Viewpoint-Agnostic Change Captioning with Cycle Consistency

Hoeseong Kim[1] Jongseok Kim[1] Hyungseok Lee[2] Hyunsung Park[2] Gunhee Kim[1,†]
[1] Seoul National University [2] AIRS Company, Hyundai Motor Group, Seoul, Korea
hsgkim@snu.ac.kr js.kim@vision.snu.ac.kr {hseokool,hyunsung}@hyundai.com gunhee@snu.ac.kr
https://github.com/hsgkim/clevr-dc

Abstract

Change captioning is the task of identifying the change and describing it with a concise caption. Despite recent advancements, filtering out insignificant changes still remains as a challenge. Namely, images from different camera perspectives can cause issues; a mere change in viewpoint should be disregarded while still capturing the actual changes. In order to tackle this problem, we present a new Viewpoint-Agnostic change captioning network with Cycle Consistency (VACC) that requires only one image each for the before and after scene, without depending on any other information. We achieve this by devising a new difference encoder module which can encode viewpoint information and model the difference more effectively. In addition, we propose a cycle consistency module that can potentially improve the performance of any change captioning networks in general by matching the composite feature of the generated caption and before image with the after image feature. We evaluate the performance of our proposed model across three datasets for change captioning, including a novel dataset we introduce here that contains images with changes under extreme viewpoint shifts. Through our experiments, we show the excellence of our method with respect to the CIDEr, BLEU-4, METEOR and SPICE scores. Moreover, we demonstrate that attaching our proposed cycle consistency module yields a performance boost for existing change captioning networks, even with varying image encoding mechanisms.

1. Introduction

With an endless stream of data in real world, it is pivotal to develop automated systems that assist human to quickly grasp the essence of the data. Consider, for example, data streams from a myriad of surveillance cameras scattered around highways. It is highly labor-intensive and almost implausible to monitor them all without automation due to the sheer amount of data and the flashing rate of change. Change detection has received much attention as one solution to this problem, along with other usages in various fields such as medical imaging and satellite imaging [21, 36]. Recent advancements even allow generating a short descriptive sentence that summarizes the detected changes, often referred to as *change captioning* [10, 19, 28].

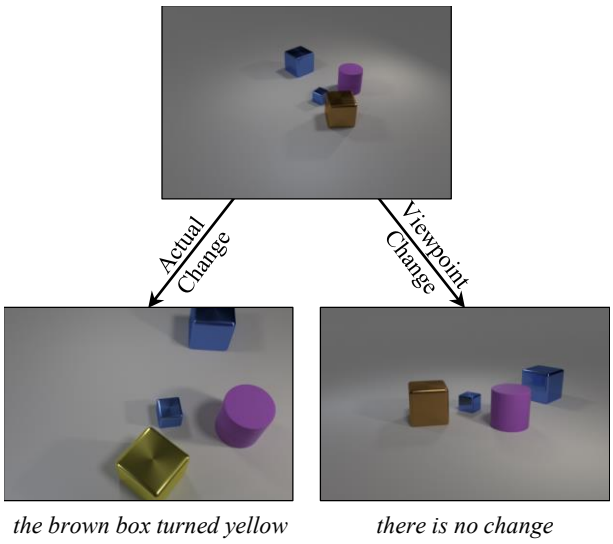

Figure 1. An example of viewpoint-agnostic change captioning. Compared to the top image, both images on the bottom are acquired from severely different viewpoints. Only the bottom left image contains an actual change, where the color of the brown box changes. Therefore, a meaningful caption should be generated only for this image ("the brown box turned yellow"), while the caption for the bottom right should indicate there is no change.

Despite such improvements in change detection and captioning, one of the most challenging aspects remains unsolved: identifying only the relevant semantic changes [26]. As shown in Figure 1, a picture of the same scene from another perspective is the epitome of an irrelevant change. Returning to the previous highway monitoring example, this can be of grave importance when aggregating data acquired

[†]Corresponding author.

from multiple cameras, as their viewpoints all differ. Collecting information from various data sources can easily happen in everyday life, especially with the prevalence of smartphones these days. Hence, the caption should only indicate what really changed while ignoring the viewpoint shift.

Although the change in perspective has been addressed in previous change captioning works, they utilize datasets with relatively small viewpoint changes [19, 28] or make use of other information [24, 25] such as depth images, point clouds, and/or ground truth camera position to compensate for greater disparities in perspectives. On the other hand, our goal is to perform change captioning using only a pair of images in any viewpoints with no additional information.

To this end, we propose a new model that can pinpoint the semantic changes in the scene even under extreme viewpoint shifts. We further improve the captioning quality by devising a cycle consistency module that builds a composite feature of the generated caption and the before image to match it with the encoded after image. Using the CLEVR engine [11], we build a synthetic dataset that simulates extreme viewpoint shifts to gauge the robustness of networks to perspective changes and the ability to isolate only the relevant differences. Finally, our contributions in this work can be summarized as follows:

1. We propose a new network for change captioning that is robust to viewpoint changes. Specifically, we tackle the problem of change captioning between pictures with extreme viewpoint shifts without relying on any extra data (*i.e*., using only one before image and one after image). To the best of our knowledge, our work is the first attempt to solve this problem under such limited conditions.

2. The technical novelties of our new network for change captioning are two-fold. First, we devise a new difference encoder that captures the change from a pair of images while being robust to the viewpoint difference. Second, we present a cycle consistency module that assesses the quality of the resultant caption by creating a composite feature of the caption and before image feature and matching it with the after image feature. The module is generalizable and can be attached to other models to improve their performance.

