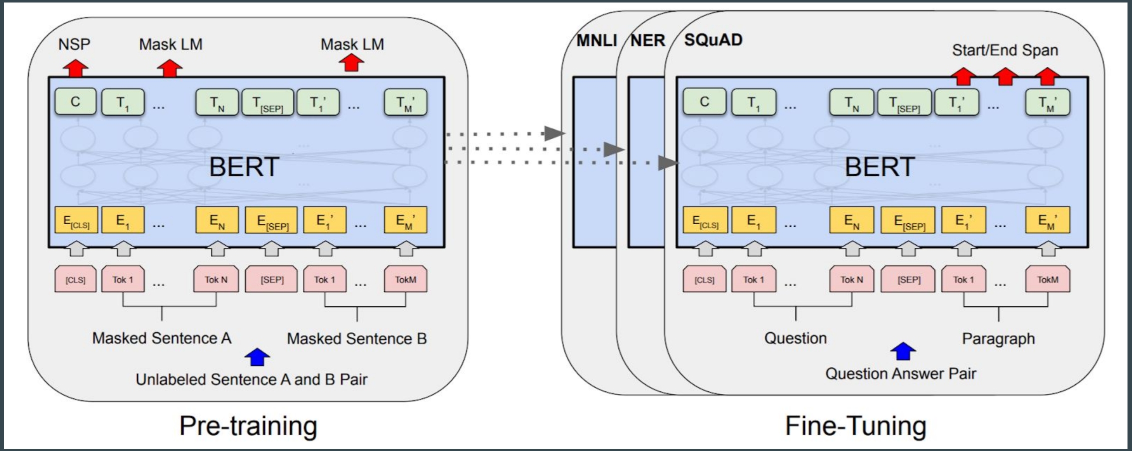
Fine-grained Image-to-Image Editing from Text Captions

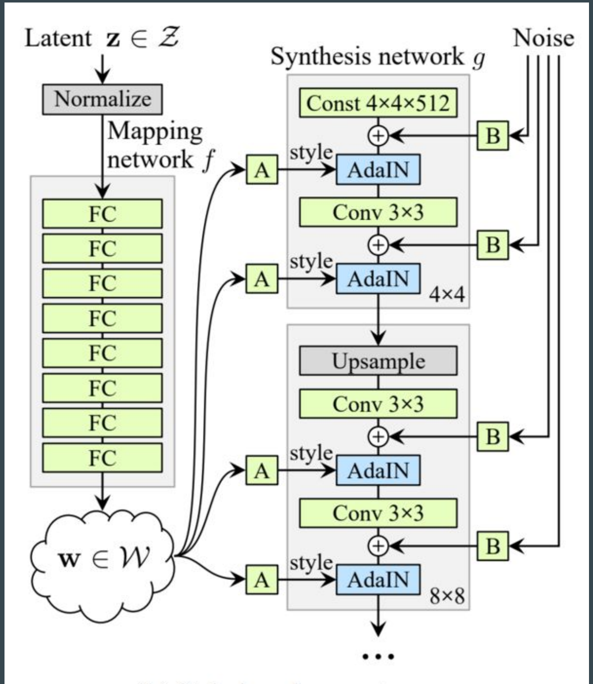
•••

Saptarashmi Bandyopadhyay, Matthew Gwilliam, David Li, Soumik Mukhopadhyay, Marco Volpi

Research Questions

1. Can BERT learn to ground style vectors?
2. Can StyleGAN2 be used to semantically edit images by algebraically manipulating style vectors?
3. Can Visual grounding with BERT improve the semantic editing of images?
4. How can we combine models trained on different modalities for multi-modal tasks?

