

Text-to-Image Generation Grounded by Fine-Grained User Attention

IEEE Winter Conference on Applications of Computer Vision (2021) (to appear)

Google Research

발표자 이은수

목차

- 목차
 - Abstract
 - Introduction
 - The TReCS System
 - Evaluation

Abstract

이 논문은 **Text-to-Image Generation task**를 다루는 논문입니다

설명

이 사진에서 우리는 풀 위에 있는 얼룩말이
풀을 먹는 것을 볼 수 있으며 배경에서는
나무를 볼 수 있습니다

원본



생성된 이미지



Abstract

이 논문에서는 **Localized Narratives** 데이터 세트를 제공합니다

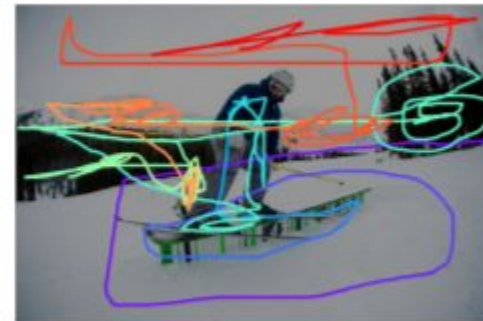
- 기존에는 이미지에 대한 설명이 짧은 문장으로만 구성되어 있던 반면, **이미지를 이야기 형태로 길게 풀어서 서술**하게 합니다
- 또한 이야기를 서술하면서 어떤 부분에 대한 서술을 하고 있는지를 **마우스 포인터로 가리키게 만들어서 수집**합니다

"A man skiing
along a rail on the
snowy hill."

Standard Caption

"In the foreground of the picture there
is snow and a railing. On the railing
there is a person in blue dress skiing.
On the right there are trees. On the
left there are trees. Behind the man
there is another person in red dress
walking. In the center there are moun-
tains. On the top sky is cloudy."

Localized Narrative



Narratives: 이야기. 이미지에 붙은 짧은 캡션과 비교해 더 긴 문장

Abstract

논문이 제안하는 TReCS 모델은 scene mask를 만들며 이로부터 이미지를 생성합니다

Composed Scene Mask



Generated Image



Introduction

Text-to-Image Generation task에서는...

1. **WordsEye**: 텍스트로 묘사된 장면을 이미지로 묘사하기 위해 데이터베이스에서 관련 3D 모델을 추출한 검색 기반 시스템
2. **GAN**: End-to-end generation model
3. **이 외 계층 구조 모델들**: Object bounding box를 만들거나, segmentation mask를 만드는 과정을 중간 단계로 추가해 현실적인 이미지 생성을 위한 representation으로 사용함

Introduction

Localized Narratives 데이터 세트는 주어진 이미지를 설명하는 동안 마우스 포인터를 기록하는 방법으로 기존의 짧은 캡션을 재구성합니다

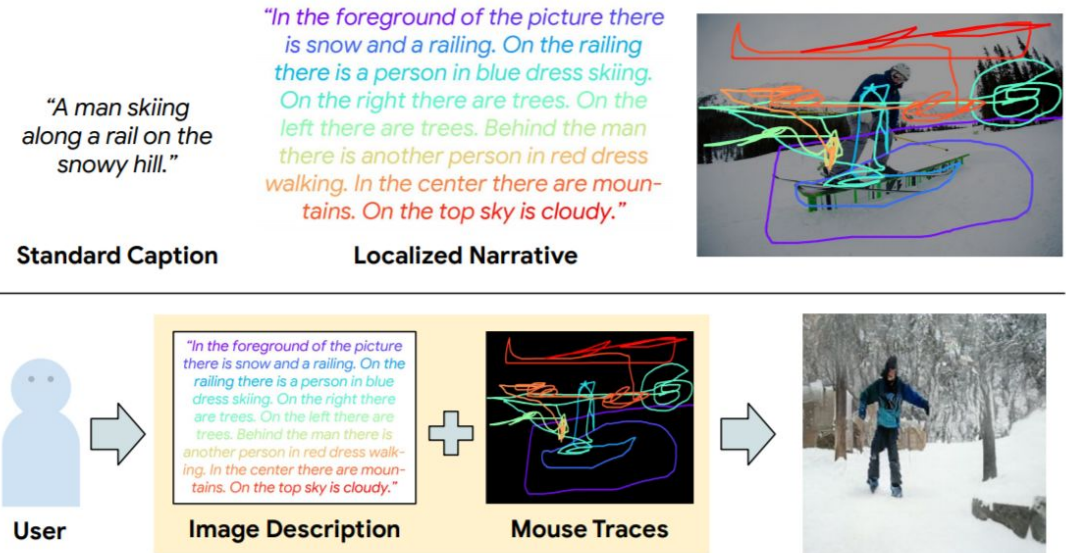


Figure 1. **[Top]**: Localized narratives comparing against standard captions. **[Bottom]**: Image synthesis from descriptions and traces.

