Domain-robust VQA with diverse datasets and methods but no target labels

Mingda Zhang Tristan Maidment Ahmad Diab Adriana Kovashka Rebecca Hwa
Department of Computer Science, University of Pittsburgh
{mzhang, kovashka, hwa}@cs.pitt.edu {tdm51, ahd23}@pitt.edu
https://people.cs.pitt.edu/~mzhang/domain-robust-vqa/

Abstract

The observation that computer vision methods overfit to dataset specifics has inspired diverse attempts to make object recognition models robust to domain shifts. However, similar work on domain-robust visual question answering methods is very limited. Domain adaptation for VQA differs from adaptation for object recognition due to additional complexity: VQA models handle multimodal inputs, methods contain multiple steps with diverse modules resulting in complex optimization, and answer spaces in different datasets are vastly different. To tackle these challenges, we first quantify domain shifts between popular VQA datasets, in both visual and textual space. To disentangle shifts between datasets arising from different modalities, we also construct synthetic shifts in the image and question domains separately. Second, we test the robustness of different families of VQA methods (classic two-stream, transformer, and neuro-symbolic methods) to these shifts. Third, we test the applicability of existing domain adaptation methods and devise a new one to bridge VQA domain gaps, adjusted to specific VQA models. To emulate the setting of real-world generalization, we focus on unsupervised domain adaptation and the open-ended classification task formulation.

1. Introduction

Visual question answering (VQA) borders on AI-completeness: it requires perception (visual and linguistic) and cognition. Despite the strong performance of recent VQA methods, they fall short of generalization and true reasoning: they are known to suffer from dataset bias [22], require domain-specific languages or domain-specific executable program annotations [34, 41], or must be trained separately for each new dataset.

Prior work in domain adaptation for object recognition examines how robust methods are when trained and tested on different datasets (domains), and further proposes techniques to bridge domain gaps. In contrast, there is a shortage of analyses of how domain-robust visual question answering methods are. Importantly, domain adaptation techniques cannot successfully be applied in the VQA setting in a straight-forward manner. First, VQA models take *inputs across multiple modalities*, each of which could contribute to the domain specificity of the trained models. Second, different VQA methods have *multiple intermediate stages* and processing steps over the inputs, which makes optimization challenging. Domain adaptation techniques could be applied at multiple of these stages, and domain adaptation can be performed jointly or separately from VQA training, with varying success. Third, *answer spaces* in different datasets are vastly different. While domain adaptation methods exist to tackle non-identical answer spaces in object recognition, this setting is not very common. Conversely, in VQA, it is the norm, since many datasets are highly specialized (for example, VizWiz [23] contains special answers "unanswerable" or "unsuitable image" because image-question pairs

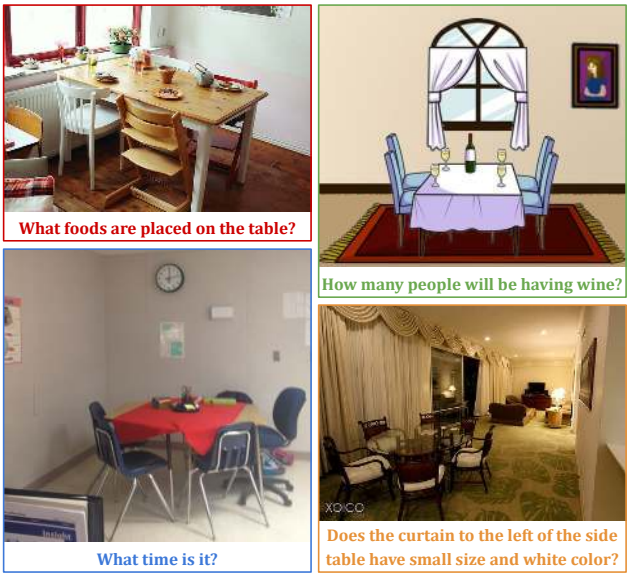


Figure 1. The same visual setting can be captured in different ways in VQA datasets, and paired with different information needs (questions). They may require deduction using visual contents, reading from a specific region of the image, or reasoning about complex spatial relationships. All examples are selected from real VQA datasets, *i.e.* VQA v2, VQA Abstract, VizWiz and GQA.

are provided by visually impaired users).

To tackle each of these challenges, we propose the following steps. First, to understand how the *multiple modalities* contribute to domain shifts, we break down and measure both visual and textual domain shifts across datasets. We disentangle shifts in image and question space by constructing synthetic dataset variants, to test how VQA methods respond to these separate shifts. To understand how the *multiple steps and mechanisms* in recent VQA models make them robust or fragile to shifts, we compare different families (classic two-stream, transformer, and neuro-symbolic methods) by exposing them to different shifts. We examine multiple mechanisms to bridge domain gaps for these methods, in the challenging setting of unsupervised adaptation where no labels from the target set are available, and discuss the differences in successful versus unsuccessful attempts. Third, to examine the contribution of *answer space differences*, we use the open-ended VQA classification formulation. Because no embedding is available for the answer options, the gap in answer spaces is more pronounced. We compare performance across datasets and observe relations between particular modality shifts and domain robustness.

In more detail, we compare image and question representations across nine datasets: VQA v1 and v2, VQA Abstract, Visual 7w, Visual Genome, COCO QA, CLEVR, GQA and VizWiz. We find there are large shifts in both visual and textual space, both at a low- and high-level (e.g. syntax and meaning). We separately apply automatic style transfer (for the visual modality) and paraphrasing (for the textual modality) to disentangle VQA methods' robustness separately to each of these artificial shifts. We also observe disparate contributions of these shifts in methods' performance across real domain gaps.

