Toward Unsupervised Realistic Visual Question Answering

Yuwei Zhang* Chih-Hui Ho* Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{yuz163,chh279, nvasconcelos}@ucsd.edu

Abstract

The problem of realistic VQA (RVQA), where a model has to reject unanswerable questions (UQs) and answer answerable ones (AQs), is studied. We first point out 2 drawbacks in current RVQA research, where (1) datasets contain too many unchallenging UQs and (2) a large number of annotated UQs are required for training. To resolve the first drawback, we propose a new testing dataset, RGQA, which combines AQs from an existing VQA dataset with around 29K human-annotated UQs. These UQs consist of both fine-grained and coarse-grained image-question pairs generated with 2 approaches: CLIP-based and Perturbation-based. To address the second drawback, we introduce an unsupervised training approach. This combines pseudo UQs obtained by randomly pairing images and questions, with an RoI Mixup procedure to generate more fine-grained pseudo UQs, and model ensembling to regularize model confidence. Experiments show that using pseudo UQs significantly outperforms RVQA baselines. RoI Mixup and model ensembling further increase the gain. Finally, human evaluation reveals a performance gap between humans and models, showing that more RVQA research is needed. Code and dataset is released on https://github.com/chihhuiho/RGQA.

1. Introduction

Visual Question Answering (VQA) is a challenging task that requires a machine to understand a question in natural language, perceive an image, and produce an answer. Despite extensive research in VQA [3, 12, 19, 36, 20, 42, 7, 54], little attention has been given to VQA robustness. In this work, we consider robustness to *unanswerable questions* (UQs), which cannot be answered by image inspection, as in Fig. 1(b). This is opposed to the traditional answerable questions (AQ), such as in Fig. 1(a).

Lack of robustness to UQs is problematic because, in the absence of image information, the VQA system frequently resorts to the answer statistically most correlated with the question. In the figure, the absence of food in (b) entices the robot to pick the answer corresponding to the "side of food" most commonly "cut" in the dataset, which happens to be the "top" (perhaps because the dataset is rich in cake images). The problem is that a decision by the robot to act on this answer would be catastrophic for the cat in the scene. More generally, the inability to reject UQs signals a deeper perceptual deficiency and exposes VQA systems to attacks.

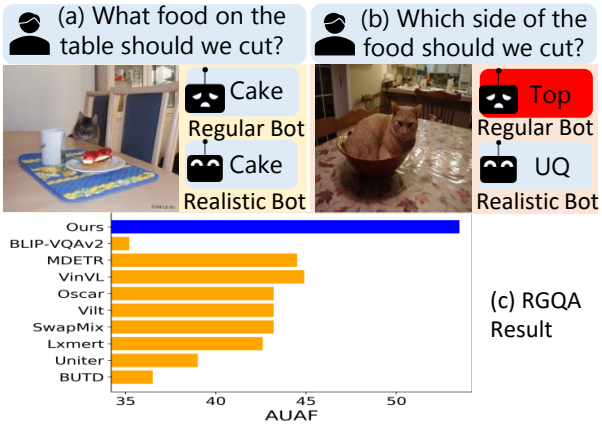


Figure 1: Realistic VQA. In VQA, a vision system answers a question by inspection of an image. However, existing approaches have no awareness if the question is *answerable* (AQ), such as in (a), or *unanswerable* (UQ), as in (b). A realistic VQA system only answers AQs. (c) RVQA performance of prior (yellow) vs. proposed (blue) models.

Vulnerability to UQs can create safety hazards for indoor robots [2] or assistants for the visually impaired [14] and reduces user trust in VQA models (see appendix for various examples from the recent large-scale BLIP model [29]). When faced with a UQ, the VQA system should refuse to answer or ask for more information. More precisely, it should assess the question, decide to (a) "accept" or "reject" it, and only (b) answer the accepted questions. Since this resembles the idea of a "realistic model" for classification [50, 48], we denote it *realistic VQA* (RVQA).

Although some prior works have addressed RVQA, existing formulations are not conducive to practical RVQA

*The first two authors contributed equally to this work.

systems, for three reasons. First, existing formulations address the *supervised* training of RVQA models. This, however, requires a significant number of annotated UQs [40, 35, 14]. The collection of a set of annotated UQs large enough to train a modern VQA network is expensive, frequently not even plausible. This is compounded by the existence of many types of UQs: training on one type does not guarantee generalization to another. Second, prior datasets generate UQs by randomly paring images and questions from an existing VQA dataset [40, 35, 21, 45]. This, however, tends to produce obviously unrelated pairs of images and questions with low semantic similarity, that are easy to reject. In the real world, RVQA models must be able to handle both simple and challenging UQs. Finally, the VQA datasets that support RVQA, such as VizWiz [14], are designed for a specific application domain, frequently containing images with few objects. This prevents the modeling of complex image-question relationships.

To address these drawbacks, we consider the problem of *unsupervised RVQA*. We start by curating a new evaluation dataset for this task, based on *testdev* set of the widely used GQA dataset [19]. The new dataset, denoted as *realistic GQA* (RGQA), is composed of $26,591$ AQs in the *testdev* set of GQA and $29,046$ additional human-annotated UQs. To penalize RVQA models that overfit on a specific type of UQs, we generate candidate UQ by two methods. *CLIP-based* UQ generation produces candidate UQs by retrieving questions sorted by CLIP [39] similarity score between image and question. *Perturbation-based* (PT-based) UQ generation perturbs the object, attribute, and relation phrases in a question. For each method, we further generate a set of easy and a set of hard candidate UQs, leading to a total of four RGQA subsets. All candidate UQs are finally annotated by humans, to guarantee they are unanswerable.

Since each AQ in RGQA is complemented by its answer, the dataset enables measuring the accuracy of both AQ/UQ detection and VQA accuracy. For this, we propose the ACC-FPR curve [9], a joint measure of VQA accuracy for AQs and UQ rejection performance. This is complemented by introducing 3 new unsupervised RVQA methods that establish a set of baselines for future RVQA work. These are classifiers with a binary output per class, which elicit a rejection when all class outputs are below a threshold. Three methods differ in training strategy and are shown capable of producing RVQA models that both reject UQs and answer AQs correctly, outperforming prior RVQA methods.

