Classification)	59.90
+ 2D RPE (15-bin Classification)	60.95
+ 3D RPE (15-bin Classification)	**62.32**

Table 2. Comparison of different weakly supervised spatial reasoning tasks on the GQA validation split.

Training Protocol and Hyperparameters. Our Fusion transformer has 5 cross-modal layers with a hidden dimension of $H = 512$. All models are trained for 20 epochs with a learning rate of $1e{-}5$, batch size of 64, using Adam [29] optimizer, on a single NVIDIA A100 40 GB GPU. The values of coefficients (α, β) in Equation 6 were chosen to be $(0.9, 0.1)$ for regression and $(0.7, 0.3)$ for classification.

Baselines. We use LXMERT jointly trained SR and GQA tasks as a strong baseline for our experiments. In addition, we also compare performance with existing non-ensemble (single model) methods on the GQA challenge, that directly learn from question-answer pairs without using external program supervision, or additional visual features. Although NSM [22] reports a strong performance on the GQA challenge, it uses stronger object detectors and top-50 object features (as opposed to top-36 used by all other baselines), rendering comparison with NSM unfair.

5.1. Results on Spatial Reasoning

We begin by evaluating the model on different spatial reasoning tasks, using various weak-supervision training methods. Table 1 and 2 summarize the results for these experiments. It can be seen that the LXMERT+SR baseline (trained without supervision from depthmaps) performs poorly for all spatial reasoning tasks. This conforms with our hypothesis, since depth information is not explicitly captured by the inputs of the current V&L methods that utilize bounding box information which contains only 2D spatial information. On average, improvements across SR tasks are correlated with improvements across the GQA task. In some cases, we observe that the method predicts the correct answer for the spatial relationship questions on the GQA task, even when it fails to correctly predict the bin-classes or object positions in the SR task. This phenomenon is observed for 18% of the correct GQA predictions. For example, the model predicts 'left' as the GQA answer and a contradictory SR output corresponding to 'right'.

Comparison of different SR Tasks. Centroid Estimation requires the model to predict the object centroid location in the unit-normalized vector space, whereas the Relative Position Estimation requires the model to determine the pairwise distance vector between the centroids. Both the tasks provide weak-supervision for spatial understanding, but we observe in Table 2 that bin-classification for the 3D RPE transfers best to the GQA accuracy.

Regression v/s Bin-Classification. Similarly, the regression version of the task poses a significant challenge for V&L models to accurately determine the polarity and the magnitude of distance between the object. The range of distances in indoor and outdoor scenes has a large variation, and poses a challenge for the model to exactly predict distances in the regression task. The classification version of the task appears to be less challenging, with the 3-way 2D relative position estimation achieving significantly high scores ($\sim$90%). The number of bins (3/15/30) also impacts performance; a larger number of bins implies that the model should possess a fine-grained understanding of distances, which is harder. We find the optimal number of bins (for both RPE and GQA) is 15.

Model	Accuracy↑	Binary↑	Open↑	Consistency↑	Validity↑	Plausibility↑	Distribution↓
Human [21]	89.30	91.20	87.40	98.40	98.90	97.20	–
Global Prior [21]	28.90	42.94	16.62	51.69	88.86	74.81	93.08
Local Prior [21]	31.24	47.90	16.66	54.04	84.33	84.31	13.98
BottomUp [2]	49.74	66.64	34.83	78.71	96.18	84.57	5.98
MAC [20]	54.06	71.23	38.91	81.59	96.16	84.48	5.34
GRN [22]	59.37	77.53	43.35	88.63	96.18	84.71	6.06
Dream [22]	59.72	77.84	43.72	91.71	96.38	**85.48**	8.40
LXMERT [51]	60.34	77.76	44.97	92.84	96.30	85.19	8.31
This Work	**62.11**	**78.20**	**47.18**	**93.13**	**96.92**	85.27	**1.10**

Table 3. Comparitive evaluation of our model with respect to existing baselines, on the GQA test-standard set, along all evaluation metrics.

