视界无限
2020

Controllable and Informative Image Captioning

金琴
中国人民大学信息学院
11/08/2020

AI·M^3

视界无限
2020

Visual Activity Understanding my two cents

金琴
中国人民大学信息学院
11/08/2020

AI·M^3

Visual Activity Understanding（VAU）

- Different level (output)
  - Action Classification
  - Activity/Event Recognition
  - Relation Detection
  - Natural Description
- Different source (input)
  - Image
  - Video
- Different Domain
  - General (coarse)
  - Specific (fine-grained)
- Different Aspect
  - Objective
  - Subjective

Action Classification

Riding horse

Applying lipstick

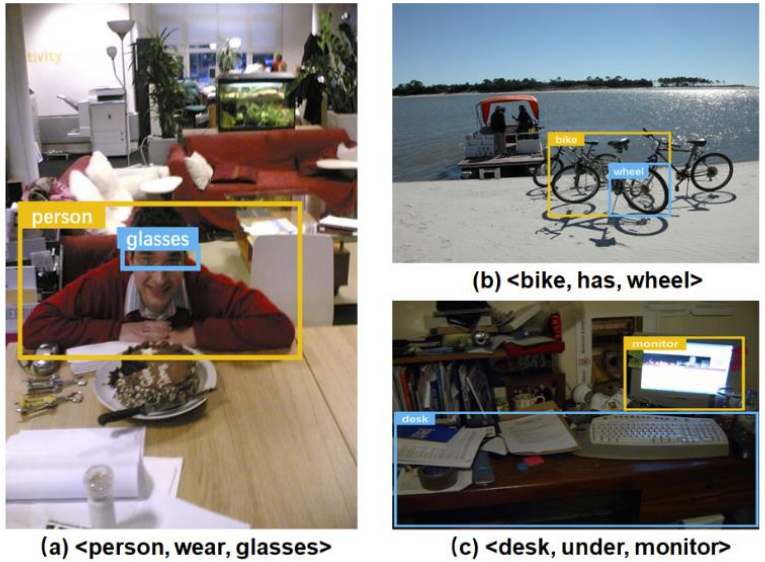
Playing Violin

Using Computer

Making a basketball dunk

- UCF101
- HMDB
- KINETICS
- ActivityNet
- AVA
- Moments-in-Time
- Something-something
- Youtube-8M
- Sports-1M
- TAPOS
- FineGym
- Figure Skating
- Aist dance video
- Let’s Dance
- MPII Cooking 2
- MERL Shopping

Visual Relation Detection



(a) <person, wear, glasses>

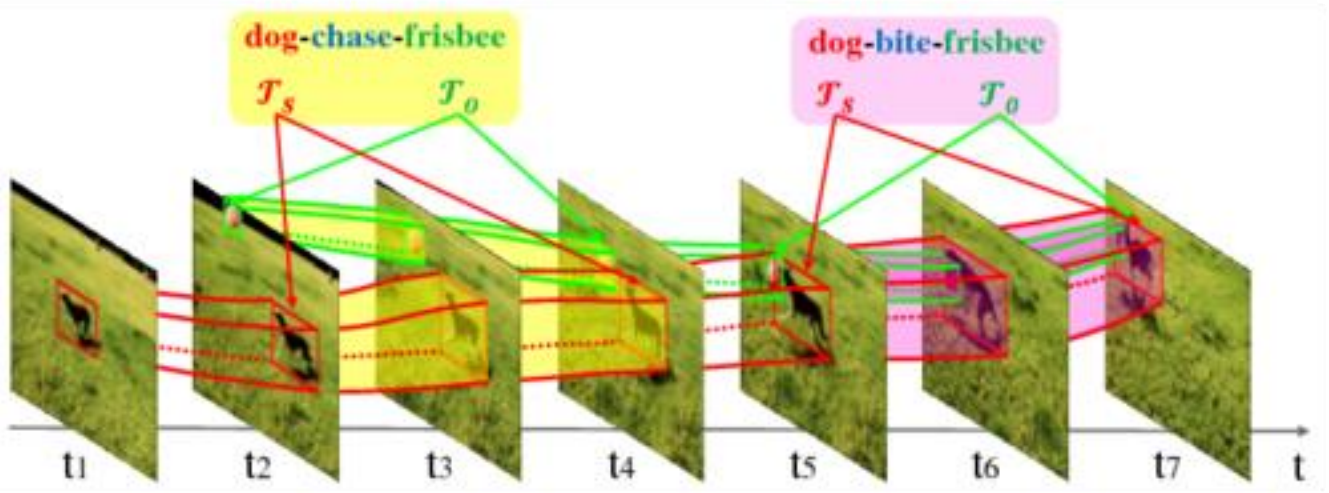
(b) <bike, has, wheel>

(c) <desk, under, monitor>

Given an image, detect objects as well as their relationships:

<subject, predicate, object>

- VRD
- Visual Genome



Given an image, detect objects as well as their relationships:

<subject, predicate, object, Ts, To>

- Vidvrd (video)
- Vidor (video)

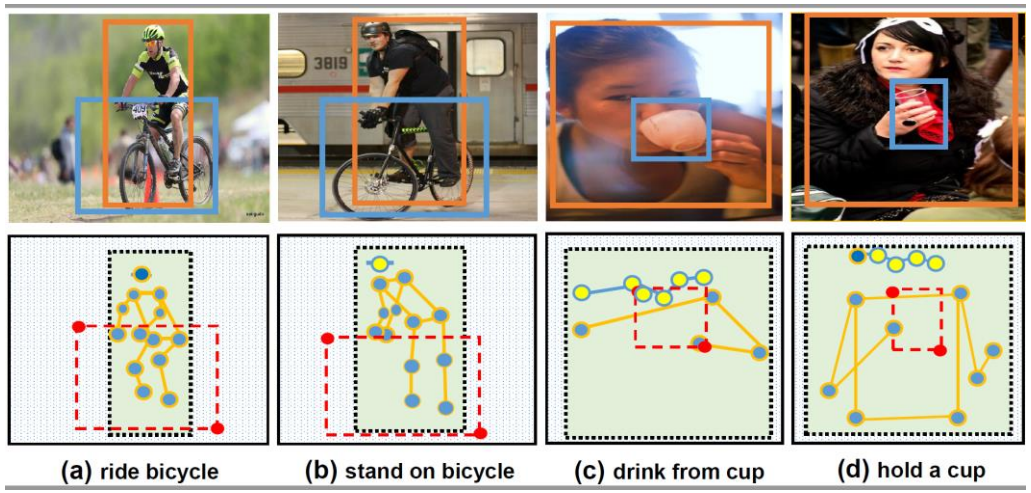
Human-Object Interaction (HOI)

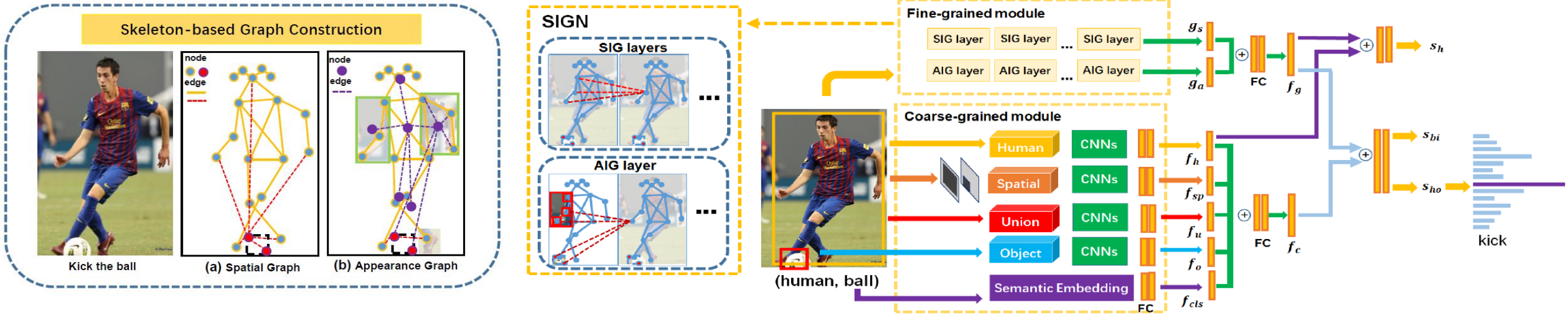
<woman, cut, cake>
<baby, cut, cake>
<woman, hold, baby>
<baby, hold, knife>
...

- HICO(HOI) 【1】
- HICO-DET(HOI) 【2】
- VG-HOI(HOI) 【3】
- HCVRD (HOI) 【4】
- HAKE (HOI) 【5】

【1】Learning to Detect Human-Object Interactions.
【2】HICO: A Benchmark for Recognizing Human-Object Interactions in Images
【3】Visual Semantic role labeling
【4】Care about you: towards large-scale human-centric visual relationship detection
【8】HAKE: Human Activity Knowledge Engine.