We find evidence that neuro-symbolic, compositional models are more robust to domain shift than others, because in those methods, perception and reasoning are more disentangled. We argue that reasoning has the potential to be domain-independent: for example, the process of reasoning about spatial relationships can in theory be abstracted away from pixel space, thus should not need retraining if the pixel space changes. Inspired by the potential of perception-reasoning disentanglement, we design a two-stage domain adaptation technique to bridge domain gaps. We show that this two-stage variant is more successful than a direct, one-stage application of [17], and a version of [47], for recovering performance lost due to domain gaps.

We are only aware of two prior works on domain adaptation for VQA [10, 38]. Both of these consider supervised domain adaptation (labels present in target dataset) while we operate in an unsupervised setting (labels on source dataset only). They work with fewer datasets (2-5) and apply domain adaptation to fewer and simpler VQA methods. Our work can be seen as a "reality check" for VQA methods, similar to prior reality checks for metric learning and weakly supervised object detection [13, 43].

To summarize, our contribution is to answer the following questions: (1) In what ways (visual, semantic, syntactic) are image-question pairs from recent VQA datasets different? (2) What kind of dataset differences most affect VQA generalization? (3) Which methods are more robust to synthetic visual shifts? (4) Which methods allow more generalization when training/testing on different VQA datasets? (5) What domain adaptation techniques most successfully bridge domain gaps? (6) What are the challenges of performing domain adaptation in unsupervised VQA?

2. Related Work

VQA method families. We consider three families of methods and their robustness to domain shifts. *Classic two-stream methods* [3, 32, 42, 55] represent the input image and question separately, then fuse the representations to obtain an answer. Perception and cognition are entangled. *Transformer methods* [12, 16, 39, 58, 65] compute multiple layers of attention between entities in each modality (e.g. words to visual regions). They often use unsupervised pre-training on massive vision-language datasets (e.g. images with text captions). Other than positional encodings, these methods have no separate relational reasoning component. *Neuro-symbolic, knowledge base, and graph methods* are conceptually distinct as they break down question-answering into modules. Some of these perform perception (e.g. recognize objects) while others perform cognition (e.g. relational reasoning about object position). Notable representatives include [1, 2, 4, 29, 31, 34, 41, 44, 61, 62]. For example, in [2, 4, 41, 60], entities are first parsed in a perception step, then reasoning takes a composable logic form, and questions are answered by verifying if objects satisfy a relationship implied by the question. [44] extract information about objects, then look up related concepts in a knowledge base, and perform reasoning using a GCN. In this paper, we show that the ability to disentangle perception and reasoning enables more domain-robust question answering.

Dataset bias in VQA. Prior work has found it is easy to introduce undesirable artifacts during dataset construction, which models can utilize to achieve misleadingly strong results. For example, [22] find that questions can be answered well using language priors (and bypassing the need for reasoning). [51] help a model cope with priors by discouraging it from producing an answer similar to that produced by an image-blind model. [56] accomplish robustness through adversarial regularization, [21] by constructing logic compositions of existing questions, [20] through semantic image mutations, and [27] by adding noise to the questions. All of these are concerned with bias or lack of robustness within a single dataset, but do not examine how datasets differ in terms of image and question compositions.

Domain adaptation (DA) and generalization (DG) cope

with domain shifts, e.g. for object recognition. Unlike generalization [9, 49, 57], adaptation [6, 17, 26, 47] assumes that some (unlabeled and/or sparsely labeled) data is available in the target domain. In the most common, classic DA setting, source and target class vocabularies overlap. Domain adaptation is challenging for VQA in that answer spaces do not overlap. This setting has also been tackled in DA for object recognition, but less commonly: in partial DA [7, 8, 67], the target class space is a subset of the source space; and in open-set DA [46, 54], the target space could have new classes not present in the source. The key idea in DA is to bridge the source and target distributions and arrive at a shared representation. Some influential methods include gradient reversal from a domain classifier to ensure domain-agnostic features [17], cycle-consistency [26], separating shared and domain-specific features [6, 38], minimizing moments of features in different domains [47], maximizing norm which correlates with transferrability [63], maximizing overlap between prototypes from different datasets [45], etc. Methods specific to particular vision tasks also exist, e.g. for object detection where gradient reversal is applied at both the image and instance (region) level [11], for semantic segmentation [68], etc. Some prior work [15, 36, 37, 40, 48, 53, 59, 66] leverages style transfer techniques to bridge domain gaps, while we use style transfer and language paraphrasing to factor our shifts in the complex multi-input setting (images and question) in VQA.

Prior work in domain-robust VQA. Our work is the first to perform fully unsupervised domain adaptation for VQA. There are only two prior works in domain-robust VQA we are aware of, but both operate in the supervised setting (i.e. some target labels are available). [10] find most of the domain shift lies in questions and answers. We consider more recent and diverse datasets, and find these contain significant image shifts as well. Further, [10] only considered a simple two-input MLP and two 2016 methods, while we consider three families of recent VQA methods. [10] is partially unsupervised; they do not use target labels to train the VQA model, but do use them to compute adaptable features. [64] only study the shift between two datasets, and only apply domain adaptation over a non-standard method for VQA. In contrast to [10, 64], we study nine datasets, and a new style transfer setting to isolate shifts in visual space.