The first is to train the classifier with pseudo UQs, obtained by randomly pairing images and questions. This suffers from the fact that pseudo UQs are noisy and not always challenging. The second improves the sampling of image-question pairs, by using a RoI Mixup strategy to encourage the model to spot fine-grained mismatches between image and question during training. The third address the limitations of random sampling at the classifier output, by ensembling multiple RVQA models. Experiments show that all strategies enhance RVQA performance and that they can be combined to achieve best results. As shown in Fig. 1(c), this combination (blue) significantly exceeds the performance of existing VQA models (yellow) under the joint objective of rejecting UQs and correctly answering AQs.

Overall, three contributions are made to VQA. First, we introduce RGQA, a new challenging testing dataset for evaluating RVQA. It contains both fine- and coarse-grained image-question pairs which better align with real-world scenarios than previous datasets. Second, we propose an unsupervised training strategy that uses free pseudo UQs, combining random sampling, RoI Mixup, and model ensembling. Finally, extensive experiments demonstrate the effectiveness of the proposed methods over prior RVQA methods. We also show that the proposed models under-perform humans, which encourages future work in the RVQA problem.

2. Related Work

In this section, we review related works. See appendix for a broader discussion of the literature.

Realistic VQA (RVQA): The study of RVQA is still in its infancy. A central question is how to assemble datasets of UQs, i.e. unrelated pairs of images and questions. Most methods start from a VQA dataset. VTFQ [40] collected a RVQA dataset by randomly pairing images and questions. QRPE [35] uses question-derived object/attribute premises. The associated image is then replaced by its Euclidean nearest neighbor in a set of images without the extracted premises. These approaches are limited by the inability of random pairing or Euclidean similarity to guarantee a fine-grained semantic mismatch between image and UQ.

VizWiz [14] is a VQA dataset from the visually impaired setting, with UQs asked by people. However, its images are of poor quality and contain one or a few objects, which prevents complex interaction between objects, scenes, and language. TDIUC [21] and C2VQA [44] are created by checking if objects mentioned in questions also appear in images. While UQ cardinality can be easily scaled up [21] by randomly paring images and questions without common objects, this assumes that the only reason for a UQ is object mismatch. In comparison, the proposed RGQA dataset considers both coarse- and fine-grained mismatches, based on stronger measures of image-question similarity. No constraints of image content are also imposed on UQ generation, producing a more challenging and diverse dataset.

All previous works address supervised RVQA, using annotated UQs, which is expensive and limits dataset sizes. For instance, [40] generates a caption per image with NeuralTalk2 [23] and measures question-caption similarity with a binary LSTM classifier. [35] further extracts the ques-

Figure 2: Examples of CLIP based (a,b,e,f) and Perturbation (PT) based UQs (c,d,g,h) in RGQA. For the PT-based UQs, the red words are modified from the original question. See appendix for more examples.

tion premise and uses the concatenated question-premise-caption triplet as classifier input. [30] uses this architecture to reject UQs in VQA. [27] uses the maximum attention score between objects and text tokens for rejection and regularizes attentions by training on UQs. In this work, we explore an unsupervised training strategy that is model-agnostic and does not rely on annotated UQs.

Out of Distribution Detection (OOD) RVQA is closely related to OOD in classification [16, 25, 33, 18, 49, 17, 28] which aims to detect samples on which a classifier has not been trained. This has been addressed by temperature scaling of classifier logits [33], using Mahalanobis distance [28] or energy scores [34] to measure the distance to the training distribution, ensembling predictions from multiple models [25, 46], or regularizing in-distribution (ID) features [8]. It is also possible to use a background dataset, with different distribution from the training dataset, during training [10, 18, 49, 32]. While background datasets can significantly improve OOD, prior works in RVQA [30, 27] show a performance degradation for AQs. We devise sampling strategies that address this problem.

The classification and OOD performance are usually reported by combining Area Under ROC curve (AUROC) and accuracy on ID samples [38, 4, 55, 51]. However, separate metrics increase the difficulty to compare models. We introduce a unified metric for the RVQA problem.

3. RGQA Dataset

In this section, we introduce the RGQA dataset for evaluating RVQA systems. It is a human-annotated dataset with $\sim 29K$ UQs and built upon the *testdev* set of GQA [19].

3.1. Dataset Curation

RGQA has a balanced coverage of AQs and UQs. AQs are image-question pairs with answers from the GQA *testdev* set. For UQs, we first generate a *candidate set* using two different approaches, *CLIP-based* and *Perturbation-based*, to mitigate potential UQ generation biases. Human annotators then decide which candidates are true UQs.

CLIP-based Candidate UQs: Leveraging recent advances in image-text pre-training, we use CLIP [39] to measure similarity between images and questions. Given an image I, we consider the set of questions $\mathcal{Q}(I)$ in the *testdev* dataset, excluding 1) existence questions (e.g. "Are there any ...?"), which can never be UQs, and 2) the questions originally paired with I. We then feed all pairs $(I, Q), Q \in \mathcal{Q}(I)$ to the CLIP model and rank the questions by similarity score. To cover the spectrum from simple to hard UQs, 85 questions sampled from the top $2,500$ are used as candidate UQs for CLIP-Hard, while the last 50 questions are used as candidate UQs for CLIP-Easy. Fig. 2 shows images from each set. The pairs of CLIP-Hard (Fig. 2 (a,b)) have more subtle mismatches than those of CLIP-Easy (Fig. 2 (e,f)).

Perturbation-based Candidate UQs: Given an AQ in GQA *testdev*, a candidate UQ counterpart is generated by

Table 1: Comparison to previous datasets. The proposed RGQA dataset has longer and more fine-grain UQs and requires a multi-task classifier to solve the RVQA problem. RGQA is only for evaluation purposes.