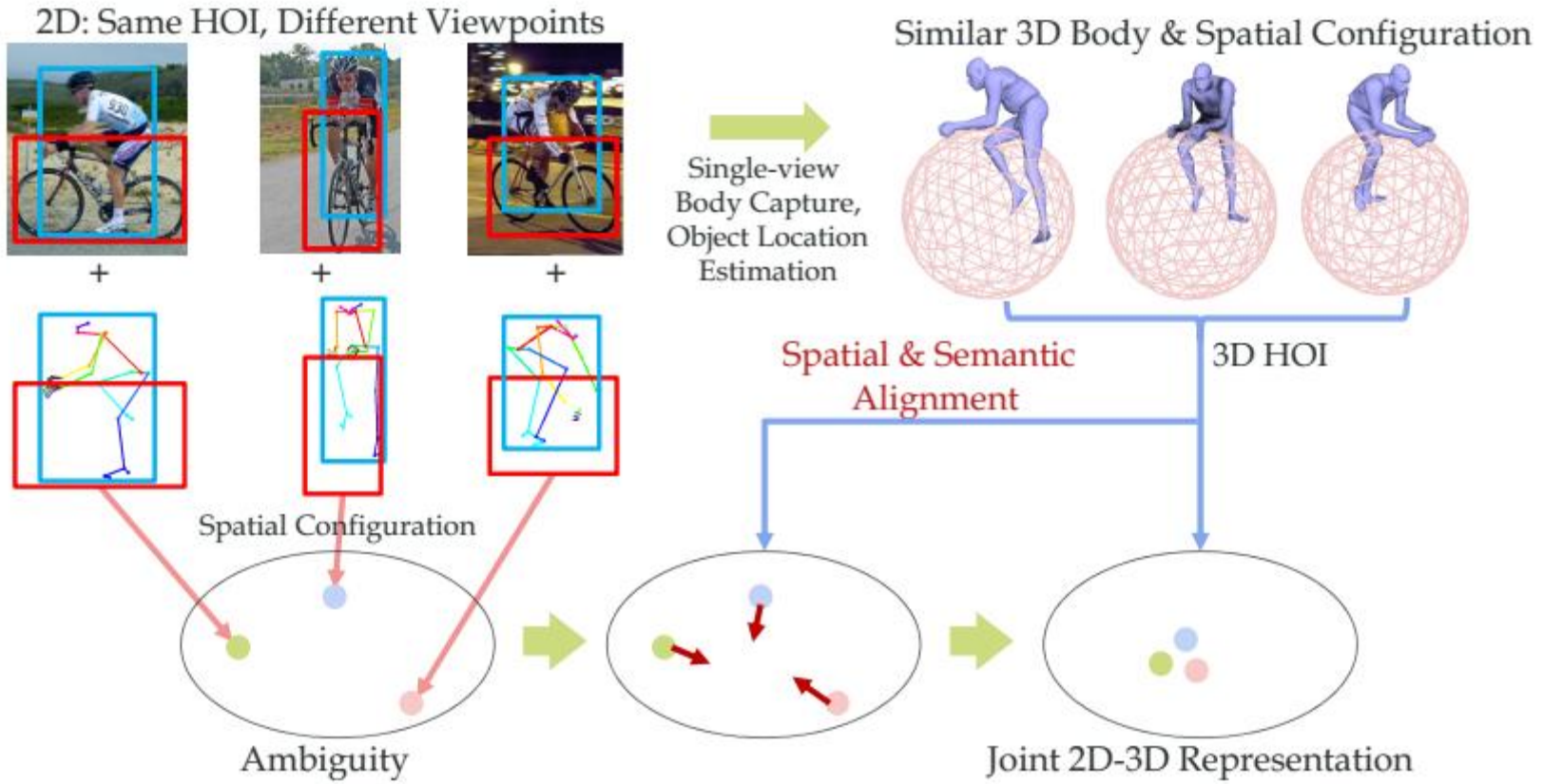
Human-Object Interaction (HOI)





S Zheng, S Chen, Q Jin. *Skeleton-based Interactive Graph Network for Human Object Interaction.* ICME 2020

Human-Object Interaction (HOI)



Y Li, X Liu, H Lu, S Wang, J Liu, J Li, C Lu. *Detailed 2D-3D Joint Representation for Human-Object Interaction. CVPR 2020.*

Image/Video Captioning

A giraffe is bowing and eating grass

A soccer player is kicking a ball into the goal

Flickr8k dataset (2013)
Flickr30k dataset (2014)
MSCOCO Captions (2015)
Visual Genome (2017)
Conceptual Captions (CC) (2018)