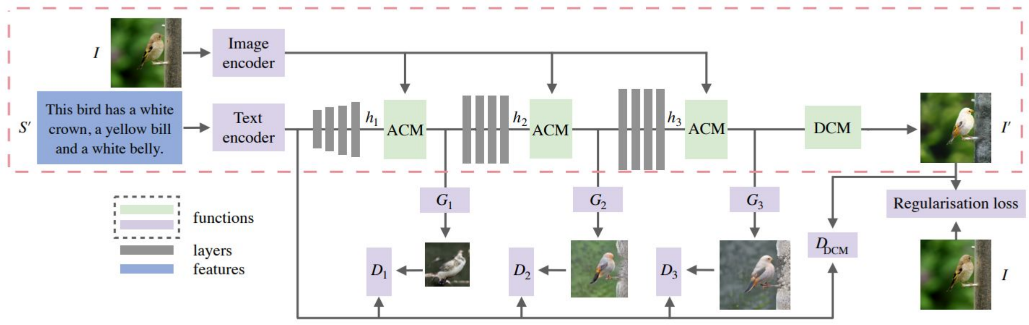




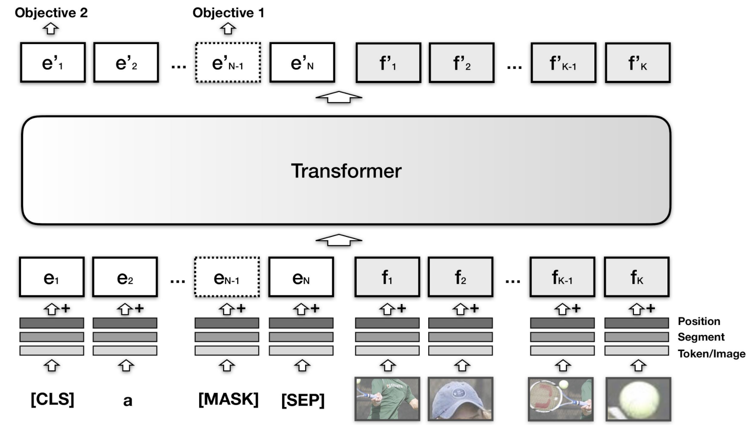
(b) Style-based generator

Related Works

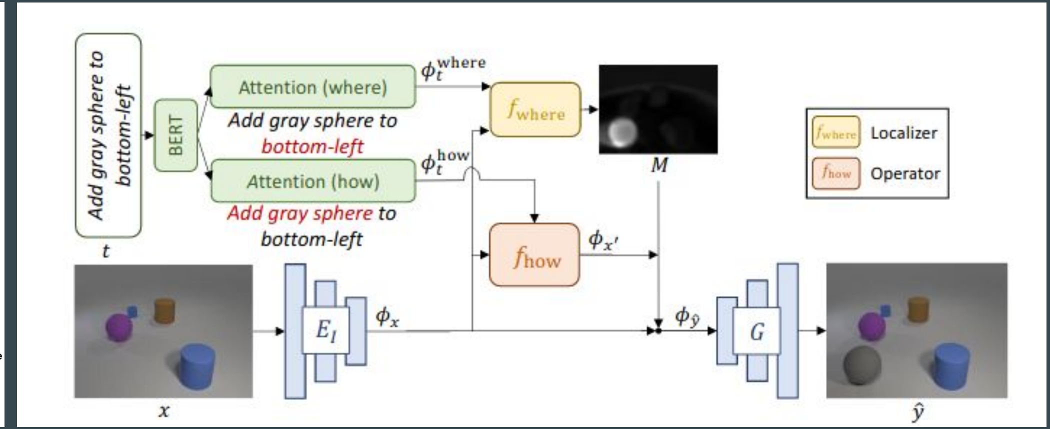
- VisualBERT
- ManiGAN
- FineGAN
- TIM-GAN