Introduction

논문이 제안하는 TReCS 모델의 이름은 **Tag-Retrieve-Compose-Synthesize**의 줄임말입니다

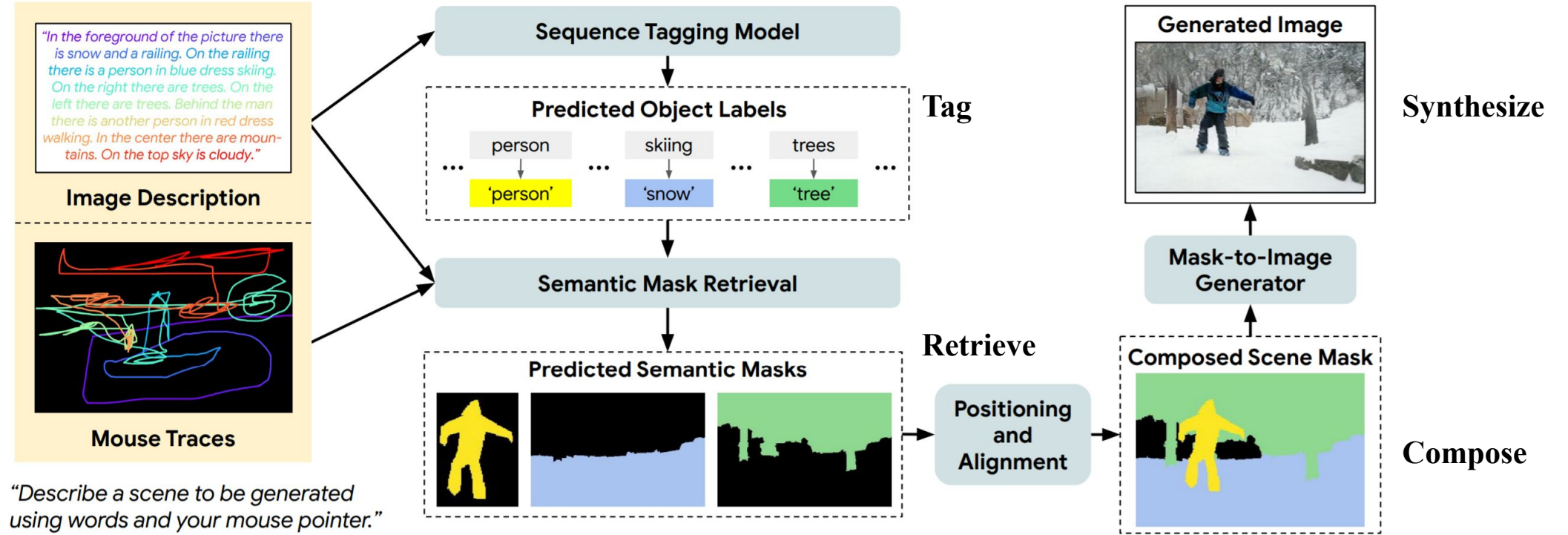


Figure 2. Multi-stage TReCS system for image synthesis using both descriptions and mouse traces.

Introduction

Contribution

1. Narratives에 대한 text-to-image generation task의 생존 가능성을 보임
2. SoTA 자연어 및 비전 기술을 사용해, 자연어와 spatial mouse traces 모두에 맞는 고품질 이미지를 생성하는 TReCS 모델을 제안함
3. 이전 SoTA를 넘는 자동 평가 및 인간 평가를 모두 수행함

The TReCS System

주어지는 입력은 다음과 같습니다

- Image description (narratives)
- Mouse traces

"In the foreground of the picture there is snow and a railing. On the railing there is a person in blue dress skiing. On the right there are trees. On the left there are trees. Behind the man there is another person in red dress walking. In the center there are mountains. On the top sky is cloudy."

