Robust Change Captioning

Dong Huk Park, Trevor Darrell, Anna Rohrbach

University of California, Berkeley

Abstract

Describing what has changed in a scene can be useful to a user, but only if generated text focuses on what is semantically relevant. It is thus important to distinguish distractors (e.g. a viewpoint change) from relevant changes (e.g. an object has moved). We present a novel Dual Dynamic Attention Model (DUDA) to perform robust Change Captioning. Our model learns to distinguish distractors from semantic changes, localize the changes via Dual Attention over "before" and "after" images, and accurately describe them in natural language via Dynamic Speaker, by adaptively focusing on the necessary visual inputs (e.g. "before" or "after" image). To study the problem in depth, we collect a CLEVR-Change dataset, built off the CLEVR engine, with 5 types of scene changes. We benchmark a number of baselines on our dataset, and systematically study different change types and robustness to distractors. We show the superiority of our DUDA model in terms of both change captioning and localization. We also show that our approach is general, obtaining state-of-the-art results on the recent realistic Spot-the-Diff dataset which has no distractors.

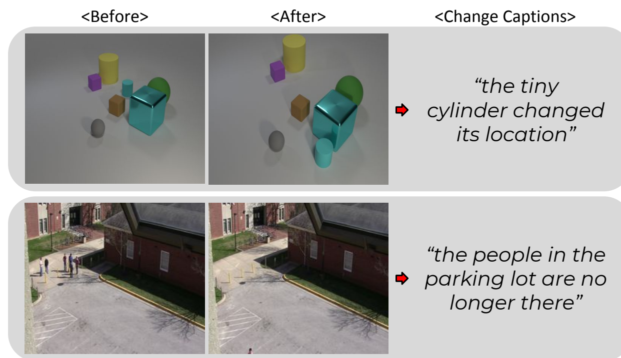

Figure 1: Robust Change Captioning requires semantic visual understanding in which scene change must be distinguished from mere viewpoint shift (top row). Not only does it require accurate localization of a change, but it also requires communicating the change via language. Our Dual Dynamic Attention Model (DUDA) demonstrates such capacity via a specialized attention mechanism.

1. Introduction

We live in a dynamic world where things change all the time. Change detection in images is a long-standing research problem, with applications in a variety of domains such as facility monitoring, medical imaging, and aerial photography [17, 46, 50]. A key challenge in change detection is to distinguish the relevant changes from the irrelevant ones [49] since the former are those that should likely trigger a notification. Existing systems aim to sense or localize a change, but typically do not convey detailed semantic content. This is an important limitation for a realistic application, where analysts would benefit from such knowledge, helping them to better understand and judge the significance of the change. Alerting a user on every detected difference likely will lead to a frustrated operator; moreover, it is desirable to have a change detection system that does not output a binary indicator of change/no-change, but instead outputs a concise description of what has changed, and where.

Expressing image content in natural language is an active area of Artificial Intelligence research, with numerous approaches to image captioning having been recently proposed [3, 12, 38, 59]. These methods have the benefit of conveying visual content to human users in a concise and natural way. They can be especially useful, when tailored to a specific task or objective, such as e.g. explaining the model's predictions [20, 45] or generating non-ambiguous referring expressions for specific image regions [40, 61].

In this work we investigate robust *Change Captioning*, where an important scene change has to be identified and conveyed using language in the presence of distractors (where only an illumination or viewpoint change occurred). We aim to generate detailed and informative descriptions that refer to the changed objects in complex scenes (see Figure 1).

To distinguish an irrelevant distractor from an actual change (e.g. an object moved), one needs to "compare" the two images and find correspondences and disagreements. We propose a *Dual Dynamic Attention Model (DUDA)* that learns to localize the changes via a specialized attention mechanism. It consists of two components: *Dual Attention* that predicts a separate spatial attention for each image in the "before"/"after" pair, and a *Dynamic Speaker* that generates a change description by semantically modulating focus among the visual features relayed from the Dual Attention. Both components are neural networks that are trained jointly with only caption-level supervision, i.e. no information about the change location is used during training.

In order to study Change Captioning in the presence of distractors, we build a *CLEVR-Change Dataset*. We rely on the image generation engine by [26], which allows us to produce complex compositional scenes. We create pairs of "before" and "after" images with: (a) only illumination/viewpoint change (distractors), and (b) illumination/viewpoint change combined with a scene change. We consider 5 scene change types (color/material change, adding/dropping/moving an object), and collect almost 80K image pairs. We augment the image pairs with automatically generated change captions based on templates (see Figure 3). Note that in the recently proposed Spot-the-Diff dataset [25], the task also is to generate change captions for a pair of images. However, their problem statement is different from ours in that: 1) they assume a change in each image pair while our goal is to be robust to distractors, 2) the images are aligned (no viewpoint shift), 3) change localization can not be evaluated as ground-truth is not available in [25].

We first evaluate our novel DUDA model on the CLEVR-Change dataset, and compare it to a number of baselines, including a naive pixel-difference captioning baseline. We show that our approach outperforms the baselines in terms of change caption correctness as well as change localization. The most challenging change types to describe are object movement and texture change, while movement is also the hardest to localize. We also show that our approach is general, applying it to the Spot-the-Diff dataset [25]. Given the same visual inputs as [25], our model matches or outperforms their approach.

2. Related Work

Here we discuss prior work on change detection, task-specific image captioning, and attention mechanism.

Change detection One popular domain for image-based change detection is aerial imagery [35, 53, 62], where changes can be linked to disaster response scenarios (e.g. damage detection) [17] or monitoring of land cover dynamics [29, 54]. Prior approaches often rely on unsupervised methods for change detection, e.g. image differencing, due to high cost of obtaining ground-truth annotations [9]. Notably, [17] propose a semi-supervised approach with human in the loop, relying on a hierarchical shape representation.

