11-695: AI Engineering
Style Transfer (cont'd)

LTI/SCS

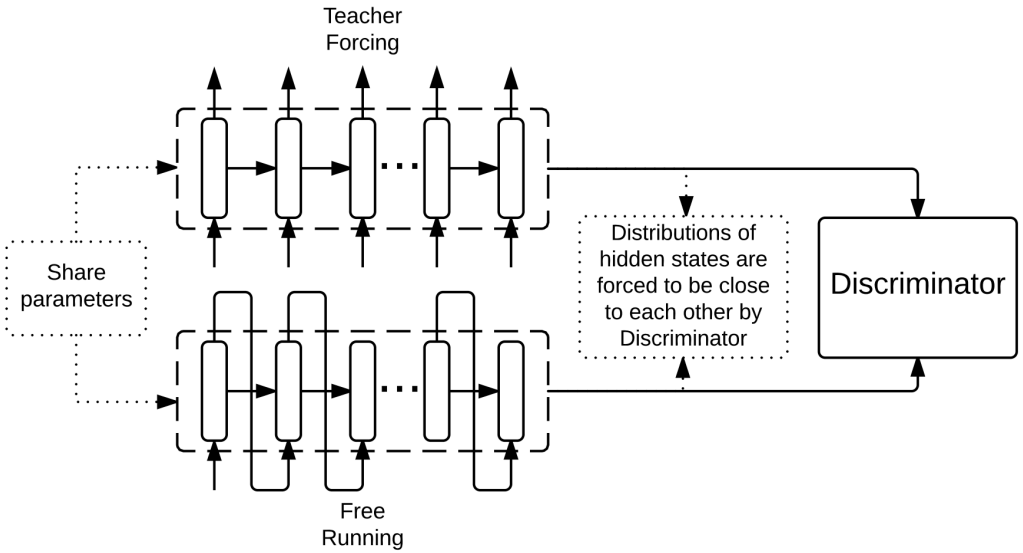
Spring 2020

Outline

1. Text Style Transfer

2. Cross-Domain Transfer/Translation
   - Image Captioning
   - Visual Question Answering
   - Image Synthesis from Text
   - Image Synthesis from Masks

Professor-Forcing[1]



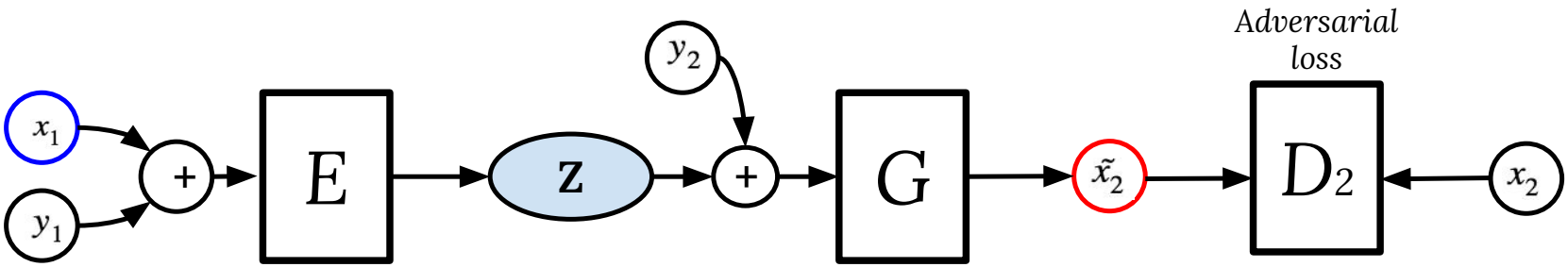
- Can not use GAN directly for text domain for discreteness
- Still want to apply adversarial training:
  - **G**: RNN as usual
  - **D**: classify *hidden states output* between free-run and teacher-forcing modes

[1] https://arxiv.org/pdf/1610.09038.pdf

Image credit: Alex Lamb *et al.* 2016

Cross-Alignment Text Style Transfer[2]

Carnegie Mellon



- Encoder + Generator = enc-RNN + dec-RNN = Seq2Seq
  - Encode: (text, style) = (x_i, y_i)
  - Embedded Representation: (z_{ci}, z_{si}), drop style z_{si},
  - combined with new style for decoding: (z_{ci}, y_j), $j = i$ or $j \neq i$
  - Result: $\hat{x}_j$ (we have transferred from style i to j).
- Professor-Forcing between 2 styles:
  - $\mathbf{D_1}$: distinguish $\hat{x}_1|z_{c1}, y_1$ and $\hat{x}_1|z_{c2}, y_1$
  - $\mathbf{D_2}$: distinguish $\hat{x}_2|z_{c2}, y_2$ and $\hat{x}_2|z_{c1}, y_2$

[2] https://arxiv.org/pdf/1705.09655.pdf

Image credit: Tianxiao Shen *et al.* 2016

Cross-Alignment TST: Results

Carnegie Mellon

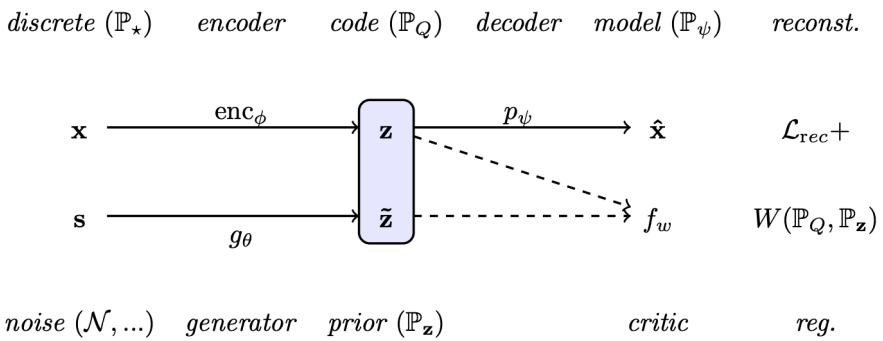
From negative to positive
consistently slow . consistently good . consistently fast .
my goodness it was so gross . my husband 's steak was phenomenal . my goodness was so awesome .
it was super dry and had a weird taste to the entire slice . it was a great meal and the tacos were very kind of good . it was super flavorful and had a nice texture of the whole side .
From positive to negative
i love the ladies here ! i avoid all the time ! i hate the doctor here !
my appetizer was also very good and unique . my bf was n't too pleased with the beans . my appetizer was also very cold and not fresh whatsoever .
came here with my wife and her grandmother ! came here with my wife and hated her ! came here with my wife and her son .

Table 3: Sentiment transfer samples. The first line is an input sentence, the second and third lines are the generated sentences after sentiment transfer by Hu et al. (2017) and our cross-aligned auto-encoder, respectively.

Image credit: Tianxiao Shen *et al.* 2016

Adversarially Regularized Autoencoder[3]

Carnegie Mellon



- EncoderRNN: $enc_\phi(z|x) = \mathbb{P}_Q$ (real), DecoderRNN: $p_\psi(\hat{x}|z)$
- Generator from noise $s \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$: $g_\theta(\tilde{z}|s) = \mathbb{P}_\mathbf{z}$ (fake)
- Discriminator: real $\mathbb{P}_Q$ vs. fake $\mathbb{P}_\mathbf{z}$
- Add binary classifier C to enforce disentanglement of y from z
- Objective: $\min_{\phi,\psi} [\mathcal{L}_{rec}(\phi, \psi) + \lambda_1 \mathbf{D}_{\text{Wasserstein}}(\mathbb{P}_Q, \mathbb{P}_\mathbf{z}) + \lambda_2 \mathcal{L}_C]$

[3] https://arxiv.org/pdf/1706.04223.pdf

Image credit: Jake (Junbo) Zhao *et al.* 2018

ARAE: Results

Carnegie Mellon

Positive	great indoor mall .
⇒ ARAE	no smoking mall .
⇒ Cross-AE	terrible outdoor urine .
Positive	it has a great atmosphere , with wonderful service .
⇒ ARAE	it has no taste , with a complete jerk .
⇒ Cross-AE	it has a great horrible food and run out service .
Positive	we came on the recommendation of a bell boy and the food was amazing .
⇒ ARAE	we came on the recommendation and the food was a joke .
⇒ Cross-AE	we went on the car of the time and the chicken was awful .
Negative	hell no !
⇒ ARAE	hell great !
⇒ Cross-AE	incredible pork !
Negative	small , smokey , dark and rude management .
⇒ ARAE	small , intimate , and cozy friendly staff .
⇒ Cross-AE	great , , , chips and wine .
Negative	the people who ordered off the menu did n't seem to do much better .
⇒ ARAE	the people who work there are super friendly and the menu is good .
⇒ Cross-AE	the place , one of the office is always worth you do a business .