Image Description

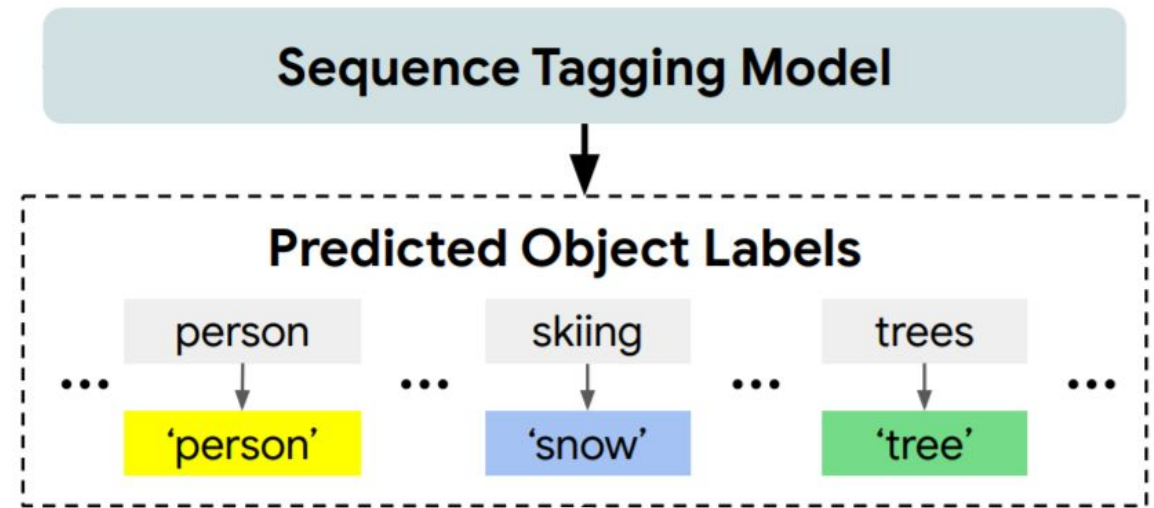


Mouse Traces

The TReCS System

Tag

- 입력으로 주어진 image description의 모든 단어마다 BERT를 사용해 object label을 예측합니다
- 이 단어가 어떤 object의 설명인지 통일하기 위해서입니다



The TReCS System

Sequence tagging

- 이미 데이터 세트에 있는 phrase-level mouse traces와, 이미지에서 추출한 segmentation mask를 합쳐 labeled traces를 만듭니다
- 주어진 description의 단어마다 labeled traces의 어떤 object에 해당하는지 예측합니다

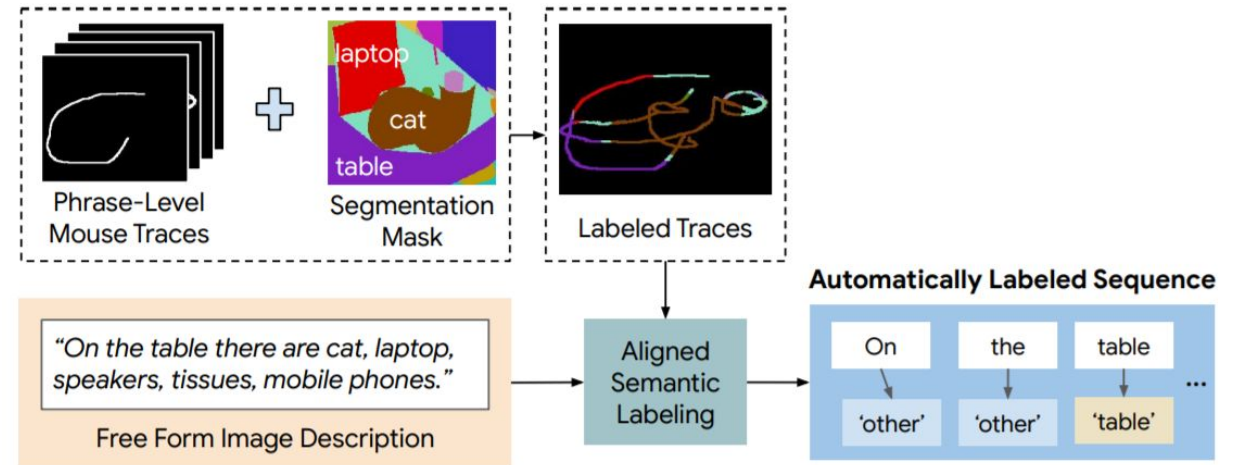


Figure 3. Overview of mouse trace sequence tagging.

