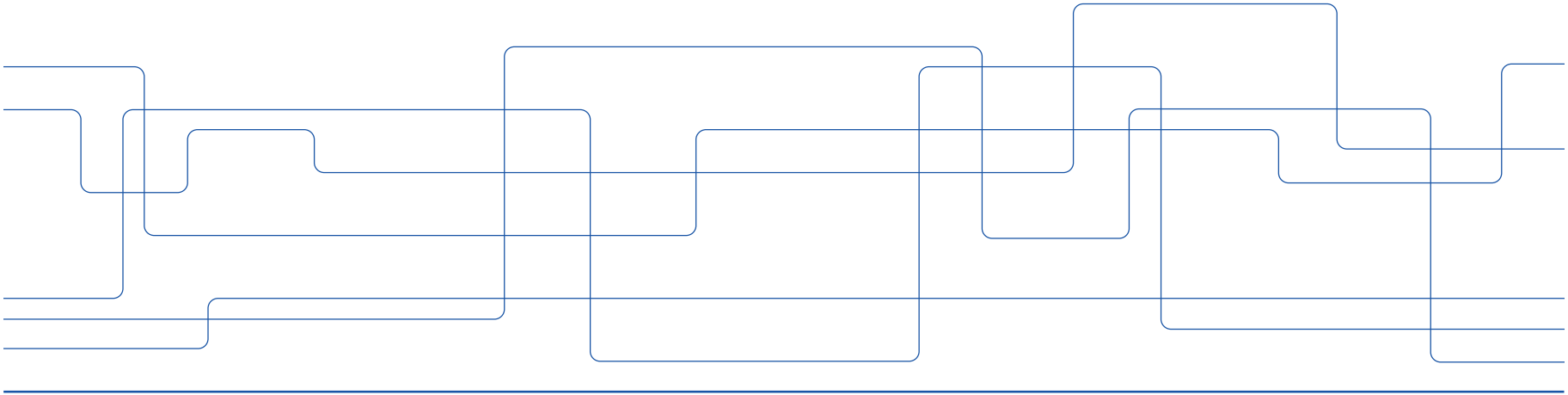
KTH ROYAL INSTITUTE OF TECHNOLOGY

Imagine this! Scripts to Compositions to Videos

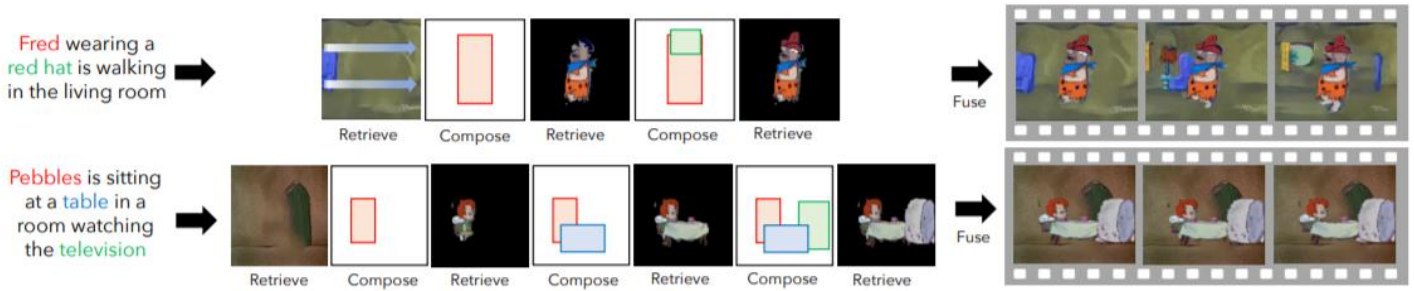
T. Gupta, D. Schwenk, A. Farhadi, D. Hoeim, A. Kembhavi.

Published at ECCV 2018.



Introduction

- Semantic Scene Generation
  - Generating scene videos from natural language descriptions
  - Requires ”Jointly modeling layout and appearance of entities…”
- https://www.youtube.com/watch?v=688Vv86n0z8&feature=youtu.be
- Generation process:

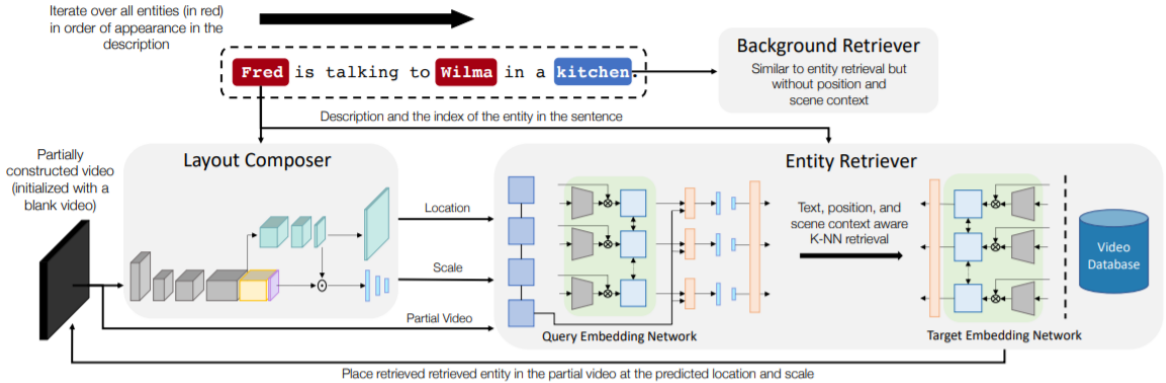


Challenges with SSG Models

- Entity Recall
  - Video must contain relevant characters, objects and background
- Layout Feasibility
  - Characters and objects must be placed in plausible locations and scales
- Appearance Fidelity
  - Entity appearance should respect the scene description
- Interaction Consistency
  - Appearance of characters and objects must be consistent with each other given the described interaction
- Language Understanding
  - Must be able to understand and translate descriptions into plausible visual instantiations

Contributions

- **C**omposition **R**etrieval **a**nd **F**usion Ne**t**work (CRAFT)



- The Flintstones Dataset
  - Densely annotated dataset based on *The Flintstones* animated series
- Outperforms standard generative model approaches on proposed metrics
  - *Visual Quality* and *Composition Consistency*

Related Work – GANs

- Generative Adversarial Networks (GANs) by I. Goodfellow et al. (2014).