Image credit: Jake (Junbo) Zhao *et al.* 2018

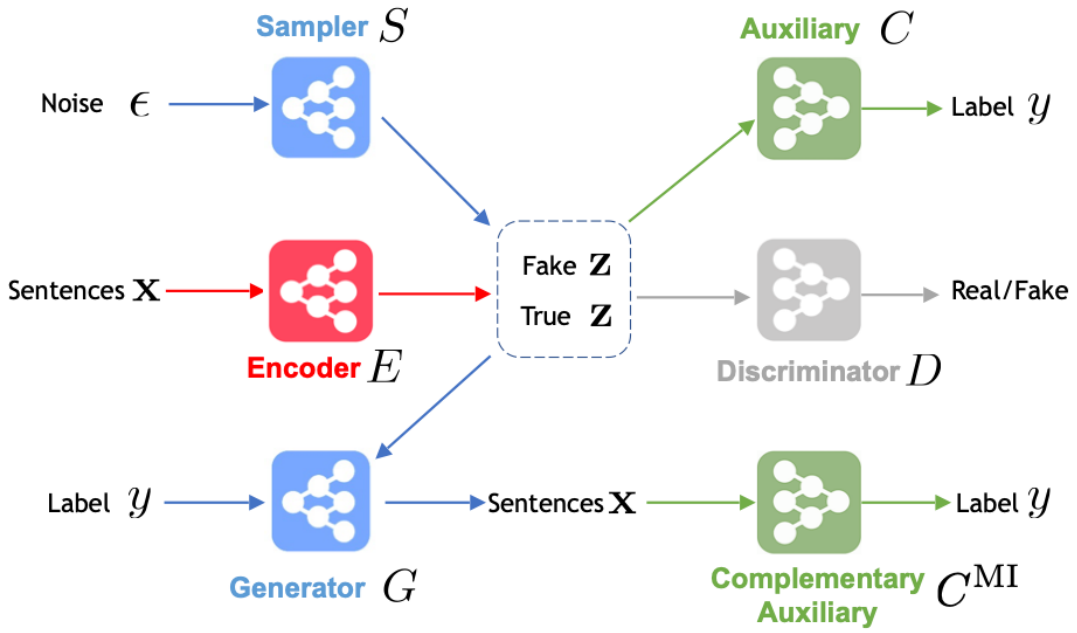
ARAE: Results

Carnegie Mellon

Science	what is an event horizon with regards to black holes ?
⇒ Music	what is your favorite sitcom with adam sandler ?
⇒ Politics	what is an event with black people ?
Science	take 1ml of hcl (concentrated) and dilute it to 50ml .
⇒ Music	take em to you and shout it to me
⇒ Politics	take bribes to islam and it will be punished .
Science	just multiply the numerator of one fraction by that of the other .
⇒ Music	just multiply the fraction of the other one that 's just like it .
⇒ Politics	just multiply the same fraction of other countries .
Music	do you know a website that you can find people who want to join bands ?
⇒ Science	do you know a website that can help me with science ?
⇒ Politics	do you think that you can find a person who is in prison ?
Music	all three are fabulous artists , with just incredible talent ! !
⇒ Science	all three are genetically bonded with water , but just as many substances , are capable of producing a special case .
⇒ Politics	all three are competing with the government , just as far as i can .
Music	but there are so many more i can 't think of !
⇒ Science	but there are so many more of the number of questions .
⇒ Politics	but there are so many more of the can i think of today .

Image credit: Jake (Junbo) Zhao *et al.* 2018

Regularized Auto-encoder (CARA[4])



- ARAE: learnable prior $\mathbb{P}_z$ from $s \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$: less posterior collapse
- Add an auxiliary classifier C^{MI} for generated sentences x
  - C^{MI} enforces conditional generation
  - Regulate the rate of C^{MI} and use BERT for encoder E

[4] http://people.ee.duke.edu/~lcarin/AAAI_LiY_6828.pdf

Image credit: Yuan Li *et al.* 2020

CARA: Results

Carnegie Mellon

	Business & Finance
ARAE	Where was the most emst adie place you apply as? Do you need a flat right now? *What is the law was a parent's length at their child's pepmed?*
$CARA_{AB}$	What is the conversion of irish money to american money? Knowing what the effect of ads are on people , why do we allow ads showing beautiful people. Where is the best place to look for a grant for a nonprofit soccer club?
	Family & Relationships
ARAE	What is the meaning of compliment? *What would you do if you just got out of heavyyme and have no job no where to* When ur a level 2 do u get 20some the same day?
$CARA_{AB}$	Why does a cheating man act like he is not cheating if he isn't interested in his Why do people think that children involved in a gay/lesiban adoption will be rebound what does it mean when someone tells you they always think about you?

Table 6: Qualitative results of conditional generation in topic-based question generation. Sentences in *Italic* form indicate their demonstrated categories do not match with their conditioned labels.

Image credit: Yuan Li *et al.* 2020

Some Other Papers

Carnegie Mellon

- A probabilistic formulation of unsupervised text style transfer[5]
- On Variational Learning of Controllable Representations for Text without Supervision[6]
- Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation[7]
- Zero-Shot Fine-Grained Style Transfer: Leveraging Distributed Continuous Style Representations to Transfer To Unseen Styles[8]
- Generating Sentences from Disentangled Syntactic and Semantic Spaces[9]
- Adapting Language Models for Non-Parallel Author-Stylized Rewriting[10]
- Structuring Latent Spaces for Stylized Response Generation[11]

[5] https://openreview.net/pdf?id=HJlA0C4tPS
[6] https://arxiv.org/pdf/1905.11975.pdf
[7] https://arxiv.org/pdf/1905.05621.pdf
[8] https://arxiv.org/pdf/1911.03914.pdf
[9] https://www.aclweb.org/anthology/P19-1602.pdf
[10] https://arxiv.org/pdf/1909.09962.pdf
[11] https://arxiv.org/pdf/1909.05361.pdf

Image credit: Yuan Li *et al.* 2020

Outline

1. Text Style Transfer

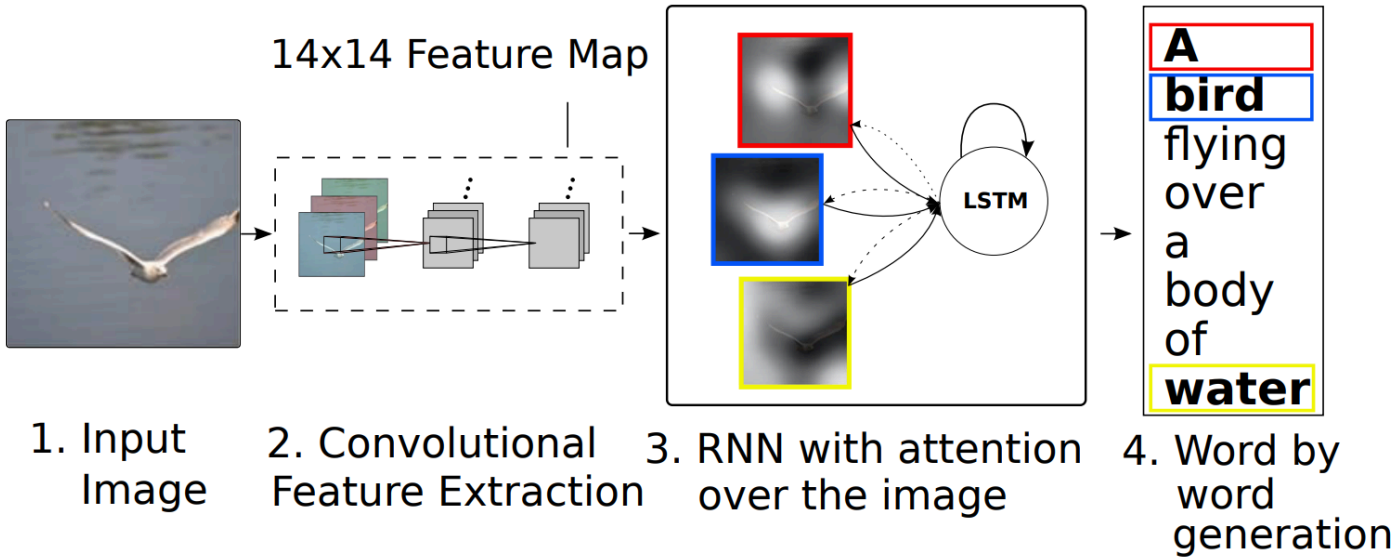
2. Cross-Domain Transfer/Translation
   - Image Captioning
   - Visual Question Answering
   - Image Synthesis from Text
   - Image Synthesis from Masks

Cross-Domain Transfer

Carnegie Mellon

- So far: Style/Attribute transfer within the same domain: Image/Video or Text
- Cross-Domain Transfer/Translation:
  - Image-to-Text: Image Captioning, VQA
  - Text-to-Image: Image Synthesis
  - Text-to-Audio: TTS (Text to Speech)
  - Audio-to-Text: ASR (Automatic Speech Recognition)
  - Any pair of modalities

Show-Attend-Tell[12]



- Still Encoder-Decoder model
  - Encode images by CNN
  - Decode to text by RNN
  - Attention to align spatial to sequential space
- Translate Image to Text, or generate Text conditioned on Image