ManiGAN



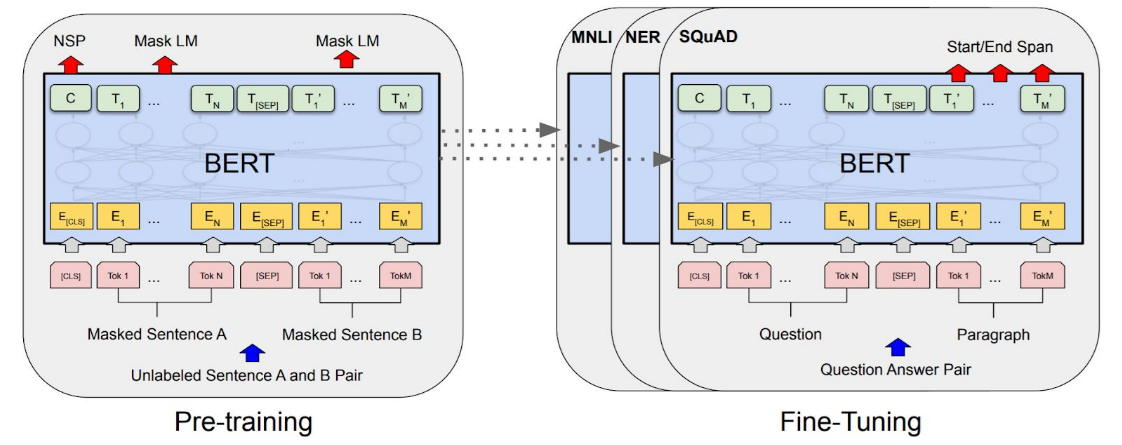
VisualBERT



TIM-GAN

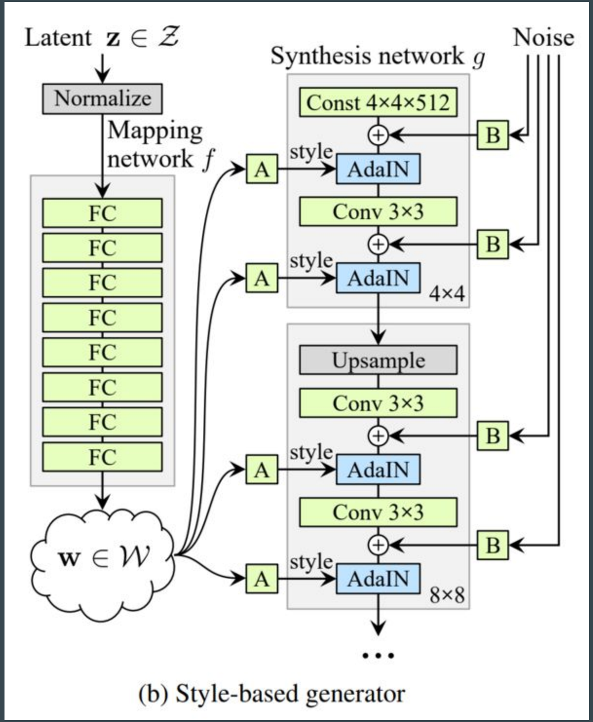
Background: Bert

- Bidirectional Encoder Representations from Transformers.
- Pre-trained NLP network by Devlin et al. (2018) from Google.
- Can be fine-tuned to perform different NLP tasks.



Background: StyleGAN

- GAN architecture from NVIDIA which uses adaptive instance normalization (AdaIN) to generate different *styles* of images.
- Consists of a mapping network (f) and an image synthesis network (g).
- Generates images from a 512-dim style vector **w** and random noise.
- We use StyleGAN2, a slightly improved implementation of StyleGAN.