3. We introduce CLEVR-DC created using the CLEVR engine [11] as a novel dataset for change captioning with extreme viewpoint shifts. We perform experiments on CLEVR-DC and two existing datasets, CLEVR-Change [19] and Spot-the-Diff [10], on all of which our method mostly outperforms multiple state-of-the-art methods.

2. Related Work

Change Captioning. Change captioning is the task of describing the difference between two visual inputs in natural language. The inputs are typically provided as images of a changing scene captured at different time steps, which respectively represent the before and after scene. As one of the earliest attempts to tackle this problem, Jhamtani *et al*. [10] approximate object-level differences by clustering pixels based on pixel-wise difference of images. Park *et al*. [19] generate an attention map for each input image ("dual dynamic attention") instead to locate the changes. Shi *et al*. [28] acquire both changed and unchanged features and feed them to the sentence decoder.

However, all these works are restricted to the inputs with no or little change in viewpoints (*e.g*. none in [10] and only small perturbations of camera coordinates in [19, 28]). On the other hand, we consider images taken from any random camera positions in 3D space (*e.g*. pictures taken even from opposite sides). Some works [24, 25] address a greater variance in viewpoints with Generative Query Network [6], but require multiple images of the same scene from different perspectives and extra information such as depth images, point clouds, and/or ground truth camera position vectors. However, our work requires only a pair of images in any viewpoint with no additional information.

Viewpoint Estimation. Viewpoint estimation locates the camera position of a given image, usually with the azimuth (θ), elevation (ϕ) and rotation angle (ψ). The approaches can be categorized according to how to formulate the problem: as regression [7, 17, 22] or as classification [5, 29, 30].

Yet, viewpoint estimation often requires ground truth viewpoint information and/or multiple images of the same scene/object. Thus, they cannot be directly applied to our problem where only a pair of images are given with no other data. Although few-shot methods have been proposed [27, 33], they assume still scenes with no change. In our work, we consider images of a changing scene from vastly different perspectives, without any ground truth viewpoint vectors or auxiliary data.

Cycle Consistency. Popularized by CycleGAN [35] for text-to-image synthesis, cycle consistency refers to evaluating the output by recreating the input with an inverse operation. This is often achieved by introducing a cycle consistency loss term to the total loss of the network. Some previous works [8, 12, 23] employ cycle consistency to construct images based on text description using GANs.

Our approach also adopts cycle consistency to rebuild the after image from the before image and the output caption. This necessitates building a multimodal composite embedding of text and image, rather than using only the text embedding to match with the image embedding. Such composition embedding has been studied in [13, 32, 34], but

they primarily focus on image retrieval, whereas our work aims at change captioning.

3. Approach

Our objective is to design a change captioning network that is robust to camera position changes (*Viewpoint-Agnostic*), scaffolded with *Cycle Consistency*. We name our approach VACC. Formally, given a pair of two images $(I_{\text{bef}}, I_{\text{aft}})$ with a significant change in viewpoint, the network generates a sentence T that captures only the semantic changes while ignoring the ones due to viewpoint movement. Figure 2 outlines the overall structure of our network, which consists of three major components: the difference encoder, the caption generator, and the cycle consistency module.

3.1. The Difference Encoder

The difference encoder identifies the difference between the two images and encodes it into features that are later utilized by other parts of the network. Formally, given two images $I_{\text{bef}}, I_{\text{aft}} \in \mathbb{R}^{C\times H\times W}$, we design a function that outputs $x_{\text{bef}}, x_{\text{aft}} \in \mathbb{R}^{D}$ that encapsulate only the semantically important difference between the two images. We first encode the images with a backbone CNN image encoder f_{CNN} and obtain $X_{\text{bef}}, X_{\text{aft}} \in \mathbb{R}^{C'\times H'\times W'}$.

Viewpoint Encoding. To mitigate the viewpoint difference, we first correlate the pixels of same objects in the before and after feature maps. For example, in Figure 2, we connect the pixels of the cyan cylinder in both images. We achieve this by obtaining the similarity map $S \in \mathbb{R}^{H'W'\times H'W'}$ that computes all pairwise similarity between the points of the two feature maps. This essentially signifies where the objects in the before image are in the after image, given that the same object viewed from a different perspective would still have similar features. Specifically, the similarity map S is obtained as:

$$S = kF_{\text{bef}}^{\top}F_{\text{aft}}, \quad \text{where } F_i = \mathcal{F}X_i \text{ for } i \in \{\text{bef}, \text{aft}\}. \quad (1)$$

$k \in \mathbb{R}$ is a learnable parameter and $\mathcal{F}$ is an operator that flattens tensors from $\mathbb{R}^{C'\times H'\times W'}$ to $\mathbb{R}^{C'\times H'W'}$. We then create two features $X_{\text{bef}|\text{aft}}, X_{\text{aft}|\text{bef}} \in \mathbb{R}^{C'\times H'\times W'}$ as:

$$X_{j|i} = \mathcal{F}^{-1}\left(F_j \alpha_{S|i}\right), \quad (2)$$

$$\alpha_{S|\text{bef}} = \text{softmax}_1(S), \quad \alpha_{S|\text{aft}} = \text{softmax}_2(S)^{\top}, \quad (3)$$

where $(i, j) \in \{(\text{bef}, \text{aft}), (\text{aft}, \text{bef})\}$. $\text{softmax}_n(S)$ applies the softmax along the n-th dimension of S, and $\mathcal{F}^{-1}$ is an operator that unflattens tensors from $\mathbb{R}^{C'\times H'W'}$ to $\mathbb{R}^{C'\times H'\times W'}$. $\alpha_{S|i} \in \mathbb{R}^{H'W'\times H'W'}$ can be regarded as an attention map that preserves the viewpoint information of X_i estimated by its object locations. Therefore, the resultant $X_{j|i}$ can be interpreted as X_j into which the viewpoint information of X_i is infused.