[12] https://arxiv.org/pdf/1502.03044v3.pdf

Image credit: Kevin Xu *et al.* 2016

Show-Attend-Tell: Results

Carnegie Mellon

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

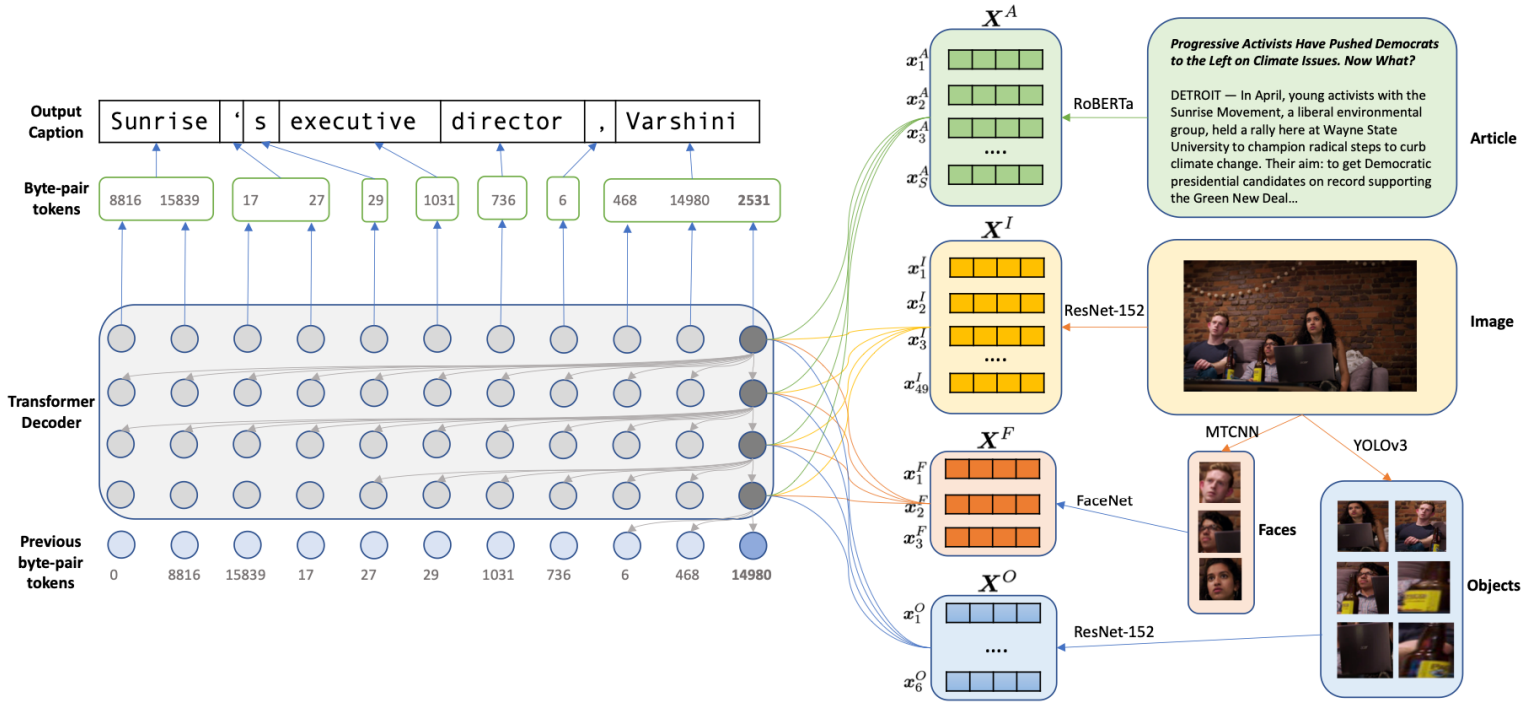
A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Image credit: Kevin Xu *et al.* 2016

Transform & Tell[13]: News Image Captioning



- Two main components:
  - Encoder: Extract 4 multimodal features $\mathbf{X}^A, \mathbf{X}^I, \mathbf{X}^F, \mathbf{X}^O$
  - Decoder: Transformer-based LM

[13] https://arxiv.org/pdf/2004.08070.pdf

Image credit: Alasdair Tran *et al.* 2020

Transform & Tell: Results

Carnegie Mellon

Japan Desperately Needs More Day Care Workers. New Mothers Need Not Apply.

TOKYO — Ever since she was a young girl, all Erica Takato wanted to do was work with small children. A few weeks into her term, she requested time off for bed rest ordered by her doctor.

....

Union officials and former teachers cite a major obstacle to the aspirations:

....

Ground-truth caption	A nursery school teacher showing a bug to his class.
Transformer + RoBERTa	Ms. Takato, who was born in Japan, was forced out of the day care program because she was pregnant.
+ image attention	Ms. Takato with her son, Kishiko, and their children, from left, Kaiti, 3, and Kaitama, 3, at a day care center in Tokyo.
+ weighted RoBERTa	Ms. Takato, with her son, Shiro, and son, at home in Tokyo. Ms. Takato, who was pregnant, said she was "so frustrated and lost hope of being able to work."
+ location-aware	A day care center in Tokyo.
+ face attention	A child care center in Tokyo. The government is eager to bring more women into the work force, and is trying to come up with enough child care for mothers.
+ object attention	A day care worker in Tokyo. The government is trying to bring more women into the work force, and the government is trying to come up with enough child care for mothers to go back to work.

Figure 4: An example article (left) and the corresponding news captions (right) from the NYTimes800k test set. The model with no access to the image makes a sensible but incorrect guess that the image is about Ms. Takato. Since the image appears in the middle of the article, only the location-aware models correctly state that the focus of the image is on a day care center.

Image credit: Alasdair Tran *et al.* 2020

Visual Question Answering (VQA[14])

Carnegie Mellon

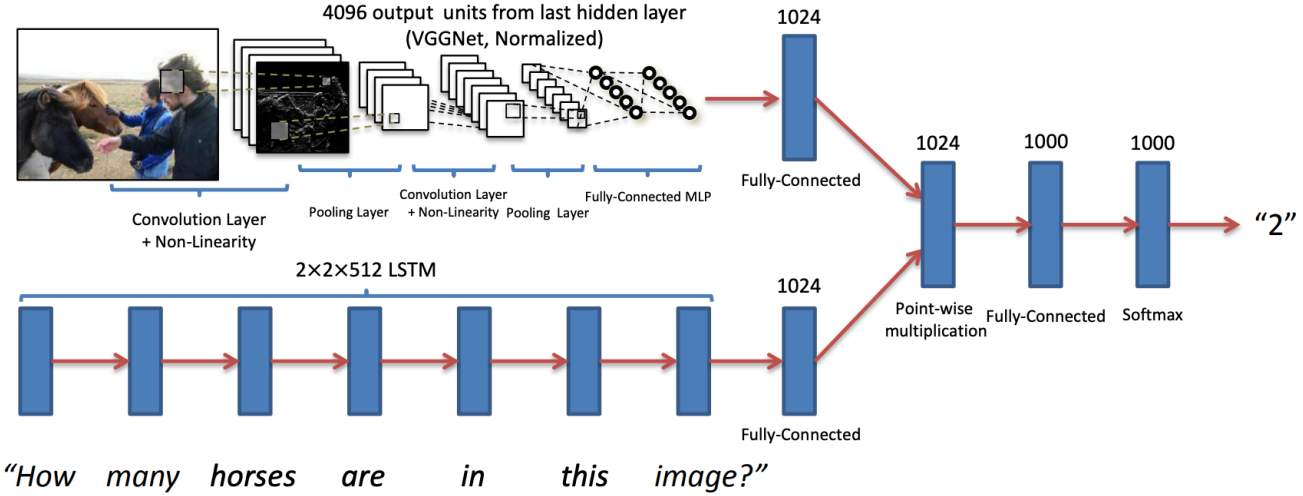


Fig. 8: Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [48] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers.

- Reference architecture for VQA
- Extracting features from both and merge/align them using point-wise multiplication

[14] https://arxiv.org/pdf/1505.00468.pdf

Image credit: Aishwarya Agrawal *et al.*

VQA: Results

Carnegie Mellon

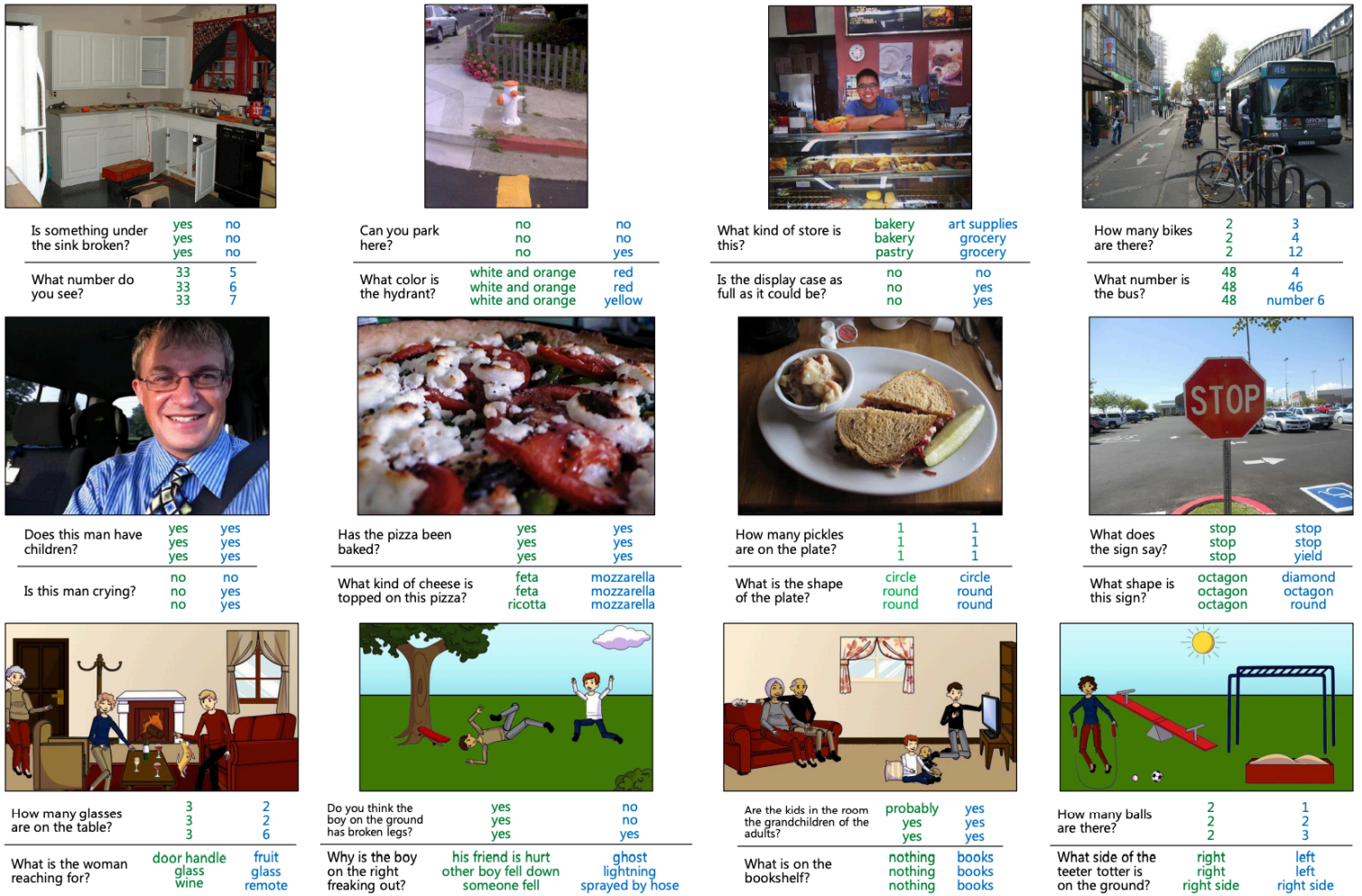


Fig. 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.