Dataset	Supervised UQ	Type	UQ Annotation	Image Source	Question Source	UQ(%)	# Test Pair	Avg. Length
VTFQ [40]	✓	UQ det.	human	MSCOCO	VQAv1	89.24	31464	7.53
QRPE [35]	✓	UQ det.	generated	MSCOCO	VQAv1	50.87	35476	7.76
C2VQA [45]	✓	UQ det.	generated	Visual Genome	Visual Genome	50.00	29106	7.10
TDIUC [21]	✓	VQA+UQ det.	generated	MSCOCO+Visual Genome	VQAv1+Visual Genome	22.17	538868	7.92
VizWiz [14]	✓	VQA+UQ det.	human	VizWiz	VizWiz	27.84	8000	8.10
RGQA	✗	VQA+UQ det.	human	GQA *testdev*	GQA *testdev*	52.22	55637	10.33

perturbing its objects and adjectives. This is implemented by first collecting a set of candidate objects and their attributes from the scene graphs of GQA *train* and *valid* set. For each AQ, objects and adjectives are extracted by POS tagging. Similar to the CLIP-based approach, both easy and hard UQs are generated by the perturbation-based approach, resulting in the subsets PT-Easy and PT-Hard. For PT-Easy, each object in the AQ is replaced by a random but different object sampled from the candidate object set. For PT-Hard, the objects in AQ are kept but their attributes are replaced by different candidate attributes of the same object. Finally, the spatial relation terms in PT-Hard are replaced by antonyms, such as "left/right" and "top/bottom". Conflicting questions, like "What color are the black shoes?" are then eliminated. Fig. 2(g,h) and Fig. 2(c,d) show examples from PT-Easy and PT-Hard, with the perturbed text in red.

Human Annotation: Human annotators analyze all image-question candidates and decide which are true UQs. Following [43, 15, 26, 5], we use 8 expert annotators with experience in visual language research. The annotator is shown an image and two questions (see interface in appendix), and asked to choose from "*valid*" (corresponding to AQs) and "*invalid*" (UQs) options for each question. We instruct the annotator to choose "*valid*" if the decision is ambiguous, due to unclear images, confusing wording, or any other reason. These annotations are discarded.

This process produced $11,264$ UQs for CLIP-Hard, $5,689$ for CLIP-Easy, $6,130$ for PT-Easy and $5,963$ for PT-Hard. The next step aimed to sample a similar number of AQs, to balance the dataset. For CLIP-Hard and CLIP-Easy, we randomly sample AQs to pair with UQs. For each UQ, we consider the associated image and retrieve the AQs originally paired with this image in GQA. We then randomly sample one of these AQs. This produced $11,158$ questions for CLIP-Easy and $20,325$ for CLIP-Hard. For PT-Easy and PT-Hard, we pair with the original AQs for each perturbed UQ which results in $12,241$ questions in total for PT-Easy and $11,913$ for PT-Hard. See appendix for more details.

3.2. Dataset Analysis

UQ Categories: RGQA covers a wide spectrum of UQs, including questions without valid answers (*e.g.* Fig. 2 (b)), with false premise at object (*e.g.* Fig. 2 (e)) or attribute level (*e.g.* Fig. 2 (d)), and underspecified questions (*e.g.* "Do the snowpants look black and long?" for Fig. 2 (f)). Many UQs also have subtle mismatches with the image, which can only be spotted via high-level understanding of image semantics. For instance, in Fig. 2(b), both the predicate "wearing" and the object "shoes" exist in the image, so the model needs to understand the semantics of "wearing" and search for their subject. Hence, beyond evaluating robustness, RGQA also measures how strongly VQA models learn semantics.

Dataset Comparison: Table 1 compares RGQA to previous VQA datasets with UQs [40, 35, 14, 45, 21]. Several of these only address UQ detection. RGQA combines this with VQA, which better matches real-world applications. It also contains higher-quality human annotations, a better balance between AQs and UQs, and longer and more complex questions (last column) than previous datasets. Overall, it poses a greater challenge to model reasoning skills.

AQs vs UQs: To gain insight on the differences between AQs and UQs, we performed an analysis from two aspects. The first is to plot the distributions of image-question CLIP similarity scores, as shown in Fig. 3. Clearly, for VTFQ [40] and QRPE [35] the scores are smaller, indicating simpler questions, and the AQ/UQ distributions have less overlap, showing that they can be easily separated. VizWiz [14], CLIP-Hard, and PT-Hard have larger scores and stronger overlap between the two distributions, indicating that their UQs have finer-grained mismatch between image and question. However, while the CLIP score measures semantic similarity, it does not capture the answerability of UQs. The second strategy addresses this limitation, by plotting the distribution of questions by the first three words (See appendix). Other than a different order for the three most popular words ("Are", "Who" and "Which") and a few changes on the proportions, there are no major differences between the AQ and UQ distributions. This shows that AQs/UQs cannot be easily separated by question structure.

3.3. Evaluation Metrics

Since UQ detection is an OOD problem, we leverage well-established OOD practices for evaluation. However, because RVQA requires jointly solving UQ detection and VQA, the common OOD practice of reporting close-set accuracy and AUROC is not satisfying. We instead proposed to use the ACC-FPR curve, introduced as CCR-FPR curve in [9], which measures the joint performance. Given a VQA classifier f and a UQ detector g, ACC is the proportion of

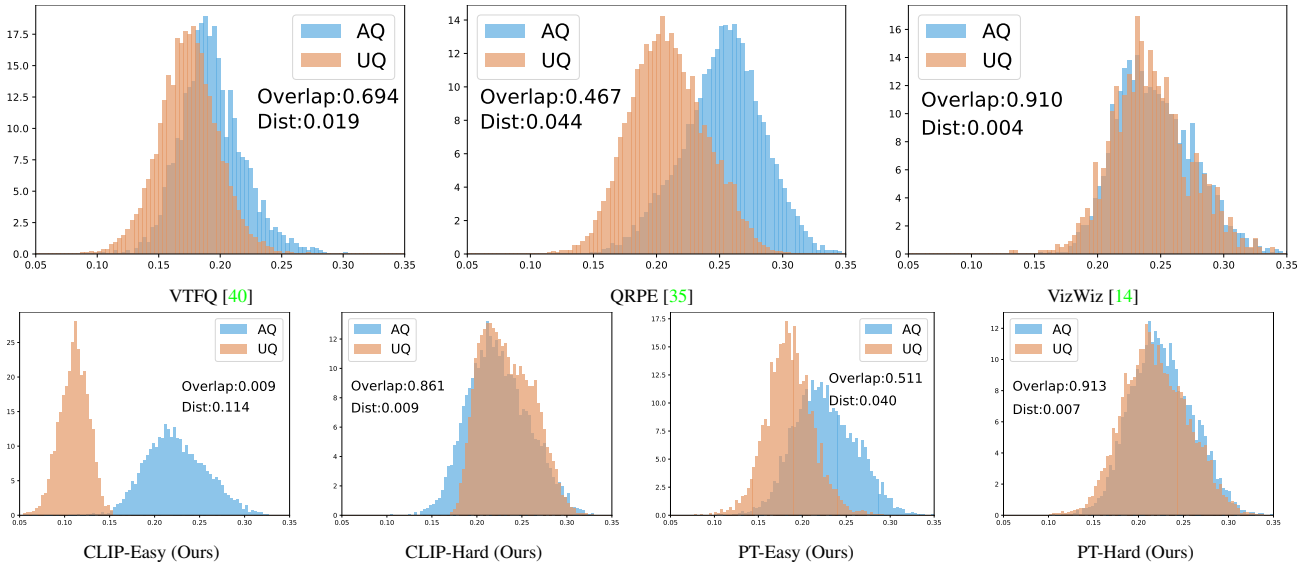


Figure 3: CLIP image-question similarity distribution of both AQs and UQs. The overlap area between 2 normalized histograms (sum of overall area=1) and the distance between the means are computed.