3. Approach

We assume we have a labeled source dataset $\mathcal{D}^S = \{\mathbf{d}_1^S, \ldots, \mathbf{d}_i^S, \ldots, \mathbf{d}_{|\mathcal{D}^S|}^S\}$, where each $\mathbf{d}_i^S$ is an image-question-answer triplet $\{\mathbf{v}_i^S, \mathbf{q}_i^S, a_i^S\}$. The image and question are inputs to the VQA model, and the ground-truth answer is the desired output. We also have an *unlabeled* target dataset $\hat{\mathcal{D}}^T = \{\mathbf{d}_1^T, \ldots, \mathbf{d}_j^T, \ldots, \mathbf{d}_{|\hat{\mathcal{D}}^T|}^T\}$ where each $\mathbf{d}_j^T$ is an image-question pair $\{\mathbf{v}_j^T, \mathbf{q}_j^T\}$, and no answers are provided even in the training set. We aim to build a VQA model using $\mathcal{D}^S$ and $\hat{\mathcal{D}}^T$, which can answer questions in $\hat{\mathcal{D}}^T$. Any two datasets $\mathcal{D}^S$ and $\hat{\mathcal{D}}^T$ have potentially large domains gaps, in terms of marginal distributions (of images, questions, or answers) or conditional distributions (e.g. answers given the images or questions). Therefore, the major challenge is to maximize the performance on $\hat{\mathcal{D}}^T$ despite the domain gaps, and our strategy is to ensure the model trained on $\mathcal{D}^S$ is as transferable to $\hat{\mathcal{D}}^T$ as possible.

We measure domain gaps for nine datasets (Sec. 3.1), describe how to construct synthetic gaps to disentangle visual and linguistic shifts (Sec. 3.2), and how to adapt domain adaptation techniques to bridge gaps (Sec. 3.3) for individual VQA methods (Sec. 3.4).

3.1. Measuring real domain gaps

The first step towards building a domain-robust VQA model is to understand the multi-faceted dataset gaps. We analyze the following datasets: (1) VQA v1 [5]; (2) VQA v2 [22]; (3) Visual Genome [35]; (4) Visual7W [69]; (5) COCO-QA [52]; (6) GQA [30]; (7) CLEVR [33]; (8) VQA Abstract [5]; (9) VizWiz [23]. We could measure shifts in the following distributions across datasets: (1) $P(\mathbf{v})$; (2) $P(\mathbf{q})$; (3) $P(a)$; (4) $P(\mathbf{q}|\mathbf{v})$; (5) $P(a|\mathbf{v})$; (6) $P(a|\mathbf{q})$; (7) $P(a|\mathbf{v},\mathbf{q})$, where $\mathbf{v}$, $\mathbf{q}$ and a represent image, question and answer respectively. Here, we focus on measuring shifts in $P(\mathbf{v})$ and $P(\mathbf{q})$, To measure how much the corresponding distribution changes across datasets, we using Maximum Mean Discrepancy (MMD):

$$\begin{aligned}\text{MMD}(\mathcal{D}^S, \hat{\mathcal{D}}^T) &= \|\mathbb{E}_{X\sim\mathcal{D}^S}[\varphi(X)] - \mathbb{E}_{Y\sim\hat{\mathcal{D}}^T}[\varphi(Y)]\|_{\mathcal{H}} \\ &= \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} k(\mathbf{y}_i, \mathbf{y}_j) \\ &- \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(\mathbf{x}_i, \mathbf{y}_j) \end{aligned} \quad (1)$$

where k represents the RBF kernel and n_s, n_t represent sample size in the source and target domains. For visual representations, we use pretrained ResNet-101 [25] to extract image embeddings $\{\mathbf{v}_i, \mathbf{v}_j\}$ for 10,000 randomly sampled images in each pair of datasets $\{\mathcal{D}^S, \hat{\mathcal{D}}^T\}$. We use the final 2048-D embedding as *high-level semantic features*, and the spatially average-pooled embedding after conv3_4 layer (512-D) as *low-level features*. For questions, we measure both semantic and syntactic gaps using two different representations. For the *semantic representation*, we choose pre-trained BERT [14] to encode 10,000 randomly sampled questions $\{\mathbf{q}_i, \mathbf{q}_j\}$ from pairwise datasets $\{\mathcal{D}^S, \hat{\mathcal{D}}^T\}$. For *syntactic features*, we follow the approach in [19] to extract 20 low-level features: question length, number of conjunctions, pronouns, prepositions, etc. We show the results in Tables 1 and 2.

3.2. Constructing synthetic shifts to isolate effects

As we can see in Tables 1 and 2, many dataset pairs differ in both their image and question distributions. Our goal is to understand precisely how different VQA methods respond to shifts in each distribution, but this is not straightforward because both modalities would affect the VQA performance. Therefore, to disentangle domain gaps arising from the image or question modality, we synthetically construct gaps in either image or question space. To do this, we use image style transfer and question paraphrasing.

Specifically, we create stylized variants of each image in $\mathcal{D}^S$. Let $F(\mathbf{v}, \mathbf{f})$ be a style transfer function which takes in a content image $\mathbf{v}$ and style image $\mathbf{f}$ and outputs the content image now with a new style, $\mathbf{v}^{\mathbf{f}}$. We choose Ada-IN [28] as our style transfer function F. We also pay extra attention to ensure colors are preserved in the style transfer process, which is important to ensure answers to color-related questions remain valid. We achieve the color preservation by converting style-transferred images into the YUV color space, and copying the UV channels from the original images. We also experimented with the color histogram matching in [18], but ultimately chose luminance-only transfer. We also control the transfer strength α in [28] to avoid losing too much information. We manually verified color and answers were preserved on a small set of images.

For questions, let G be a paraphrasing function, $G(\mathbf{q}, \mathbf{g}) = \mathbf{q}^{\mathbf{g}}$, where $\mathbf{q}$ is a question and $\mathbf{g}$ is a reference "style". We finetuned a massively pretrained sequence-to-sequence generative T5 model [50] on Quora duplicate questions[1], to shift the question $\mathbf{q}$ to a different style.