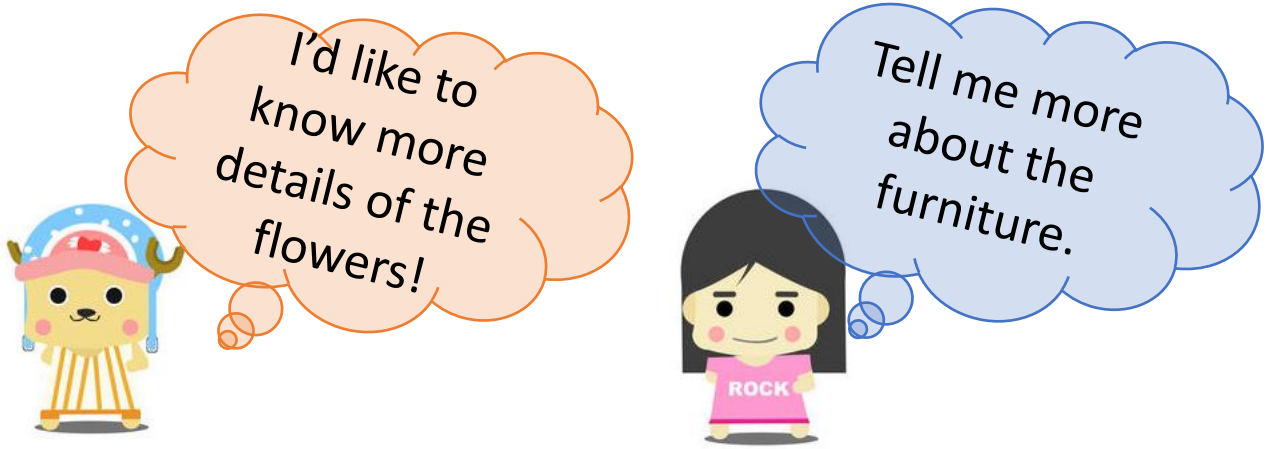
MSVD (2011)
MSR-VTT (2016, 2017)
TGIF (2016)
Vatex (2019)
ActivityNet (2015)

- Dense captioning
- Paragraph captioning
- Novelty captioning
- Discriminative captioning
- Controllable captioning
- Informative captioning

Image Captioning

- **Intention-agnostic Image Captioning**
  - Passively generate image descriptions

- A couple of chairs sitting next to a table with flowers.
- A couple of chairs sitting next to each other on a table.
- A couple of white chairs sitting on top of a wooden table.



☹ Fail to realize different user intentions

☹ Lack diversity

Image Captioning

- **Controllable Image Captioning**
  - Describe user interested image contents via control signals

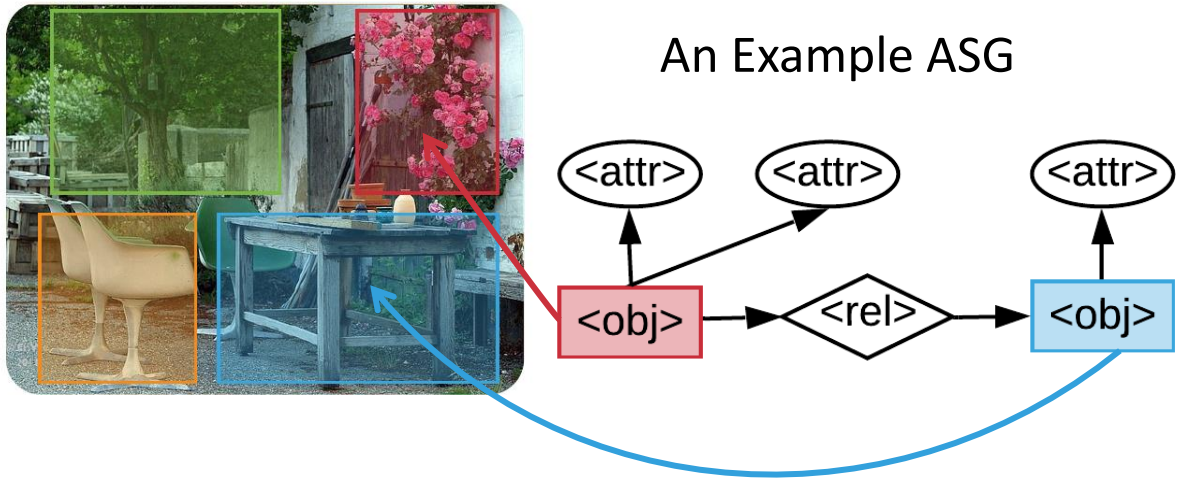
- Single object / region
  -  : Two white chairs sitting next to each other.
  -  : A tree is in the background.
- A set / sequence of objects
  -  : Two white chairs are in front of a tree.

☹ Cannot control at more **fine-grained** level

- ❑ What objects and relationships should be described?
- ❑ How many details should be generated for an object?
- ❑ What is the descriptive order?

ASG: Fine-grained Control Signal

- Abstract Scene Graph (ASG)
  - Directed graph of abstract nodes
    - object, attribute, relationship
    - Nodes are grounded but without semantic labels
- Advantages
  - Represent user desired contents at a fine-grained level
  - Easy to construct
    - Created automatically
    - Designated by users



A cluster of pink flowers are in front of a wooden table.

<---attr---> <attr> <obj> <-------rel------> <-attr-> <-obj->

Chen, Shizhe, et al. "Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs." *CVPR*, 2020.