Image credit: Aishwarya Agrawal *et al.*

Up-Down Attention for VQA[15]

Carnegie Mellon

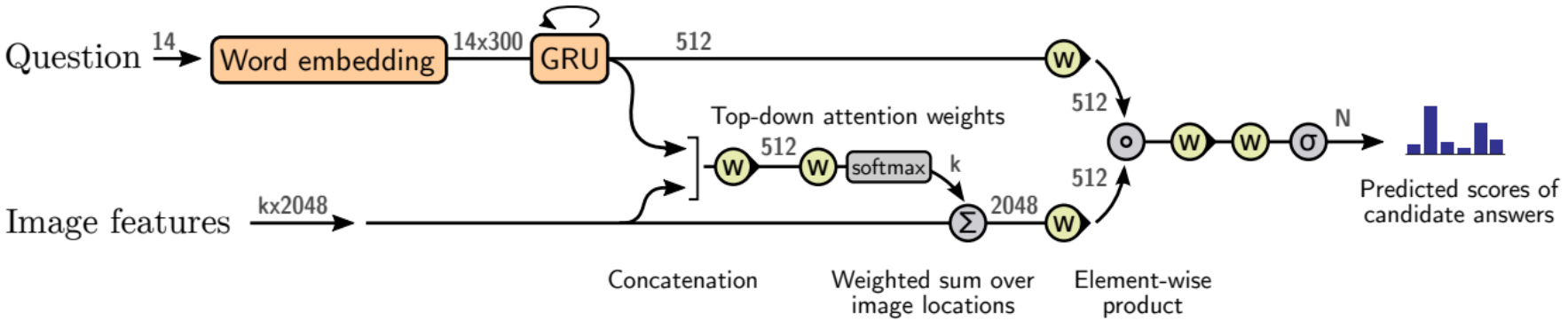


Figure 4. Overview of the proposed VQA model. A deep neural network implements a joint embedding of the question and image features $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_k\}$. These features can be defined as the spatial output of a CNN, or following our approach, generated using bottom-up attention. Output is generated by a multi-label classifier operating over a fixed set of candidate answers. Gray numbers indicate the dimensions of the vector representations between layers. Yellow elements use learned parameters.

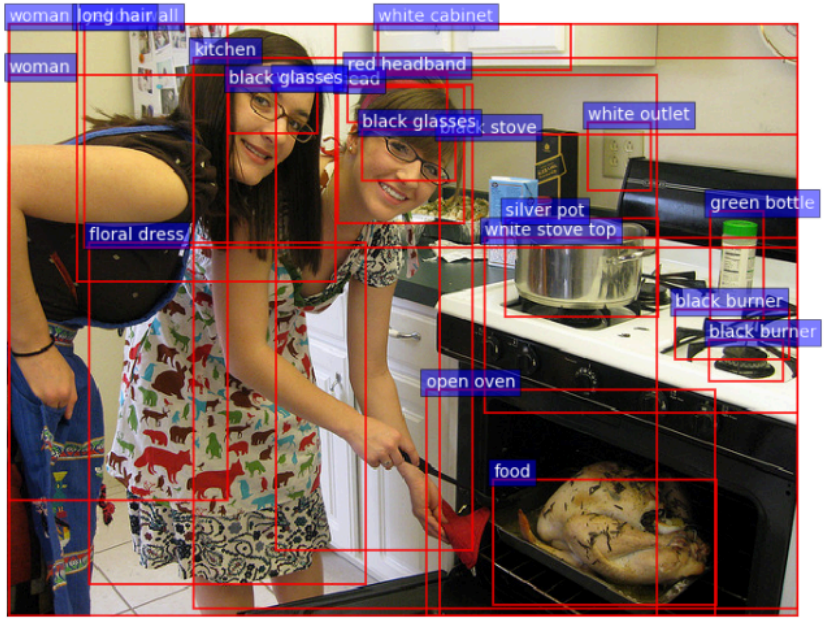
- Still based on the reference model
- With up-down attention

[15] https://arxiv.org/pdf/1707.07998.pdf

Image credit: Peter Anderson *et al.* 2017

Up-Down Attention for VQA

Carnegie Mellon



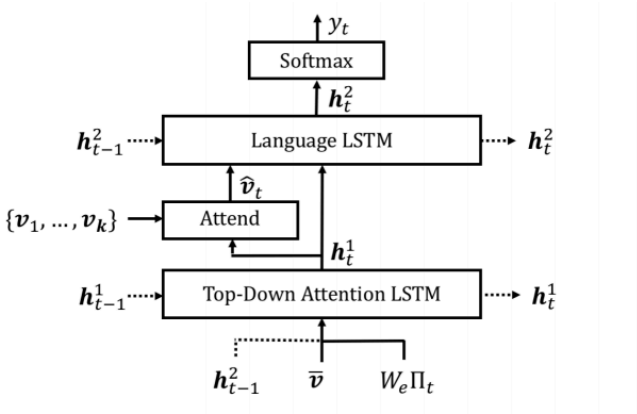


Figure 3. Overview of the proposed captioning model. Two LSTM layers are used to selectively attend to spatial image features $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$. These features can be defined as the spatial output of a CNN, or following our approach, generated using bottom-up attention.

- Faster RCNN to extract region vectors
- With up-down attention: using 2 LSTMs to align image-text

Image credit: Peter Anderson *et al.* 2017

Up-Down Attention for VQA: Results

Carnegie Mellon

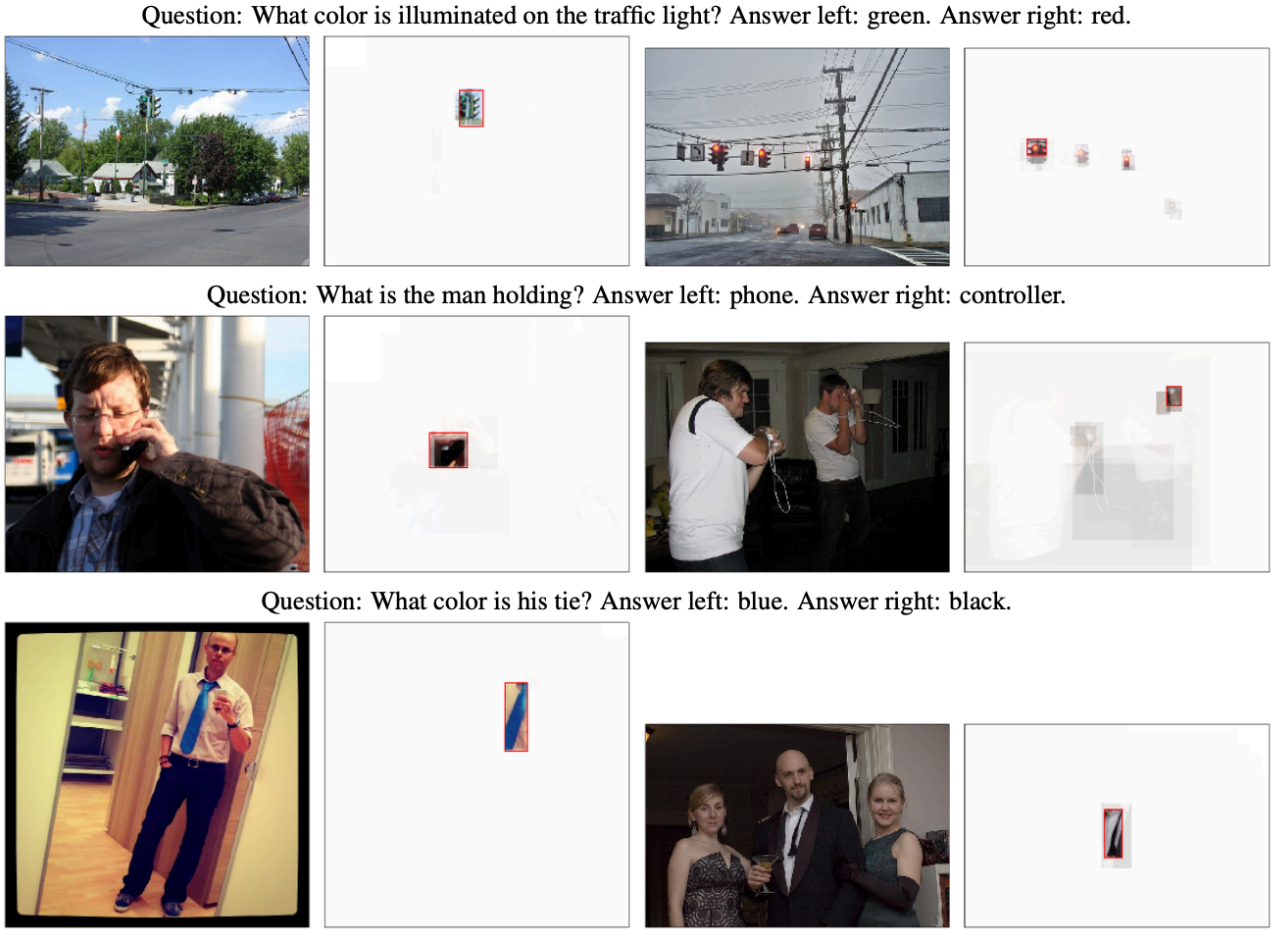
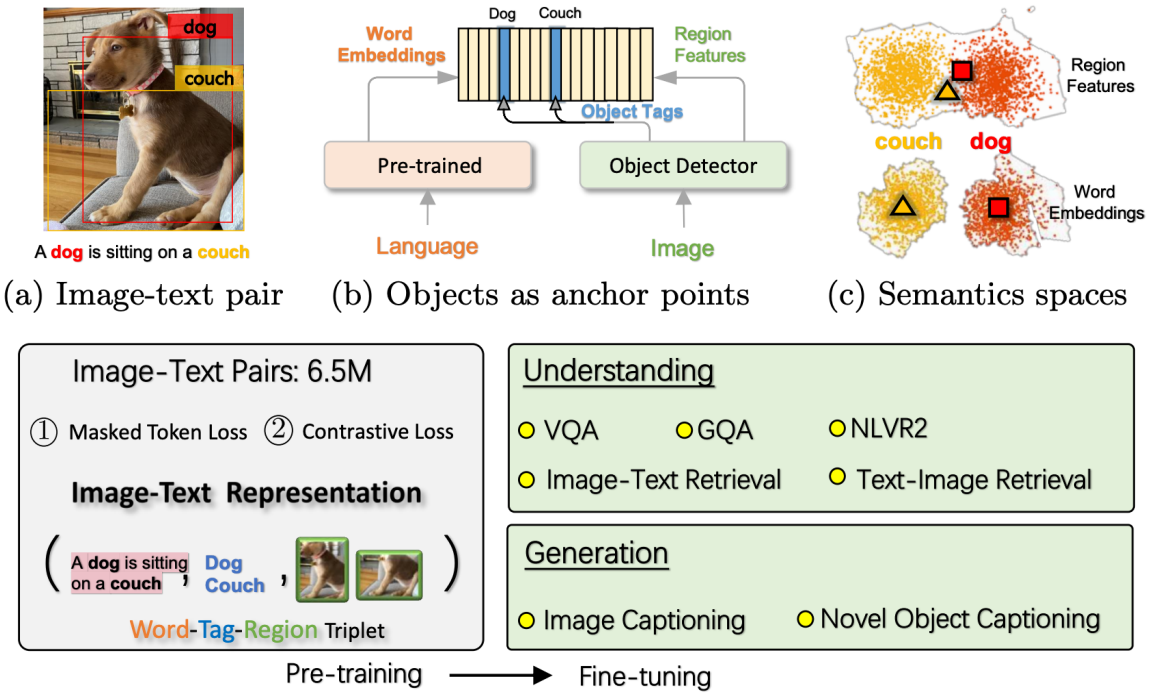