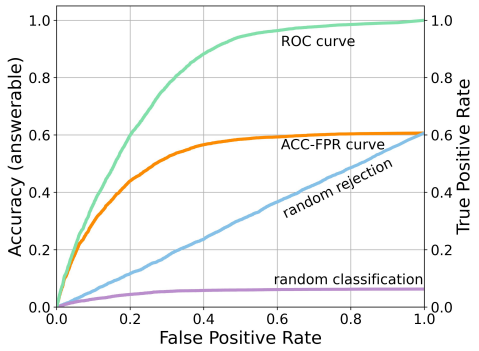


Figure 4: Comparison between ROC curve (green; right axis) and ACC-FPR curve (orange; left axis). See text for details.

AQs with correct VQA prediction and accepted as AQ, i.e.

$$\text{ACC} = \frac{|\{x_i | f(x_i) = a_i, g(x_i) = \text{AQ}, (x_i, a_i) \in \mathcal{D}^{aq}\}|}{|\mathcal{D}^{aq}|}, \tag{1}$$

where $x_i = (v_i, q_i)$ denotes image-question pair, a_i is the corresponding VQA answer and $\mathcal{D}^{aq}$ is the dataset of AQs. FPR is the proportion of UQs falsely accepted as AQ, i.e.

$$\text{FPR} = \frac{|\{x_i | g(x_i) = \text{AQ}, x_i \in \mathcal{D}^{uq}\}|}{|\mathcal{D}^{uq}|}, \tag{2}$$

where $\mathcal{D}^{uq}$ is the dataset of UQs. The ACC-FPR curve is drawn by connecting ACCs (y-axis) at different FPRs (x-axis) as in Fig. 4. We define the maximum value of the curve on the y axis (best accuracy the model can achieve on AQs) as full accuracy (FACC).

A RVQA model with a strong VQA classifier f and a UQ detector g (orange line) has higher FACC than a model with the same g but random f (purple line). On the other hand, a model with the same f but random g (blue line) has the same FACC but underperforms the RVQA model for all FPRs less than 1. Note that the ROC curve (green line) is the special case of ACC-FPR curve with FACC$=1$. As a single evaluation metric, we use *Area Under* ACC-FPR curve (AUAF), for joint performance, FPR at 95% FACC (FF95) for rejection, and FACC for classification.

4. Unsupervised RVQA Learning

In this section, we introduce unsupervised RVQA and three model-agnostic methods for unsupervised training.

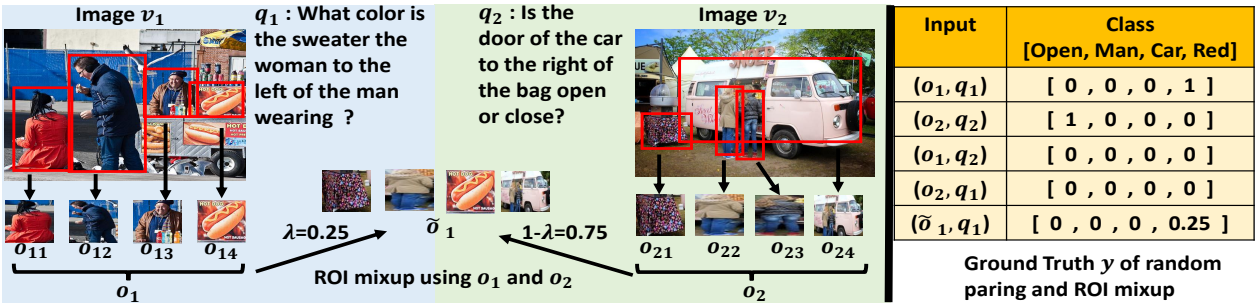
4.1. Unsupervised RVQA

Unsupervised RVQA learns a model, VQA classifier f and UQ detector g, from a dataset of AQs $\mathcal{D}^{aq}_{tr} = \{(x_i, a_i)\}$, *without* annotated UQs. At testing, $g(x)$ decides whether a pair x is accepted. If so, $f(x)$ then predicts an answer.

4.2. Training with Pseudo UQ

Inspired by recent OOD works using an auxiliary background dataset [9, 18, 32, 37] for training, we investigate training the RVQA model with a background dataset. For image classification, choosing a background dataset of reasonable scale and effective performance is non-trivial [32]. However, this is much simpler for RVQA: a simple and natural choice is to randomly pair images and question $\{(v_i, q_i)\}$ already available in the VQA dataset. Given an image v_i, we randomly sample a question q_k belonging to a different image $v_k \neq v_i$ to form a ***pseudo*** unanswerable image-question pair (v_i, q_k). Fig. 5 illustrates an example of this random paring, where image ν_1 is paired with question q_2. Like this example, most randomly sampled pairs are unanswerable[1]. The ***pseudo UQs*** are used to construct

[1]We inspected 100 pairs on GQA *train* and found 77% to be UQs.



Input	Class [Open, Man, Car, Red]
(o_1, q_1)	[0 , 0 , 0 , 1]
(o_2, q_2)	[1 , 0 , 0 , 0]
(o_1, q_2)	[0 , 0 , 0 , 0]
(o_2, q_1)	[0 , 0 , 0 , 0]
$(\tilde{o}_1, q_1)$	[0 , 0 , 0 , 0.25]

Ground Truth y of random paring and ROI mixup

Figure 5: Illustration of the pseudo UQ and RoI Mixup. The right table shows the label for different visual question inputs.

the unsupervised background dataset $\hat{\mathcal{D}}_{tr}^{uq}$.