Synthetic dataset pairs: We apply the image style transfer and question paraphrasing separately, to construct new pairs of VQA datasets that *only have domain shift in one modality*. For example, by experimenting on $\mathcal{D}^S = \{\mathbf{v}, \mathbf{q}, a\}$ and $\hat{\mathcal{D}}^T = \{\mathbf{v}^{\mathbf{f}}, \mathbf{q}\}$, the results would reveal the model's robustness on image domain shift. If we choose $\hat{\mathcal{D}}^T = \{\mathbf{v}, \mathbf{q}^{\mathbf{g}}\}$, then similar experiments would show the impacts from question domain shift. Note that in both settings, the answers are kept unchanged thus the impacts from answer space shift will be eliminated. We do not use the answers on the target domain to train, even though they are identical to those in the source domain.

3.3. Bridging domain gaps

Our goal is to ensure high accuracy on $\hat{\mathcal{D}}^T$, even though we have no ground-truth answers in the target domain as supervision. Thus, we minimize a loss of this type:

$$L(\mathcal{D}^S, \hat{\mathcal{D}}^T; \theta) = L_{ce}(\mathcal{D}^S; \theta) + \lambda L_{fd}(\hat{\mathcal{D}}^T, \mathcal{D}^S; \bar{\theta}) \quad (2)$$

In the above, θ refers to the parameters of a VQA model, to be defined in Sec. 3.4. L_{ce} is cross-entropy loss (computed on the source dataset only), and L_{fd} is a loss that computes the discrepancy between the **f**eature **d**istributions of the source and target domains, computed over images and/or questions. The bar in $\bar{\theta}$ refers to the model component over which we apply L_{fd} (see Sec. 3.4).

For L_{fd}, we consider two domain adaptation strategies from object recognition, and a new variant of one of them. First, we adapt an adversarial domain classifier as described in DANN [17], and reverse its gradient. The idea is to learn features that prevent the model $\bar{\theta}$ from being able to successfully distinguish between source and target domains. To successfully adapt DANN, we have to consider the differences between DA for object recognition and DA for VQA. In particular, DANN can be applied over both image and question inputs (or over intermediate representations that depend on both). We describe how we adapt DANN for each VQA method, in Sec. 3.4. Second, we use a simplified single-source version of Moment Matching [47] which minimizes moment-related distances to reduce domain gaps.

We treat answering as 1000-way open-ended classification, and ensure the output space is the same for all datasets; we provide details in Sec. 4. Alternatives include answering as generation (which is challenging for automatic evaluation) or as a multiple-choice task (which may introduce biases due to the choice of the incorrect answers [10]).

3.4. Adaptation for VQA models

VQA models: We analyze domain robustness of VQA models from different families: (1) Classic two-stream methods (RelNet [55]); (2) Neuro-symbolic methods (NSCL [41]); and (3) Transformer methods (LXMERT [58]). We also test MAC [29] and TbD [42], which are hybrids of classic and neuro-symbolic methods.

Challenges: Applying domain adaptation is challenging in the unsupervised open-ended classification setting. The first challenge is the *lack of labels on the target dataset, in the setting we assume.* To the best of our knowledge, only two prior works [10, 64] tried to tackle the domain adaptation problem for VQA. However, *one leveraged multiple-choice options [10], and both leveraged labels in the target domains, which are not available in our setting.* More specifically, Chao *et al*. [10] minimize Jensen-Shannon Divergence (JSD) to achieve domain adaptation in the multiple-choice VQA task. All datasets they investigated are derived from COCO so there is little visual domain shift, thus they only focused on dealing with question and answer/decoys shift. We noticed their improvements mostly come from minimizing JSD over answer/decoys (*i.e*. minimizing JSD over *questions* brings negligible $< 0.4\%$ performance boost). In addition, [10]'s feature transformation method (impoverished VQA model without image inputs) requires labels from the target dataset. However, this is not applicable under the open-ended setting because we assume