Another prominent domain is street scenes [1, 28]. Notably, [50] propose a Panoramic Change Detection Dataset, built off Google Street View panoramic images. In their follow-up work, [51] propose an approach to change detection which relies on dense optical flow to address the difference in viewpoints between the images. In a recent work, [43] rely on 3D models to identify scene changes by re-projecting images on one another. Another line of work targets change detection in video, e.g. using a popular CD-net benchnmark [16, 58], where background subtraction is a successful strategy [8]. Instead of relying on costly pixel-level video annotation, [30] propose a weakly supervised approach, which estimates pixel-level labels with a CRF.

Other works address a more subtle, fine-grained change detection, where an object may change its appearance over time, e.g. for the purpose of a valuable object monitoring [14, 24]. To tackle this problem, [52] estimate a dense flow field between images to address viewpoint differences.

Our DUDA model relies on an attention mechanism rather than pixel-level difference or flow. Besides, our task is not only to detect the changes, but also to describe them in natural language, going beyond the discussed prior works.

Task-specific caption generation While most image captioning works focus on a generic task of obtaining image relevant descriptions [3, 12, 57], some recent works explore pragmatic or "task-specific" captions. Some focus on generating textual explanations for deep models' predictions [19, 20, 45]. Others aim to generate a discriminative caption for an image or image region, to disambiguate it from a distractor [4, 10, 40, 39, 55, 61]. This is relevant to our work, as part of the change caption serves as a referring expression to put an object in context of the other objects. However, our primary focus is to correctly describe the scene changes.

The most related to ours is the work of [25], who also address the task of change captioning for a pair of images. While we aim to distinguish distractors from relevant changes, they assume there is always a change between the two images. Next, their pixel-difference based approach assumes that the images are aligned, while we tackle viewpoint change between images. Finally, we systematically study different change types in our new CLEVR-Change Dataset. We show that our approach generalizes to their Spot-the-Diff dataset in subsection 5.3.

Attention in image captioning Attention mechanism [6] over the visual features was first used for image captioning by [59]. Multiple works have since adopted and extended this approach [15, 36, 47], including performing attention over object detections [3]. Our DUDA model relies on two forms of attention: *spatial* Dual Attention used to localize changes between two images, and *semantic* attention, used by our Dynamic Speaker to adaptively focus on "before", "after" or "difference" visual representations.

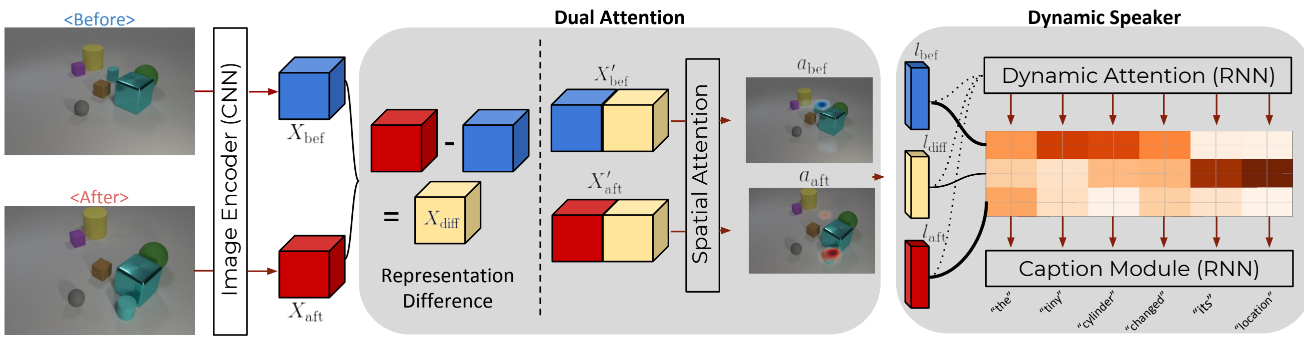

Figure 2: Our Dual Dynamic Attention Model (DUDA) consists of two main components: Dual Attention (subsection 3.1) and Dynamic Speaker (subsection 3.2).

3. Dual Dynamic Attention Model (DUDA)

We propose a *Dual Dynamic Attention Model (DUDA)* for change detection and captioning. Given a pair of "before" and "after" images (I_{bef} and I_{aft}, respectively), our model first detects whether a scene change has happened, and if so, locates the change on both I_{bef} and I_{aft}. The model then generates a sentence that not only correctly describes the change, but also is spatially and temporally grounded in the image pair. To this end, our model includes a Dual Attention (localization) component, followed by a Dynamic Speaker component to generate change descriptions. An overview of our model is shown in Figure 2.

We describe the implementation details of our Dual Attention in subsection 3.1, and our Dynamic Speaker in subsection 3.2. In subsection 3.3, we detail our training procedure for jointly optimizing both components using change captions as the only supervision.

3.1. Dual Attention

Our Dual Attention acts as a change localizer between I_{bef} and I_{aft}. Formally, it is a function $f_{\text{loc}}(X_{\text{bef}}, X_{\text{aft}}; \theta_{\text{loc}}) = (l_{\text{bef}}, l_{\text{aft}})$ parameterized by θ_{loc} that takes X_{bef} and X_{aft} as inputs, and outputs feature representations l_{bef} and l_{aft} that encode the change manifested in the input pairs. In our implementation, $X_{\text{bef}}, X_{\text{aft}} \in \mathbb{R}^{C\times H\times W}$ are image features of $I_{\text{bef}}, I_{\text{aft}}$, respectively, encoded by a pretrained ResNet [18].