To further harness the similarity map for later use, we obtain a latent feature $s_i \in \mathbb{R}^{D\times H'\times W'}$ as an embedding of the similarity map. We develop on the empirical observation of $\alpha_{S|i}$ that salient object matches tend to form visibly noticeable clusters, whereas background matches are scattered around the map. Since we mainly require information about the foreground objects, we define a feature s_i that accentuates the object information in S_i:

$$s_i = \frac{1}{H'W'} \sum_{H',W'} \text{conv}_1(\text{MaxPool}(\text{conv}_2(\alpha_{S|i}))) \quad (4)$$

where $i \in \{\text{bef}, \text{aft}\}$. $\text{conv}_{1/2}$ denotes 2D convolutions with a 3×3 kernel. Convolutions are applied after reshaping (and transposing if necessary) so that only dimensions corresponding to i are affected. For example, for $i = \text{aft}$, $\alpha_{S|i} \in \mathbb{R}^{H'_1W'_1\times H'_2W'_2}$ is reshaped to $\mathbb{R}^{H'_2W'_2\times 1\times H'_1\times W'_1}$ (subscripts added only for clarity). Max pooling allows the network to preserve the salient object features while reducing the background information via downsampling.

Difference Encoding. We then define the feature embedding that encodes the difference between the viewpoint encoded image features $X_{j|i}$. One straightforward way is to simply subtract the two features, which yet falters when the viewpoint is not aligned. To be more robust to viewpoint changes, we adopt a variant of the Fused Difference module [13]. Specifically, we fuse the obtained features and compute the difference feature as follows:

$$\widetilde{X}_{j|i} = \text{ReLU}(\text{conv}_3([X_{\text{bef}|i} \odot X_{\text{aft}|i}; X_{j|i}])) \quad (5)$$

$$X_{\text{diff}|i} = \widetilde{X}_{\text{aft}|i} - \widetilde{X}_{\text{bef}|i} \quad (6)$$

where $(i, j) \in \{(\text{bef}, \text{aft}), (\text{aft}, \text{bef})\}$ and $\widetilde{X}_{i|i} = X_i$. The notation $[;]$, $\odot$, and conv_3 indicate concatenation, Hadamard product, and the 2D convolution with a 1×1 kernel, respectively. The network can refer to this fused feature that contains the information of both images, and decide which difference should be underscored, thereby creating a more robust feature compared to simple subtraction.

Finally, inspired by [19], we acquire the attention map and embedding for each image as:

$$\alpha_i = \sigma(\text{conv}_4(\text{ReLU}(\text{conv}_5([X_i; X_{\text{diff}|i}; s_i])))) \quad (7)$$

$$x_i = \sum_{H',W'} \alpha_i \odot X_i \quad (8)$$

where $i \in \{\text{bef}, \text{aft}\}$. σ, conv_4, and conv_5 are the sigmoid function and 2D convolutions with a 1×1 and 3×3 kernel, respectively. We use a 3×3 kernel for conv_5 since it facilitates the network to also represent larger objects adequately.

3.2. The Caption Generator

The caption generator creates a caption T that describes the changes detected. Consider how humans identify what

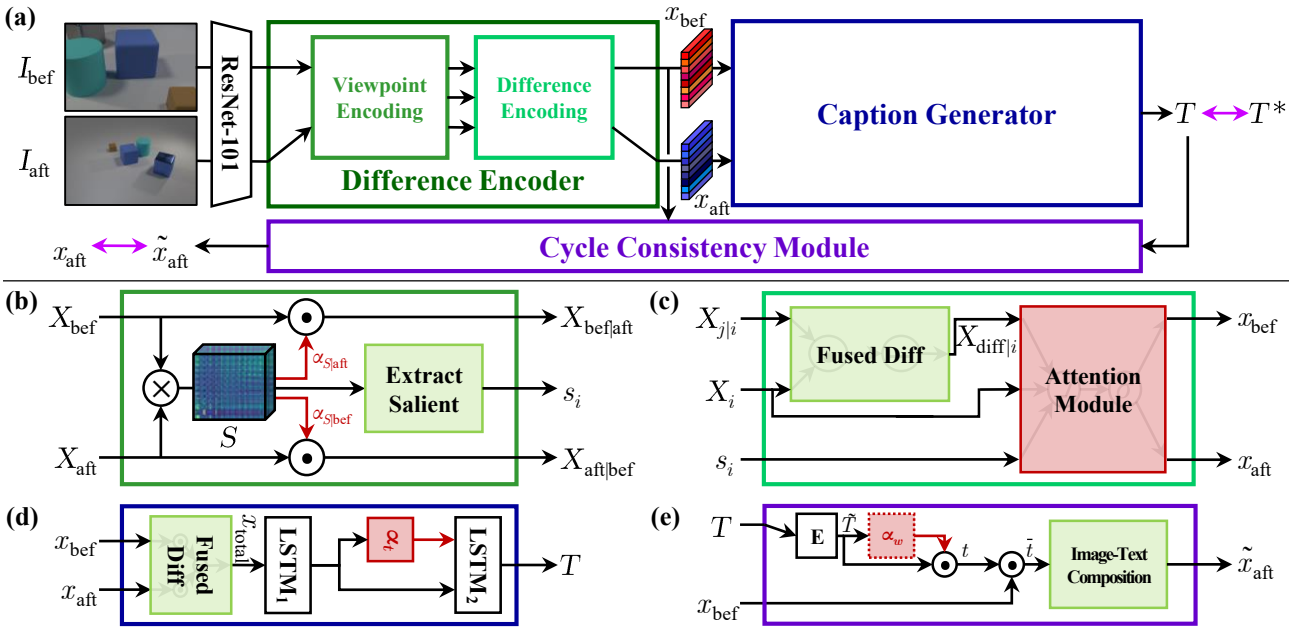

Figure 2. (a) An overview of our network architecture. Given a pair of two images (I_{bef}, I_{aft}) with an extreme viewpoint shift, the network generates a sentence T that captures only the semantic changes. The network largely comprises three parts: (b), (c) the difference encoder, (d) the caption generator, and (e) the cycle consistency module. Crimson modules and arrows indicate attention mechanism, and violet double-sided arrows represent cross-entropy loss.