Image credit: Peter Anderson *et al.* 2017

Object-Semantics Aligned Pre-training (OSCAR[16])



(a) Image-text pair (b) Objects as anchor points (c) Semantics spaces

Fig. 1: OSCAR pipeline. The model takes a triple as input, is pre-trained with two losses (a masked token loss over words & tags, and a contrastive loss between tags and others), and fine-tuned for 5 understanding and 2 generation tasks (detailed in Sec. 4).

- Pre-training are now important for vision-language tasks
- Make use of object detector's *tags* for alignment

[16] https://arxiv.org/pdf/2004.06165v3.pdf

Image credit: Xiujun Li *et al.* 2020

OSCAR: Transformer and Pre-training

Carnegie Mellon

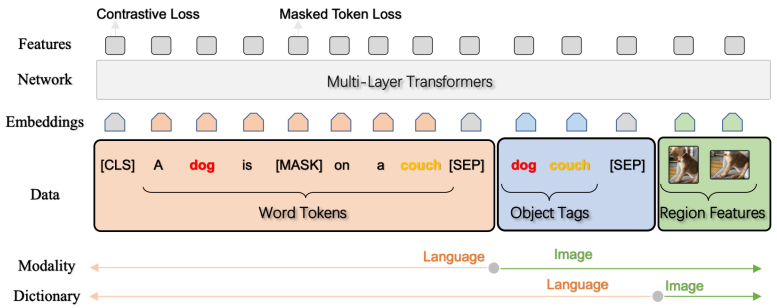


Fig. 3: Illustration of OSCAR. We represent the image-text pair as a triple [word tokens , object tags , region features], where the object tags (*e.g.*, "dog" or "couch") are proposed to align the cross-domain semantics; when removed, OSCAR reduces to previous VLP methods. The input triple can be understood from two perspectives: a *modality* view and a *dictionary* view.

- Pre-training: BERT for words, Faster-RCNN for obj detection, uses multiple datasets: $[w, q, v]$ are [word, tag, visual features]
  - Randomly mask 15% input $h = [w, q]$ and minimize

$$\mathcal{L}_{MTL} = -\mathbf{E}_{(w,q)\sim\mathcal{D}} \log p(h_i | h_{\neq i}, v)$$

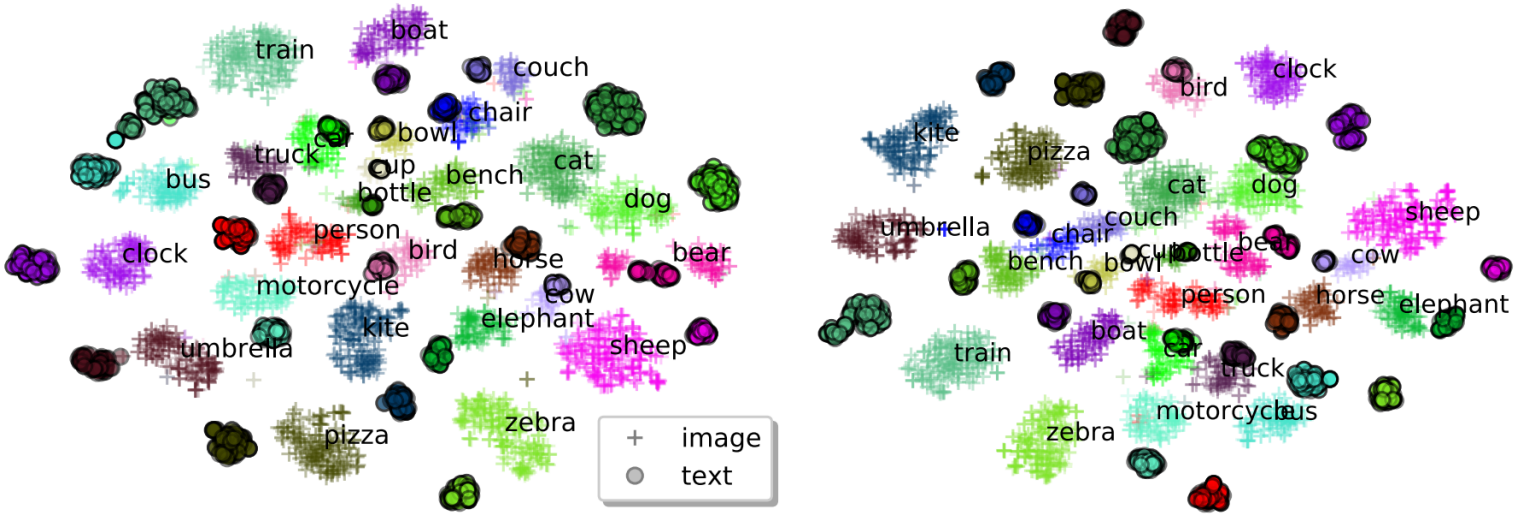
  - Randomly pollute 50% q, use binary classifier $y = \mathrm{FC}([q', v], v)$ to have contrastive loss:

$$\mathcal{L}_{MTL} = -\mathbf{E}_{(w,q',v)\sim\mathcal{D}} \log p\left(y | \mathrm{FC}([q', v], v)\right)$$

Image credit: Xiujun Li *et al.* 2020

OSCAR: Embedded Representation

Carnegie Mellon



(a) OSCAR (b) Baseline (No tags)

Fig. 4: 2D visualization using t-SNE. The points from the same object class share the same color. Please refer Appendix for full visualization.

- Pre-trained representations are adapted to specific tasks
- State-of-the-art for many tasks: Image Retrieval, Image Captioning, VQA-v2, ...

Image credit: Xiujun Li *et al.* 2020

OSCAR: Image Captioning Results

Carnegie Mellon

Oscar: a small train on a city street with people near by .
Baseline: a train that is sitting on the side of the road .

GT: a small train on a city street with people near by .
A black and red small train in shopping area.
A group of people near a small railroad train in a mall .

Tags: sign, tree, sidewalk, train, woman, person, trees, street, bus, stairs, store, man, balcony, building, people

Oscar: a red rose and white flowers in a vase .
Baseline: a vase filled with red and white flowers .

GT: A red rose in a glass vase on a table
beautiful red rose and white flowers are in a vase .
The bouquet has one red rose in it.