With $\mathcal{D}_{tr}^{aq}$ and $\hat{\mathcal{D}}_{tr}^{uq}$, the VQA classifier f and binary UQ detector g can be trained to minimize the risk

$$\begin{aligned}\mathcal{R} &= E_{(x_i,a_i)\in\mathcal{D}_{tr}^{aq}}\mathbb{I}(f(x_i)\neq a_i) \\ &+ E_{x_i\in\mathcal{D}_{tr}^{aq}}\mathbb{I}(g(x_i)\neq \text{AQ}) + E_{x_i\in\hat{\mathcal{D}}_{tr}^{uq}}\mathbb{I}(g(x_i)\neq \text{UQ}),\end{aligned} \tag{3}$$

where the first term is the classification error and the last two are the detection error. Different from most OOD methods, which use softmax outputs [18], VQA models are usually trained as multi-label models. Let $\mathcal{Y} = \{1, \ldots, K\}$ be the set of possible answers. Then, the ground truth for i^{th} example $x_i = (v_i, q_i)$ and k^{th} answer is a binary variable, $y_{i,k} \in \{0, 1\}$, with $y_{i,k} = 1$ if the answer holds for x_i and $y_{i,k} = 0$ otherwise. The VQA model f has K binary outputs, where $f_k(x)$ is the predicted probability for k^{th} answer, implemented with sigmoid functions and trained with the *binary cross entropy* (BCE) loss

$$l_i = \sum_{k=1}^{K} y_{i,k} \log f_k(x_i) + (1 - y_{i,k}) \log(1 - f_k(x_i)). \tag{4}$$

In Sec. 5.2.1, several configurations of models f and g are ablated. Best results were obtained with an *integrated* model, where both f and g share the network according to

$$g(x) = \mathbb{I}(\max_k f_k(x) > \theta) \rightarrow y^* = \arg\max_k f_k(x), \tag{5}$$

where $\rightarrow$ means that the second equation is only implemented if $g(x) = 1$. The rejection step first checks that there is at least one f_k above threshold θ. If so, VQA is performed. Otherwise, the example x is identified as a UQ and rejected. This model minimizes (3) by simply assigning labels $y_{i,k} = 0, \forall k \in \mathcal{Y}$ to each UQ x_i, leading to

$$\mathcal{L}^{rvqa} = \frac{1}{N_{tr}^{aq} + N_{tr}^{uq}} \sum_{i=1}^{N_{tr}^{aq}+N_{tr}^{uq}} l_i, \tag{6}$$

where N_{tr}^{aq}, N_{tr}^{uq} is the size of $\mathcal{D}_{tr}^{aq}$ and $\hat{\mathcal{D}}_{tr}^{uq}$, respectively.

4.3. RoI Mixup

While random pairing is effective for constructing a background dataset of UQs, it tends to produce coarse-grained UQs, where (see Fig. 5) image and question are weakly related. To increase the coverage of fine-grained mismatches, we propose an additional sampling strategy denoted as *RoI Mixup*, motivated by mixup data augmentation [53, 52, 6]. Most VQA models have an object-based architecture [42, 7, 11, 31, 54], where image v_i is represented as a set of M (usually fixed) objects features $o_i = \{o_{i,m}\}_{m=1}^{M}$ detected by a pre-trained object detector [41]. In training, RoI Mixup randomly replaces a portion $1 - \lambda$, where $\lambda \in (0, 1)$, of the objects in image v_i with objects from another image $v_j \neq v_i$. This leads to a new and mixed set of objects $\tilde{o}_i$

$$\tilde{o}_i = \{o_{i,m}\}_{m=1}^{\lambda M} \bigcup \{o_{j,n}\}_{n=1}^{(1-\lambda)M} \tag{7}$$

with a new target one-hot vector $\tilde{y}_i = \lambda y_i$. Note that y_i can either be a correct answer, for AQs, or a zero vector, for UQs. Intuitively, by reducing the percentage λ of original objects, the probability of the question being AQ should also shrink by λ. Fig. 5 illustrates the mixing of two sets of visual features o_1 and o_2 with $\lambda = 0.25$ to synthesize the object set $\tilde{o}$. Following [53], λ is sampled as $\lambda \sim \text{Beta}(1, \beta)$ where β is a tunable hyper-parameter.

4.4. Model Ensembling

Random pairing and RoI Mixup are sampling strategies to create a background UQ dataset with a mix of coarse- and fine-grained UQs. It is also possible to improve the performance by regularizing the model output. As in the calibration literature [25, 46], we achieve this with model ensembles. Given C models $\{f^c\}_{c=1}^{C}$, model f^c predicts the probability of answer y_k as $p^c(y_k = 1|x) = f_k^c(x)$. Assuming the predictions of different models are independent, the probability predicted by the ensemble is $p^E(y_k = 1|x) = f_c^E(x) = \prod_{c=1}^{C} f_k^c(x)$. Model ensembling is then implemented by replacing f with f^E in (5), which produces more conservative predictions and rejects more UQs.

5. Experiments

In this section, we discuss a set of experiments that leverage the proposed RGQA dataset and metrics to evaluate the RVQA performance of both existing VQA models and proposed unsupervised RVQA training techniques. In what follows, "RP" means the model is trained with pseudo UQs,"Mix" means RoI Mixup examples are also used, and "Ens" is the ensembling of RP and Mix.

Table 2: RVQA comparison of recent VQA models, using MSP for the UQ detector g. * indicates that the model is not finetuned on GQA dataset. Larger AUAF and smaller FF95 are better.

	CLIP-Easy			CLIP-Hard			PT-Easy			PT-Hard			Avg.
Classifiers	AUAF	FF95↓	FACC	AUAF	FF95↓	FACC	AUAF	FF95↓	FACC	AUAF	FF95↓	FACC	AUAF
BUTD [7]	38.45	64.75	53.50	36.13	79.14	53.08	37.83	66.05	53.02	33.60	83.11	51.31	36.50
Uniter [7]	40.03	73.15	57.08	39.42	80.48	57.10	41.45	61.76	56.82	35.17	83.52	55.08	39.01
LXMERT [42]	42.39	76.25	0.87	42.60	78.92	60.49	47.30	61.79	59.94	38.12	85.14	58.76	42.60
SwapMix [13]	46.31	71.98	61.05	42.44	78.41	60.10	46.19	62.27	59.77	37.78	82.73	58.37	43.18
Vilt [24]	46.17	69.62	58.91	40.66	79.21	57.39	48.06	60.54	60.64	37.93	82.40	57.63	43.21
Oscar [31]	45.51	72.14	62.09	41.76	80.04	61.72	46.38	64.27	**63.44**	39.16	83.15	60.20	43.2
VinVL [54]	49.86	69.87	**64.89**	46.36	78.16	**64.61**	41.68	84.27	63.38	41.67	84.26	**63.37**	44.89
MDETR [22]	47.81	70.32	62.91	43.86	78.94	62.05	47.14	70.04	62.93	39.04	84.11	60.30	44.46
BLIP-VQAv2* [29]	35.93	69.39	51.67	34.94	82.10	51.13	37.44	69.33	52.49	32.62	86.91	49.79	35.23

Table 3: Comparison between different RVQA approaches on AUAF. Cells with light cyan background denote training with pseudo UQs. See appendix for full table with FF95 and FACC.