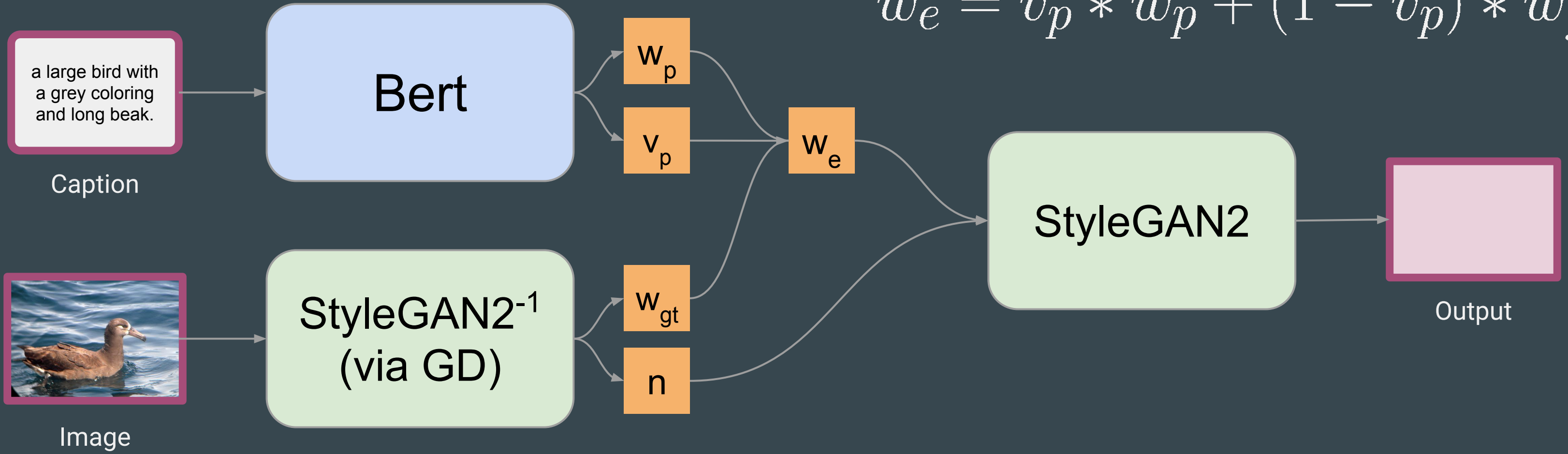


(b) Style-based generator

Architecture: Bert + StyleGAN2

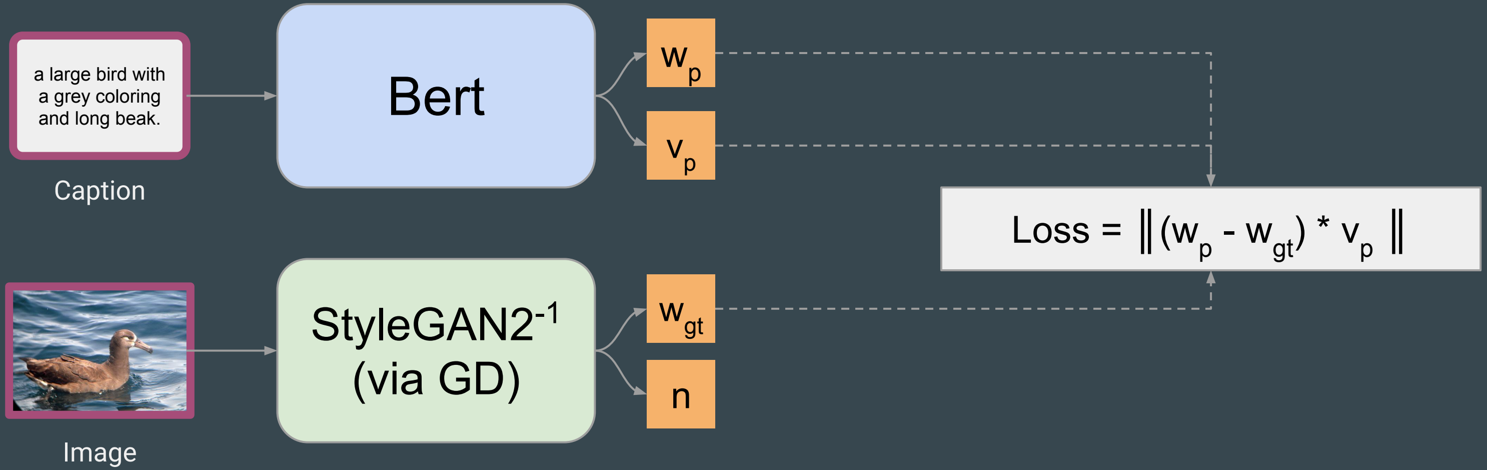
- Use Bert to predict style vectors based on image captions.
- Have StyleGAN2 synthesis network generate images.

$$w_e = v_p * w_p + (1 - v_p) * w_{gt}$$


a large bird with a grey coloring and long beak.
Caption
Image
Bert
StyleGAN2^{-1} (via GD)
w_p
v_p
w_{gt}
n
w_e
StyleGAN2
Output

Method 1: Weighted Style Prediction

- Make Bert predict a style edit (w_p) and a style weights (v_p).
- Style weights are used to ground the style vector based on the sentence.
- Style edit represents the predicted style.



a large bird with a grey coloring and long beak.

Caption

Bert

w_p

v_p

Loss = $\|(w_p - w_{gt}) * v_p\|$

Image

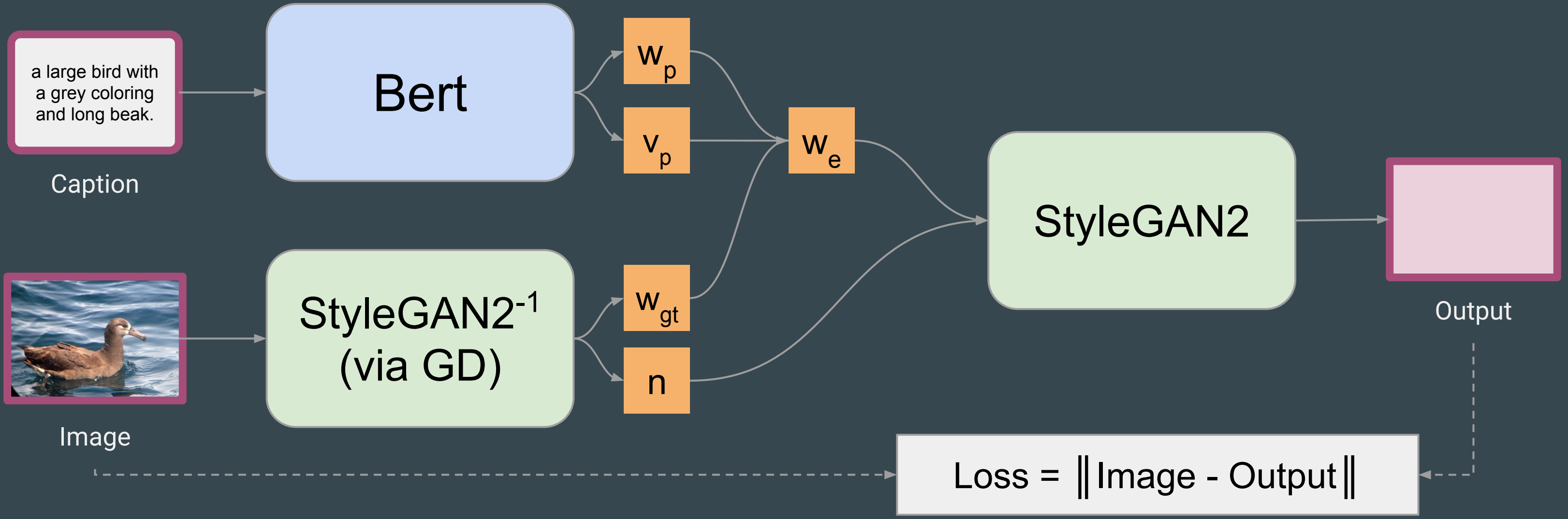
StyleGAN2^{-1}
(via GD)