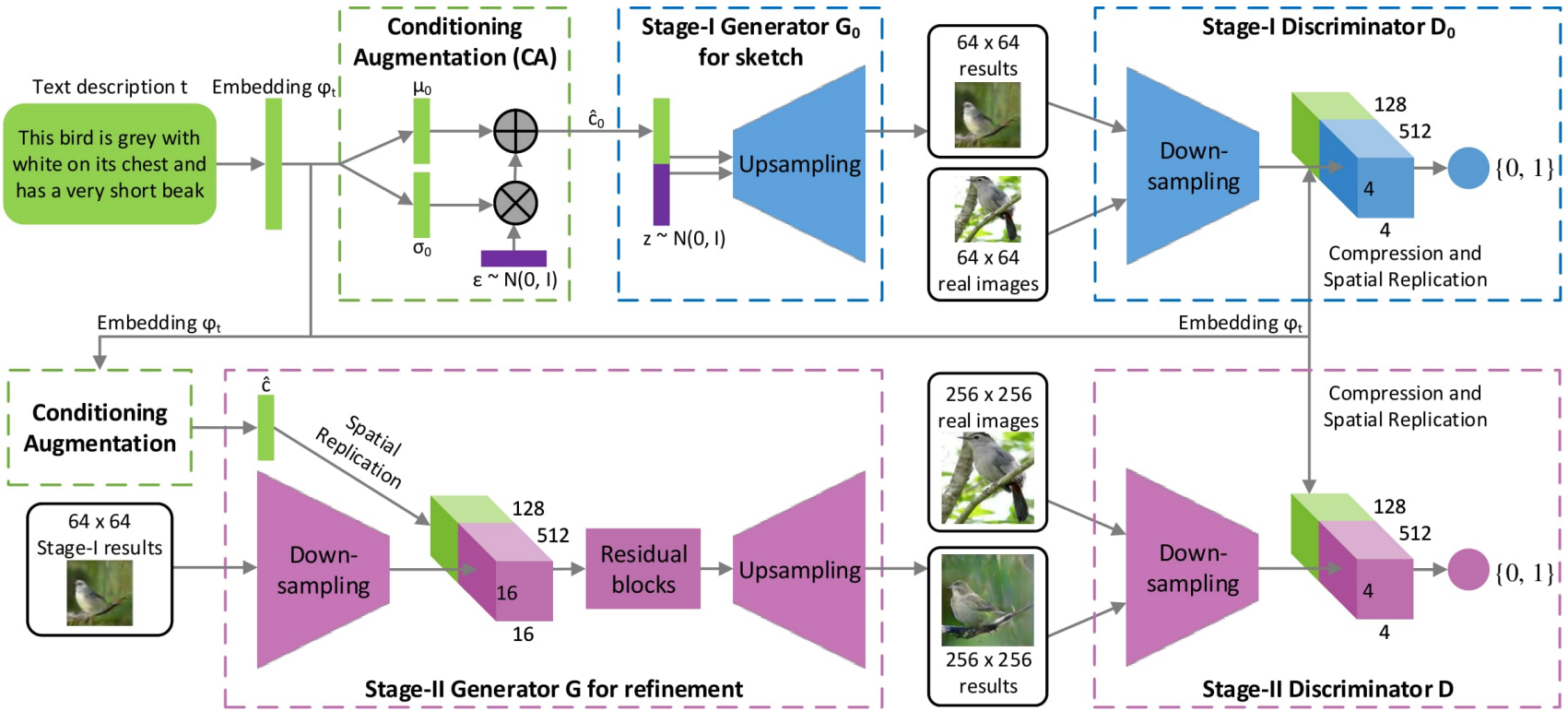
Tags: leaf, bouquet, flowers, stem, table, rose, flower, leaves, vase, plant

Fig. 5: Examples of image captioning. Objects are colored, based on their appearance against the groud-truth (GT): all , OSCAR & tags , tags only .

Image credit: Xiujun Li *et al.* 2020

StackGAN[17]

Carnegie Mellon



- Two-phase generation, stage 2 is conditioned on stage 1
- Regularization for text embedding on G: $\mathbf{D}_{KL}\left(\mathbf{N}(\mu_0, \sigma_0) \| \mathbf{N}(\mathbf{0}, \mathbf{I})\right)$

[17] https://arxiv.org/pdf/1612.03242.pdf

Image credit: Han Zhang *et al.* 2017

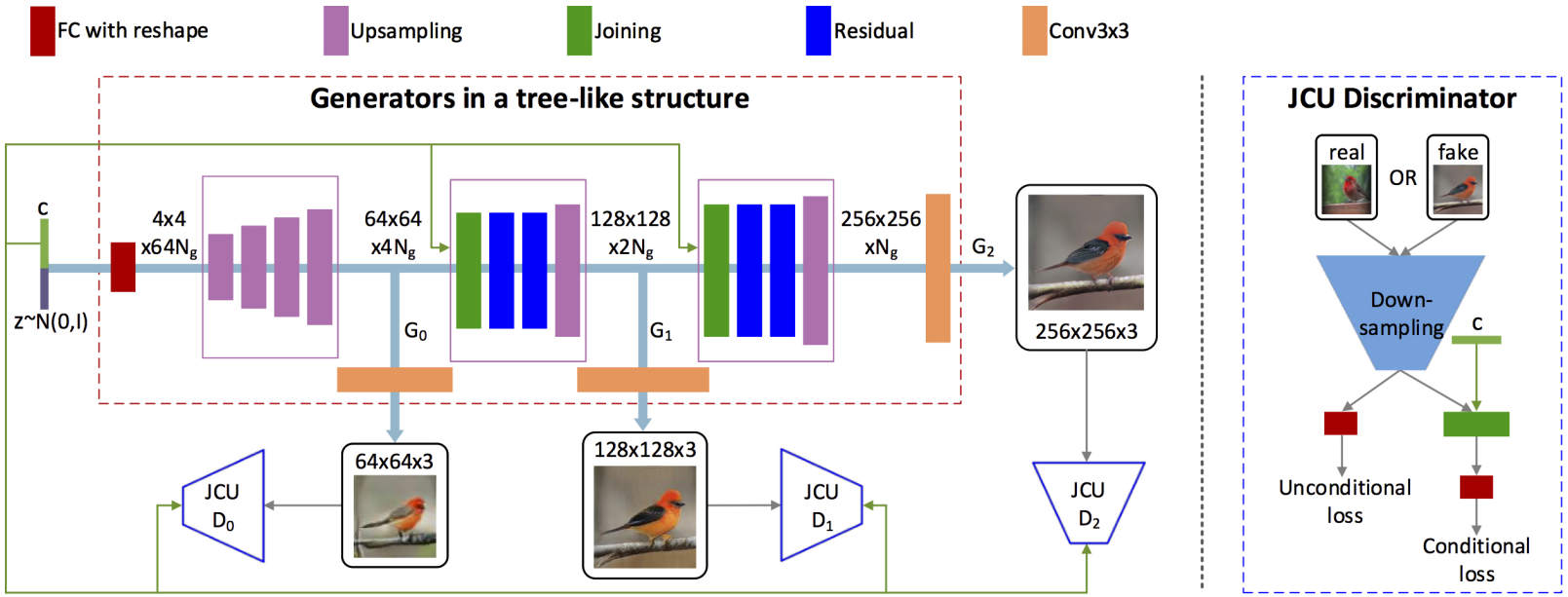
StackGAN: Results

Carnegie Mellon

Image credit: Han Zhang *et al.* 2017

StackGAN++[18]

Carnegie Mellon



- Instead of having 2 phases, now model multi-scale resolutions
- Multiple Gs and Ds with tree-like structure
  - G: $h_0 = F_0(c, z)$ and $h_i = F_i(h_{i-1}, c)$ and outputs $s_i = G_i(h_i)$
  - D_i: distinguishes s_i with real image x_i