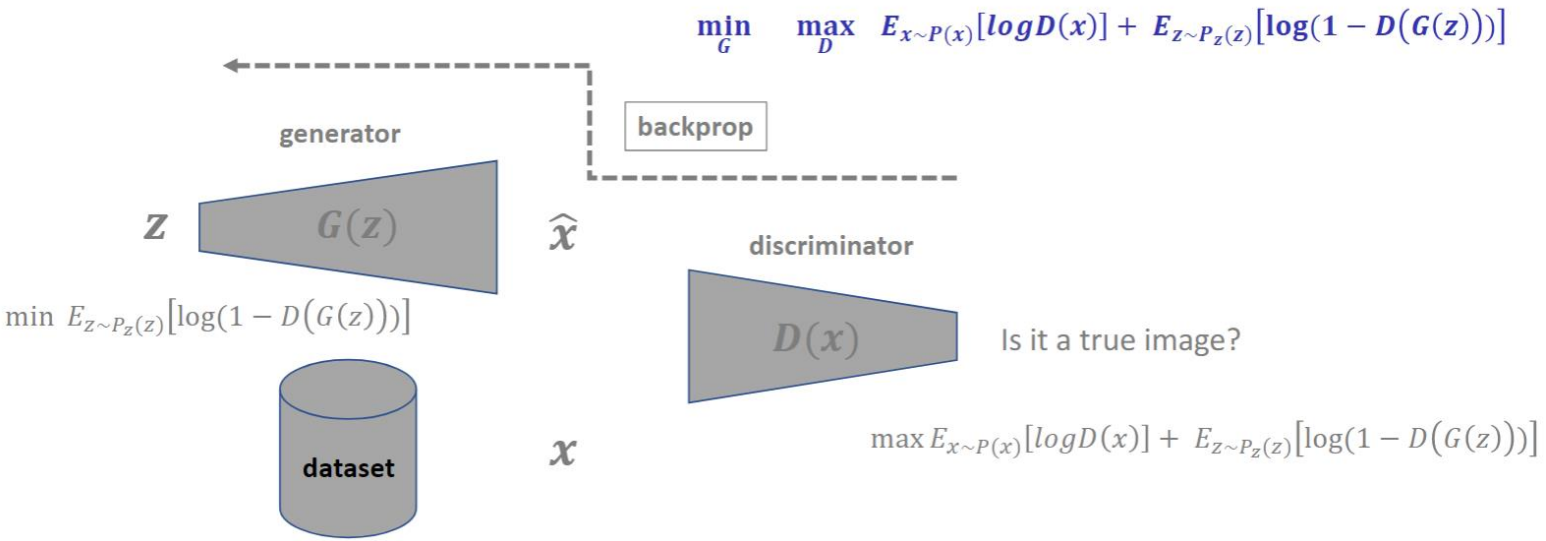


Figure courtesy: Slides ”Generative Models” by Hossein Azizpour in course ”Advanced Deep Learning” (FDD3412).

Related Work – High quality images by GANs

- “Progressive Growing of GANs for Improved Quality, Stability, and Variation” by T. Karras et al, ICLR 2018.
- Generation of high-resolution images with generator architecture using skip-connections.

- “Conditional Image Synthesis with Auxiliary Classifier GANs” by A. Odena et al, ICML 2017.
- Class-conditional GAN for high-resolution images. “High-resolution images provides class information not present in low resolution images.”

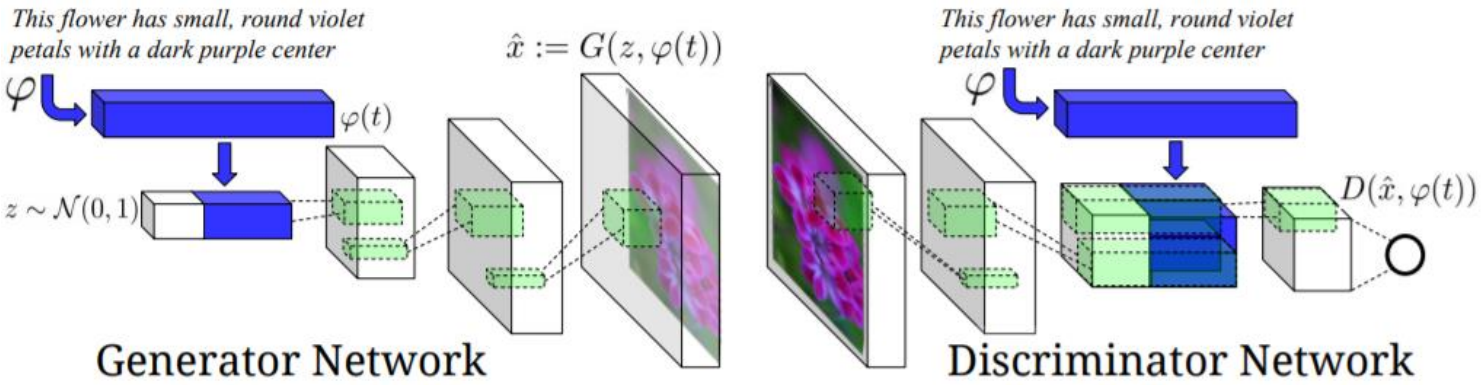
goldfinch

daisy

redshank

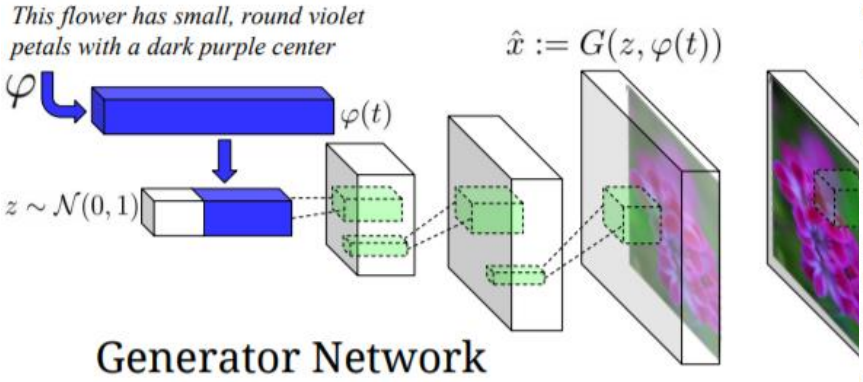
Related Work – Text to Image GANs

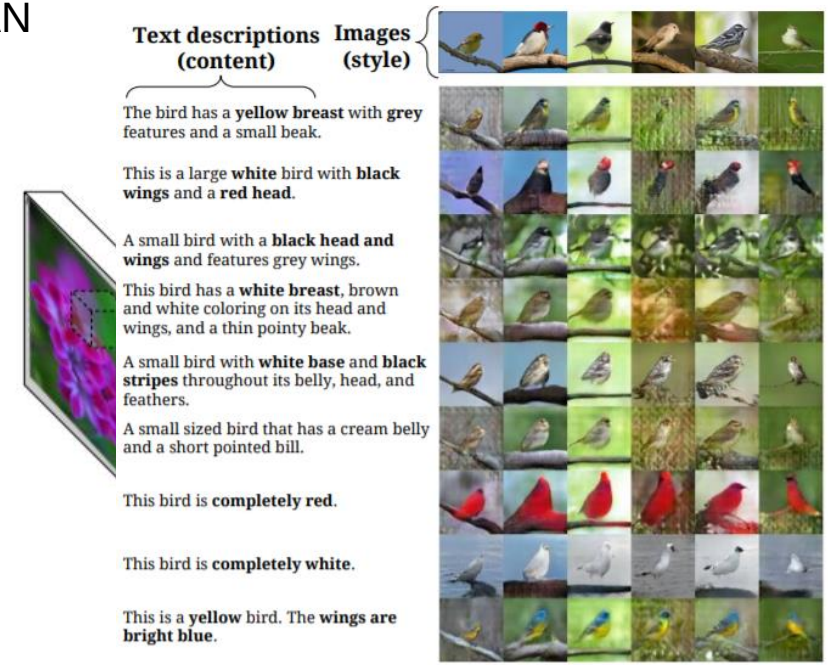
- “Generative Adversarial Text to Image Synthesis” by S. Reed et al. ICML 2016.
- Text-conditional convolutional GAN