changed; we analyze not only the two pictures separately but also both of them holistically. For example, for the pair in Figure 1, "the brown box", "yellow" and "turned" are acknowledged by respectively focusing on the before image, the after image and the entire pair to capture the difference. Based on this intuition, we compute $x_{diff} \in \mathbb{R}^D$ that heavily highlights the difference, and finally $x_{total} \in \mathbb{R}^D$ that encompasses all information as follows:

$$\tilde{x}_i = \text{ReLU}(\text{FC}([x_{bef} \odot x_{aft}; x_i])) \quad (9)$$

$$x_{diff} = \tilde{x}_{aft} - \tilde{x}_{bef} \quad (10)$$

$$x_{total} = \text{ReLU}(\text{FC}([x_{bef}; x_{diff}; x_{aft}])) \quad (11)$$

where $i \in \{bef, diff, aft\}$. The "holistic analysis" is realized as $x_{bef} \odot x_{aft}$ in Eq. 9, which has not been included in previous works. The principal idea behind x_{total} is to concatenate all three features so that the network learns to decide which dimensions are salient for capturing the difference.

Then, we opt for a variant of the top-down captioning model [2, 19] and sample the words for the final caption:

$$h_t^1 = \text{LSTM}_1\left([x_{total}; h_{t-1}^2], h_{t-1}^1\right) \quad (12)$$

$$\alpha_t = \text{softmax}(\text{FC}(h_t^1)) \quad (13)$$

$$h_t^2 = \text{LSTM}_2\left(\left[\sum_i \alpha_t[i] \cdot x_i; \mathbf{E}w_{t-1}\right], h_{t-1}^2\right) \quad (14)$$

$$w_t \sim \text{softmax}(\text{FC}(h_t^2)) \quad (15)$$

where $i \in \{bef, diff, aft\}$, $\mathbf{E}$ is the word embedding matrix, h_t^1 and h_t^2 are the hidden states of LSTMs, and w_t is the word sampled at time t.

3.3. The Cycle Consistency Module

The cycle consistency module verifies the resultant caption correctly explains the difference of the two images. We achieve this by generating a feature that represents the after image using the before image and change caption. Concretely, with caption $T = [w_1, w_2, \cdots, w_l]$ and image feature x_{bef}, we devise a function that computes $\tilde{x}_{aft} \in \mathbb{R}^C$ that is later matched with x_{aft}.

We first encode T with the word embedding matrix $\mathbf{E}$ and obtain $\tilde{T} = [\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_l] \in \mathbb{R}^{l \times E}$ (*i.e.*, $\tilde{w}_i = \mathbf{E}w_i$). Captions typically include relatively unimportant words, and thus we assign different weights to each word. The caption embedding $t \in \mathbb{R}^C$ is formulated as follows:

$$\alpha_w = \text{softmax}(\text{FC}(\text{ReLU}(\text{FC}(\mathbf{v} \odot \tilde{w})))) \quad (16)$$

$$t = \text{FC}\left(\sum_l \alpha_w[l] \cdot \tilde{w}_l\right) \quad (17)$$

where $\mathbf{v}$ is a learnable parameter.

Inspired by [13, 32], we employ a variant of Text Image Residual Gating to create a composite feature that integrates

the before image feature and change caption:

$$\tilde{t} = [t; x_{\text{bef}} \cdot t] \tag{18}$$

$$f_g = \sigma(\text{FC}(\text{ReLU}([x_{\text{bef}}; \tilde{t}\,]))) \odot x_{\text{bef}} \tag{19}$$

$$f_r = \text{FC}(\text{ReLU}(\text{FC}(\text{ReLU}([x_{\text{bef}}; \tilde{t}\,])))) \tag{20}$$

$$\tilde{x}_{\text{aft}} = w_g f_g + w_r f_r \tag{21}$$

where w_g and w_r are learnable parameters. The intuition here is to compute a rough gated feature first then fine-tune it by perturbing it with a residual connection to obtain the final composite feature.

3.4. Training

All components of the network are jointly trained end-to-end by minimizing the distance between the generated caption T and target ground truth caption T^*. Formally, the distance is evaluated as the cross-entropy loss as:

$$\mathcal{L}_{\text{XE}} = -\sum_t \log(p_\theta(w_t^* | w_{1:t-1}^*)) \tag{22}$$

where θ denotes all parameters in the network.