Comparison of different methods. The Early Fusion with Image Patches method, which uses both the relative position distance vectors and the pyramidal patch features with the fusion transformer, achieves the best performance across all spatial tasks and the GQA task. It can be observed from Table 1 that both of these additional inputs improve performance in 3D RPE. These performance improvements can be attributed to the direct relation between the distance-vector features and prediction targets. On the other hand, patch features implicitly possess this spatial relationship information, and utilizing both the features together results in the best performance. However, even with a direct correlation between the input and output, the model is far from achieving perfect performance on the harder 15/30-way bin-classification or regression tasks, pointing to a scope for further improvements.

Early v/s Late Fusion. We can empirically conclude that Early fusion performs better than Late fusion through our experiment results in Table 1. We hypothesize that the Fusion Transformer layers are more efficient than Late Fusion at extracting the spatial relationship information from the projected relative position distance vectors.

Effect of Patch Sizes. We study the effect of different image patches' grid sizes, such as 3×3, 5×5, 7×7, and 9×9 and several combinations of such sets of patch-features. We observe the best performing feature combination to be the entire image and a set of patches with grids in 3×3, 5×5 and 7×7. Adding smaller patches such as 9×9 grid did not lead to an increase in performance. Extracting features from ResNet101 also leads to minor gains ($+0.05\%$).

5.2. Results on GQA

Tables 3 and 4 summarize our results on the GQA and GQA-OOD visual question answering tasks. Our best method, LXMERT with Early Fusion and Image Patches, jointly trained with weak-supervision on 15-way bin-classification Relative Position Estimation task improves over the baseline LXMERT, by 1.77% and 1.3% respectively on GQA and GQA-OOD, achieving a new state-of-the-art. It performs slightly better than LXMERT (72.9%) on VQA-v2. The most significant improvement is observed on the open-ended questions (2.21%). We can observe that weak-supervision and joint end-to-end training of SR and question answering using the transformer architecture can train systems to be consistent in spatial reasoning tasks and to better generalize in spatial VQA tasks.

Model	Uses Image	Acc-All↑	Acc-Tail↑	Acc-Head↑
Question Prior [26]	No	21.6	17.8	24.1
LSTM [3]	No	30.7	24.0	34.8
BottomUp [2]	Yes	46.4	42.1	49.1
MCAN [62]	Yes	50.8	46.5	53.4
BAN4 [28]	Yes	50.2	47.2	51.9
MMN [8]	Yes	52.7	48.0	55.5
LXMERT [51]	Yes	54.6	49.8	57.7
This Work	Yes	**55.9**	**50.3**	**59.4**

Table 4. Comparison of several VQA methods on the GQA-OOD test-dev splits. Acc-tail: OOD settings, Acc-head: accuracy on most probable answers (given context), scores in %.

OOD Generalization. We also study generalization to distribution shifts for GQA, where the linguistic priors seen during training, undergo a shift at test-time. We evaluate our best method on the GQA-OOD benchmark and observe that we improve on the most frequent head distribution of answers by 1.7% and also the infrequent out-of-distribution (OOD) tail answer by 0.5%. This leads us to believe that training on SR tasks with weak-supervision might allows the model to reduce the reliance on spurious linguistic correlations, enabling better generalization abilities.

Few-Shot Learning. We study the effect of the weakly supervised RPE task in the few-shot setting on open-ended questions, with results shown in Figure 5. We can observe that even with as low as 1% and 5% of samples, joint training with relative position estimation improves over LXMERT trained with same data by 2.5% and 5.5%, respectively, and is consistently better than LXMERT at all other fractions. More importantly, with only 10% of the

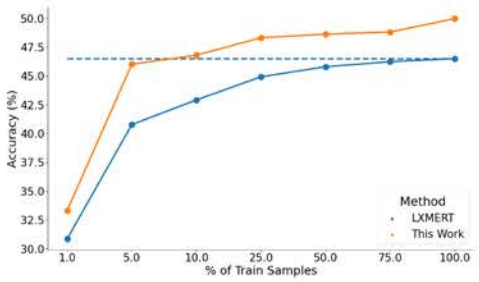

Figure 5. Performance of our best method, when trained in the few-shot setting and evaluated on open-ended questions from the GQA-testdev split, compared to LXMERT.

training dataset our method achieves a performance close to that of the baseline LXMERT trained with the entire (100%) dataset. Most spatial questions are answered by relative spatial words, such as "left", "right", "up", "down" or object names. Object names are learned during the V&L pre-training tasks, whereas learning about spatial words can be done with few spatial VQA samples and a proper supervision signal that contains spatial information.