w_{gt}

n

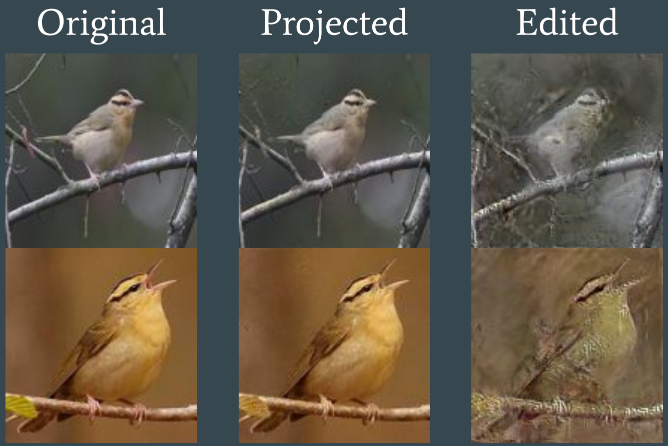
Method 2: Reconstruction

- Train Bert to refine the style vector w_{gt} based on the caption.



Method 1 & 2 Results

- We've pretrained a StyleGAN2 model on CUB birds and implemented our pipeline using Bert + StyleGAN2 in PyTorch.
- Attempting to add a red body yields a blurry image out of StyleGAN:



Background: FineGAN

- Disentanglement

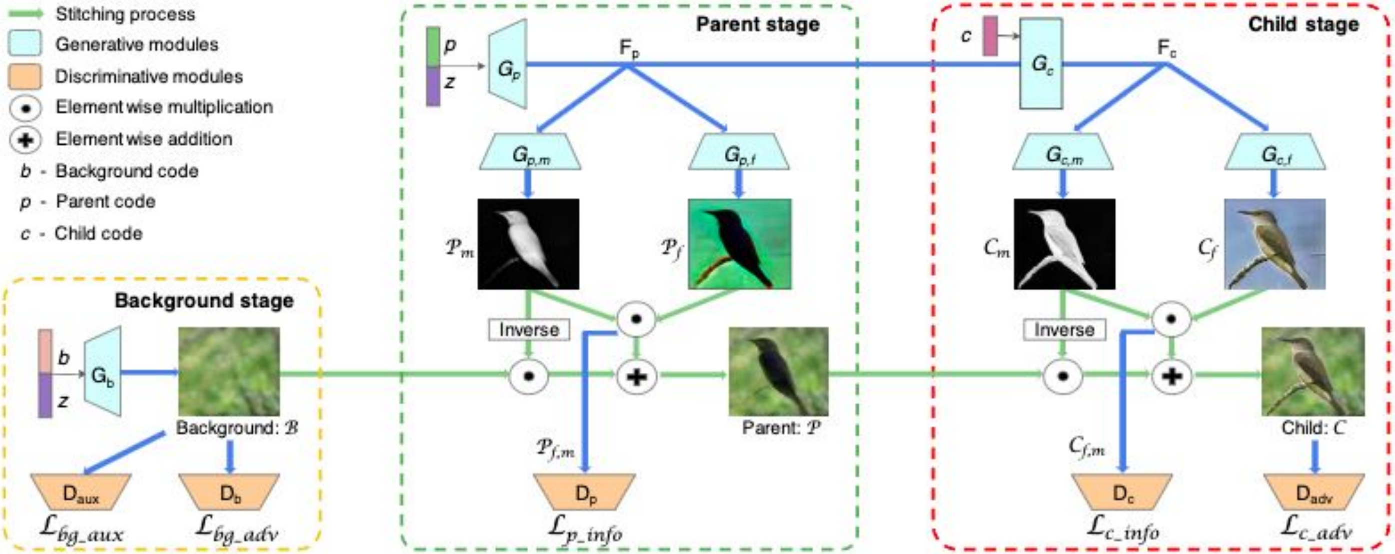


Figure 2. **FineGAN architecture** for hierarchical fine-grained image generation. The background stage, conditioned on random vector z and background code b, generates the background image B. The parent stage, conditioned on z and parent code p, uses B as a canvas to generate parent image P, which captures the shape of the object. The child stage, conditioned on c, uses P as a canvas to generate the final child image C with the object's appearance details stitched into the shape outline.