In addition, we leverage the cycle consistency module by imposing the network to align the resultant feature $\tilde{x}_{\text{aft}}$ with x_{aft}. Concretely, for a mini-batch of size B with ground truth pairs $\{(\tilde{x}_{\text{aft},i}, x_{\text{aft},i})\}_{i=1}^B$, we also minimize the cross-entropy loss $\mathcal{L}_{\text{cycle}}$ calculated as:

$$\mathcal{L}_{\text{cycle}} = -\frac{1}{B}\sum_i \log \frac{\exp(\tilde{x}_{\text{aft},i} \cdot x_{\text{aft},i})}{\sum_j \exp(\tilde{x}_{\text{aft},i} \cdot x_{\text{aft},j})}. \tag{23}$$

Finally, we add an L_1 regularization term $\mathcal{L}_{\text{reg}}$ of attention maps α_i in Eq. 7 to suppress unnecessary activations:

$$\mathcal{L}_{\text{reg}} = \frac{1}{B}\sum_{\text{all}} |\alpha_{\text{bef}}| + \frac{1}{B}\sum_{\text{all}} |\alpha_{\text{aft}}|. \tag{24}$$

We ultimately minimize the following total loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{XE}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{25}$$

where λ_{cycle} and λ_{reg} are hyperparameters.

4. Experiments

4.1. Datasets

For evaluation, we test with three datasets: two existing benchmarks and one modified from an existing benchmark by adding significant viewpoint changes.

CLEVR-DC. CLEVR-DC is a synthetic dataset we create to simulate extreme viewpoint shifts. Using the CLEVR engine [11], we generate 48,000 pairs of 480×320 before and after images, including 85% for training, 5% for validation and 10% for test, respectively. We generate 8,000 image pairs each simulating the color change, texture change, addition, removal and relocation of an object, and perspective changes only. The coordinates (x, y, z) of camera positions are randomly sampled by $x, y \overset{iid}{\sim} \mathcal{U}(-11, 11)$ and $z \sim \mathcal{U}(2, 11)$. The camera is repositioned for all after images in the same way.

We also provide natural language ground truth captions that describe the change in each image pair (avg 8.9 captions/pair). The captions are generated with predefined templates that first specify the object affected and then explain the change (*e.g.*, the yellow cube has disappeared). We determine the relative positions of objects with respect to the before image except object addition pairs, and explicitly include which image is used in the caption (*e.g.*, in front of the rubber ball *in the perspective of the before image*). We ensure that the object can be uniquely determined from the caption. For image pairs only with viewpoint shifts, five captions are sampled from the predefined sentences stating there is no change (*e.g.*, nothing has changed). Please refer to Appendix for details.

CLEVR-Change. CLEVR-Change [19] is a dataset similar to CLEVR-DC, also created with the CLEVR engine [11]. Around 8K images each are prepared for COLOR, TEXTURE, ADD, DROP and MOVE changes, paired with a distractor that only contains a small viewpoint change. All after images include a relatively small random camera position jitter. We use the same split as done in [19].

Spot-the-Diff. Spot-the-Diff [10] is a dataset with a total of 12,562 real-world images sampled from VIRAT Ground Video Dataset [16] and manually obtained human annotated captions. Image viewpoints are mostly well-aligned between image pairs. We follow the same split as [19] for a fair comparison.

4.2. Experiment Settings

For the backbone image encoder, we select ResNet-101 [9] pretrained on ImageNet [4]. We use the features extracted from the third residual block with the output shape of $1024 \times 14 \times 14$. The hidden state dimensions of all LSTMs are 512, and words are embedded as 300-dim vectors ($E = 300$). The caption generator generates the total feature x_{total} of dimension $D = 512$, and the cycle consistency module models $\tilde{x}_{\text{aft}}$ as a 512-dim vector ($C = 512$). The hyperparameters in the total loss $\mathcal{L}$ are set to $\lambda_{\text{cycle}} = 0.001$ and $\lambda_{\text{reg}} = 0.00125$. The network is trained for 20,000 iterations with a batch size of $B = 128$. We adopt the Adam Optimizer [14] with a learning rate of 0.00075. We implement our model using PyTorch [20].

We evaluate the model performance using four most popular automatic language metrics [15]: CIDEr [31], BLEU-4 [18], METEOR [3] and SPICE [1].

Model	CIDEr	BLEU-4	METEOR	SPICE
Without Cycle Consistency				
DUDA [19]	56.7	40.3	27.1	16.1
M-VAM [28]	60.1	40.9	27.1	15.8
VACC – CC	**70.0**	**44.5**	**29.2**	**17.1**
With Cycle Consistency				
DUDA + CC	62.0	41.7	27.5	16.4
M-VAM + CC	60.6	41.0	27.2	15.7
VACC (ours)	**71.7**	**45.0**	**29.3**	**17.6**
With *Incorrect* Cycle Consistency				
M-VAM (U)	57.1	39.2	26.3	14.5
VACC (I)	69.0	44.2	29.0	17.2

Table 1. Quantitative results on the test split of CLEVR-DC. CC refers to our cycle consistency module, + and – indicate the model was trained with and without CC, respectively. For models with incorrect cycle consistency configuration, U and I denote that the unchanged features and the backbone image encoder outputs X_{bef}, X_{aft} were utilized, respectively.

4.3. Baseline Models

Here we compare and contrast our model with the baseline models tested. We first point out that the cycle consistency module is a novel component not included in any previous models.

DUDA. [19] Dual Dynamic Attention Model (DUDA) models the difference using pixel-wise difference ($X_{\text{aft}} - X_{\text{bef}}$) and localizes the change using "dual attention," namely one attention map for each image. Then the "dynamic speaker" module dynamically selects which feature to attend to and generates the caption. Our model differs in that we use the fused difference and we incorporate a viewpoint encoding module.

M-VAM. [28] Mirrored Viewpoint-Adapted Matching (M-VAM) framework derives changed and unchanged features with respect to both before and after scenes (hence mirrored). Despite the semblance with our model in that the similarity map is acquired, how the map is utilized differs greatly. M-VAM uses the similarity map to obtain the unchanged and changed probability maps, and apply these directly as attention maps to model the difference. By contrast, our approach employs the similarity map as a reference to embed the viewpoint information, and derive viewpoint-encoded image features for resolving the difference encoding problem in a later step.