We first subtract X_{bef} from X_{aft} in order to capture semantic difference in the representation space. The resulting tensor X_{diff} is concatenated with both X_{bef} and X_{aft} which are then used to generate two separate spatial attention maps $a_{\text{bef}}, a_{\text{aft}} \in \mathbb{R}^{1\times H\times W}$. Following [41], we utilize elementwise *sigmoid* instead of *softmax* for computing our attention maps to avoid introducing any form of global normalization. Finally, a_{bef} and a_{aft} are applied to the input features to do a weighted-sum pooling over the spatial dimensions:

$$X_{\text{diff}} = X_{\text{aft}} - X_{\text{bef}} \tag{1}$$

$$X'_{\text{bef}} = [X_{\text{bef}}\ ;\ X_{\text{diff}}],\ X'_{\text{aft}} = [X_{\text{aft}}\ ;\ X_{\text{diff}}] \tag{2}$$

$$a_{\text{bef}} = \sigma(\text{conv}_2(\text{ReLU}(\text{conv}_1(X'_{\text{bef}})))) \tag{3}$$

$$a_{\text{aft}} = \sigma(\text{conv}_2(\text{ReLU}(\text{conv}_1(X'_{\text{aft}})))) \tag{4}$$

$$l_{\text{bef}} = \sum_{H,W} a_{\text{bef}} \odot X_{\text{bef}},\ l_{\text{bef}} \in \mathbb{R}^C \tag{5}$$

$$l_{\text{aft}} = \sum_{H,W} a_{\text{aft}} \odot X_{\text{aft}},\ l_{\text{aft}} \in \mathbb{R}^C \tag{6}$$

where $[;]$, conv, σ, and $\odot$ indicate concatenation, convolutional layer, elementwise *sigmoid*, and elementwise multiplication, respectively. See Figure 2 for the visualization of Dual Attention component.

This particular architectural design allows the system to attend to images differently depending on the type of a change and the amount of a viewpoint shift, which is a capability crucial for our task. For instance, to correctly describe that an object has moved, the model needs to localize and match the moved object in *both* images; having single attention that locates the object only in one of the images is likely to cause confusion between e.g. moving vs. adding an object. Even if there is an attribute change (e.g. color) which does not involve object displacement, single attention might not be enough to correctly localize the changed object under a viewpoint shift. Unlike [60, 42, 37, 31, 45], DUDA utilizes Dual Attention to process *multiple* visual inputs separately and thereby addresses Change Captioning in the presence of distractors.

3.2. Dynamic Speaker

Our Dynamic Speaker is based on the following intuition: in order to successfully describe a change, the model should not only learn *where* to look in each image (*spatial* attention, predicted by the Dual Attention), but also *when* to look at each image (*semantic* attention, here). Ideally, we would like the model to exhibit dynamic reasoning, where it learns when to focus on "before" (l_{bef}), "after" (l_{aft}), or "difference" feature ($l_{\text{diff}} = l_{\text{aft}} - l_{\text{bef}}$) as it generates a sequence of words. For example, it is necessary to look at the "after" feature (l_{aft}) when referring to a new object added to a scene. Figure 2 illustrates this behaviour.

To this end, our Dynamic Speaker predicts an attention $\alpha_i^{(t)}$ over the visual features l_i's at each time step t, and obtains the dynamically attended feature $l_{\text{dyn}}^{(t)}$:

$$l_{\text{dyn}}^{(t)} = \sum_i \alpha_i^{(t)} l_i \tag{7}$$

where $i \in (\text{bef}, \text{diff}, \text{aft})$. We use the attentional Recurrent Neural Network [5] to model this formulation.

Our Dynamic Speaker consists of two modules, namely the dynamic attention module and the caption module. Both are recurrent models based on LSTM [21]. At each time step t, the LSTM decoder in the dynamic attention module takes as input the previous hidden state of the caption module $h_c^{(t-1)}$ and some latent projection v of the visual features l_{bef}, l_{diff}, and l_{aft} to predict attention weights $\alpha_i^{(t)}$:

$$v = \text{ReLU}(W_{d_1}[l_{\text{bef}} \; ; \; l_{\text{diff}} \; ; \; l_{\text{aft}}] + b_{d_1}) \tag{8}$$

$$u^{(t)} = [v \; ; \; h_c^{(t-1)}] \tag{9}$$

$$h_d^{(t)} = \text{LSTM}_d(h_d^{(t)} | u^{(t)}, h_d^{(0:t-1)}) \tag{10}$$

$$\alpha^{(t)} \sim \text{Softmax}(W_{d_2} h_d^{(t)} + b_{d_2}) \tag{11}$$

where $h_d^{(t)}$ and $h_c^{(t)}$ are LSTM outputs at decoder time step t for dynamic attention module and caption module, respectively, and W_{d_1}, b_{d_1}, W_{d_2}, and b_{d_2} are learnable parameters. Using the attention weights predicted from Equation (11), the dynamically attended feature $l_{\text{dyn}}^{(t)}$ is obtained according to Equation (7). Finally, $l_{\text{dyn}}^{(t)}$ and the embedding of the previous word w_{t-1} (ground-truth word during training, predicted word during inference) are input to the LSTM decoder of the caption module to begin generating distributions over the next word:

$$x^{(t-1)} = E \mathbb{1}_{w_{t-1}} \tag{12}$$

$$c^{(t)} = [x^{(t-1)} \; ; \; l_{\text{dyn}}^{(t)}] \tag{13}$$

$$h_c^{(t)} = \text{LSTM}_c(h_c^{(t)} | c^{(t)}, h_c^{(0:t-1)}) \tag{14}$$

$$w_t \sim \text{Softmax}(W_c h_c^{(t)} + b_c) \tag{15}$$

where $\mathbb{1}_{w_{t-1}}$ is a one-hot encoding of the word w_{t-1}, E is an embedding layer, and W_c, b_c are learned parameters.