[1] https://www.kaggle.com/c/quora-question-pairs

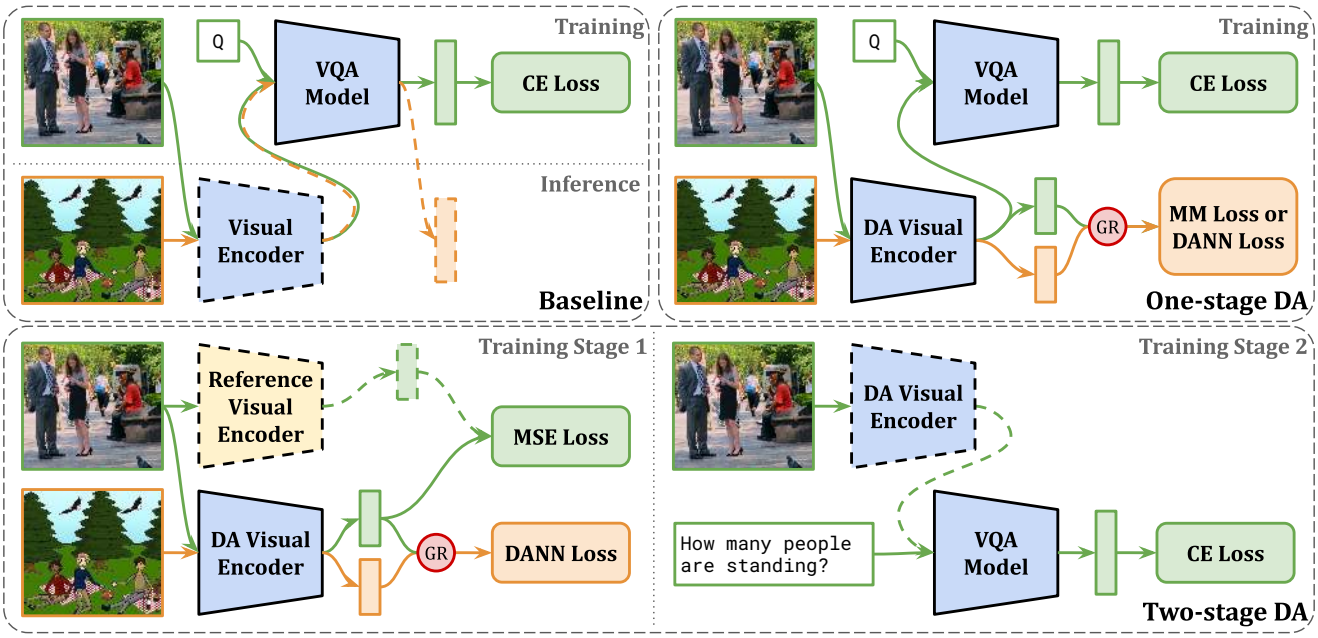


Figure 2. Illustration of the domain adaption strategies as described in Section 3.4. We show both training and inference stages for the baseline where no domain adaptation is applied (top left), and the training procedure for one-stage domain adaption with DANN or Moment Matching (top right). In the bottom, we show the training procedure of our proposed two-stage DANN approach. Specifically, we first train a domain-adaptive visual feature extractor in the first stage, with a MSE loss to encourage preserving semantics, and a domain confusion loss (DANN) to reduce domain gaps. Next, using extracted features from the domain-invariant extractor, we train a VQA model on source data. The gradient reversal layer (GR) [17] is only used with DANN. Dashed lines indicate no gradients due to module being frozen or for inference only.

no answers and decoys for the target dataset.

A second challenge is that *joint optimization of VQA with a domain adaptation objective (Eq. 2) is unstable* because the VQA loss and DA loss may compete, making optimization difficult. This is especially true for complex, state-of-the-art VQA models. To cope with the challenge of applying domain adaptation over VQA, [10] break up adaptation and VQA training into two stages; they primarily use a simple MLP, while we evaluate DA with recent VQA models. [10] empirically use a GAN-like approach to estimate JSD, which makes their training computationally intensive and hard to adapt to more complicated VQA models. [64] also reports similar challenges in training a complex multitask (VQA+DA) method, and they handle it by carefully tuning the scalars for their multitask loss. Notably, they make the scalars corresponding to the unsupervised feature alignment very small (*e.g.*, 0.003, 0.025), and the multiplier for the source classifier is also small (0.001 vs 1 for the supervised target loss). This highlights the challenge of leveraging transfer from the source domain without target labels.

Baseline and one-stage approaches: We report the performance of two reference models: (1) the accuracy on the source dataset, which indicates model capacity, and (2) the accuracy on the target dataset assuming target labels are fully available, which serves as an empirical upper bound for domain adaptation. The simplest baseline is directly applying a model trained on a source dataset, on test data from the target domain, without any domain adaptation. The training and inference procedure is illustrated in Fig. 2. As another baseline, we also investigated an end-to-end pipeline to combine the DANN training with VQA training, shown as "One-stage DA". Specifically, we added the domain discrimination loss and reversed its gradients to update the visual representations. However, it is non-trivial to find the best place for applying domain discrimination for different VQA methods. For example, for MAC we added a linear classifier to distinguish the domains and applied the DANN loss on the visual embedding before feeding them into the MAC unit. In addition to DANN, we experimented with moment matching [47] where the first- and second-order moments are enforced to align across domains. In this case the gradient reversal layer is no longer needed.

Proposed two-stage DA approach: To better cope with the challenges, we also propose a two-stage approach to build a domain-invariant feature extractor and VQA module sequentially. A figurative illustration of the process is shown in Fig. 2 (bottom). The motivation for breaking up domain adaptation and VQA modeling is to stabilize the training for greater robustness. The idea is partially inspired by neuro-symbolic methods, which separate perception (in this case, feature extraction) and reasoning (the VQA model after feature extraction). Our two-stage strategy is summarized as:

	Visual 7W	VG	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abs.	GQA Bal.	VizWiz
Visual 7W	–	0.04	0.18	0.18	0.56	0.88	0.18	0.46	0.25
VG	0.01	–	0.16	0.16	0.54	0.87	0.16	0.44	0.27
VQA v1	0.06	0.07	–	0.00	0.44	0.81	0.03	0.34	0.28
VQA v2	0.06	0.07	0.00	–	0.44	0.81	0.03	0.35	0.28
COCO QA	0.20	0.20	0.15	0.15	–	0.69	0.44	0.26	0.58
CLEVR	0.22	0.22	0.17	0.17	0.19	–	0.81	0.58	0.76
VQA Abs.	0.06	0.06	0.02	0.02	0.15	0.19	–	0.34	0.27
GQA Bal.	0.10	0.11	0.06	0.06	0.15	0.13	0.07	–	0.43
VizWiz	0.06	0.06	0.10	0.10	0.23	0.22	0.10	0.12	–

Table 1. Domain gaps in question space; red shading is MMD over 768-D BERT embeddings, blue is MMD over 20-D syntax statistics.