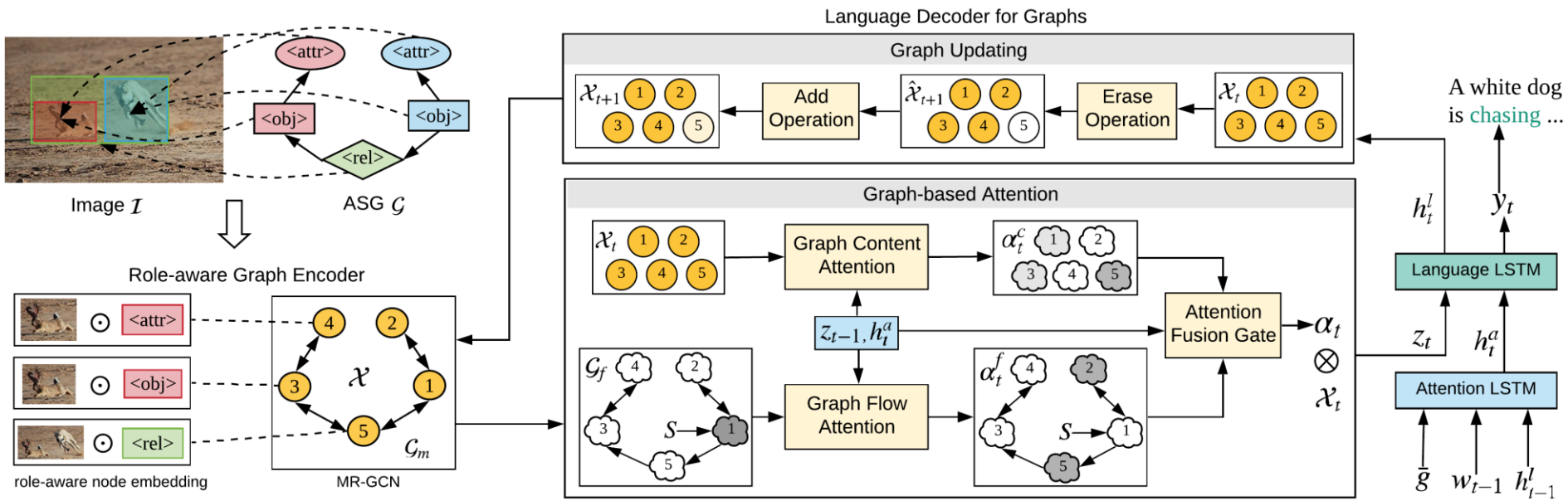
ASG2Caption Model

Role-aware Graph Encoder

- Role-aware node embedding
  - To differentiate fine-grained intentions of nodes
- Multi-relational GCN
  - To improve semantic representations with contexts

Language Decoder for Graphs

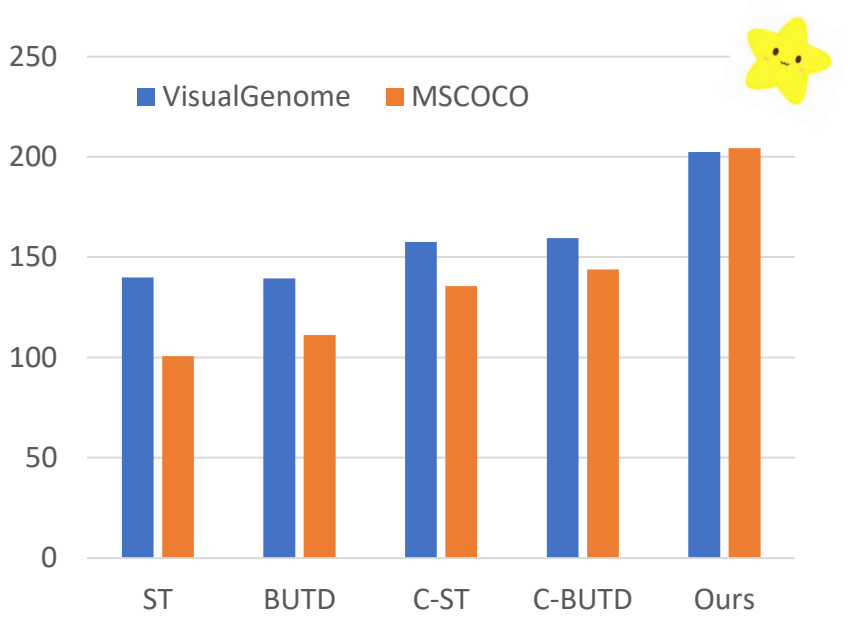
- Graph-based attention
  - to capture both semantics of nodes and graph flow structures
- Graph updating mechanism
  - to cover all information in ASG without omission or repetition