	DI	C	T	A	D	M	All
# Img Pairs	39,803	7,958	7,963	7,966	7,961	7,955	79,606
# Captions	199,015	58,850	58,946	59,198	58,843	58,883	493,735
# Bboxes	-	15,916	15,926	7,966	7,961	15,910	64,679

Table 1: CLEVR-Change Dataset statistics: number of image pairs, captions, and bounding boxes for each change type: DISTRACTOR (DI), COLOR (C), TEXTURE (T), ADD (A), DROP (D), MOVE (M).

3.3. Joint Training

We jointly train the Dual Attention and the Dynamic Speaker end-to-end by maximizing the likelihood of the observed word sequence. Let θ denote all the parameters in DUDA. For a target ground-truth sequence $(w_1^*, \ldots, w_T^*)$, the objective is to minimize the cross entropy loss:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(w_t^* | w_1^*, \ldots, w_{t-1}^*)) \tag{16}$$

where $p_\theta(w_t | w_1, \ldots, w_{t-1})$ is given by Equation (15). Similar to [41], we apply L_1 regularization to the spatial attention masks generated by our Dual Attention in order to minimize unnecessary activations. We also use an entropy regularization over the attention weights generated by our Dynamic Speaker to encourage exploration in using visual features. The final loss function we optimize is as follows:

$$L(\theta) = L_{XE} + \lambda_{L_1} L_1 - \lambda_{ent} L_{ent} \tag{17}$$

where L_1 and L_{ent} are L_1 and entropy regularization, respectively, and λ_{L_1} and λ_{ent} are hyperparameters. Note, that the Dual Attention component receives no direct supervision for change localization. The only available supervision is obtained through the Dynamic Speaker, which then directs the Dual Attention towards discovering the change.

4. CLEVR-Change Dataset

Given a lack of an appropriate dataset to study Change Captioning in the presence of distractors, we build the CLEVR-Change Dataset, based on the CLEVR engine [26]. We choose CLEVR, inspired by many works that use it to build diagnostic datasets for various vision and language tasks, e.g. visual question answering [26], referring expression comprehension [22, 34], text-to-image generation [13] or visual dialog [33]. As Change Captioning is an emerging task we believe our dataset can complement existing datasets, e.g. [25], which is small, always assumes the presence of a change and lacks localization ground-truth.

First, we generate random scenes with multiple objects in them, which serve as "before" images. Note, that in domains such as satellite imagery [35, 53, 62] or surveillance/street scenes [1, 28, 43], typical distractors include

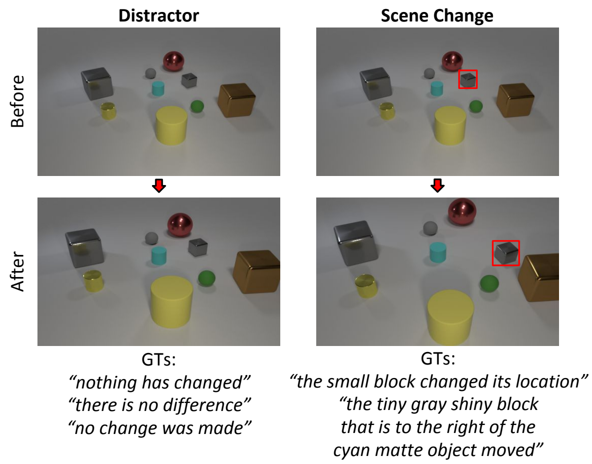

Figure 3: CLEVR-Change examples: distractors vs. scene changes, ground-truth captions and bounding boxes.

changes in camera position/zoom or illumination. Motivated by these applications we approach distractor construction accordingly. For each "before" image we create two "after" images. In the first one, we change the camera position leading to a different angle, zoom, and/or illumination. We have a specific allowed range for the transformation parameters: for each (x, y, z) camera location, we randomly sample a number from the range between -2.0 and 2.0, and jitter the original coordinates by the sampled amount. In the second "after" image, we additionally introduce a *scene change*. We consider the following types of scene changes: (a) an object's color is changed, (b) an object's texture is changed, (c) a new object is added, (d) an existing object is dropped, (e) an existing object is moved. In the following we refer to these as: COLOR, TEXTURE, ADD, DROP, MOVE, and DISTRACTOR for no scene change. In total, we generate $39,803$ "before" images with respectively $79,606$ "after" images. We make sure that the number of data points for each scene change type is balanced. The dataset is split into $67,660$, $3,976$, and $7,970$ training/validation/test image pairs, respectively.

Based on the created "before" and "after" scenes, we further augment them with change captions. Each caption is automatically constructed with two parts: the *referring* part (e.g. "A large blue sphere to the left of a red object") and the *change* part (e.g. "has appeared"). Note that for all the change types except ADD, the referring part is generated based on the "before" image, while for ADD, the "after" image is used. To get the change part, we rely on a set of change specific templates (see supplemental for details). However, note that the proposed DUDA model is not limited to templated language as we further demonstrate on a dataset with natural language descriptions.

Finally, we obtain spatial locations of where each scene change took place, so that we can evaluate the correctness of change localization. Specifically, we obtain bounding boxes for all the objects affected by a change, either in one image or in both ("before"/"after"), depending on the change type. The overall dataset statistics are shown in Table 1, and some examples of distractors vs. scene changes with their descriptions and bounding boxes are shown in Figure 3.

5. Experiments

In this section, we evaluate our DUDA model on the Change Captioning task against a number of baselines. First, we present quantitative results for the ablations and discuss their implications on our new CLEVR-Change Dataset. We also provide qualitative analysis of the generated captions, examine attention weights predicted by DUDA, and assess its robustness to viewpoint shift. Finally, we test the general effectiveness of our approach on the Spot-the-Diff [25], a realistic dataset with no distractors.