	Visual 7W	VG	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abs.	GQA Bal.	VizWiz
Visual 7W	–	0.00	0.01	0.01	0.01	0.10	0.08	0.00	0.04
VG	0.01	–	0.00	0.01	0.01	0.10	0.08	0.00	0.04
VQA v1	0.02	0.02	–	0.00	0.00	0.10	0.08	0.01	0.04
VQA v2	0.03	0.02	0.01	–	0.00	0.10	0.08	0.01	0.03
COCO QA	0.04	0.04	0.03	0.03	–	0.10	0.08	0.01	0.03
CLEVR	0.54	0.54	0.54	0.54	0.54	–	0.10	0.10	0.09
VQA Abs.	0.36	0.36	0.36	0.36	0.36	0.59	–	0.08	0.08
GQA Bal.	0.03	0.03	0.03	0.03	0.04	0.54	0.36	–	0.04
VizWiz	0.22	0.22	0.21	0.21	0.21	0.52	0.42	0.22	–

Table 2. Domain gaps in image space; red shading is MMD over ResNet-101 2048-D features, blue is MMD over conv3_4 512-D features.

1. Extract features for images in the source dataset, as defined in the VQA method (*e.g.* use pre-trained ResNet).
2. Train a domain-invariant feature extractor with both source and target datasets (without labels), using (a) an MSE loss which encourages the extracted features on source dataset to preserve semantics, and (b) a BCE loss with gradient reversal layer to prevent distinguishing the source and target domains.
3. Apply the backbone from step 2 to extract visual features and train a VQA model on the source dataset.
4. Take the visual feature extractor from step 2 and VQA model from step 3, then feed in the target dataset and evaluate the performance.

VQA method specifics: Each VQA method extracts features in a particular way, resulting in small variances in our two-stage DA implementation. For MAC and TbD, feature extraction is executed with ResNet-101 prior to training the VQA model, following the methodology outlined previously. NSCL uses ResNet-34 to extract features from different regions in the image (region proposals via a pre-trained Mask R-CNN [24]), and allows for NSCL to fine-tune ResNet-34 during training. To most closely follow our two-stage methodology, we replace the pretrained ResNet-34 with a frozen ResNet-34 backbone trained in step 2. RelNet uses 4 convolutional layers to extract features from the images. We used a pre-trained set of these convolutional layers to export the source features for VQA and DA. For LXMERT, the initial visual features are from pre-trained Faster R-CNN and processed by vision-only transformer layers. We kept the Faster R-CNN backbone untouched and fine-tuned the transformer layers to be domain-invariant.

4. Experimental Validation

We show four groups of results: shifts in image and question space for nine datasets (Sec. 4.1), robustness of five methods to synthetic shifts in visual or textual space using the CLEVR dataset (Sec. 4.2), different ways to apply unsupervised domain adaptation using MAC on three datasets (Sec. 4.3), and finally robustness of two methods using eight real dataset pairs (Sec. 4.4).

4.1. Domain shifts in nine datasets

Tables 1 and 2 show how the questions and images in nine datasets differ. Each table is a composition of two triangles. In Table 1, the lower triangle contains Maximum Mean Discrepancy (MMD) statistics using BERT embeddings, while the upper triangle shows MMD statistics using syntax features (Sec. 3.1). MMD computes how different two distributions are, with higher values indicating larger difference. The shading ranges from white to red/blue, with darker, more vivid colors indicating larger values.

In the lower triangle of Table 1, we observe that Visual 7W and Visual Genome (VG) are similar, and VQA v1 and v2 are similar, as expected. GQA is similar to VQA v1/v2 in terms of semantics (captured through BERT), but it is different in terms of syntax. VQA Abstract is much more similar to VQA v1/v2 in terms of syntax than to other datasets (blue triangle), but in terms of semantic content (red triangle), it is also fairly similar to Visual 7W and VG. COCO QA and CLEVR stand out from other datasets both in terms of semantics and syntax (both rows/columns for COCO QA and CLEVR have high values except on diag-

Method / Type	Source Acc.	Target Acc. (direct)	Target Acc. (2-stage DANN)	Target Acc. (10% scratch)	Target Acc. (10% finetune)	Target Acc. (full)
NSCL (NS)	**98.0**	**59.7**	**68.6**	60.0	75.8	**95.9**
MAC (NS/CL)	93.4	**62.6**	**65.2**	**84.6**	**82.1**	88.6
TbD (NS/CL)	**99.1**	36.3	41.3	**72.5**	**84.2**	**95.3**
RelNet (CL)	93.7	44.8	47.2	61.5	77.1	91.4
LXMERT (TR)	94.8	58.0	–	60.9	65.9	91.3

Table 3. Method robustness on CLEVR, using style transfer of the original images (domain shift in image space). We bold the best two results per column. The most important columns are Target (direct) and Target (2-stage DANN) as they require no supervision on the target. We observe neuro-symbolic methods are most robust. – means performance degraded on LXMERT with DANN.

Methods	Q	I1	I1+Q	I2	I2+Q
NSCL (NS)	–	**71.0**	–	60.6	–
MAC (NS/CL)	52.2	45.9	28.1	60.9	37.9
TbD (NS/CL)	**52.9**	55.7	**36.1**	**70.4**	**42.6**
RelNet (CL)	49.6	20.5	19.1	46.2	31.6
LXMERT (TR)	**53.4**	50.6	**36.6**	58.0	40.5

Table 4. Method robustness on CLEVR. We show performance under artificial *Question* shifts, followed by *Image* shifts with two styles (resulting in I1 and I2), and two settings where both Image and Question shifts are applied (I+Q). – means we were unable to test on NSCL since their semantic parser is not open-sourced. We bold the best result and those within 1% of the best.