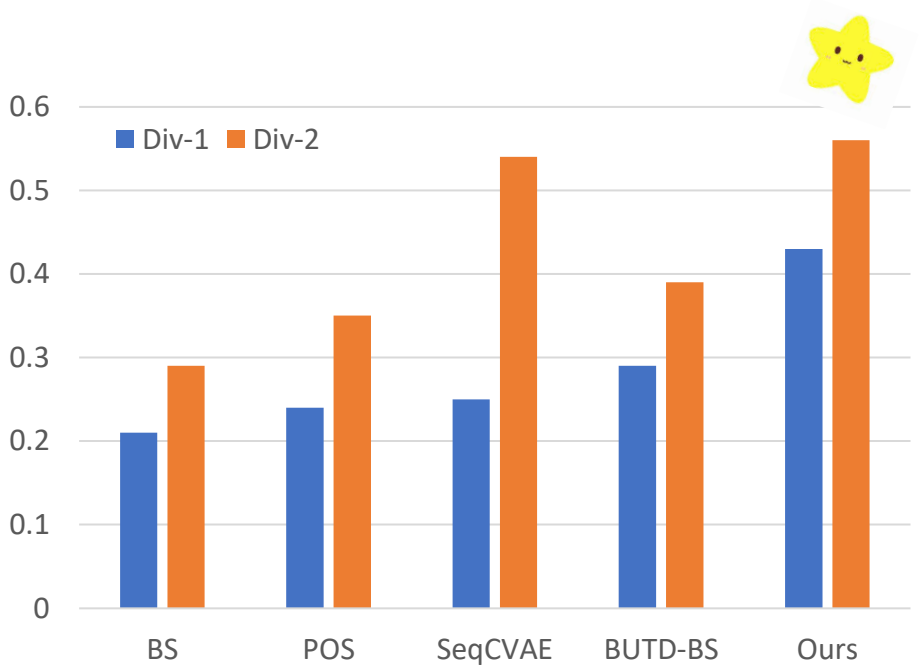
Chen, Shizhe, et al. "Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs." *CVPR,* 2020.

Quantitative Evaluation

- Controllability Evaluation



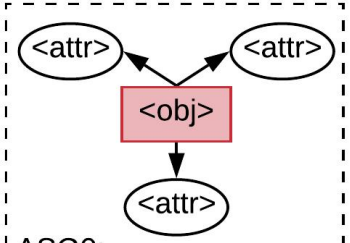
- Diversity Evaluation



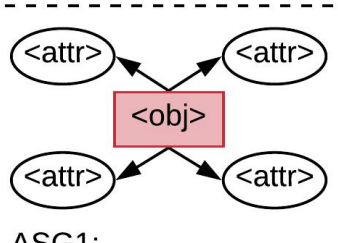
Codes and datasets are available at: https://github.com/cshizhe/asg2cap

Chen, Shizhe, et al. "Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs." *CVPR*, 2020.

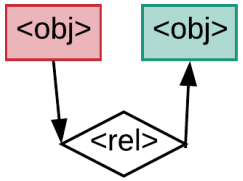
Qualitative Evaluation



ASG0:
a beautiful young woman is sitting down.

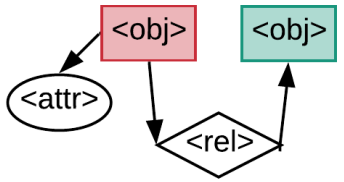


ASG1:
a beautiful brown haired woman is sitting down.

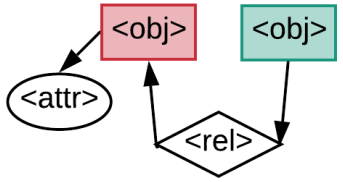


ASG2:
a woman sitting on a bench.

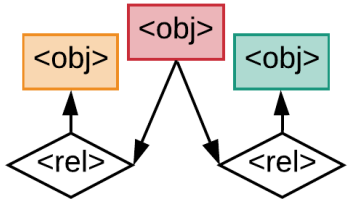
More attributes: + details on hair color



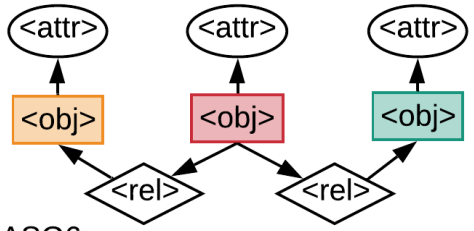
ASG3:
a young woman sitting on a bench.



ASG4:
a bench with a young woman sitting on it.



ASG5:
a woman sitting on a bench by a tree.