[18] https://arxiv.org/pdf/1710.10916.pdf

Image credit: Han Zhang *et al.* 2018

StackGAN++: Results

Carnegie Mellon

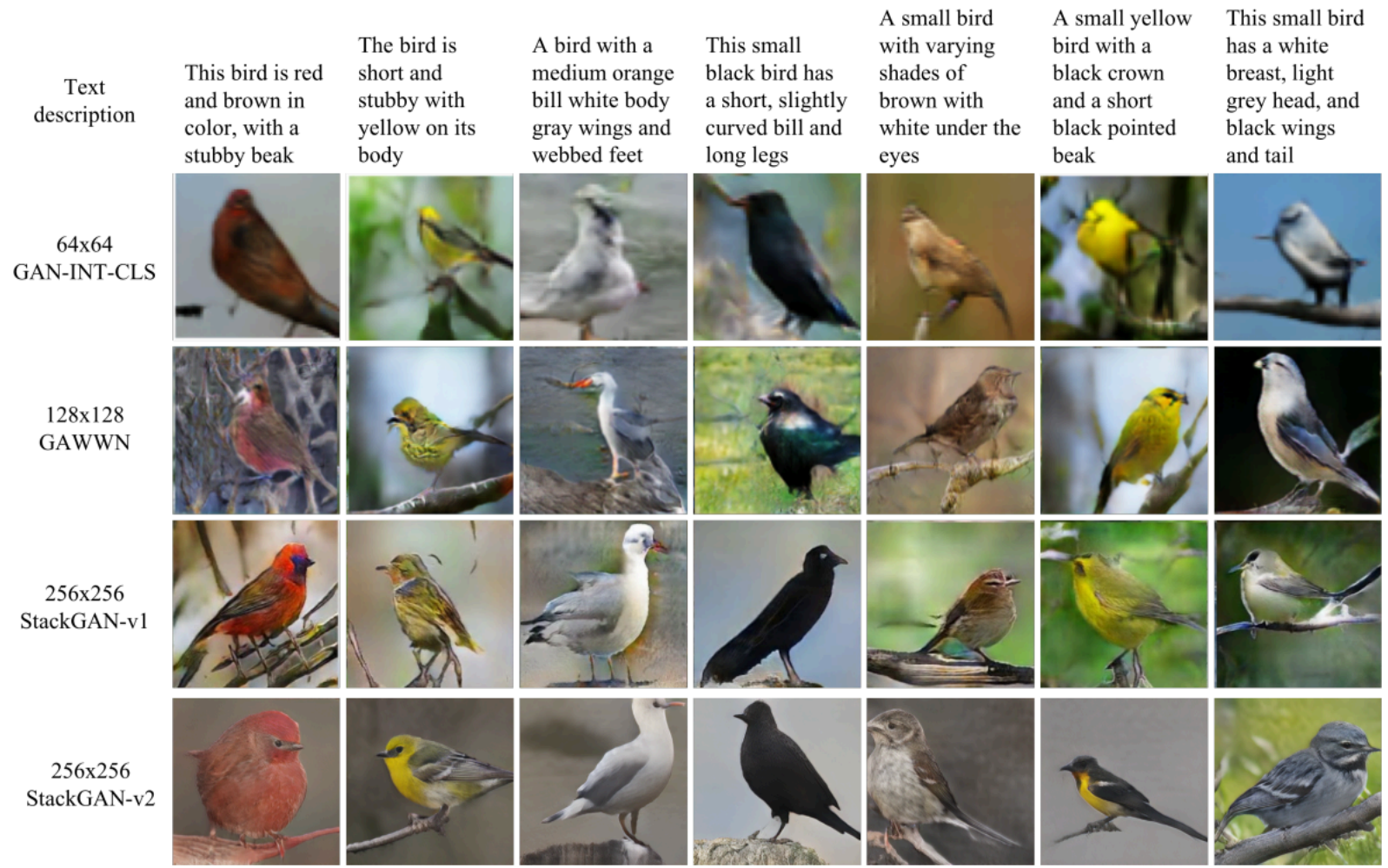
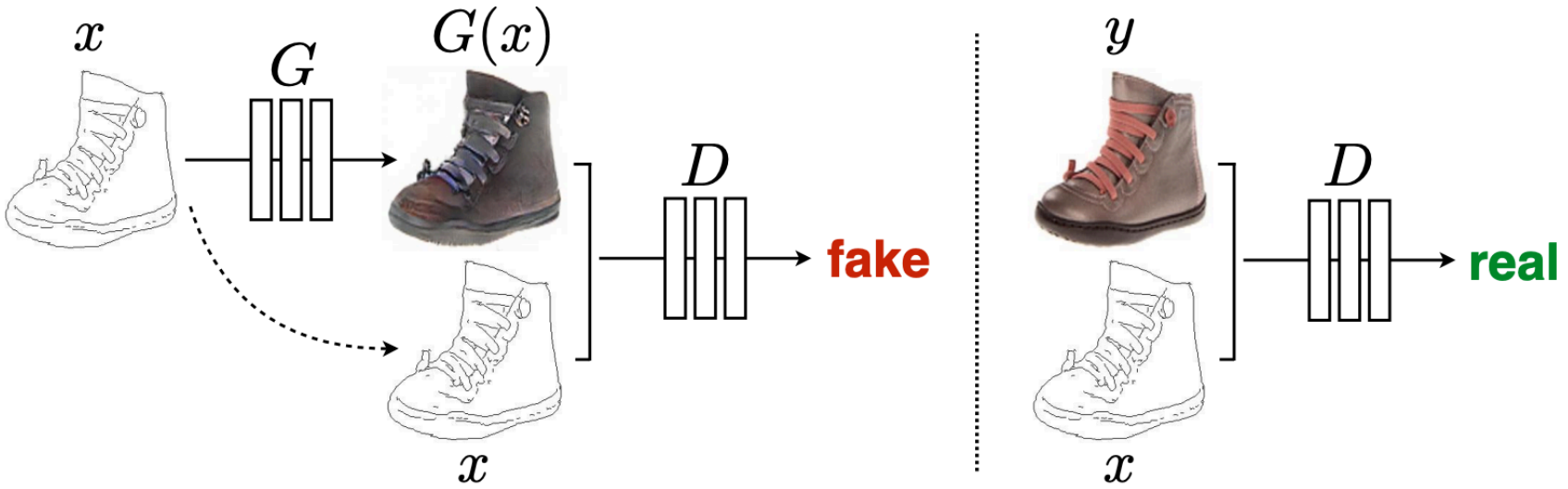


Fig. 3: Example results by our StackGANs, GAWWN [33], and GAN-INT-CLS [35] conditioned on text descriptions from CUB test set.

Image credit: Han Zhang *et al.* 2018

Pix2Pix[19]



- Use CGAN to transfer from edges to photos
- So the edge maps are the labels fed to both D and G

[19] https://arxiv.org/pdf/1611.07004.pdf

Image credit: Phillip Isola *et al.* 2016

Pix2Pix: Results

Carnegie Mellon

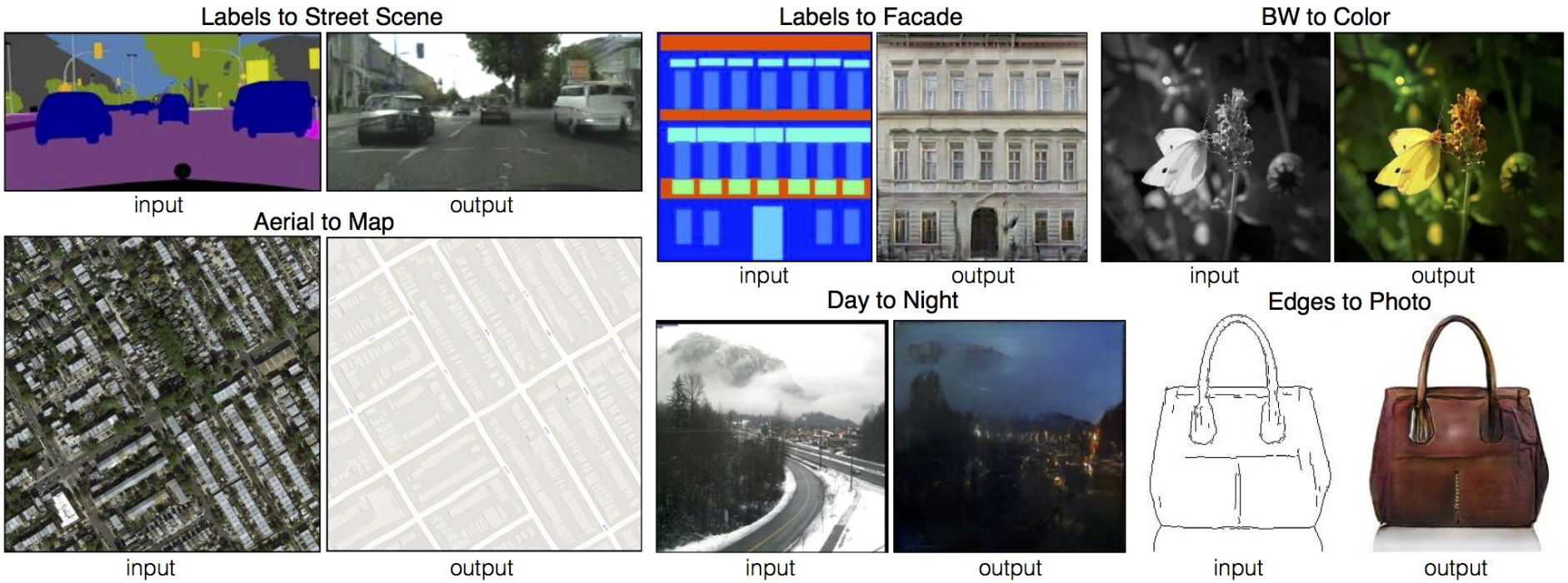


Image credit: Phillip Isola *et al.* 2016

Semantically Multimodal Image Synthesis (SMIS[20])

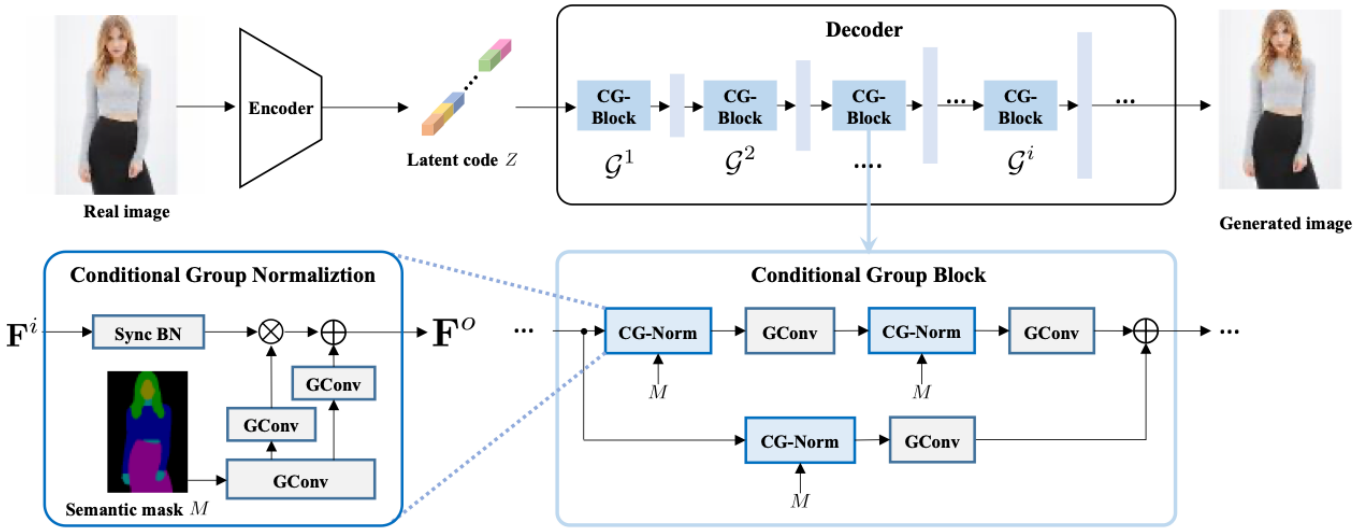


Figure 2: Architecture of our generator (GroupDNet). "GConv" means group convolution and "Sync BN" represents synchronized batch normalization. $\mathcal{G}^i$ is the group number of i-th layer. Note normally $\mathcal{G}^i \geq \mathcal{G}^{i+1}$ for $i \geq 1$ for GroupDNet.