5.3. Error Analysis

We perform three sets of error analyses to understand the different aspects of the weakly-supervised SR task, the consistency between the relative SR task and the VQA task, and the errors made in the VQA task.

Spatial Reasoning Tasks. SR-Regression appears to be the most challenging version, as the system needs to reconstruct the relative object distances from the input image to a 3D unit-normalized vector space. The classification variant has a higher recall and better polarity, i.e., an object to the "right" is classified correctly in the 'right' direction regardless of magnitude, i.e. the correct distance bin-class, compared to the regression task. The majority of errors ($\sim 60\%$) are due to the inability to distinguish between close objects.

Consistency between SR and VQA. The baseline LXMERT trained only on weak-supervision tasks without patch features or relative position distance vectors predicts 18% of correct predictions with wrong spatial relative positions. This error decreases to 3% for the best method that uses early fusion with image patches, increasing the faithfulness or consistency between the two tasks. We manually analyze 50 inconsistent questions and observe 23 questions contain ambiguity, i.e., multiple objects can be referred by the question and lead to different answers.

Manual Analysis. We analyze 100 cases of errors from the GQA test-dev split and broadly categorize them as follows, with percentage of error in parentheses:

1. predictions are **synonyms or hypernyms** of ground-truth; for example, "curtains–drapes", "cellphone–phone", "man–person", etc. (8%)
2. predictions are **singular/plural** versions of the gold answer, such as, "curtain-curtains", "shelf-shelves". (2%)
3. **Ambiguous questions** can refer to multiple objects leading to different answers; for example, in an image with two persons having black and brown hair standing in front of a mirror, a question is asked: "Does the person in front of the mirror have black hair?". (5%)
4. **Errors in answer annotations.** (5%)
5. **Wrong predictions**. Examples of this include predicting "right" when the true answer is "left" or the prediction of similar object classes as the answer, such as "cellphone–remote control", "traffic-sign–stop sign". In many cases, the model is able to detect an object, but unable to resolve its relative location with respect to another object; this could be attributed to either spurious linguistic biases or the model's lack of spatial reasoning. (80%)

This small-scale study concludes that 20% of the wrong predictions could be mitigated by improved evaluation of subjective, ambiguous, or alternative answers. Luo *et al.* [34] share this observation and suggest methods for a more robust evaluation of VQA models.

6. Discussion

The paradigm of pre-trained models that learn the correspondence between images and text has resulted in improvements across a wide range of V&L tasks. Spatial reasoning poses the unique challenge of understanding not only the semantics of the scene, but the physical and geometric properties of the scene. One stream of work has approached this task from the perspective of sequential instruction-following using program supervision. In contrast, our work is the first to jointly model geometric understanding and V&L in the same training pipeline, via weak supervision from depth estimators. We show that this increases the faithfulness between spatial reasoning and visual question answering, and improves performance on the GQA dataset in both fully supervised and few-shot settings. While in this work, we have used depthmaps as weak supervision, many other concepts from physics-based vision could further come to the aid of V&L reasoning. Future work could also consider spatial reasoning in V&L settings without access to bounding boxes or reliable object detectors (for instance in bad weather and/or low-light settings). Challenges such as these could potentially reveal the role that geometric and physics-based visual signals could play in robust visual reasoning.