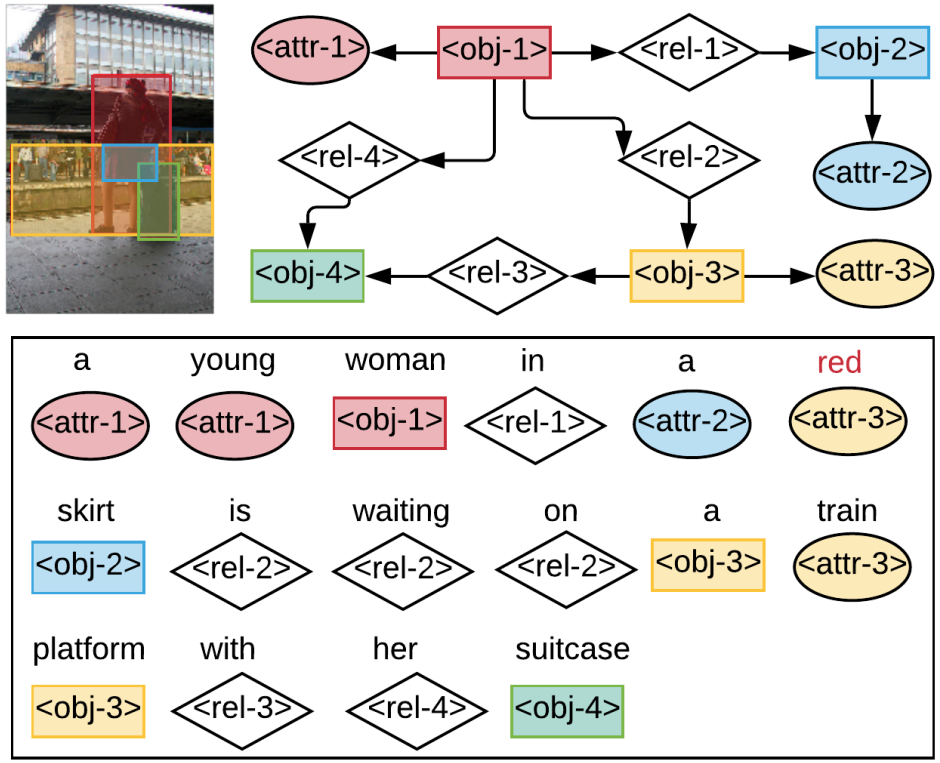


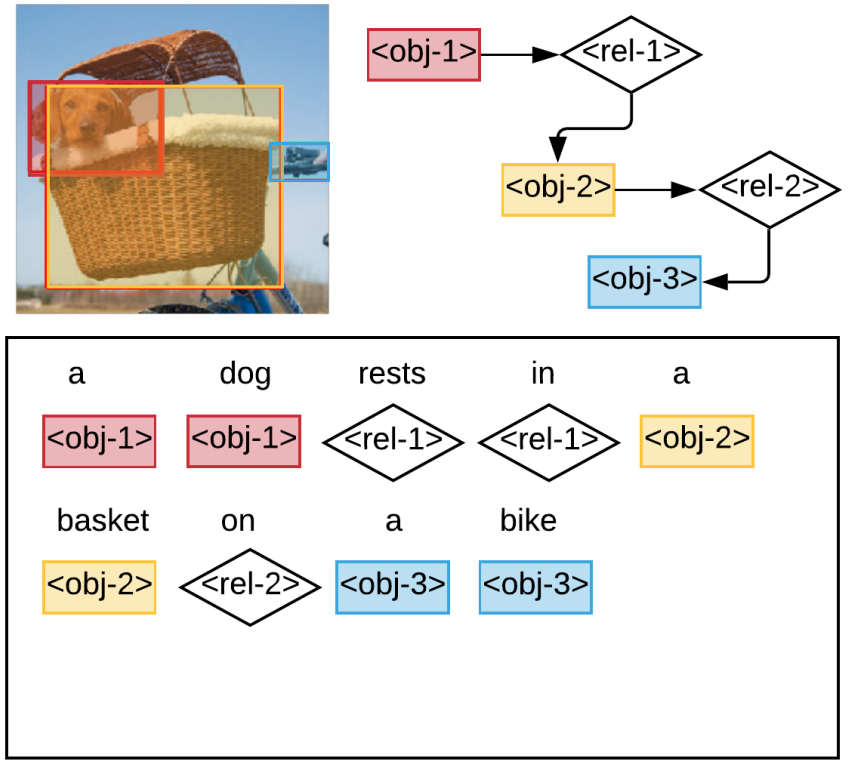
ASG6:
a young woman sitting on a wooden bench next to a large tree.