4.4. Results on CLEVR-DC

Quantitative Results. Table 1 summarizes the quantitative results of our model and baselines. Our model outperforms all tested baseline models in all four metrics. The difference is the most apparent in CIDEr, where our model surpasses baselines by up to 15.

We also report the results of baseline models to which our cycle consistency module is attached. All models benefit from the cycle consistency module, with the improvement of up to 5.3 in CIDEr. This demonstrates the generalizability of our cycle consistency module, as the models with difference encoding mechanisms all exhibit performance gains.

Ablation	CIDEr	BLEU-4	METEOR	SPICE
VP	65.6	43.1	28.0	16.7
VP + CC	67.7	43.3	28.3	16.6
Diff	69.1	44.0	28.6	16.8
Diff + CC	69.7	44.1	28.8	17.2
VP + Diff	70.0	44.5	29.2	17.1
VP + Diff + CC	**71.7**	**45.0**	**29.3**	**17.6**

Table 2. Ablation studies of our model on the test split of CLEVR-DC. VP, Diff and CC indicate using our viewpoint encoding mechanism, difference modeling mechanism, and cycle consistency module, respectively.

The variants with incorrectly configured cycle consistency modules provide further insight into how the cycle consistency module operates. M-VAM outputs both changed and unchanged features from the image encoder. The correct configuration (M-VAM + CC) uses changed features for cycle consistency, whereas the incorrect one uses unchanged features. As the latter attempts to recreate the *unchanged* part of the after image with the *unchanged* part of the before image and the caption of what *changed*, the cycle consistency module fails to function and a performance decrease is observed. For VACC (I), we use X_{bef} and X_{aft} instead of x_{bef} and x_{aft} for cycle consistency. $X_{\text{bef/aft}}$ contain both the viewpoint change and semantic difference, whereas the change caption only summarizes the semantic difference. Therefore, the cycle consistency module is unable to fully reconstruct X_{aft} as it is not supplied with the new camera position in X_{aft}, and again the scores decrease.

Ablation Studies. Table 2 shows the results from ablation studies of our model. VP, Diff, and CC indicate using our viewpoint encoding, difference encoding, and cycle consistency module, respectively. Comparing the ablative variants without the cycle consistency module, enabling both the viewpoint encoding and the difference encoding prove to be the most effective. This shows that the two embeddings interplay synergistically to extract the viewpoint robust differences in the image pair. We can also draw the same conclusion from the ablative variants that have the cycle consistency module.

Ablative results also indirectly proves that our cycle consistency module is generalizable. The results show that attaching the cycle consistency module to VP and Diff also yields an extra performance boost. Interpreting them as two different image encoding mechanisms, we can also conclude that the cycle consistency module is transferable to

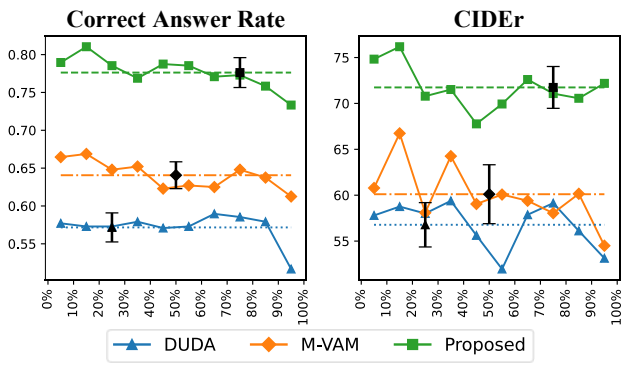

Figure 3. Correct answer rates and CIDEr scores by the cosine distance between before and after camera positions. Values are acquired for every 10th percentile (*i.e.*, 0%-10%, 10%-20%, ...). Dotted lines and error bars show the average and standard deviation, respectively. Best viewed in color.

other change captioning networks from these results.

Viewpoint Changes. We further investigate how robustly the network performs with respect to drastic viewpoint changes. To quantify the degree of change between a pair of images, we use the cosine distance between two camera positions. It is based on the fact that zooming in or out merely changes the sizes of the objects, whereas rotating the camera alters both the relative positions and sizes.

Figure 3 plots the variation of correct answer rates and CIDEr scores according to the cosine distance of camera positions. The correct answer rate refers to the ratio of correct predictions of the change type (*e.g.*, COLOR, TEXTURE, ADD, DROP, MOVE and DISTRACTOR). The correct answer rate graphs display downward trends for all models as the cosine distance increases, since the difficulty increases as the viewpoint change becomes greater. Our model scores the highest correct answer rates for all percentiles of cosine distances, and its standard deviation is comparable to that of other models (all about 0.02). In the CIDEr graphs, our model is the only approach with no clear downward trend and maintains the standard deviation to be the lowest (2.42, 3.21 and 2.28 for DUDA, M-VAM and ours, respectively) while having the highest CIDEr scores. These results suggest that our network is more robust to viewpoint changes compared to existing baseline models.

Qualitative Examples. Figure 4 represents three qualitative examples for all tested models. Our model generates the most accurate captions compared to other baseline models. Additionally, the visualized attention weights also signify that the network attends to the adequate parts of the input. For all images, the image attention maps (α_{bef}, α_{aft} in Eq. 7) have the highest weight around where the change has actually occurred. In Figure 4c, our network can contrast all objects in the images and conclude that no change is made. The text attention also behaves as predicted: in Figures 4a and 4b, α_t attend to x_{bef} when referring to the object in the before image, x_{aft} when checking the final state after the change, and x_{diff} when identifying the type of change. In Figure 4c, the network holistically compares all channels to determine that there is no change. α_w for the cycle consistency module also successfully attends to the words that are important for changes (*e.g.*, red changed green, yellow missing, seem identical).