Related Work – Text to Image GANs

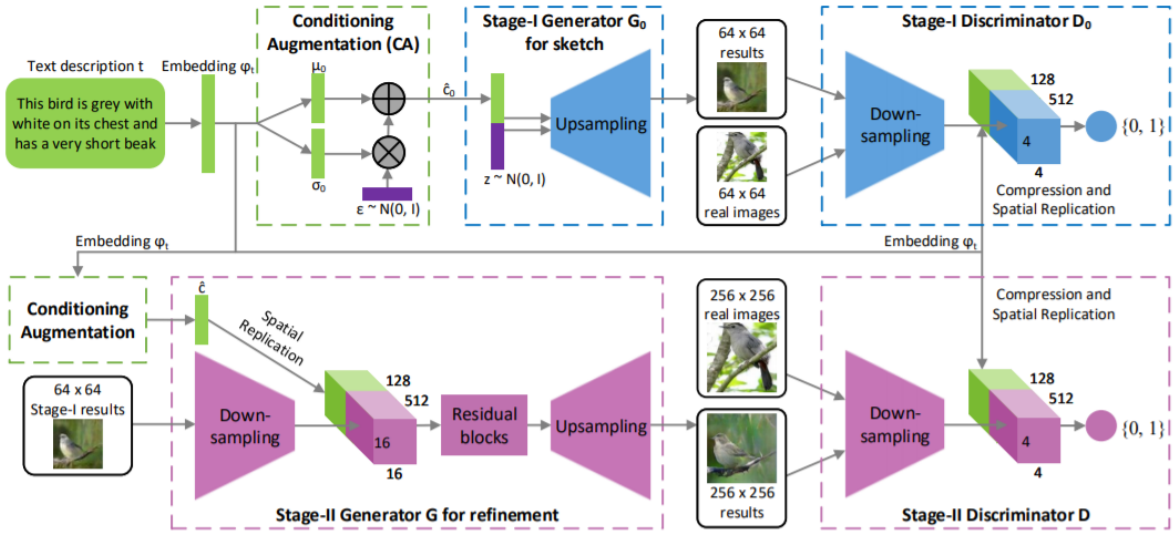
- “Generative Adversarial Text to Image Synthesis” by S. Reed et al. ICML 2016.
- Text-conditional convolutional GAN

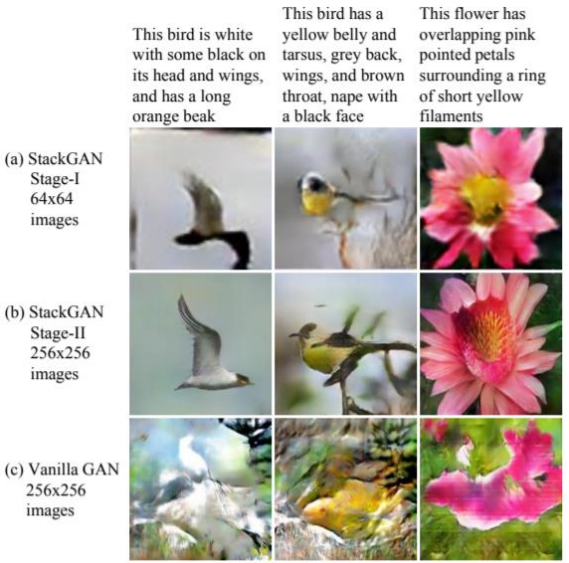




Related Work – Text to Image GANs

- "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks" by H. Zhang et al. ICCV 2017.
- Generate object shape + background then the details by stacking two GANs

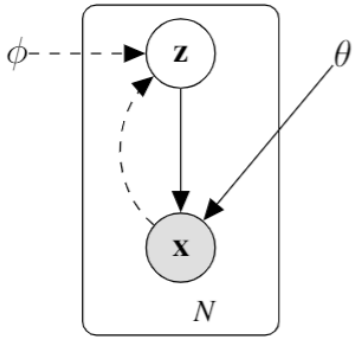




Related Work – Variational Autoencoders

- “Auto-Encoding Variational Bayes” by D. Kingma and M. Welling, ICLR 2014.
- Tractable inference by optimizing the ELBO.

$$\log p_\theta(\mathbf{x}) \geq E_{z\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]$$



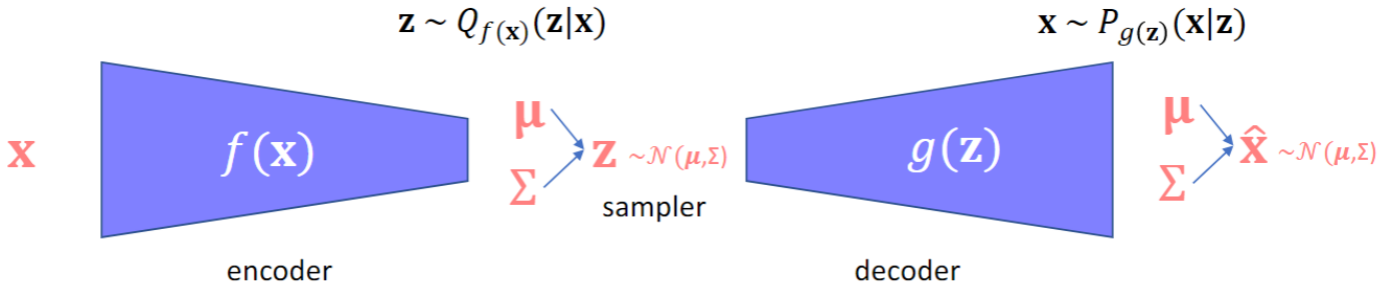
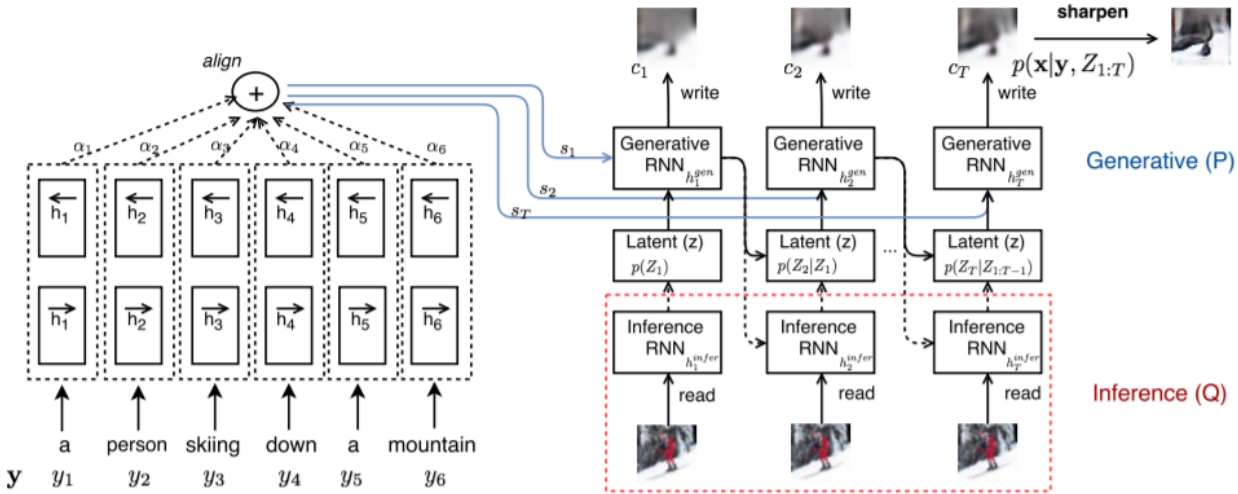


Figure courtesy: Slides ”Generative Models” by Hossein Azizpour in course ”Advanced Deep Learning” (FDD3412).