	VQA v2	CLEVR	GQA Bal.
Source Accuracy	54.0	95.8	44.6
Target (direct)	41.0	45.9	37.3
Target (1-stage DANN)	42.2	45.7	37.4
Target (1-stage MM)	42.6	46.6	**38.6**
Target (2-stage DANN)	**42.8**	**46.7**	38.5
Target (full)	49.1	90.0	42.1

Table 5. Different DA methods on MAC (NS/CL), image shift.

onal), but CLEVR's syntax (darker blue) stands out more than COCO QA's syntax (lighter blue), while in terms of semantics they are similarly unique. GQA and VizWiz are also relatively unique, but less so than CLEVR. In Sec. 4.4, we show how these shifts affect cross-dataset performance.

Some dataset pairs that were distinct in terms of questions are similar in terms of images, and vice versa, as shown in Table 2. COCO QA is now fairly similar to other datasets (in terms of images), but VQA Abstract and VizWiz become more unique (darker shading) than in Table 1; they are two of the three rows/columns with high values, in addition to CLEVR. Results are generally consistent in the lower/upper triangles (from ResNet layers closer to the output or input, respectively) except that in higher dimensions (lower triangle), absolute MMD scores are larger.

4.2. Methods' robustness to synthetic domain shifts

Tables 3 and 4 show how robust different VQA methods are to synthetic shifts on the CLEVR dataset. In Table 3, we show robustness to visual shifts. We evaluate performance by the method on the original CLEVR dataset, performance of the model trained on CLEVR and applied in the shifted setting (*e.g.* style-transferred images) directly, target performance with unsupervised domain adaptation (specifically, 2-stage DANN), and three supervised settings for comparison – two that use 10% of the target training data, and one that uses 100% of the target training data. We use the default recommended hyperparameters without exhaustive search. We observe all methods' performance drops in the Target setting compared to Source, as expected. However, in Target (direct) and Target (2-stage DANN), both of which do not use labels on the target, NSCL and MAC (both neuro-symbolic or NS hybrid) retain the best performance. Using a small amount of target data for fine-tuning, MAC and TbD (both NS hybrids) perform best.

Table 4 demonstrates each method's change in performance when evaluated with paraphrased questions (first column), style-transferred images using two separate styles, I1 and I2 (second and fourth columns), and combined question and image shifts (third and fifth columns). LXMERT is most robust to question shifts, likely due to its extensive pre-training on language data, followed by TbD. Neuro-symbolic or hybrid methods (NSCL or TbD) are most robust to image shifts, consistent with our hypothesis.

4.3. Domain adaptation for synthetic shifts

In Table 5, we evaluate different domain adaptation strategies with MAC as the backbone model on VQA v2, CLEVR, and GQA Balanced, where artificial domain shifts are created in the image space. By comparing Source and Target (full) accuracy, we deduce the image style transfer preserves the information required for VQA as accuracy only drops slightly. However, in all datasets, we see quite significant performance drop if a trained model is directly applied to the corresponding target dataset. The domain adaptation strategies (1-stage DANN [17] and Moment Matching [47], and our 2-stage DANN) help to different degree. Our proposed 2-stage DANN is always significantly better than then 1-stage DANN, and better than the 1-stage MM on two of three datasets. Note that differences between methods are significant in that the range between Target (direct) and Target (full) is very small for two of the three datasets. It is worth mentioning that training the 1-stage DANN baseline is highly unstable as the optimization is more difficult. We repeated the experiments multiple

	Datasets		Accuracy (%)			
	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{A} \rightarrow \mathcal{B}$	$\mathcal{B} \rightarrow \mathcal{A}$
MAC	VQA v2	CLEVR	53.3	95.9	29.8	18.7
		GQA Bal.		44.4	32.0	35.6
		VQA Abs.		48.3	33.6	31.7
		VG		33.3	26.2	23.1
LXMERT	VQA v2	CLEVR	67.6	84.9	31.6	34.8
		GQA Bal.		58.2	50.5	51.5
		VQA Abs.		56.3	34.3	34.6
		VG		41.0	36.7	31.4

Table 6. Robustness across VQA datasets; best viewed in color.

Datasets	Image		Question	
	Appearance	Semantic	Syntactic	Semantic
CLEVR	High	High	High	High
GQA Bal.	Low	Low	Med. High	Medium
VQA Abs.	Med. High	Med. High	Low	Low
VG	Low	Low	Med. Low	Medium

Table 7. Summary of shifts, VQA-v2 $\leftrightarrow$ selected datasets.

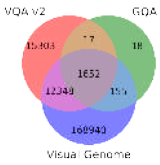
times and only preserved the 1-stage DANN models that did not collapse. Because of the challenges mentioned, on real dataset shifts, we only achieved marginal gains using domain adaptation, over directly applying the source model, consistent with prior work [10, 64].

4.4. Generalization under real domain shifts

Table 6 shows the robustness of two recent VQA methods among five datasets: VQA v2, CLEVR, GQA Balanced, VQA Abstract and Visual Genome. These datasets have different answer spaces, as shown in Fig. 3. Since the final classification layer is coupled with the answer vocabulary, models trained on one dataset cannot be directly applied to another. To mitigate this issue, we obtain a shared 1000-class answer space by computing the 1000 most common answers across all five selected datasets. We report training and evaluating a model on the same dataset (*i.e.* Acc of $\mathcal{A}$ and $\mathcal{B}$), and training on one and evaluating on the other (*e.g.* Acc of $\mathcal{A} \rightarrow \mathcal{B}$ denotes training on $\mathcal{A}$ and evaluating on $\mathcal{B}$). The accuracy is calculated on the validation split for individual datasets (except for GQA where we use testdev split as recommended), and is obtained by matching the top-1 prediction with the ground-truth answer(s).