5.1. Experimental setup

Here, we detail our experimental setup in terms of implementation and evaluation schemes.

Implementation Details. Similar to [23, 27, 48], we use ResNet-101 [18] pretrained on ImageNet [11] to extract visual features from the images. We use features from the convolutional layer right before the global average pooling, obtaining features with dimensionality of 1024 x 14 x 14. The LSTMs used in the Dynamic Speaker have a hidden state dimension of 512. The word embedding layer is trained from scratch and each word is represented by a 300-dim vector. We train our model for 40 epochs using the Adam Optimizer [32] with a learning rate of 0.001 and a batch size of 128. The hyperparameters for the regularization terms are $\lambda_{L_1} = 2.5e^{-03}$ and $\lambda_{ent} = 0.0001$. Our code and dataset will be made publicly available at `github.com/Seth-Park/RobustChangeCaptioning`.

Evaluation. To evaluate change captioning, we rely on BLEU-4 [44], METEOR [7], CIDEr [56], and SPICE [2] metrics which measure overall sentence fluency and similarity to ground-truth. For change localization, we rely on the Pointing Game evaluation [63]. We use bilinear interpolation to upsample the attention maps to the original image size, and check whether the point with the highest activation "falls" in the ground-truth bounding box.

5.2. Results on CLEVR-Change Dataset

Pixel vs. representation difference [25] utilize pixel difference information when generating change captions under the assumption that the images are aligned. To obtain insights into whether a similar approach can still be effective when a camera position changes, we introduce the following baselines: *Capt-Pix-Diff* is a model that directly utilizes

Approach	Total				Scene Change				Distractor			
	B	C	M	S	B	C	M	S	B	C	M	S
Capt-Pix-Diff	30.2	75.9	23.7	17.1	21.9	36.2	17.7	7.9	43.4	98.2	38.9	26.3
Capt-Rep-Diff	33.5	87.9	26.7	19.0	26.0	51.8	21.1	10.1	49.4	105.3	41.7	27.8
Capt-Att	42.7	106.4	32.1	23.2	38.3	87.2	27.9	18.0	53.5	106.6	43.2	28.4
Capt-Dual-Att	43.5	108.5	32.7	23.4	38.5	89.8	28.5	18.2	56.3	108.9	44.0	28.7
DUDA (Ours)	**47.3**	**112.3**	**33.9**	**24.5**	**42.9**	**94.6**	**29.7**	**19.9**	**59.8**	**110.8**	**45.2**	**29.1**

Table 2: Change Captioning evaluation on our CLEVR-Change Dataset. Our proposed model outperforms all baselines on BLEU-4 (B), CIDEr (C), METEOR (M), and SPICE (S) in each setting (i.e. Total, Scene Change, Distractor).

Approach	CIDEr						METEOR						SPICE					
	C	T	A	D	M	DI	C	T	A	D	M	DI	C	T	A	D	M	DI
Capt-Pix-Diff	4.2	16.1	30.1	27.1	18.0	98.2	7.4	16.0	24.4	20.9	18.2	38.9	1.3	6.8	11.4	10.6	9.2	26.3
Capt-Rep-Diff	44.5	21.9	50.1	49.7	26.5	105.3	19.2	18.2	25.7	23.5	18.9	41.7	8.2	8.8	12.1	12.0	9.6	27.8
Capt-Att	112.1	75.9	91.5	98.4	49.6	106.6	30.5	25.4	30.2	31.2	22.2	43.2	17.9	16.3	19.0	**22.3**	14.5	28.4
Capt-Dual-Att	115.8	82.7	85.7	103.0	52.6	108.9	32.1	26.7	29.5	**31.7**	22.4	44.0	19.8	17.6	16.9	21.9	14.7	28.7
DUDA (Ours)	**120.4**	**86.7**	**108.2**	**103.4**	**56.4**	**110.8**	**32.8**	**27.3**	**33.4**	31.4	**23.5**	**45.2**	**21.2**	**18.3**	**22.4**	22.2	**15.4**	**29.1**

Table 3: A Detailed breakdown of Change Captioning evaluation on our CLEVR-Change Dataset by change types: Color (C), Texture (T), Add (A), Drop (D), Move (M), and Distractor (DI).

pixel-wise difference in the RGB space between "before" and "after" images. We use pyramid reduce downsampling on the RGB difference to match the spatial resolution of the ResNet features. The downsampled tensor is concatenated with the ResNet features on which we apply a series of convolutions and max-pooling. The resulting feature, which combines "before", "after", and "pixel difference" information, is input to an LSTM for sentence generation. On the other hand, *Capt-Rep-Diff* relies on representation difference (i.e. X_{diff}) instead of pixel difference. A series of convolutions and max-pooling are applied to the representation difference and then input to an LSTM decoder. As shown in the first two rows of Table 2, *Capt-Rep-Diff* outperforms *Capt-Pix-Diff* in all settings, indicating that representation difference is more informative than pixel difference when comparing scenes under viewpoint shift. We believe this is because the subtraction operation introduces a more useful inductive bias when done in a highly semantic space captured by visual representations with large receptive fields than in pixel space. As a result, we deliberately use representation difference in all subsequent experiments.