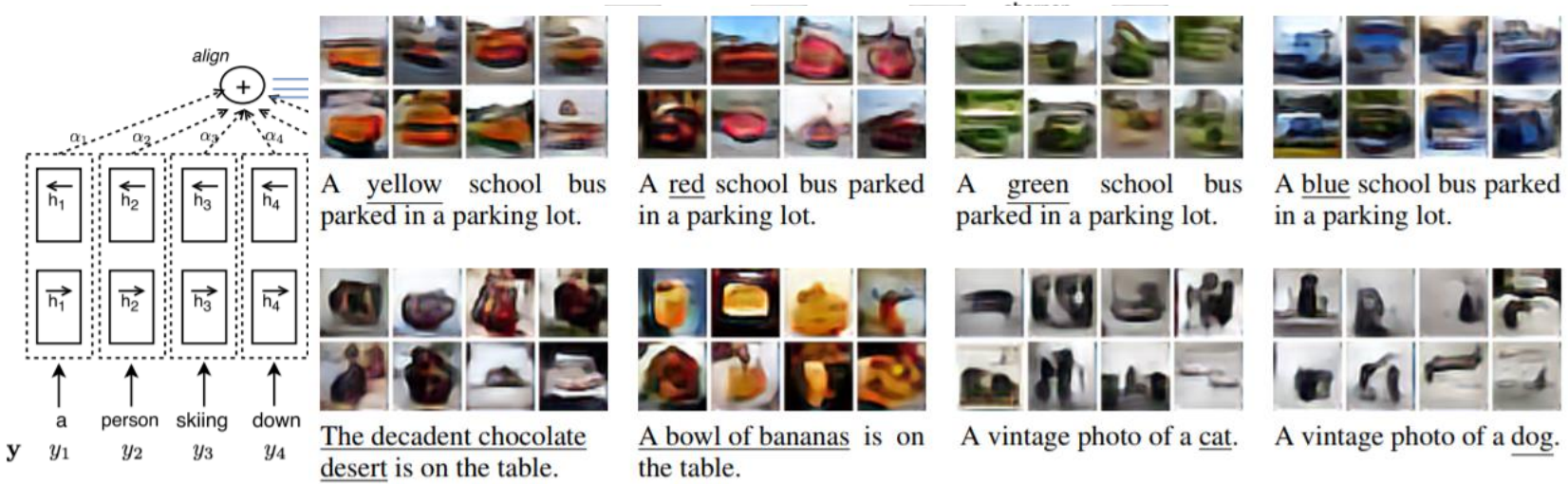
Related Work – Text to Image VAE

- “Generating Images from Captions with Attention” by E. Mansimov et al. ICLR 2016.



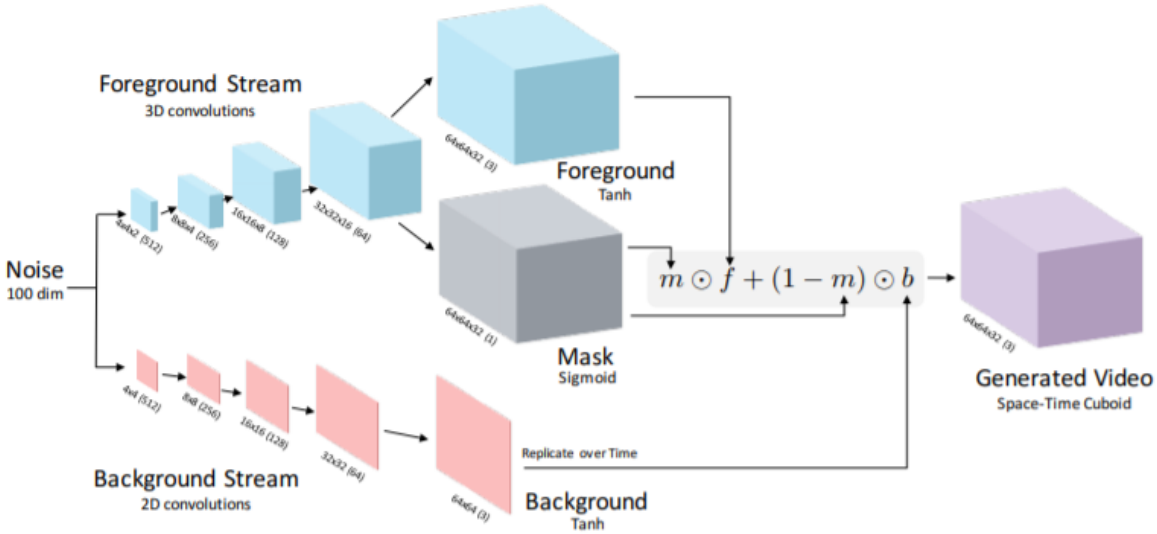
Related Work – Text to Image VAE

- “Generating Images from Captions with Attention” by E. Mansimov et al. ICLR 2016.



Related Work – Video Generation

- “Generating Videos with Scene Dynamics” by C. Vondrinck et al. NIPS 2016.
- Videos: https://github.com/cvondrick/videogan



Related Work – Video Generation

- “Attentive Semantic Video Generation using Captions” by T. Marwah et al. ICCV 2017.
- Videos: https://github.com/Singularity42/cap2vid