Acknowledgments. The authors acknowledge support from the NSF via grants #1750082 and #1816039, DARPA SAIL-ON program #W911NF2020006, and ONR award #N00014-20-1-2332.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018. 3

[2] 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*, June 2018. 3, 7

[3] 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. 2, 7

[4] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. WeaQA: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3420–3435, Online, Aug. 2021. Association for Computational Linguistics. 3, 4

[5] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of computer vision*, 72(3):239–257, 2007. 2

[6] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. *arXiv preprint arXiv:2011.14141*, 2020. 2, 3

[7] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 2

[8] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 655–664, 2021. 7

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 3, 4

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 27:2366–2374, 2014. 3

[12] Zhiyuan Fang, Shu Kong, Zhe Wang, Charless Fowlkes, and Yezhou Yang. Weak supervision and referring attention for temporal-textual association learning. *arXiv preprint arXiv:2006.11747*, 2020. 3

[13] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 1

[14] Siddha Ganju, Olga Russakovsky, and Abhinav Gupta. What's in a question: Using visual questions as a form of supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 241–250, 2017. 3

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[16] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 1, 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[19] Lisa-Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 3

[20] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018. 2, 7

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 1, 2, 5, 7

[22] Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*, 2019. 2, 6, 7

[23] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 2

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1

[25] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Weak supervision helps emergence of word-object alignment and improves vision-language tasks. *arXiv preprint arXiv:1912.03063*, 2019. 3

[26] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? *CVPR*, 2021. 2, 5, 7

[27] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised

instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 3

[28] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 7

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 4, 5

[31] Jun Li, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017. 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 1, 3, 4

[34] Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 'just because you are right, doesn't mean I am wrong': Overcoming a bottleneck in development and evaluation of open-ended VQA tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2766–2771, Online, Apr. 2021. Association for Computational Linguistics. 8

[35] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019. 2

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc. 3

[37] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019. 3

[38] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, 2011. 4

[39] Rene Ranftl, Vibhav Vineet, Qifeng Chen, and Vladlen Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4058–4066, 2016. 2

[40] 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. 3, 4

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 2015. 5

[42] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[43] Ashutosh Saxena, Sung H Chung, Andrew Y Ng, et al. Learning depth from single monocular images. In *NIPS*, volume 18, pages 1–8, 2005. 3

[44] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002. 2

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the ACL*, pages 2556–2565, 2018. 4

[46] Nitesh Shroff, Ashok Veeraraghavan, Yuichi Taguchi, Oncel Tuzel, Amit Agrawal, and Rama Chellappa. Variable focus video: Reconstructing depth and video for dynamic scenes. In *2012 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2012. 2

[47] 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. 3

[48] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *International Conference on Machine Learning*, pages 1611–1619. PMLR, 2014. 3

[49] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 1

[50] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019. 1

[51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP 2019*, 2019. 1, 3, 4, 7

[52] Huixuan Tang, Scott Cohen, Brian Price, Stephen Schiller, and Kiriakos N Kutulakos. Depth from defocus in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2740–2748, 2017. 2

[53] Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. In *International Conference on Learning Representations*, 2018. 3

[54] Hoa Trong Vu, Claudio Greco, Aliia Erofeeva, Somayeh Jafaritazehjani, Guido Linders, Marc Tanti, Alberto Testoni, Raffaella Bernardi, and Albert Gatt. Grounded textual entailment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2354–2368, 2018. 1

[55] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 2

[56] M. Watanabe and S.K. Nayar. Rational Filters for Passive Depth from Defocus. *International Journal on Computer Vision*, 27(3):203–225, May 1998. 2

[57] Masahiro Watanabe and Shree K Nayar. Telecentric optics for computational vision. In *European Conference on Computer Vision*, pages 439–451. Springer, 1996. 2

[58] Reg G Willson. Modeling and calibration of automated zoom lenses. In *Videometrics III*, volume 2350, pages 170–186. International Society for Optics and Photonics, 1994. 2

[59] Xiaofeng Yang, Guosheng Lin, Fengmao Lv, and Fayao Liu. Trrnet: Tiered relation reasoning for compositional visual question answering. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 414–430. Springer, 2020. 2

[60] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pages 1039–1050, 2018. 2

[61] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 1

[62] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 7

[63] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019. 1, 2

[64] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4233–4241, 2017. 3

[65] Handong Zhao, Quanfu Fan, Dan Gutfreund, and Yun Fu. Semantically guided visual question answering. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1852–1860. IEEE, 2018. 3

[66] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[67] Changyin Zhou, Stephen Lin, and Shree K Nayar. Coded aperture pairs for depth from defocus and defocus deblurring. *International journal of computer vision*, 93(1):53–72, 2011. 2