Model	CIDEr	BLEU-4	METEOR	SPICE
Capt-Pix-Diff [19]	75.9	30.2	23.7	17.1
Capt-Rep-Diff [19]	87.9	33.5	26.7	19.0
Capt-Att [19]	106.4	42.7	32.1	23.2
Capt-Dual-Att [19]	108.5	43.5	32.7	23.4
DUDA [19]	112.3	47.3	33.9	24.5
M-VAM [28]	**114.9**	<u>50.3</u>	<u>37.0</u>	<u>30.5</u>
VACC (ours)	<u>114.2</u>	**52.4**	**37.5**	**31.0**

Table 3. Quantitative results on the test split of CLEVR-Change [19]. Best scores are in boldface, and the second bests are underlined.

Model	CIDEr	BLEU-4	METEOR	ROUGE
DDLA [10]	32.8	8.5	12.0	28.6
DUDA [19]	34.0	8.1	11.5	28.3
M-VAM [28]	<u>38.1</u>	**10.1**	<u>12.4</u>	<u>31.3</u>
VACC (ours)	**41.5**	<u>9.7</u>	**12.6**	**32.1**

Table 4. Quantitative results on the test split of Spot-the-Diff [10]. Presented the same way as Table 3.

4.5. CLEVR-Change

We also evaluate our model on a similar dataset but with much limited viewpoint changes. Table 3 shows the quantitative results of all models, and a qualitative example is provided in Figure 5a. The results suggest our model outperforms all baselines in most metrics and has comparable CIDEr results. The image and text attention maps in the qualitative example also indicate that our attention modules act in a predictable manner.

4.6. Spot-the-Diff

We also test the real-life applicability of our approach with a dataset with real-life images and human annotated captions. Table 4 lists the quantitative results of all models tested. Our method excels all baseline models in most metrics with the exception of BLEU-4, which is still on par with that of the state-of-the-art. A qualitative example is given in 5b, and the behavior of the attention modules is again congruent with our expectations.

5. Conclusion

In this work, we propose a novel neural network for change captioning that is resilient against viewpoint

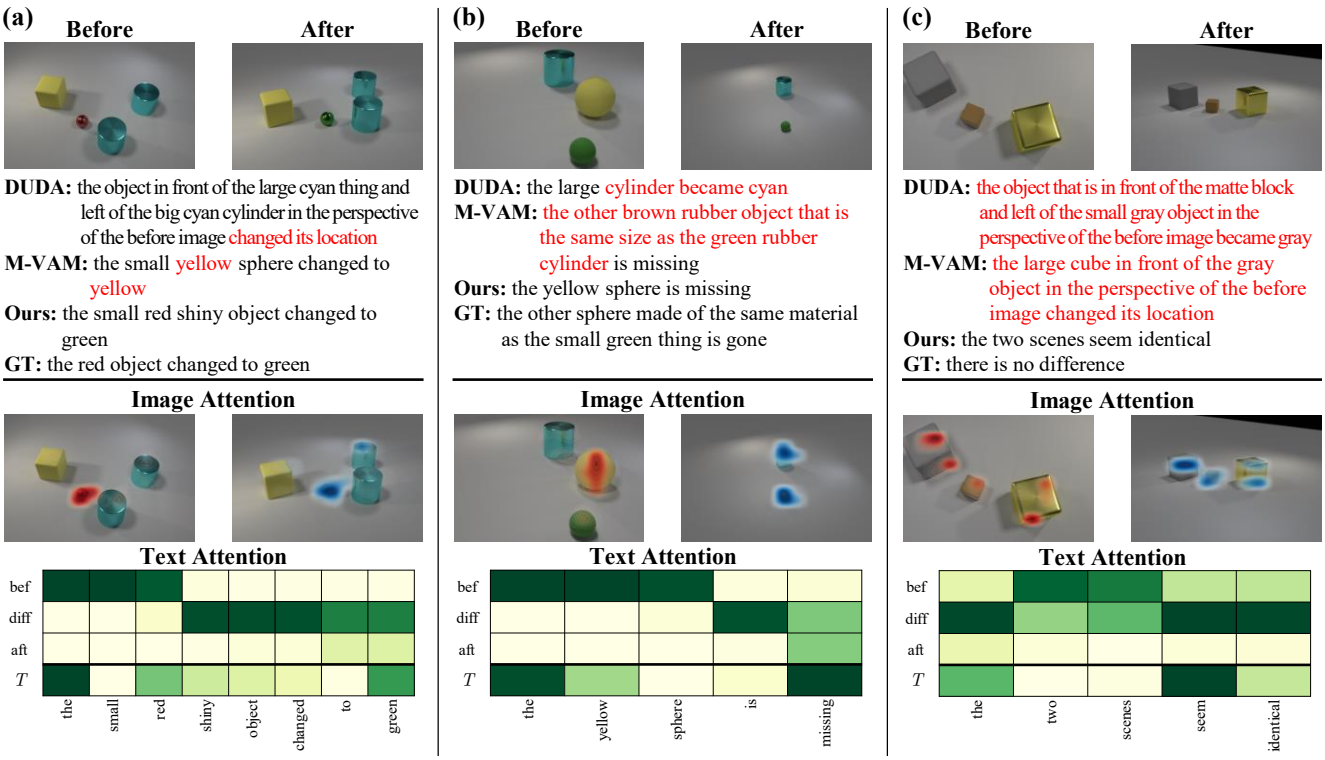

Figure 4. Qualitative examples on the test split of CLEVR-DC. For each image pair, we report the captions obtained by our method and baselines along with the ground truth. Incorrect parts of the captions are in red. The attention weights in Eqs. 7, 13 and 16 are also visualized for analysis. Image Attention displays the attention maps α_{bef} and α_{aft} in the difference encoder. In Text Attention, the rows labeled bef, diff and aft represent the attention weights α_t in the caption generator, and the row labeled T denotes the attention weight α_w in the cycle consistency module. Darker shades indicate higher attention weights.