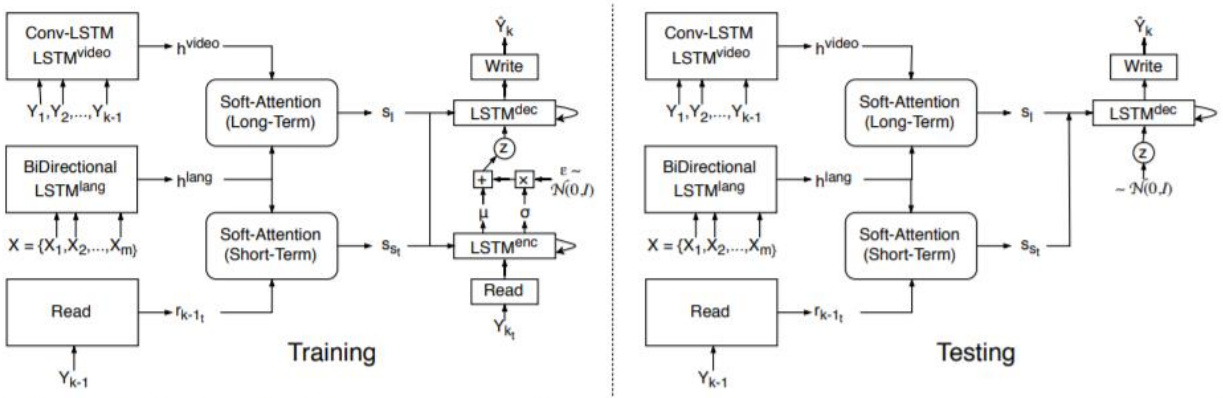


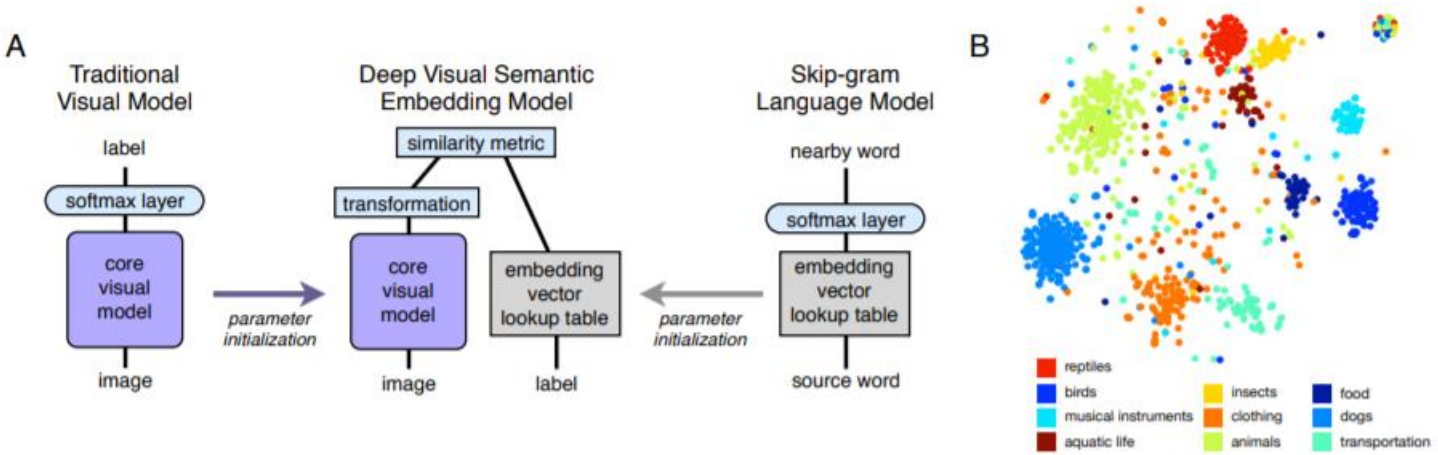
Figure 2. Proposed network architecture for attentive semantic video generation with captions. $Y = \{Y_1, Y_2, \cdots, Y_k\}$ denotes the video frames generated by the architecture, while $X = \{X_1, X_2, \cdots, X_m\}$ denotes the set of words in the caption.

Related Work – Other Video Generation Works

- “Stochastic Video Generation with a Learned Prior” by E. Denton and R. Fergus et al. ICML 2018.
- “Stochastic Variational Video Prediction” by M. Babaeizadeh et al. ICLR 2018.
- “Stochastic Adversarial Video Prediction” by A. Lee et al. arxiv preprint 2018.
- Videos: https://alexlee-gk.github.io/video_prediction/

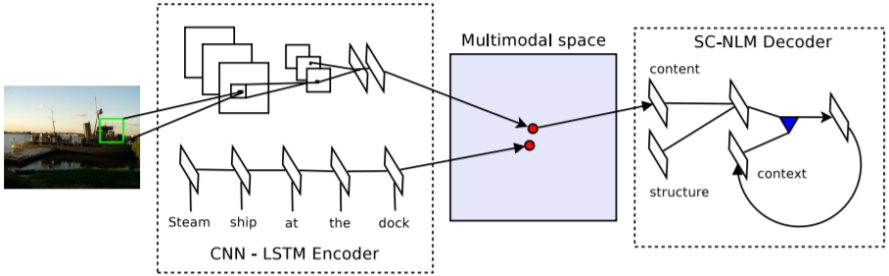
Related Work – Visual-Semantic Embeddings

- “DeViSE: A Deep Visual-Semantic Embedding Model” by A. Frome et al. NIPS 2013.

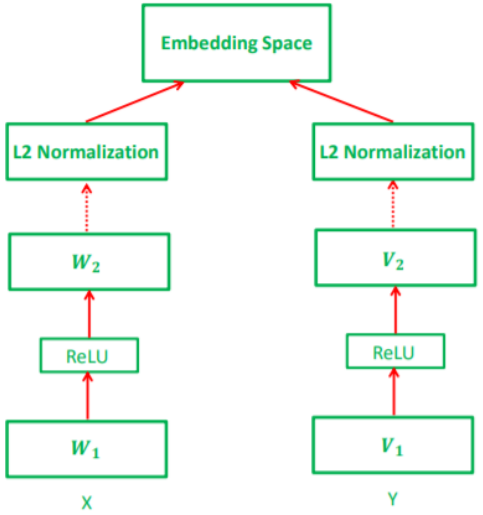


Related Work – Visual-Semantic Embeddings

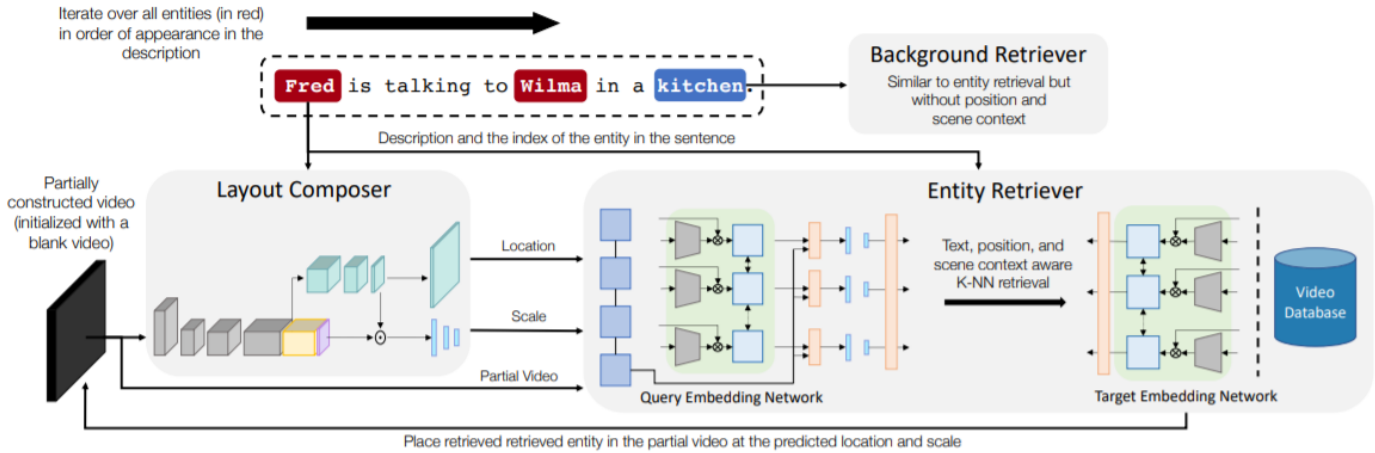
- “Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models” by R. Kiros et al, NIPS 2014 Workshop.



- “Learning Deep Structure-Preserving Image-Text Embeddings” by L. Wang et al, CVPR 2016.