Role of localization To understand the importance of localization for change description, we compare models with and without spatial attention mechanism. *Capt-Att* is an extension of *Capt-Rep-Diff* which learns a single spatial attention which is applied to both "after" and "before" features. The attended features are subtracted and input to an LSTM decoder. We observe that *Capt-Att* significantly outperforms *Capt-Rep-Diff*, indicating that the capacity to explicitly localize the change has a high impact on the caption quality in

	C	T	A	D	M	Total
Capt-Att	46.68	57.90	22.84	47.80	17.57	39.37
Capt-Dual-Att	40.97	46.55	54.33	45.67	19.89	39.35
DUDA (Ours)	**54.52**	**65.75**	**48.68**	**50.06**	**22.77**	**48.10**

Table 4: Pointing game accuracy results. We report per change-type performance (Color (C), Texture (T), Add (A), Drop (D), Move (M)) as well as the total performance. The numbers are in %.

general. Note, that the improvements are more pronounced for scene changes (i.e. C, T, A, D, M) than for distractors (DI), see Table 3, which is intuitive since the localization ability matters most when there actually is a scene change.

Dual attention Using multiple spatial attentions has been shown to be useful for many purposes including multi-step/hierarchical reasoning [60, 42, 37] and model interpretability [31, 45]. To this extent, we train a model that deploys Dual Attention and evaluate its application to Change Captioning in the presence of distractors. *Capt-Dual-Att* is an extension of *Capt-Att* which learns two separate spatial attentions for the pair of images. Compared to *Capt-Att*, *Capt-Dual-Att* achieves higher performance overall according to Table 2. However, the improvements are limited in the sense that the margin of increase is small and not all change types improve (see Table 3). A similar issue can be seen in the Pointing Game results in Table 4. We speculate that without a proper inductive bias, it is difficult to learn how to utilize two spatial attentions effectively; a more

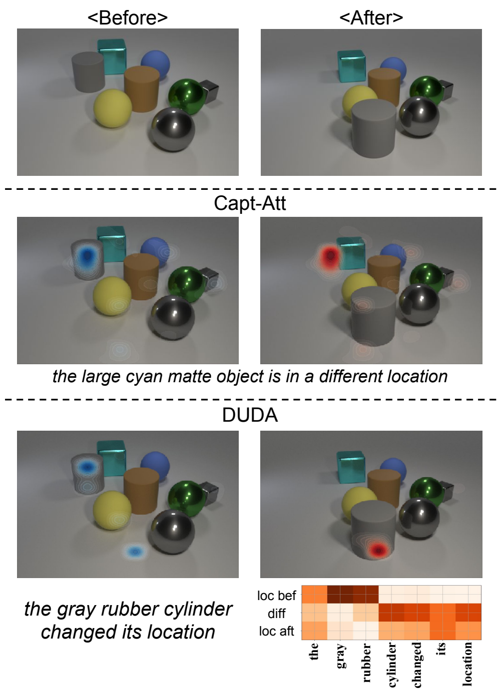

Figure 4: Qualitative results comparing *Capt-Att* and DUDA. The blue and red attention maps are applied to "before" and "after", respectively. The blue and red attention maps are the same for *Capt-Att* whereas in DUDA they are separately generated. The heat map on the lower-right is the visualization of the dynamic attention weights where the rows represent the amount of attention given to each visual feature (e.g. loc bef, diff, loc aft) per word.

complex speaker that enforces the usage of multiple visual signals might be required, leading to the development of our Dynamic Speaker.

Dynamic speaker Our final model with the Dynamic Speaker outperforms all previously discussed baselines not only in captioning (Table 2, Table 3) but also in localization (Table 4), supporting our intuition above. In Figure 4, we compare results from *Capt-Att* and DUDA. We observe that a single spatial attention used in *Capt-Att* cannot locate and associate the moved object in "before" and "after" images, thus confusing the properties of the target object (i.e. large *cyan* matte). On the other hand, our model is able to locate and match the target object in both scenes via Dual Attention, and discover that the object has moved. Moreover, it can be seen that our Dynamic Speaker predicts the attention weights that reveal some reasoning capacity of our model, where it first focuses on the "before" when addressing the changed object and gradually shifts attention to "diff" and "after" when mentioning the change.

Measuring robustness to viewpoint shift The experiments above demonstrate the importance of Dual and Dynamic Attention modules in improving robustness as the Dual Attention learns the spatial correspondence between two images w.r.t. the changed object and the Dynamic Speaker facilitates the learning of the Dual Attention. We now further validate such robustness by analyzing the performance under varying degrees of viewpoint shift. To measure the amount of viewpoint shift for a pair of images, we use the following heuristics: for each object in the scene, *excluding* the changed object, we compute the IoU of the object's bounding boxes across the image pair. We assume the more the camera changes its position, the less the bounding boxes will overlap. We compute the mean of these IoUs and sort the test examples based on this (lower IoU means higher difficulty). The performance breakdown in terms of change captioning and localization is shown in Figure 5. Our model outperforms the baselines on both tasks, including the more difficult samples (to the left). We see that both captioning and localization performance degrades for the baselines and our model (although less so) as viewpoint shift increases, indicating that it is an important challenge to be addressed on our dataset.

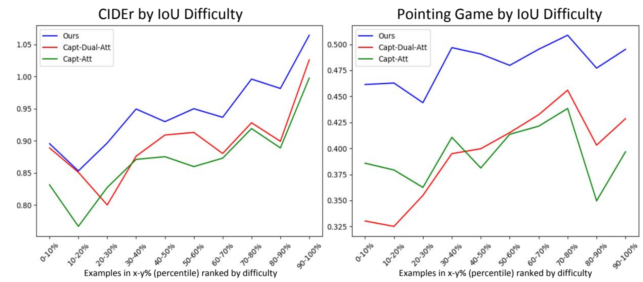

Figure 5: Change captioning and localization performance breakdown by viewpoint shift (measured by IoU).

Figure 6 illustrates two examples with large viewpoint changes, as measured by IoU. The overlaid images show that the scale and location of the objects may change significantly. The left example is a success, where DUDA is able to tell that the object has disappeared. Interestingly, in this case, it rarely attends to the "difference" feature. The right example illustrates a failure, where DUDA predicts that no change has occured, as a viewpoint shift makes it difficult to relate objects between the two scenes. Overall, we find that most often the semantic changes are confused with the distractors (no change) rather than among themselves, while MOVE suffers from such confusion the most.