The TReCS System

Noises

Object label을 예측하는 과정에서 noise가 발생할 수 있기 때문에 본 논문에서는 3가지 고전적인 방법을 사용합니다

1. TF-IDF weighting
2. IBM Model
3. Hidden Markov Model

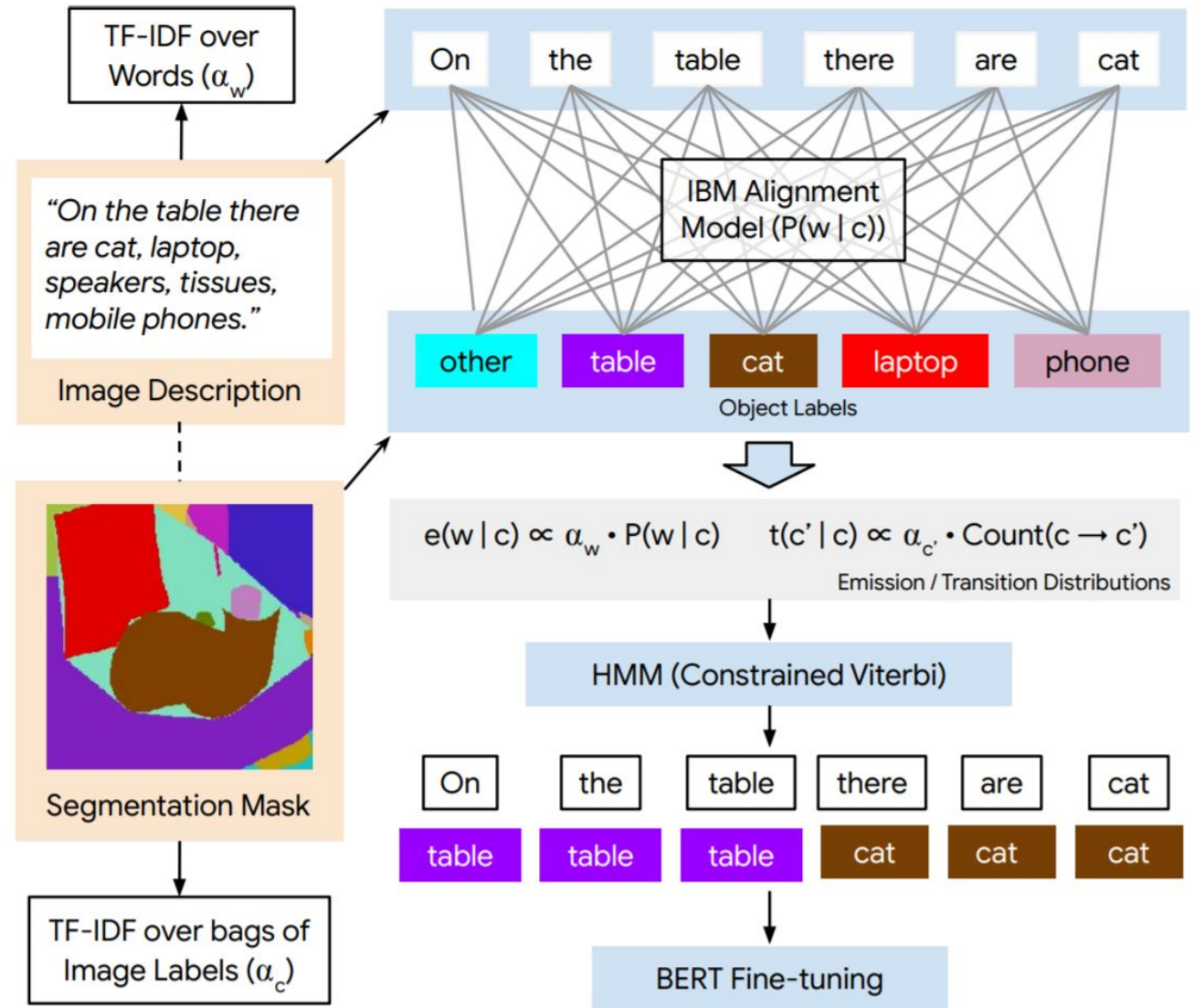


Figure 4. Semantic label refinement process.