Since source/target datasets have different upper bounds (*i.e.* $\mathcal{B}$ Acc), we normalize the transferred accuracy by dividing by $\mathcal{B}$, and illustrate the relative normalized performance using the intensity of shading: darker background of a cell indicates higher ratio of the transferred accuracy and the source/target accuracy. Blue backgrounds measure how well a transferred model $\mathcal{A} \rightarrow \mathcal{B}$ performs compared to its upper bound, as they are all transferring from the same source $\mathcal{A}$, while red backgrounds measure how well different source models $\mathcal{B} \rightarrow \mathcal{A}$ transfer to the same target $\mathcal{A}$.

By comparing the accuracy on the training and evaluation datasets, we see that in most cases LXMERT (TR)



Dataset	Most Frequent Answers
VQA v2	yes, no, 2, 1, white, 3, 4, ...
CLEVR	no, yes, 1, 0, small, rubber, ...
GQA Bal.	no, yes, left, right, man, ...
VQA Abs.	yes, no, 2, 1, red, 3, white, ...
VG	1, white, 2, daytime, black, ...

Figure 3. Venn diagram of answer vocabulary of three datasets. A large portion of answers are not shared across datasets, and the distribution (*e.g.* most frequent answers) may differ as well.

generalizes better across datasets than MAC (NS/CL). We hypothesize that transformer-based methods like LXMERT benefit from their massive pre-training (which includes disjoint GQA and VQA v2 data). We also observe that GQA and Visual Genome are more useful sources when transferring knowledge to VQA v2, compared to CLEVR. This observation is consistent with our statistical analysis in Tables 1 and 2, and for simplicity we extracted relevant information in Table 7. We see that GQA Balanced and Visual Genome are similar to VQA v2 in multiple aspects. We also note that GQA Balanced has *smaller semantic shifts than syntactic shifts* with respect to VQA v2, while VG has smaller syntactic than semantic shifts with VQA v2. This makes GQA Balanced more helpful as a source dataset (darker shading for GQA$\rightarrow$VQA-v2 than for VG$\rightarrow$VQA-v2 in Table 6, for both MAC and LXMERT). Finally, the only case MAC is more robust than LXMERT (in terms of shading) is VQA-v2$\leftrightarrow$VQA-Abstract, which is the dataset with largest visual shifts after CLEVR. One possibility is that LXMERT is better suited to deal with question shifts and MAC with visual shifts, because of its neuro-symbolic nature and dedicated perception module.

5. Conclusion

We showed domain differences between VQA datasets can come from the visual and linguistic space; different methods are more susceptible to visual or linguistic shifts, and high-level semantic shifts make methods more fragile than syntactic ones. We found neuro-symbolic methods are more robust to synthetic visual-only domain shifts and some real dataset shifts, but transformer methods handle real linguistic and some visual shifts better due to pretraining. We demonstrated that while unsupervised domain adaptation in VQA is challenging, better gains can be made through a two-stage DANN which shares similar intuition as neuro-symbolic methods. In the future, we will explicitly handle shifts in answer space, and develop DA techniques that can flexibly choose how much to adapt over each modality,

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grant No. 1718262. It was also supported by the Univ. of Pittsburgh Momentum Fund, and Google/Amazon/Adobe gifts. We thank the reviewers and AC for their suggestions.

References

[1] Somak Aditya, Yezhou Yang, and Chitta Baral. Explicit reasoning over end-to-end neural architectures for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[2] Saeed Amizadeh, Hamid Palangi, Oleksandr Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning". In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 2

[3] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[5] 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 (ICCV)*, 2015. 3

[6] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 3

[7] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[8] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[9] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[10] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4, 5, 8

[11] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[12] 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. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[13] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 3

[15] Aysegul Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *arXiv preprint arXiv:1807.09384*, 2018. 3

[16] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. 2, 3, 4, 5, 7

[18] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[19] Katy Gero, Chris Kedzie, Jonathan Reeve, and Lydia Chilton. Low level linguistic controls for style transfer and content preservation. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, 2019. 3

[20] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[21] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[22] 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 (CVPR)*, 2017. 1, 2, 3

[23] 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 (CVPR)*, 2018. 1, 3

[24] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 3

[26] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alyosha Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 3

[27] Jia-Hong Huang, Cuong Duc Dao, Modar Alfadly, and Bernard Ghanem. A novel framework for robustness analysis of visual qa models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2

[28] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[29] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 4

[30] 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 (CVPR)*, 2019. 3

[31] Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[32] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[33] 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 (CVPR)*, 2017. 3

[34] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[35] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 2017. 3

[36] Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P Xing. Semantic-aware grad-gan for virtual-to-real urban scene adaption. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 3

[37] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[38] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[39] 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 (NeurIPS)*, 2019. 2

[40] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[41] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 4

[42] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4

[43] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[44] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[45] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[46] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[47] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4, 5, 7

[48] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 3

[49] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2020. 4

[51] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[52] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 3

[53] Adrian L Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 3

[54] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[55] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 4

[56] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[57] Yi-Zhe Song. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[58] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 2, 4

[59] Chris Thomas and Adriana Kovashka. Artistic object recognition by unsupervised style adaptation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018. 3

[60] Ben-Zion Vatashsky and Shimon Ullman. Vqa with no questions-answers training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[61] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural symbolic models for interpretable visual question answering. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. 2

[62] 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 (TPAMI)*, 2018. 2

[63] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[64] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multi-modal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. 3, 4, 5, 8

[65] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[66] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[67] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[68] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[69] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3