5.3. Results on Spot-the-Diff Dataset

We also evaluate our DUDA model on the recent Spot-the-Diff dataset [25] with real images and human-provided descriptions. This dataset features mostly well aligned image pairs from surveillance cameras, with one or more changes between the images (no distractors). We evaluate our model in a single change setting, i.e. we generate a single change description, and use all the available human descriptions as references, as suggested by [25].

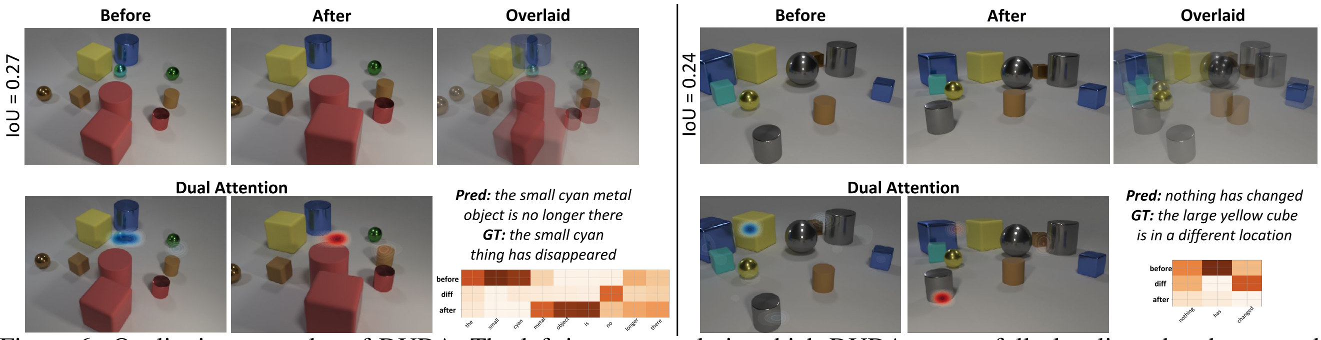

Figure 6: Qualitative examples of DUDA. The left is an example in which DUDA successfully localizes the change and generates correct descriptions with proper modulations among "before", "diff", and "after" visual features. The right example is a failure case. We observe that significant viewpoint shift leads to incorrect localization of the change, thus confusing the dynamic speaker.

Approach	B	C	M	R
DDLA* [25]	0.081	0.340	0.115	0.283
DUDA*	0.081	0.325	0.118	0.291

Table 5: We evaluate our approach on the Spot-the-Diff dataset [25]. * We report results averaged over two runs,for DDLA [25], we use the two sets of results reported by the authors. See text for details.

We present our results in Table 5. The DDLA approach of [25] relies on precomputed spatial clusters, obtained using pixel-wise difference between two images, assuming that the images are aligned. For a fair comparison we rely on the same information: we extract visual features from both "before" and "after" images using the spatial clusters. We apply Dual Attention over the extracted features to learn which clusters should be relayed to the Dynamic Speaker. The rest of our approach is unchanged. As can be seen from Table 5, DUDA matches or outperforms DDLA on most metrics. We present qualitative comparison in Figure 7. As can be seen from the examples, our DUDA model can attend to the right cluster and describe changes corresponding to the localized cluster.

Despite the usage of natural images and human descriptions, the Spot-the-Diff dataset is not the definitive test for robust change captioning as it does not consider the presence of distractors. That is, one does not have to establish whether the change occurred as there is always a change between each pair of images, and the images are mostly well-aligned. We advocate for a more practical setting of robust change captioning, where determining whether the change is by itself relevant is an important part of the problem.

6. Conclusion

In this work, we address robust Change Captioning in the general setting that includes distractors. We propose the novel Dual Dynamic Attention Model to jointly localize and describe changes between images. Our dynamic attention scheme is superior to the baselines and its visualization provides an interpretable view on the change caption generation mechanism. Our model is robust to distractors in the sense that it can distinguish relevant scene changes from illumination/viewpoint changes. Our CLEVR-Change Dataset is a new benchmark, where many challenges need to be addressed, e.g. establishing correspondences between the objects in the presence of viewpoint shift, resolving ambiguities and correctly referring to objects in complex scenes, and localizing the changes in the scene amidst viewpoint shifts. Our findings inform us of important challenges in domains like street scenes, e.g. "linking" the moved objects in before/after images, as also noted in [25]. Our results on Spot-the-Diff are complementary to those we have obtained on the larger CLEVR-Change dataset. While Spot-the-Diff is based on real images, there are minimal or no distractor cases in the dataset. This suggests that valuable future work will be to collect real-image datasets with significant semantic and distractor changes.

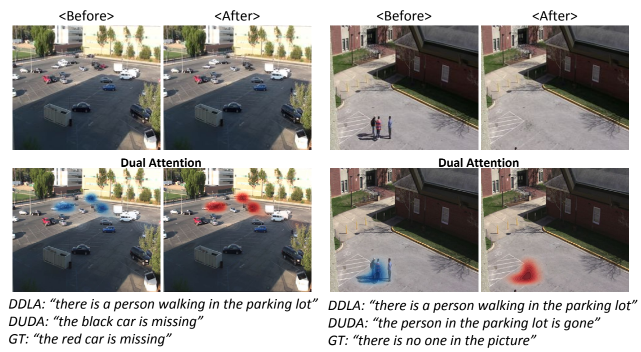

Figure 7: Example outputs of our model on the Spot-the-Diff dataset [25]. We visualize clusters with the maximum dual attention weights. We also show results from the DDLA [25] and the ground-truth captions.