	BUTD [1]					UNITER [7]					LXMERT [42]				
RVQA Approaches	CLIP easy	CLIP hard	PT easy	PT hard	Avg.	CLIP easy	CLIP hard	PT easy	PT hard	Avg.	CLIP easy	CLIP hard	PT easy	PT hard	Avg.
FRCNN	33.58	30.73	31.43	26.94	30.67	35.81	33.09	33.67	28.82	32.84	38.43	35.22	35.73	31.00	35.09
MSP	38.45	36.13	37.83	33.60	36.50	40.03	39.42	41.45	35.17	39.01	42.39	42.60	47.30	38.12	42.60
ODIN	38.47	36.14	37.80	33.60	36.50	40.04	39.43	41.45	35.16	39.02	42.41	42.59	47.33	38.12	42.61
Maha	30.05	25.75	25.34	23.93	26.26	37.52	33.74	35.87	31.68	34.70	57.68	44.96	49.44	39.25	47.83
Energy	38.47	36.19	37.77	33.67	36.52	40.10	39.42	41.41	35.19	39.03	38.76	42.11	47.00	37.90	41.44
Q-C	53.04	36.20	47.14	29.06	41.36	56.61	38.67	50.12	30.93	44.08	60.39	41.31	53.11	33.18	46.99
Resample	40.25	37.73	39.54	34.78	38.07	58.66	48.08	53.65	39.84	50.05	60.47	50.80	55.74	42.18	52.29
RP w/ hard UQ	43.74	43.27	37.62	36.17	40.2	44.92	47.14	41.89	37.92	42.96	53.60	51.39	46.95	42.96	48.72
RP(Ours)	56.31	44.09	50.51	37.18	47.02	58.35	48.37	54.42	40.27	50.35	60.51	51.49	56.08	42.53	52.65
Mix(Ours)	56.85	44.70	51.27	37.59	47.60	59.08	49.00	54.63	41.50	51.05	60.79	51.91	56.83	43.56	53.27
Ens(Ours)	**57.25**	**45.46**	**51.95**	**38.46**	**48.28**	**59.62**	**49.65**	**55.79**	**42.14**	**51.8**	**61.03**	**52.42**	**56.90**	**43.75**	**53.52**

5.1. Experimental Set-up

An RVQA model consists of a VQA model f and a UQ detector g. RVQA methods vary along three dimensions: VQA model f, RVQA architecture, which determines how f and g are combined, and RVQA approach, which uses the architecture to implement the RVQA method. We consider several models, architectures, and approaches.

VQA models: We consider the nine VQA models [1, 7, 42, 24, 22, 31, 54, 29, 13] listed in Table 2. These represent a sampling of the literature, ranging from smaller models like BUTD [1] to recent large scale models, like VinVL [54]. All models are finetuned on GQA [19], except BLIP [29] whose finetuning requirements exceed our resources. BUTD/UNITER/LXMERT were trained for 1/7/7 epochs, respectively, with the original hyperparameters. For MDETR/OSCAR/VinVL/SwapMix, we used VQA checkpoints fine-tuned on GQA from the authors' githubs. Since Vilt [24] does not have a GQA checkpoint, it was fine-tuned on GQA using its pre-trained weights and fine-tuning procedure from prior works [22, 31]. See appendix for details.

RVQA approaches: We group RVQA approaches into two categories. *Post-hoc, training free methods* use the fine-tuned VQA model f directly, implementing g with post-hoc operations. These frequently involve thresholding a confidence score derived from the model predictions, a popular approach in the OOD literature . *Training based methods* re-train the VQA model, using unlabeled data (pseudo-UQs), to learn g. The proposed RP, Mix, and Ens methods are of this type. We considered the following approaches.

Post-hoc, training free methods.

MSP [16]: Confidence score is the largest probability output by VQA model; **ODIN [33].** Extension of MSP that uses temperature scaling and input processing to boost performance. For RVQA, input processing is only applied to visual features. The temperature is $1e5$ and the noise $1e-4$ for all datasets; **Maha [28].** Estimate class-conditional Gaussian distribution of the VQA model features and use the Mahalanobis distance with respect to the closest class as confidence score. **Energy [34, 47].** Energy scoring method, initially proposed for Softmax based models [34] and recently adapted to multi-label models [47]. We find that only considering the top-2 classes improves performance. **FR-CNN.** A rule-based method, which compares object names detected by Faster-RCNN [41] with the nouns in the question. All object names and nouns are converted into word stems. If there exist nouns that are not in the object names, the question is declared as UQ.

Training based methods.

Resample [32]. An OOD method that performs iterative adversarial weighting of background examples (i.e. pseudo UQs), assigns higher weights to harder examples and the reweighted dataset is trained. **Q-C [40].** A caption is generated per image and its similarity to the question is measured. While [40] adopts NeuralTalk2 [23], we use BLIP [29] captions. To measure similarity, we finetune a BERT model

that takes the concatenation of a caption and a question and predicts whether the two match, with a binary score.

RVQA architectures: Several configurations of model f and detector g were considered. **Integrated:** sequential implementation of g and f as in (5); **Branched:** a common backbone with decoupled classifier heads for f and g; **Multi-branched:** generalizes Branched by taking features from multiple layers; **Separated:** trains g and f separately, with different models [30]. $\mathbf{K+1}$: [55] defines UQs as an additional $(K+1)^{th}$ VQA class and trains f as a $K+1$-class classifier. The integrated approach is applicable to all methods discussed above. The remaining architectures are only possible for training-based methods since they require pseudo-UQs to train separate g heads, models, or classes.