Model - CRAFT

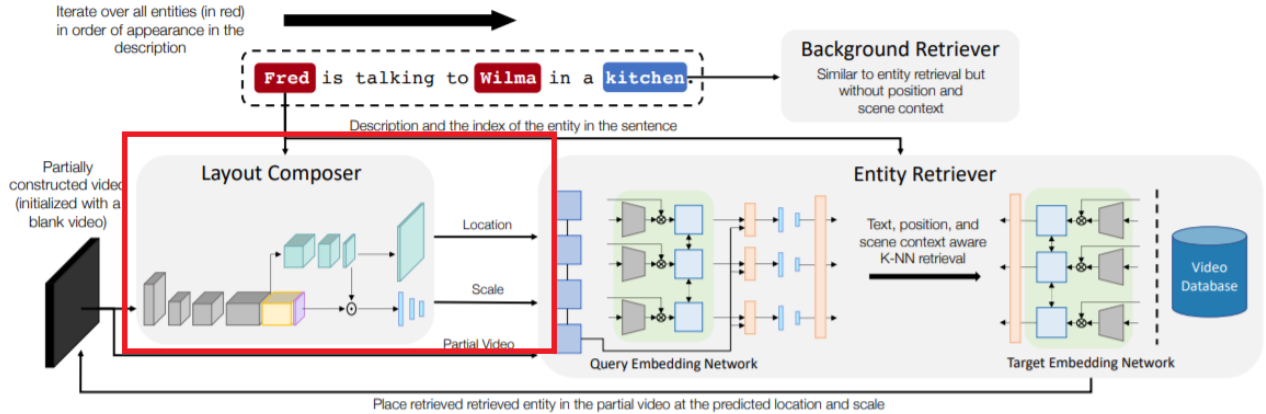


- Each part is a NN trained independently using ground truth supervision
- At prediction time:
  - Begin with empty video and add entities based on order of appearance in description

Model - Notation

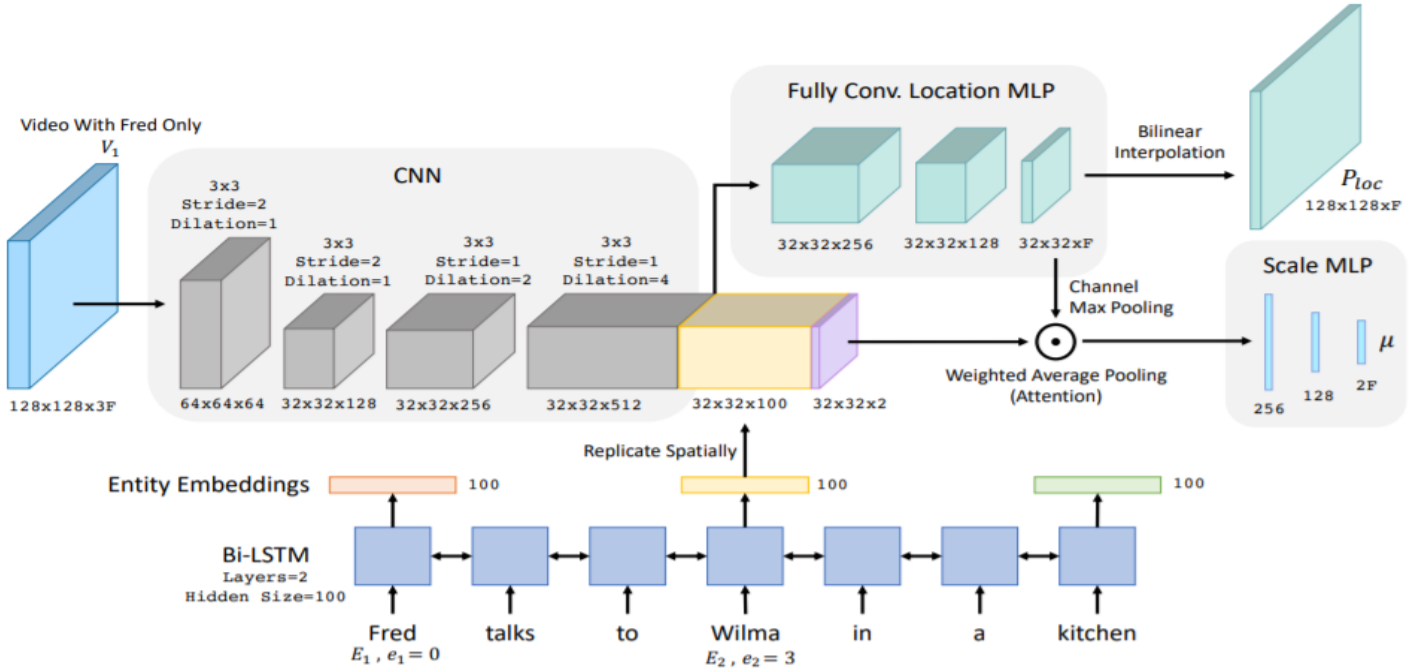
Caption	
T	Caption with length $\|T\|$
$\{E_i\}_{i=1}^{n}$	n entities in T in order of appearance
$\{e_i\}_{i=1}^{n}$	entity noun positions in T
Video	
F	number of frames in a video
$\{(l_i, s_i)\}_{i=1}^{n}$	position of entities in the video
l_i	entity bounding box at each frame ($\{(x_{if}, y_{if}, w_{if}, h_{if})\}_{f=1}^{F}$)
s_i	entity pixel segmentation mask at each frame
V_{i-1}	partially constructed video with entities $\{E_j\}_{j=1}^{i-1}$
$V\ (= V_n)$	full video containing all entities
$\{(V^{[m]}, T^{[m]})\}_{m=1}^{M}$	training data points, where M = number of data points

Model – Layout Composer



- Generates plausible layout of the scene
  - Predicts locations and scales of each character and object
- Needs to include spatial knowledge
  - A person being talked to faces the person speaking
  - A couch goes under a person sitting on it

Model – Layout Composer

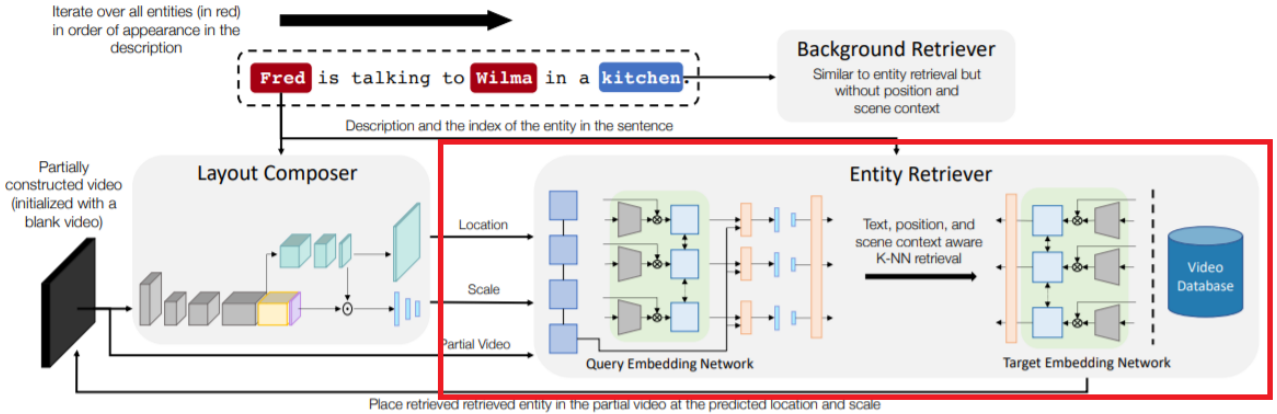


- Goal: Predict location and scale of characters and objects from partially created video