Acknowledgements

This work was in part supported by the US DoD, Berkeley Deep Drive (BDD), and Berkeley Artificial Intelligence Research (BAIR) Lab.

References

[1] Pablo F Alcantarilla, Simon Stent, German Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42(7):1301–1322, 2018. 2, 4

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. Springer, 2016. 5

[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. 2018. 1, 2

[4] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. 2016. 2

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014. 4

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. 2015. 2

[7] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 5

[8] Simone Bianco, Gianluigi Ciocca, and Raimondo Schettini. How far can you get by combining change detection algorithms? In *International Conference on Image Analysis and Processing*, pages 96–107. Springer, 2017. 2

[9] Lorenzo Bruzzone and Diego F Prieto. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote sensing*, 38(3):1171–1182, 2000. 2

[10] Reuben Cohn-Gordon, Noah Goodman, and Chris Potts. Pragmatically informative image captioning with character-level reference. 2018. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. Ieee, 2009. 5

[12] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. 2015. 1, 2

[13] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Keep drawing it: Iterative language-based image generation and editing. 2018. 4

[14] Wei Feng, Fei-Peng Tian, Qian Zhang, Nan Zhang, Liang Wan, and Jizhou Sun. Fine-grained change detection of misaligned scenes with varied illuminations. pages 1260–1268, 2015. 2

[15] Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. 39(12):2321–2334, 2017. 2

[16] Nil Goyette, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Prakash Ishwar, et al. Changedetection. net: A new change detection benchmark dataset. 2012. 2

[17] Lionel Gueguen and Raffay Hamid. Large-scale damage detection using satellite imagery. pages 1321–1328, 2015. 1, 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016. 3, 5

[19] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. Springer, 2016. 2

[20] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. 2018. 1, 2

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[22] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. pages 53–69, 2018. 4

[23] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. 2017. 5

[24] Rui Huang, Wei Feng, Zezheng Wang, Mingyuan Fan, Liang Wan, and Jizhou Sun. Learning to detect fine-grained change under variant imaging conditions. 2017. 2

[25] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. 2018. 2, 4, 5, 7, 8

[26] 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. 2017. 2, 4

[27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Fei-Fei Li, C. Lawrence Zitnick, and Ross B. Girshick. Inferring and executing programs for visual reasoning. 2017. 5

[28] Hirokatsu Kataoka, Soma Shirakabe, Yudai Miyashita, Akio Nakamura, Kenji Iwata, and Yutaka Satoh. Semantic change detection with hypermaps. *arXiv:1604.07513*, 2016. 2, 4

[29] Salman H Khan, Xuming He, Fatih Porikli, and Mohammed Bennamoun. Forest change detection in incomplete satellite images with deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5407–5423, 2017. 2

[30] Salman H Khan, Xuming He, Fatih Porikli, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Learning deep structured network for weakly supervised change detection. 2017. 2

[31] Jinkyu Kim and John F Canny. Interpretable learning for self-driving cars by visualizing causal attention. 2017. 3, 6

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[33] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. 2019. 4

[34] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. 2019. 4

[35] Zhunga Liu, Gang Li, Gregoire Mercier, You He, and Quan Pan. Change detection in heterogenous remote sensing images via homogeneous pixel transformation. *IEEE Transactions on Image Processing*, 27(4):1822–1834, 2018. 2, 4

[36] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. 2017. 2

[37] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. 2016. 3, 6

[38] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. pages 7219–7228, 2018. 1

[39] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. 2018. 2

[40] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. 2017. 1, 2

[41] 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*, 2018. 3, 4

[42] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *arXiv:1611.00471*, 2016. 3, 6

[43] Emanuele Palazzolo and Cyrill Stachniss. Fast image-based geometric change detection given a 3d model. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6308–6315. IEEE, 2018. 2, 4

[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. pages 311–318. Association for Computational Linguistics, 2002. 5

[45] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. 2018. 1, 2, 3, 6

[46] Julia Patriarche and Bradley Erickson. A review of the automated detection of change in serial imaging studies of the brain. *Journal of digital imaging*, 17(3):158–174, 2004. 1

[47] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. 2016. 2

[48] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. 2018. 5

[49] Richard J Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005. 1

[50] Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using cnn features and superpixel segmentation. pages 61–1, 2015. 1, 2

[51] Ken Sakurada, Weimin Wang, Nobuo Kawaguchi, and Ryosuke Nakamura. Dense optical flow based change detection network robust to difference of camera viewpoints. *arXiv:1712.02941*, 2017. 2

[52] Simon Stent, Riccardo Gherardi, Björn Stenger, and Roberto Cipolla. Precise deterministic change detection for smooth surfaces. pages 1–9. IEEE, 2016. 2

[53] Jiaojiao Tian, Shiyong Cui, and Peter Reinartz. Building change detection based on satellite stereo imagery and digital surface models. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):406–417, 2014. 2, 4

[54] Corina Vaduva, Teodor Costachioiu, Carmen Patrascu, Inge Gavat, Vasile Lazarescu, and Mihai Datcu. A latent analysis of earth surface dynamic evolution using change map time series. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4):2105–2118, 2013. 2

[55] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. 2017. 2

[56] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. 2015. 5

[57] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. pages 3156–3164, 2015. 2

[58] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. pages 387–394, 2014. 2

[59] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. pages 2048–2057, 2015. 1, 2

[60] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. pages 21–29, 2016. 3, 6

[61] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speakerlistener-reinforcer model for referring expressions. 2017. 1, 2

[62] Massimo Zanetti and Lorenzo Bruzzone. A generalized statistical model for binary change detection in multispectral images. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 3378–3381. IEEE, 2016. 2, 4

[63] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. 126(10):1084–1102, 2018. 5