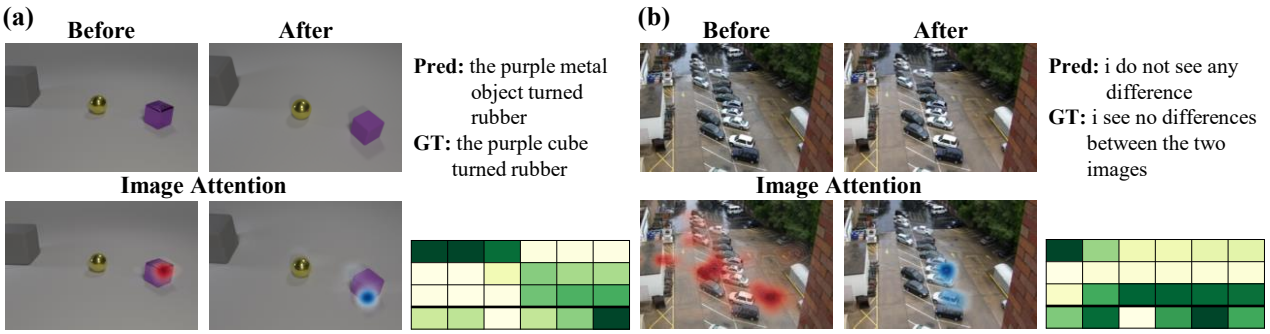

Figure 5. Qualitative examples on the test split of (a) CLEVR-Change and (b) Spot-the-Diff. Presented the same way as Figure 4.

changes. We design a new viewpoint encoding and difference modeling mechanism and tackle the problem utilizing only one image each from the before and after scene. Furthermore, we devise a cycle consistency module that evaluates the quality of caption by fusing the generated text and before image feature and contrasting it with the after image feature. We show the excellence of our network on three datasets, including one novel dataset we present to address drastic viewpoint changes. The results indicate our cycle consistency module can be merged with other existing models for extra performance gains.

Acknowledgments. We thank SNUVL members for the fruitful discussions, especially Youngjae Yu for the insight on cycle consistency. This work was supported by AIR Lab (AI Research Lab) in Hyundai Motor Company through HMC-SNU AI Consortium Fund, and the ICT R&D program of MSIT/IITP (No. 2019-0-01309, Development of AI technology for guidance of a mobile robot to its goal with uncertain maps in indoor/outdoor environments; and No. 2019-0-01082, SW StarLab).

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 5

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 4

[3] 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

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5

[5] Gilad Divon and Ayellet Tal. Viewpoint estimation—insights & model. In *ECCV*, pages 252–268, 2018. 2

[6] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 2

[7] Renaud Marlet Francisco Massa and Mathieu Aubry. Crafting a multi-task cnn for viewpoint estimation. In *BMVC*, pages 91.1–91.12, 2016. 2

[8] Satya Krishna Gorti and Jeremy Ma. Text-to-image-to-text translation using cycle consistent adversarial networks. *arXiv*, 2018. 2

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

[10] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *EMNLP*, pages 4024–4034, 2018. 1, 2, 5, 7

[11] 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 *CVPR*, pages 2901–2910, 2017. 2, 5

[12] KJ Joseph, Arghya Pal, Sailaja Rajanala, and Vineeth N Balasubramanian. C4synth: Cross-caption cycle-consistent text-to-image synthesis. In *WACV*, pages 358–366. IEEE, 2019. 2

[13] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *AAAI*, volume 35, pages 1771–1779, 2021. 2, 3, 4

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 5

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5

[16] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, pages 3153–3160. IEEE, 2011. 5

[17] Margarita Osadchy, Yann Le Cun, and Matthew L Miller. Synergistic face detection and pose estimation with energy-based models. *JMLR*, 8(5), 2007. 2

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 5

[19] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *ICCV*, 2019. 1, 2, 3, 4, 5, 6, 7

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv*, 2019. 5

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

[22] Hugo Penedones, Ronan Collobert, Francois Fleuret, and David Grangier. Improving object classification using pose information. Technical report, Idiap, 2012. 2

[23] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, pages 1505–1514, 2019. 2

[24] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. 3d-aware scene change captioning from multiview images. *IEEE RA-L*, 5(3):4743–4750, 2020. 2

[25] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. Indoor scene change captioning based on multimodality data. *Sensors*, 20(17):4761, 2020. 2

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

[27] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE TPAMI*, 31(5):824–840, 2008. 2

[28] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *ECCV*, 2020. 1, 2, 6, 7

[29] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *ICCV*, pages 2686–2694, 2015. 2

[30] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *CVPR*, pages 1510–1519, 2015. 2

[31] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 5

[32] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, pages 6439–6448, 2019. 2, 4

[33] Bram Wallace and Bharath Hariharan. Few-shot generalization for single-image 3d reconstruction via priors. In *ICCV*, pages 3818–3827, 2019. 2

[34] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv*, 2019. 2

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2

[36] Zhe Zhu. Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS P&RS*, 130:370–384, 2017. 1