Model – Layout Composer Summary

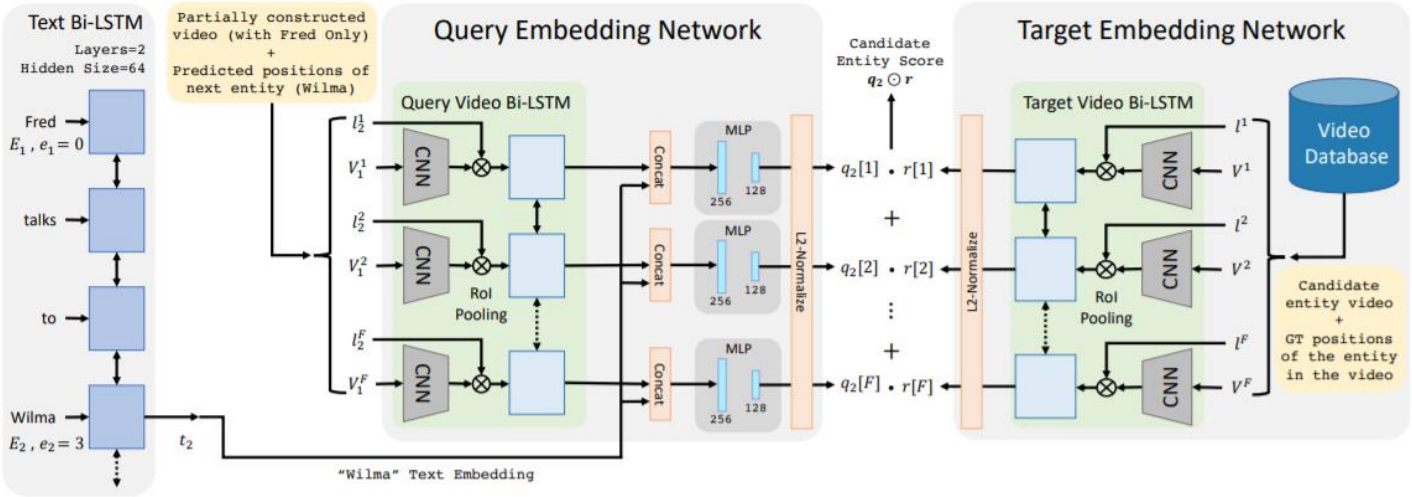
- Feature computation backbone
  - CNN processing the partially constructed video
  - Bi-LSTM encoding the description up to the entity to be generated
  - 2-D grid coordinates. What is the purpose?!
- Location predictor
  - MLP with softmax output to produce scores for each location
- Scale predictor
  - MLP producing mean of scale for each position (covariance is constant)
- Feature sharing and multitask training
  - Location network produces F probability maps in a single forward-pass
  - Max-pooling over the F spatial maps from location network to get attention map. Combine attention map together with feature backbone using weighted average pooling and feed into scale MLP.

Model – Entity Retriever



- Find spatio-temporal patch within a database
  - Must match the description and be consistent with video constructed so far
- Consists of two neural networks: Query and Target Embedding Network
- Needs to respect implicit relational constraints and context
  - E.g. *Fred is talking to Wilma*, Wilma has to face Fred in the correct direction

Model – Entity Retriever



- Goal: Retrieve relevant entities by comparing embeddings of partially constructed video and description with embeddings from a video database

Model – Entity Retriever Losses

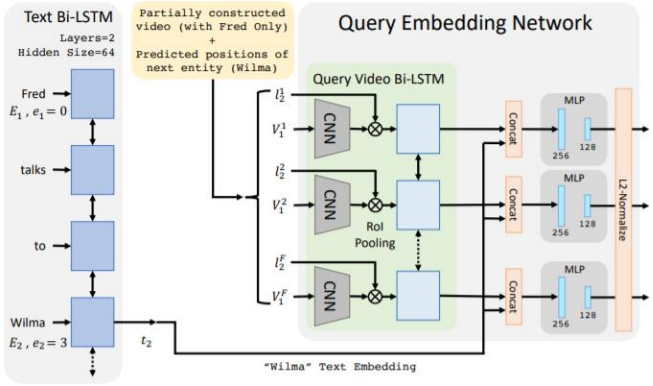
- Loss function for matching the embeddings
  - Use true video and entity positions and match them with partially generated video
  - Embedding score from same video should be larger than embedding scores between all other videos in mini-batch

$$\mathcal{L}_{triplet} = \frac{1}{B \cdot (B-1)} \sum_{b=1}^{B} \sum_{b^- \in \delta_b} \Big[\max(0, \gamma + q_{i_b}^{[m_b]} \odot r_{i_{b^-}}^{[m_{b^-}]} - q_{i_b}^{[m_b]} \odot r_{i_b}^{[m_b]}) + \max(0, \gamma + q_{i_{b^-}}^{[m_{b^-}]} \odot r_{i_b}^{[m_b]} - q_{i_b}^{[m_b]} \odot r_{i_b}^{[m_b]}) \Big]$$

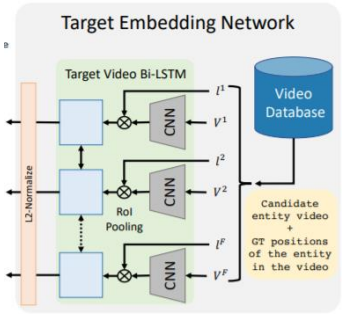
- Auxiliary multi-label classification loss
  - Use multi-label classifier to predict nouns, adjectives and verbs given the entity embedding
  - Uses a constructed vocabulary of all nouns, adjectives and verbs in the training data
  - Prediction is performed with both query and target embeddings

Model – Entity Retriever Networks

- Query embedding network
  - RoIAlign to crop out per-frame feature maps using predicted bounding box from Layout Composer
  - LSTM over averaged RoIAlign features to capture temporal context

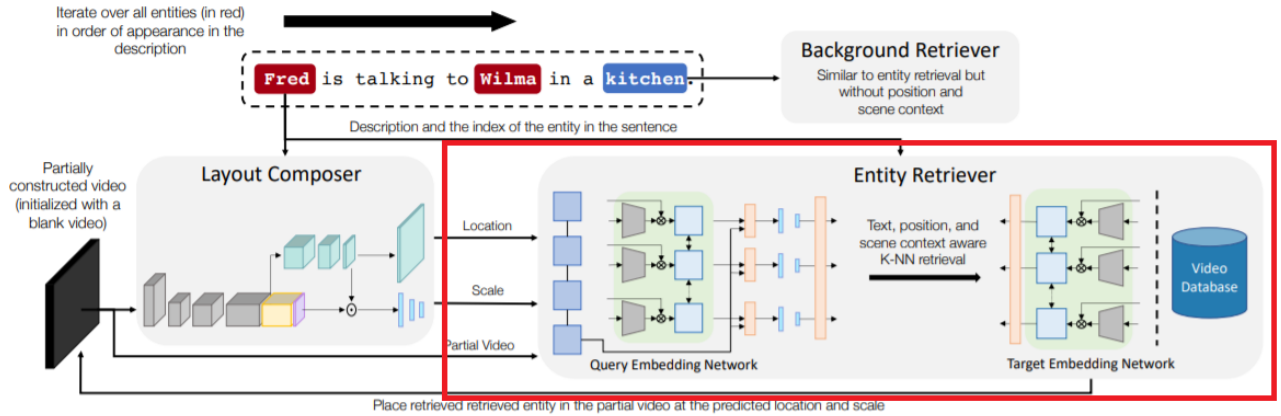


- Target embedding network
  - No text annotations
  - No concatenation of 2-D coordinate features and per-frame video features from CNN



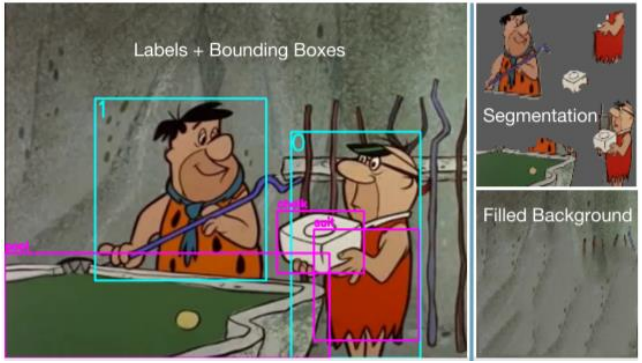
Model – Entity Retriever Question

- How do we retrieve the image patches at test time? I think it's like
  - Predict query embeddings as during training
  - Target embeddings are the same during test
  - K-NN retrieval in the embedding space?