Background: Conditional Similarity Networks

- Conditional contrastive learning

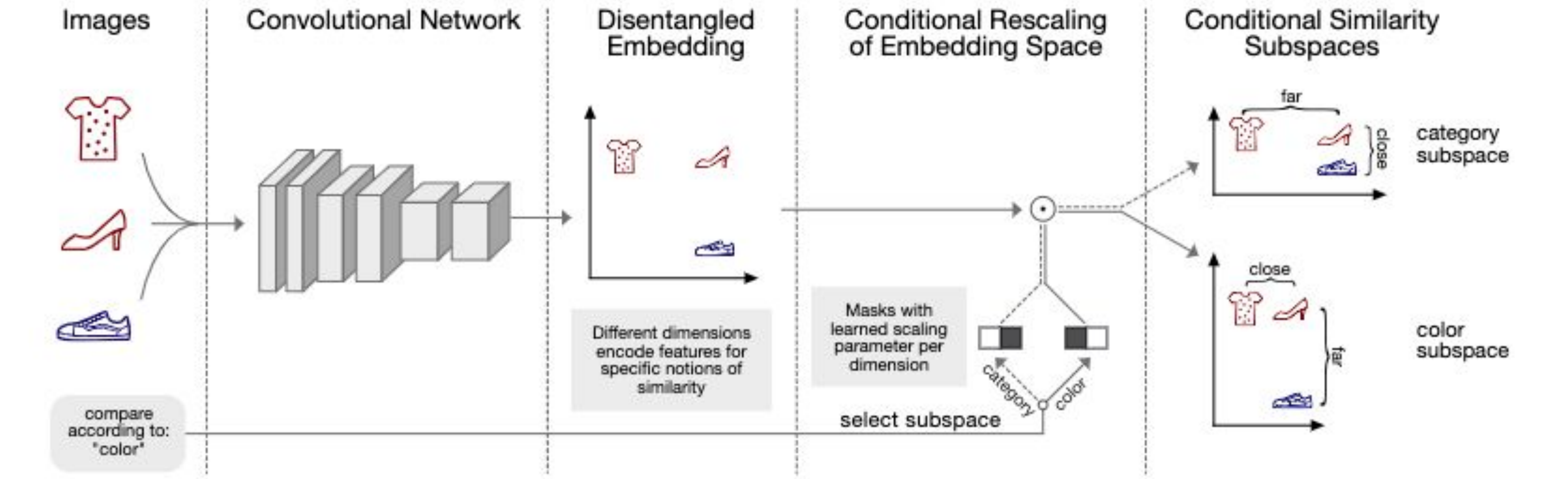
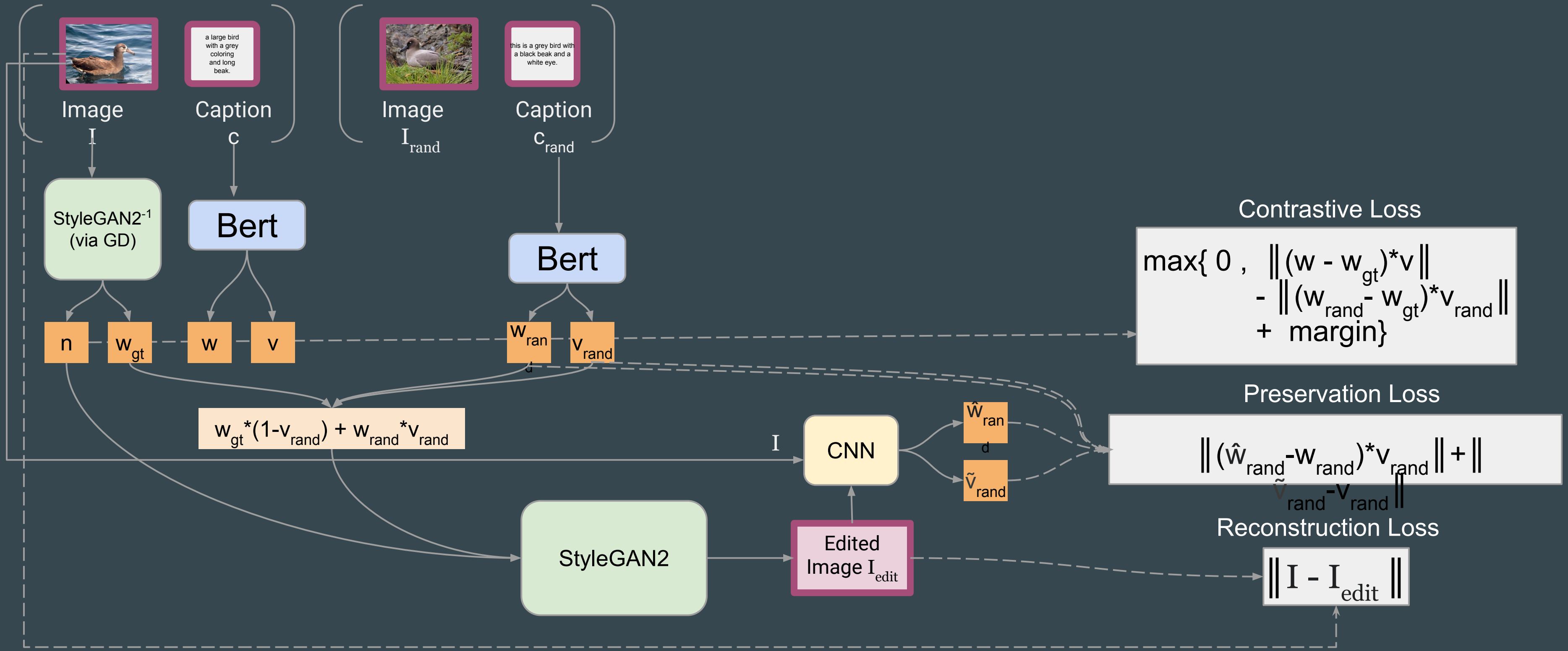