- Encoder:
  - Class-specific inputs: $X_c = M_c \circ X$ and latent output Z_c
  - Concat them before feeding to Encoder: $S = \text{concat}_c X_c$
- Decoder generates conditioned on classes

[20] https://arxiv.org/pdf/2003.12697.pdf

Image credit: Zhen Zhu *et al.* 2020

SMIS: Tasks

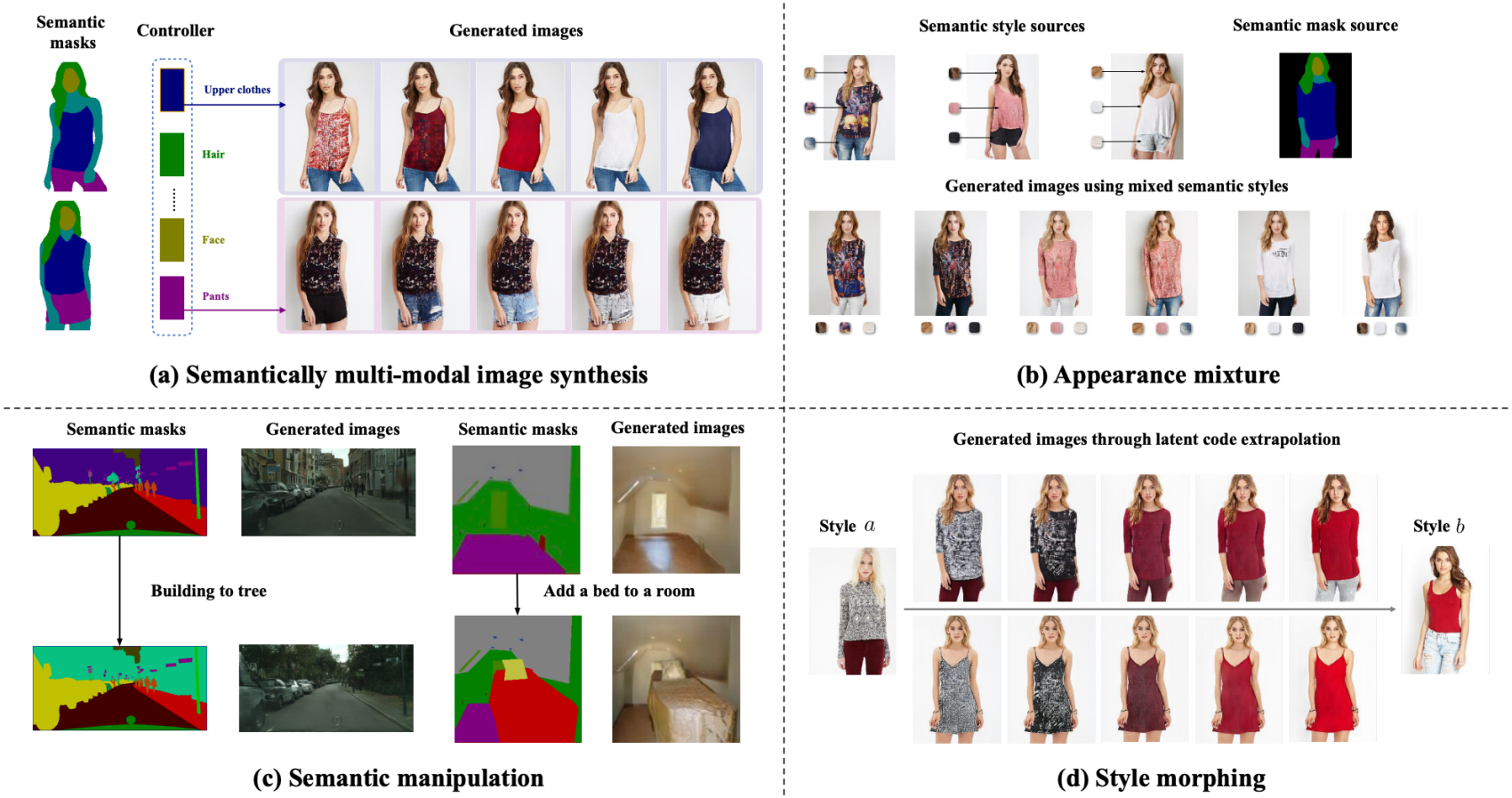


Figure 6: Exemplar applications of the proposed method. (a) Demonstration of the semantically multi-modal image synthesis (SMIS) task. (b) Application of our SMIS model in appearance mixture. Our model extracts styles of different semantic classes from different sources and generates a mixed image by combining these semantic styles with the given semantic mask. (c) Application of our SMIS model in semantic manipulation. (d) Application of our SMIS model in image extrapolation. **Zoom in** for better details.

- Applicable for diverse tasks

Image credit: Zhen Zhu *et al.* 2020

SMIS (GroupDNet): Comparison

Carnegie Mellon

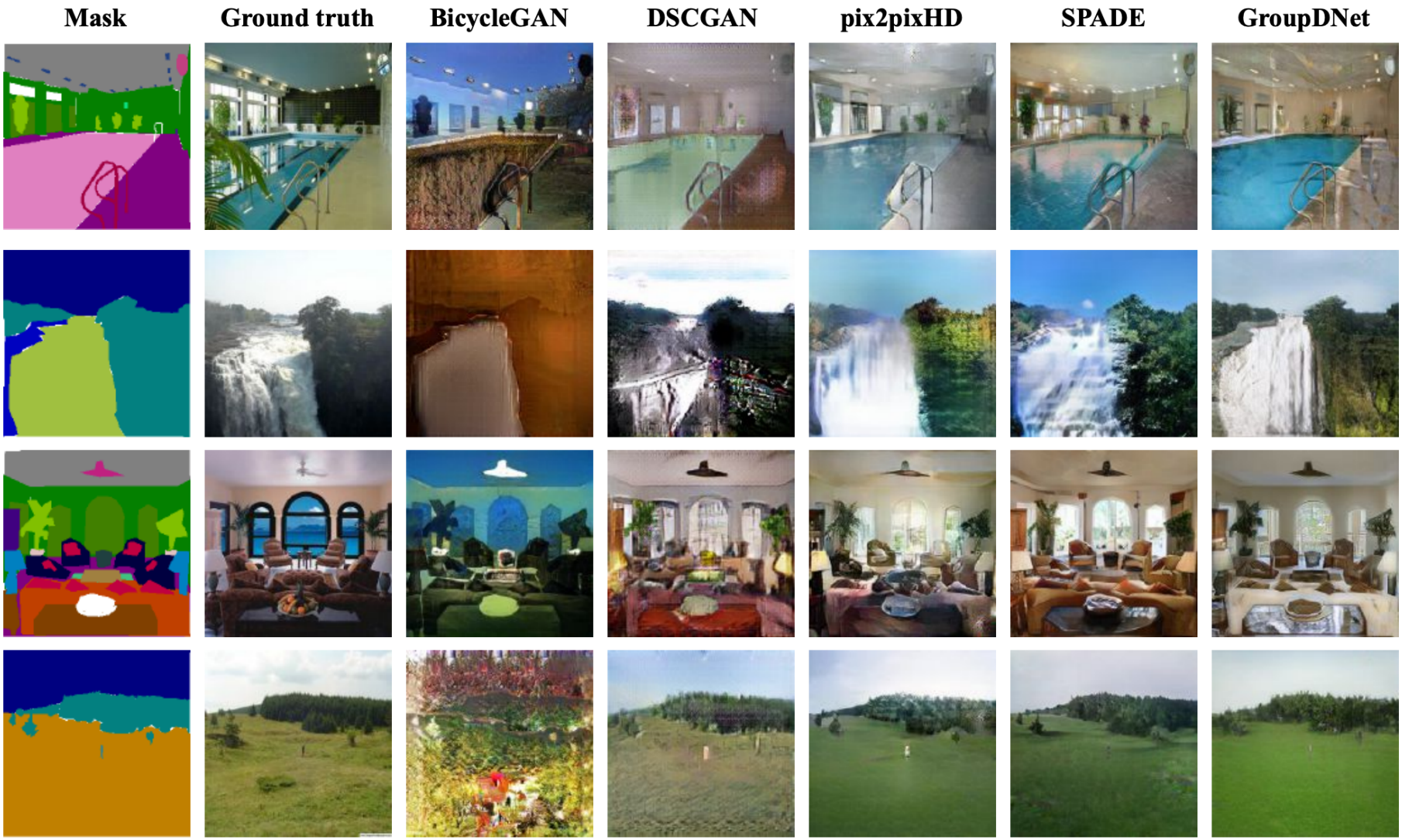


Figure 14: Qualitative comparison of our model with several label-to-image methods on the ADE20K dataset.

Image credit: Zhen Zhu *et al.* 2020