5.2. Quantitative Results

The combinatorial space of RVQA methods, VQA models, and RVQA architectures makes a comparison of all possibilities infeasible. We instead use a factorial experiment: start by ablating the architecture given a model and method, then compare models given the best architecture, and finally compare different methods for a few models.

5.2.1 RVQA Architecture

We started by investigating if the multiple architectures possible for trained models have any benefit over the integrated architecture of (5), which can be universally used. These experiments used the LXMERT VQA model and RP training. Fig. 6 left compares the averaged AUAF of the different architectures on RGQA. The *integrated* architecture has top performance, followed by Separated that, besides not being universal, doubles parameter sizes and inference time. We use the integrated architecture in the following experiments.

5.2.2 VQA Model

We next compared the UQ robustness of the different VQA models, using the MSP RVQA approach. Table 2 shows that all models are quite vulnerable to UQs, with average AUAF across datasets below 45. This shows that there is significant room for improvement. Interestingly, larger and more recent models do not fare significantly better than smaller models. Despite their superior AQ performance (FACC), they have similar FF95 and AUAF to the smaller models at the top of the table. Since the smaller models are much easier to train, we use them in the remaining experiments.

5.2.3 RVQA Approach

We finally compared the proposed RP, Mix, and Ens to all prior RVQA approaches discussed above. In these experiments, all approaches use BUTD, UNITER or LXMERT

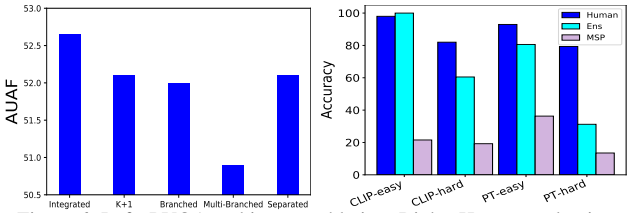


Figure 6: Left: RVQA architecture ablation. Right: Human evaluation.

models. Non-trainable approaches use models learned from AQs alone, trainable methods leverage a background dataset of pseudo UQs. For Mix, we empirically find the best β value per model and use it for all subsets. See appendix for more details. Table 3 (see appendix for full table with FACC and FF95) summarizes the performance of all models on the 4 RGQA subsets. The last column is the averaged AUAF across subsets. The table allows several conclusions.

Post-hoc approaches do not help. While MSP outperforms FRCNN, post-hoc approaches like ODIN, Maha, and Energy, which do not leverage pseudo-UQ, fail to improve on MSP. Surprisingly, these approaches have similar performance for CLIP-Easy and CLIP-Hard, even though CLIP-Easy has much coarser-grained image-question pairs.

Pseudo UQs are effective. The cyan cells of Table 3 show that training based methods, which leverage pseudo UQs, have significantly better RVQA performance (AUAF) than methods that do not. This is mainly due to a decrease of FF95 without sacrificie of FACC (see all metrics in appendix). Q-C consistently improves upon MSP by $5-10$ pts. Resample further improves performance for most models. However, the proposed RP improves on both, outperforming Q-C by $\sim$5.9 pts and Resample by $\sim$3.4 pts on average. This is somewhat surprising, since Resample is a more sophisticated sampling strategy. We hypothesize that Resample is unsuitable for the noisy background data generated by random pairing, likely applying larger weights to noisy examples (AQs) and hurting RVQA performance. The proposed Mix and Ens approaches have additional gains, producing the best results across VQA models. Finally, unlike prior RVQA works [30, 27], RP, Mix, and Ens do not harm VQA performance, even improving FACC. See appendix for GQA test set performance.

Impact of VQA model. Comparing the 3 models of Table 3, shows that RVQA approaches are more beneficial for models of higher VQA accuracy (FACC). For instance, for MSP on CLIP-Hard, from BUTD to LXMERT a FACC increase from 53.08 to 60.49 (shown in appendix) is accompanied by an AUAF increase from around 36 to 42. This shows that better VQA reasoning skills help the model detect UQs. However, note that these gains saturate quickly, as shown in Table 2. Together, the two tables show that RVQA benefits more from pseudo-UQ than from large models.

UQ Diversity. Most approaches achieve higher AUAF on CLIP-Easy and PT-Easy, because these 2 subsets have either low CLIP score or object level mismatch between

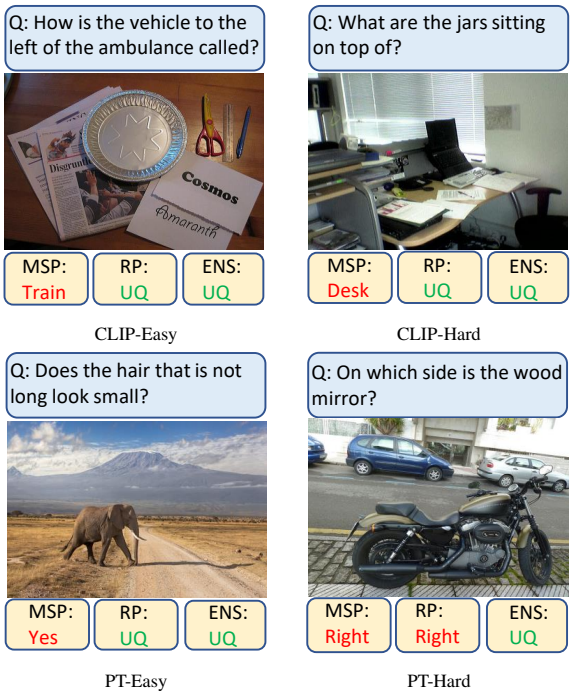


Figure 7: Qualitative examples for a threshold such that all models achieve 55% accuracy.

image and question. Conversely, most approaches underperform on CLIP-Hard and PT-Hard, where UQs have subtle mismatches at attribute or relation level. This trend holds across VQA models and subsets. We also consider RP training only on hard pseudo UQs, selected by CLIP score, (RP w/ hard UQs in Table 3), which produced a weaker AUAF than standard RP, especially on CLIP-Easy and PT-Easy. These results show the importance of UQ diversity.