Images
Convolutional Network
Disentangled Embedding
Conditional Rescaling of Embedding Space
Conditional Similarity Subspaces
Different dimensions encode features for specific notions of similarity
Masks with learned scaling parameter per dimension
category
color
select subspace
compare according to: "color"
far
close
category subspace
close
far
color subspace

Figure 2. The proposed Conditional Similarity Network consists of three key components: First, a learned convolutional neural network as feature extractor that learns the disentangled embedding, i.e., different dimensions encode features for specific notions of similarity. Second, a condition that encodes according to which visual concept images should be compared. Third, a learned masking operation that, given the condition, selects the relevant embedding dimensions that induce a subspace which encodes the queried visual concept.

Method 3: Contrast + Reconstruction + Preservation


a large bird with a grey coloring and long beak.
this is a grey bird with a black beak and a white eye.
Image
I
Caption
c
Image
I_{rand}
Caption
c_{rand}
StyleGAN2^{-1} (via GD)
Bert
Bert
n
w_{gt}
w
v
w_{ran} d
v_{rand}
w_{gt}*(1-v_{rand}) + w_{rand}*v_{rand}
I
CNN
$\hat{w}_{ran}$ d
$\tilde{v}_{rand}$
StyleGAN2
Edited Image I_{edit}

Contrastive Loss

$$\max\{ 0 , \ \| (w - w_{gt})*v \| - \| (w_{rand} - w_{gt})*v_{rand} \| + \ \text{margin}\}$$

Preservation Loss

$$\| (\hat{w}_{rand} - w_{rand})*v_{rand} \| + \| \tilde{v}_{rand} - v_{rand} \|$$

Reconstruction Loss

$$\| I - I_{edit} \|$$

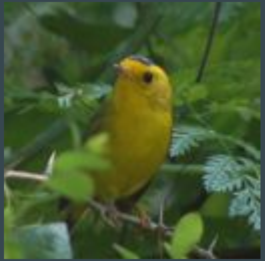
Method 3: Contrast + Reconstruction + Preservation

- Building on top of the previous method
- 2 input pairs of (Image, Caption) - (I,c) and (I_{rand},c_{rand})
- Idea is to edit I based on c_{rand} to give I_{edit}.
- 3 losses -
  - **Contrastive Loss** - To make sure that the BERT generated style from c is close to ground truth style (w_{gt}) while the BERT generated style from c_{rand} is far from w_{gt}.
  - **Preservation Loss** - To make sure that the information of style generated from c_{rand} is preserved in I_{edit}.
  - **Reconstruction Loss** - To make sure that I_{edit} remains structurally close to I.
- StyleGAN2 is pre-trained and non-trainable.

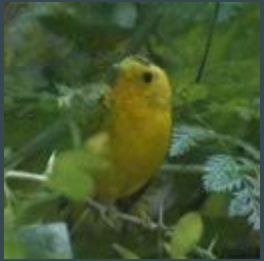
Method 3 Results

- We tested with reconstruction loss of the edited image against only the original image and with both the original and random image.