The TReCS System

Retrieve

- 듀얼 인코더를 통해 object label마다 semantic mask를 생성합니다
- 듀얼 인코더는 서술과 가장 잘 일치하는 상위 k개의 이미지를 검색하고, 해당 이미지에서 탐지된 클래스의 COCO-Stuff mask를 선택합니다

Semantic Mask Retrieval

Predicted Semantic Masks



The TReCS System

듀얼 인코더는 BERT와 Inception V3를 사용하는 구조입니다
각 배치마다 이미지 및 캡션을 검색하도록 시뮬레이션됩니다

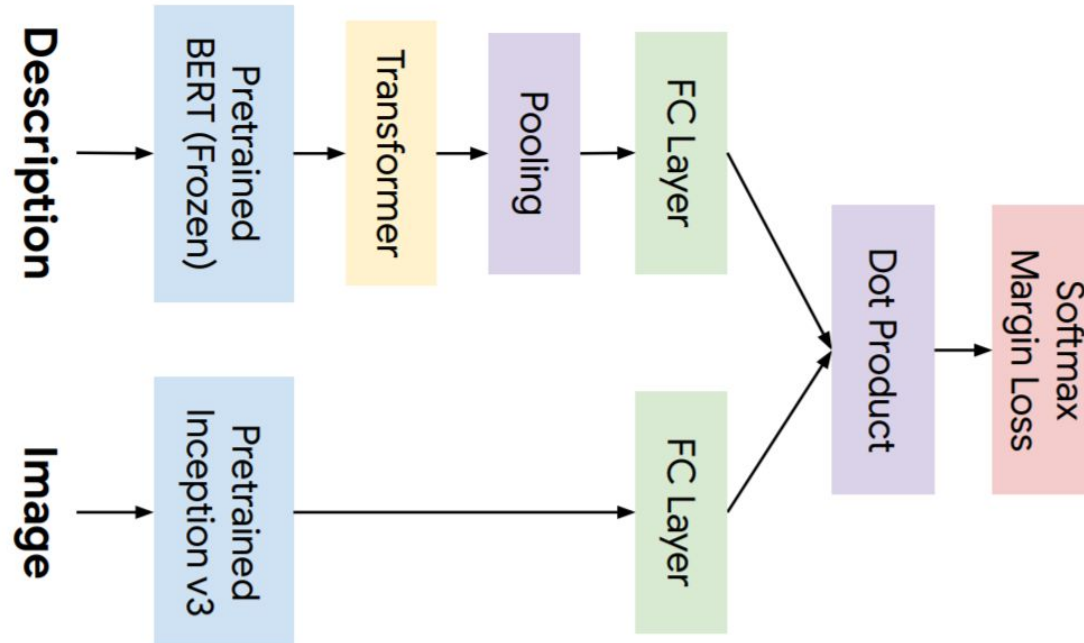


Figure 5. Dual-encoder approach for learning aligned image-text representations.

The TReCS System


Query Description	Retrieved Image #1	Retrieved Image #2	Retrieved Image #3
In this picture we can see food and spoon in the plate.			
In the image there is a donkey truck on the side of the road beside it there is a caution board, on back ...			
In a room people are seated on wooden chairs. In the center there is a rectangular dining table ...			

Figure 10. Dual encoder retrieval examples.

듀얼 인코더 결과물

The TReCS System

Composition

- 배경 레이어(background)는 종종 사람, 고양이, 비행기 등 객체 클래스를 구성하는 객체 레이어(foreground)를 덮어버리기 때문에 둘을 나눕니다
- 객체 레이어는 labeled traces의 중심으로 위치를 조정합니다