5.3. Qualitative results

Confidence score distribution: Fig. 8 compares the confidence score distribution of the post-hoc MSP approach to the proposed RP and Ens training-based methods. It shows that MSP tends to be over-confident for both AQs (blue) and UQs (orange), while RP and Ens have higher (lower) scores for AQs (UQs). MSP is also not able to capture fine-grained mismatches. For instance, it assigns to UQ C a higher score than to AQ A. Finally, the confidence scores of AQ B show that RP and Ens can even detect incorrect annotations in the original GQA dataset.

Model prediction: Fig. 7 shows some qualitative examples from the four subsets of RGQA. The rejection threshold is set such that all models have accuracy of 55%. Ens correctly rejects all UQs, and RP three of the four, while MSP fails in all cases. Note that, for the fine-grained mismatches of the hard subsets, the VQA system tends to respond by statistical association -the missing jars are "sitting on the desk" and the nonexistent wood mirror is on the "right," which is the side of the bike closest to the camera.

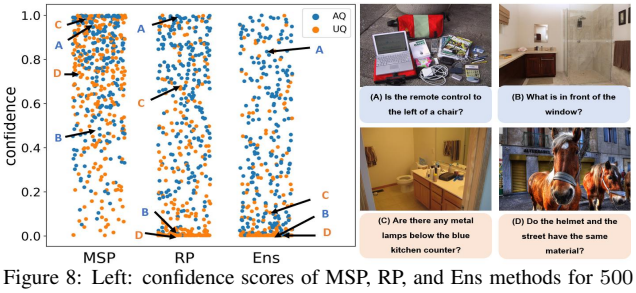


Figure 8: Left: confidence scores of MSP, RP, and Ens methods for 500 random samples. Right: qualitative examples. AQs/UQs are shown in blue/orange. B is an annotation error in the original GQA dataset.

5.4. Human Evaluation

To assess the challenge posed by the UQs in RGQA dataset, we conducted a human evaluation on MTurk. Workers were asked to perform the binary rejection on 50 AQs and 50 UQs for each subset. Fig. 6 right shows the rejection accuracy on UQs, comparing to models thresholded so as to achieve the same true positive rate on AQs. As expected, annotators found CLIP-Hard and PT-Hard more challenging. While Ens approaches human performance on the easier subsets, the gap on harder subsets is large.

6. Conclusion

We studied the problem of realistic VQA (RVQA) that aims to both reject UQs and answer AQs. Prior RVQA methods assume labeled UQs for training. It was argued that prior datasets are insufficient because they contain poor-quality images or lack of UQ diversity. To address this, we assembled the RGQA dataset, using 2 approaches to generate candidate UQs for human annotation. This allowed RGQA to cover broader granularities in image-question mismatch. A combination of pseudo UQs, RoI Mixup, and model ensembles was then proposed for unsupervised training of RVQA models. Experiments show that the resulting models outperform RVQA baselines for both easy and hard UQs. Comparison to human performance shows that more research is needed in RVQA.

7. Acknowledgments

This work was partially funded by NSF awards IIS-1924937 and IIS-2041009, a gift from Amazon, a gift from Qualcomm, and NVIDIA GPU donations. We also acknowledge and thank the use of the Nautilus platform for some of the experiments discussed above.

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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 7

[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*, pages 3674–3683, 2018. 1

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1

[4] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358, 2021. 3

[5] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19098–19107, June 2022. 4

[6] Jie-Neng Chen, Shuyang Sun, Ju He, Philip Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 6

[7] 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, 6, 7

[8] Jiacheng Cheng and Nuno Vasconcelos. Learning deep classifiers consistent with fine-grained novelty detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1673, 2021. 3

[9] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2, 4, 5

[10] Akshay Raj Dhamija, Manuel Günther, and Terrance E. Boult. Reducing network agnostophobia. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9175–9186, Red Hook, NY, USA, 2018. Curran Associates Inc. 3

[11] 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. 6

[12] 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. *International Journal of Computer Vision*, 127:398–414, 2017. 1

[13] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Loddon Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5068–5078, 2022. 7

[14] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1, 2, 4, 5

[15] Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P. Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *Conference and Labs of the Evaluation Forum*, 2018. 4

[16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 3, 7

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 3

[18] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019. 3, 5, 6

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 7

[20] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017. 1

[21] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 1965–1973, 2017. 2, 4

[22] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770, 2021. 7

[23] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2, 7

[24] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 7

[25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. 3, 6

[26] Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. QED: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics*, 9:790–806, 2021. 4

[27] Doyup Lee, Yeongjae Cheon, and Wook-Shin Han. Regularizing attention networks for anomaly detection in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1845–1853, 2021. 3, 8

[28] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 3, 7

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 7

[30] Mengdi Li, Cornelius Weber, and Stefan Wermter. Neural networks for detecting irrelevant questions during visual question answering. In *International Conference on Artificial Neural Networks*, pages 786–797. Springer, 2020. 3, 8

[31] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 6, 7

[32] Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 5, 7

[33] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 3, 7

[34] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020. 3, 7

[35] Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. The promise of premise: Harnessing question premises in visual question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 926–935, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 2, 4, 5

[36] 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. 1

[37] Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*. PMLR, 2022. 5

[38] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2307–2316, 2019. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3

[40] Arijit Ray, Gordon Christie, Mohit Bansal, Dhruv Batra, and Devi Parikh. Question relevance in VQA: Identifying non-visual and false-premise questions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 919–924, Austin, Texas, Nov. 2016. Association for Computational Linguistics. 2, 4, 5, 7

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6, 7

[42] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 1, 6, 7

[43] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 4

[44] Andeep S Toor and Harry Wechsler. Biometrics and forensics integration using deep multi-modal semantic alignment and joint embedding. *Pattern Recognition Letters*, 113:29–37, 2018. 2

[45] Andeep S. Toor, Harry Wechsler, and Michele Nappi. Question part relevance and editing for cooperative and context-aware vqa (c2vqa). In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, CBMI '17, New York, NY, USA, 2017. Association for Computing Machinery. 2, 4

[46] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *ECCV*, 2018. 3, 6

[47] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 2021. 7

[48] Pei Wang and Nuno Vasconcelos. Towards realistic predictors. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 37–53, Cham, 2018. Springer International Publishing. 1

[49] Hang Wu and May D. Wang. Training confidence-calibrated classifier via distributionally robust learning. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 295–304, 2020. 3

[50] Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving long-tailed recognition with deep realistic taxonomic classifier. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[51] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021. 3

[52] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 6

[53] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 6

[54] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021. 1, 6, 7

[55] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, pages 4401–4410, 2021. 3, 8