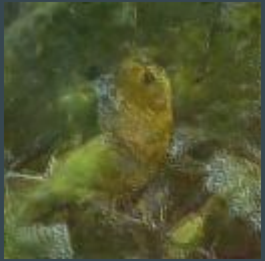
Input Image



Reconstruction with Input

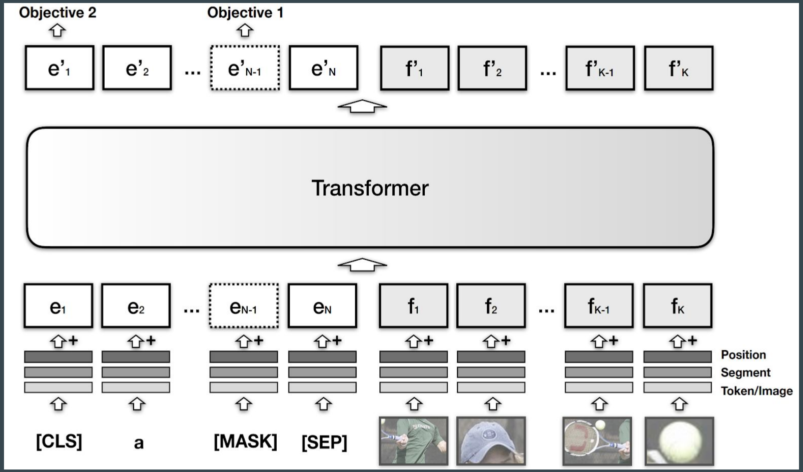


Reconstruction against both inputs



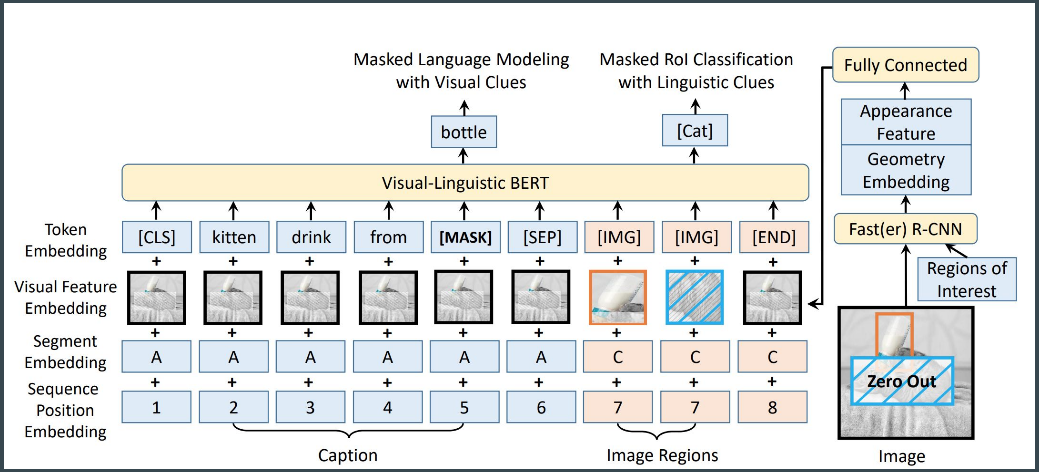
Background: Implicit Visual Grounding

- VisualBERT
  - embedding is Σ of a visual feature representation, a segment embedding and a position embedding
  - grounds visual concepts as seen by examining the corresponding attention matrices
- VL-BERT
  - caption segments consists of both a token embedding as well as a visual feature embedding



Objective 2
Objective 1
e'₁ e'₂ ... e'N-1 e'N f'₁ f'₂ ... f'K-1 f'K
Transformer
e₁ e₂ ... eN-1 eN f₁ f₂ ... fK-1 fK
Position
Segment
Token/Image
[CLS]
a
[MASK]
[SEP]

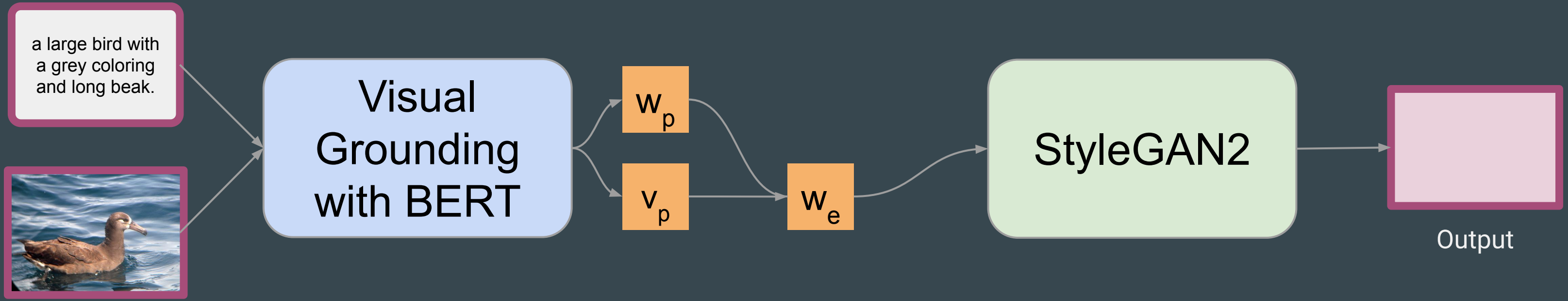
VisualBERT



VL-BERT

Method 4 Implicit Visual Grounding

1. Efficient editing by:
    a. aligning text to be replaced with image sub-region
    b. identifying the associated attention matrices from the dictionary
    c. replace the older attention weights in the image sub-region with the new ones essentially creating the new image aligned to the new word.
2. Quantitative evaluation:
    a. Calculate similarity on the non-aligned image sub-regions with the edited text
3. Use VisualBERT instead of BERT to predict the style vectors


a large bird with a grey coloring and long beak.
Visual Grounding with BERT
w_p
v_p
w_e
StyleGAN2
Output

Execution Issues of Visual Grounding

- Pre Trained models limited to large datasets like COCO
- Limited Downstream tasks like RefCOCO, VCR and VQA.
- Limited storage for downloading large datasets (~70G)
- Limited resources for fine tuning the pre-trained model
- Incomplete documentation of public repositories

Conclusion & Future Work

- Semantically grounding the latent space of a pre-trained StyleGAN2 model
- Using BERT and different training losses to predict style vectors
- BERT+StyleGAN2 architecture failed to edit images efficiently
- Visual Grounding with BERT
  - Efficient editing
  - Quantitative evaluation

Thank You