**Positioning
and
Alignment**

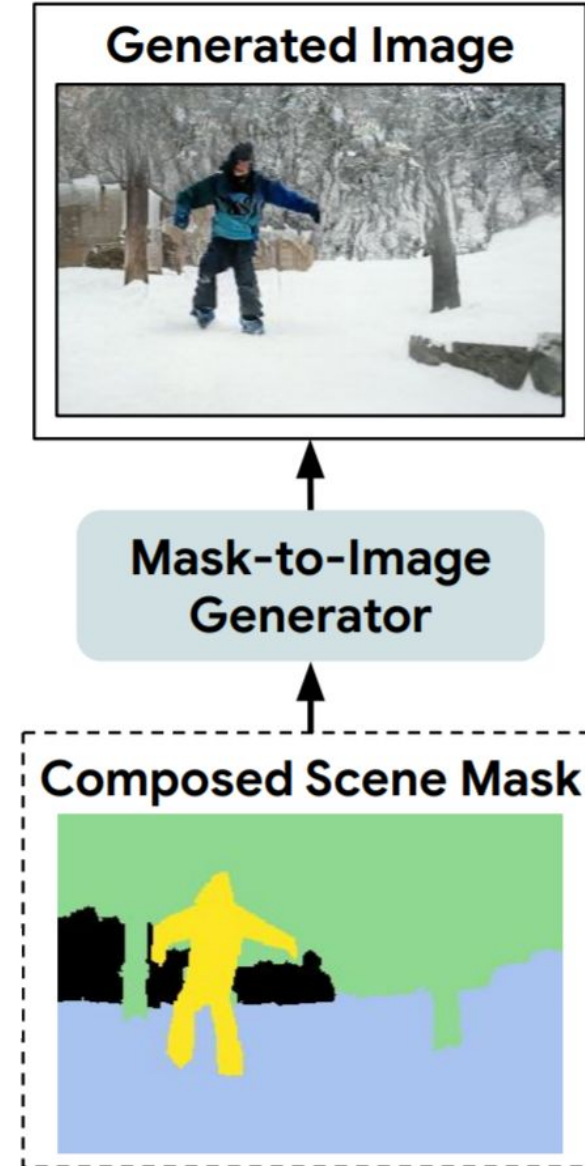
Composed Scene Mask



The TReCS System

Synthesize

- Scene mask를 외부 Mask-to-Image Translation 모델(예: SPADE, **CC-FPSE** 등)에 입력으로 넣어 이미지를 얻습니다
- 논문에서는 CC-FPSE를 사용하는 것이 가장 성능이 좋았다고 말하고 있습니다



Evaluation

이미지 품질

- 어떤 모델의 품질이 더 좋은지 선택하는 **인간 평가**, Inception V3의 예측 분포 점수를 측정하는 **Inception Score(IS)**, 실제 이미지의 분포와 생성된 이미지의 분포간 유사성을 측정하는 **Freechet Inception Distance(FID)** 3가지 항목으로 측정합니다

이미지 정렬 품질

- 설명에 맞게 이미지가 생성되었는지 확인하기 위해, 생성된 이미지를 Image Captioning Model에 넣어 나온 설명과, 원래 설명의 **BLEU, METEOR, CIDER** 점수를 측정합니다

Evaluation

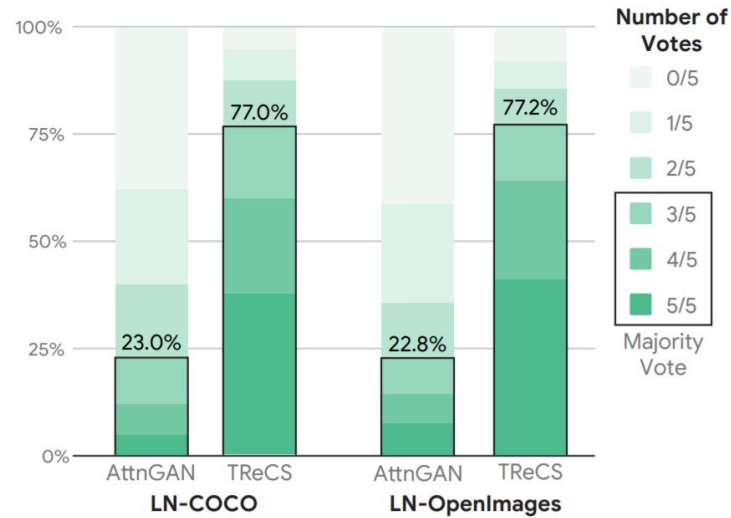


Figure 6. Human evaluation of image quality on LN-COCO validation set and LN-OpenImages test set. Models were fine-tuned on the LN-COCO training set. Of the decisions with 5/5 votes (indicating unanimous preference), TReCS was selected 88.3% of the time on LN-COCO, compared to 11.7% for AttnGAN. On LN-OpenImages, TReCS was selected unanimously 84.1%, compared to 15.9% for AttnGAN.