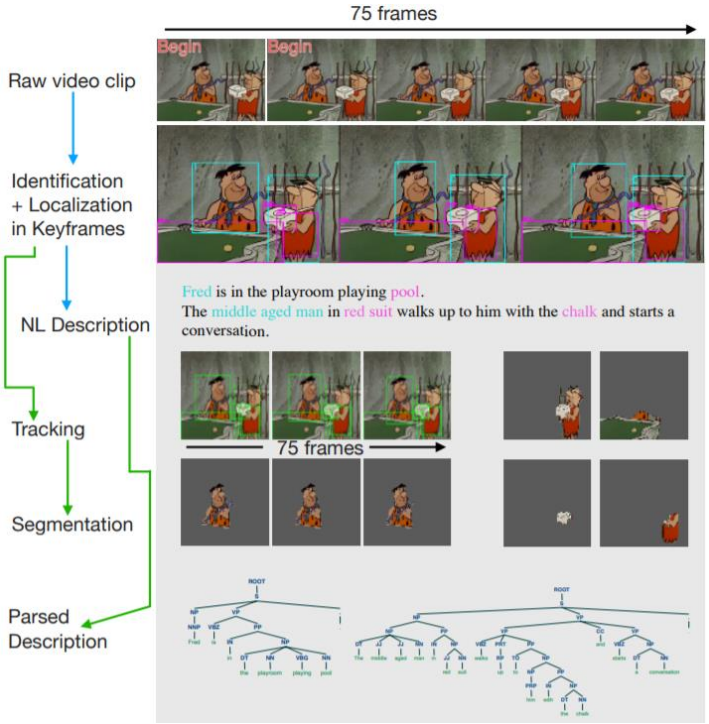
Model – Background Retriever

- First construct a database with backgrounds
  - Remove characters from all videos
  - Hole filling with PatchMatch
- Query embedding network
  - Pass description through text Bi-LSTM and use output at background word location
- Target embedding network
  - Pass video through video Bi-LSTM
- No conditioning on position or entities in scene



Flintstones Dataset

- Clips are 3 seconds = 75 frames long
- Annotation process using crowd-sourcing
  - Identify and localize characters with labelling and bounding boxes
  - Provide 1-2 word of the clip's setting (e.g. living room, park)
  - Provide 1-4 sentence descriptions of clips
  - Annotate important objects and give them bounding boxes
- Segmentation masks of characters and objects
  - Segmentation using Simple Linar Iterative Clustering (SLIC) and GrabCut
  - Retrieve scene backgrounds in this stage as well



Experiments - Evaluation

- Different metrics for each model component
- Ablation studies
- Human evaluations
- Modelling pixels distributions vs Entity retrieval

Experiments – Layout Composer Evaluation

- Negative log-likelihood of GT entity positions under predicted distribution
  - Captures both location and scale
- Averaged normalized pixel distance of GT from the predicted entity locations
  - Only measures location accuracy
- “Feature backbone” ablation study

Text	Scene Context	2D Coord. Grid	Dil. Conv	NLL	Pixel Dist.
Uniform Distribution				>9.704	>0.382
✗	✓	✓	✓	9.845	0.180
✓	✗	✓	✓	8.167	0.185
✓	✓	✗	✓	8.250	0.287
✓	✓	✓	✗	7.780	0.156
✓	✓	✓	✓	**7.636**	**0.148**

Experiments – Entity Retriever Evaluation

- Measure noun, adjective and verb recall (@1 and @10) averaged across entities in test set
- Feature ablation with text, location and scene features

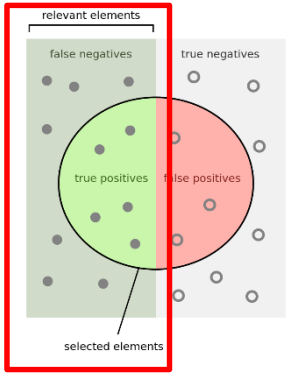


Figure courtesy: https://en.wikipedia.org/wiki/Precision_and_recall

Query Features			Recall@1			Recall@10		
Text	Context	Location	Noun	Adj.	Verb	Noun	Adj.	Verb
✗	✓	✓	24.88	3.04	9.48	55.22	19.39	37.18
✓	✗	✓	60.54	9.5	11.2	**77.71**	39.92	43.58
✓	✓	✗	56.14	8.56	11.34	73.03	39.35	41.48
✓	✓	✓	**61.19**	**12.36**	**14.77**	75.98	**47.72**	**46.86**

Experiments – Entity Retriever Evaluation

- Effect of auxiliary loss
  - Classifying nouns, adjectives and verbs given query and/or target embeddings

	Auxiliary Loss		Recall@1			Recall@10		
Triplet	Query	Target	Noun	Adj.	Verb	Noun	Adj.	Verb
✗	✓	✓	35.75	7.79	8.83	63.62	43.35	33.12
✓	✗	✓	51.68	3.8	8.66	67.86	25.28	39.46
✓	✓	✗	50.54	4.94	9.94	66.36	28.52	39.5
✓	✗	✗	48.59	3.04	9.34	65.64	20.15	37.95
✓	✓	✓	**61.19**	**12.36**	**14.77**	**75.98**	**47.72**	**46.86**

Experiments – Generalization to unseen videos

- Embedding approach can use any unseen video database
  - Transfer learning to new domains of videos
  - Requires a model that generalizes very well
- Compare using the seen training videos as target database vs unseen test videos and measure entity recall during test

Video Database	Recall@1			Recall@10		
	Noun	Adj.	Verb	Noun	Adj.	Verb
Seen (Train)	61.19	12.36	14.77	75.98	47.72	46.86
Unseen (Test)	50.52	11.98	10.4	69.1	41.25	42.57

Experiments – Human Evaluation

- Two metrics using 0-4 scale:
  - Compositional consistency of entities given descriptions
  - Overall visual quality independent of description
- Baseline: CRAFT model with generator network replacing the target embedding network in the entity retriever (Pixel Generation L1 in table below)

	Composition Consistency			Visual Quality		
	Position	Rel. Size	Interact.	FG	BG	Sharpness
Pixel Generation L1	0.69	0.65	0.55	0.96	1.44	1.07
Ours (GT Position)	1.69	1.69	1.34	1.49	1.65	**2.16**
Ours	**1.78**	**1.86**	**1.46**	**1.98**	**1.95**	1.82

Experiments – Results

- https://www.youtube.com/watch?v=688Vv86n0z8&feature=youtu.be

Discussion

- Future directions?
  - Model pixel distributions with GANs and VAEs or go with retrieval methods?
- Some parts are a bit poorly explained
  - Mention the requirements, but don’t elaborate on why their approach works
  - But they are humble about their results and show failure cases!
- Q1: 2-D grid coordinates in feature backbone part of Layout Composer
  - Provides ”spatial awareness” to features. What do they mean?
- Q2: How is the database used during test time in the Entity Retriever?
  - K-NN retrieval in embedding space?
- Q3: Are the nouns, adjectives and verbs in the descriptions annotated by crowd-sourcers?