이미지 품질 인간 평가

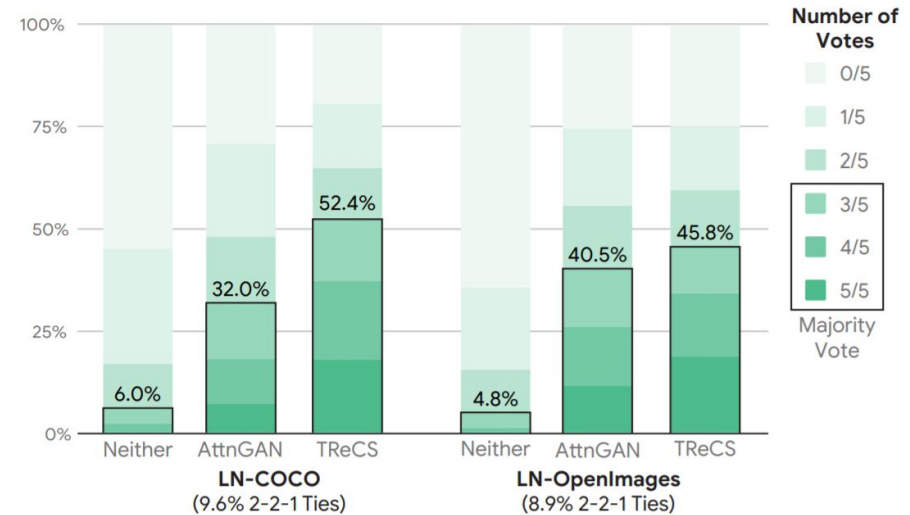


Figure 7. Human evaluation of image-text alignment on LN-COCO validation and LN-OpenImages test sets. Models were fine-tuned on the LN-COCO training set. Of the decisions with 5/5 votes (indicating unanimous preference), TReCS was selected 71.0% of the time on LN-COCO, compared to 27.8% for AttnGAN. On LN-OpenImages, TReCS was selected unanimously 61.4%, compared to 37.9% for AttnGAN.

이미지 정렬 품질 인간 평가

Evaluation

Dataset	Method	IS \uparrow	FID \downarrow
LN-COCO	Obj-GAN [†]	16.5	66.5
	AttnGAN [†]	17.4	59.4
	AttnGAN	20.8	51.8
	TRECS	<u>21.3</u>	<u>48.7</u>
LN-OpenImages	AttnGAN	<u>15.3</u>	<u>56.6</u>
	TRECS	14.7	61.9

Table 1. Image quality scores on LN-COCO validation and LN-OpenImages test sets. \uparrow (\downarrow) indicates that a higher (lower) number is better performance. [†] indicates models pretrained on the original COCO, but not fine-tuned on the LN-COCO training set.

이미지 품질 자동 평가

Evaluation

A person wearing jacket, helmet is on a ski board holding ski sticks. There is snow. On the back there is a banner, stand, ropeway ...



Group of people standing and we can see kites in the air and sky with clouds. A far we can see trees. This is grass.



In this picture we can see one boy is holding a bat and playing a game, he is keeping a cap. soundings there is ...



In the picture there is a road on the road there are many vehicles there are many poles on the road there are many trees ...



Evaluation

The image is outside of the city. In the image in middle there are few bags, on right side we can see a person standing ...



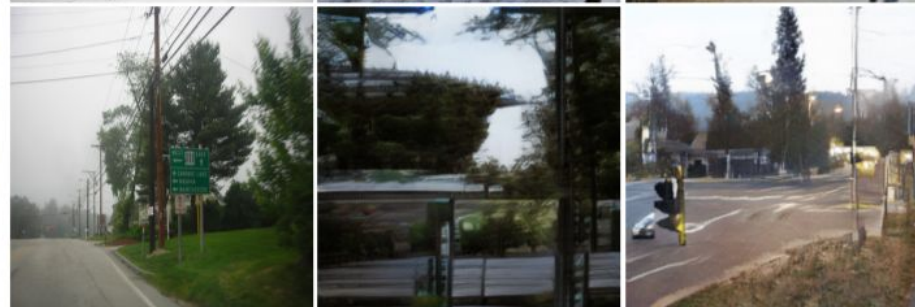
In this image we can see three fire engines on the road. In the background there are trees, houses and sky.



In this image i can see few benches and the ground covered with the snow. In the background i can see few ...



In this image I can see the road ... In the back there are signal lights and the vehicles. I can see many trees, building ...



Thank you