Inverse direction of edges: change descriptive order

Qualitative Evaluation

More Grounded and Explainable Caption Generation Process





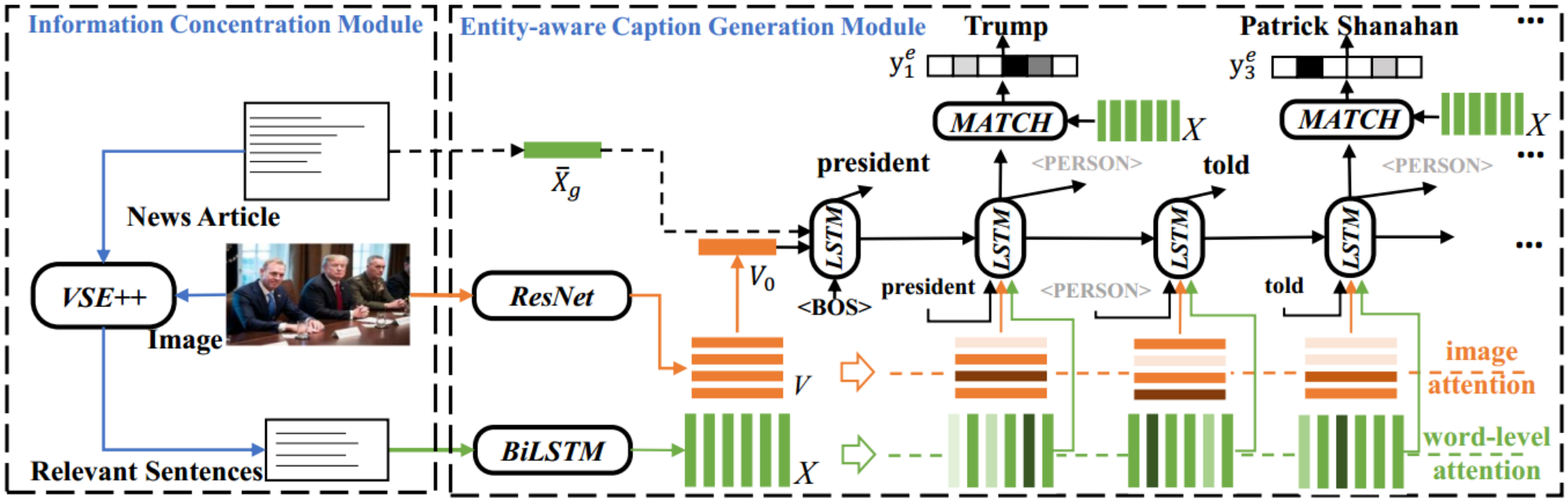
Chen, Shizhe, et al. "Say As You Wish: Fine-grained Control of Image Caption Generation with Abstract Scene Graphs." *CVPR*, 2020.

Informative Image Captioning

General caption:
a young girl in a white tank top is holding a stick
News caption:
Aisling Donnelly holding the snake discovered in a back garden in County Tyrone

News article:
A family in County Tyrone have been paid a surprise visit by a snake. Aisling Donnelly of Clonoe just outside Coalisland said her brother and his friend came across the large snake in their back garden…Aisling Donnelly told BBC Radio Ulster that the family were not sure about how to handle or deal with the surprise visitor... However Aisling's sister was brave enough to lift the snake and then the whole family held it after the initial fear disappeared. The snake has not yet been named, although Aisling said the whole country has come round to see it … Although the Donnelly family are not planning on keeping the snake, Aisling added, 'I wouldn't mind one actually'.

Entity-aware News Image Captioning



ICECAP: **I**nformation **C**oncentrated Entity-aware image **CAP**tioning.

Information Concentration Module: selects relevant sentences from the article.

Entity-aware Caption Generation Module: generates the entity-aware image caption.

Hu, Anwen, et al. "ICECAP: Information Concentrated Entity-aware image CAPtioning." *ACM Multimedia,* 2020.

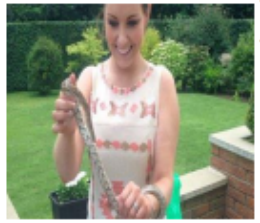
Entity-aware News Image Captioning

Examples

Biten et al.[3] + SAttIns: Keira Knightley (√) and Keira Knightley (×) in Learning to Drive
ICECAP: Problematic Keira Knightley (√) and Benedict Cumberbatch (√) in the film The Imitation Game The Imitation Game (×)
Ground: Keira Knightley and Benedict Cumberbatch in The Imitation Game

Relevant Sentences: (1) Problematic Keira Knightley <PERSON> and Benedict Cumberbatch <PERSON> in The Imitation Game The Imitation Game <WORK-OF-ART> a biopic about mathematician and World War Two codebreaker Alan Turing <PERSON> won the top prize at the Toronto International Film Festival…

(a)



Biten et al.[3] + SAttIns: Aisling (×) was found dead in Coalisland (×)
ICECAP: Aisling Donnelly (√) was found at the park in a field in his farm in County (×)
Ground: Aisling Donnelly holding the snake discovered in a back garden in County Tyrone

Relevant Sentences: (1) A family in County<GEOGRAPHIC PLACE> Tyrone<PERSON> have been paid a surprise visit by a snake. … (2) Aisling Donnelly<PERSON> told BBC Radio Ulster that the family were not sure about how to handle or deal with the surprise visitor' … (3) However Aisling<PERSON> 's sister was brave enough to lift the snake and then the whole family held it after the initial fear disappeared …

(b)

Fine-grained Understanding in Specific Domain

- FineGym
- Figure Skating
- Aist dance video
- Let’s Dance
- MPII Cooking 2
- MERL Shopping

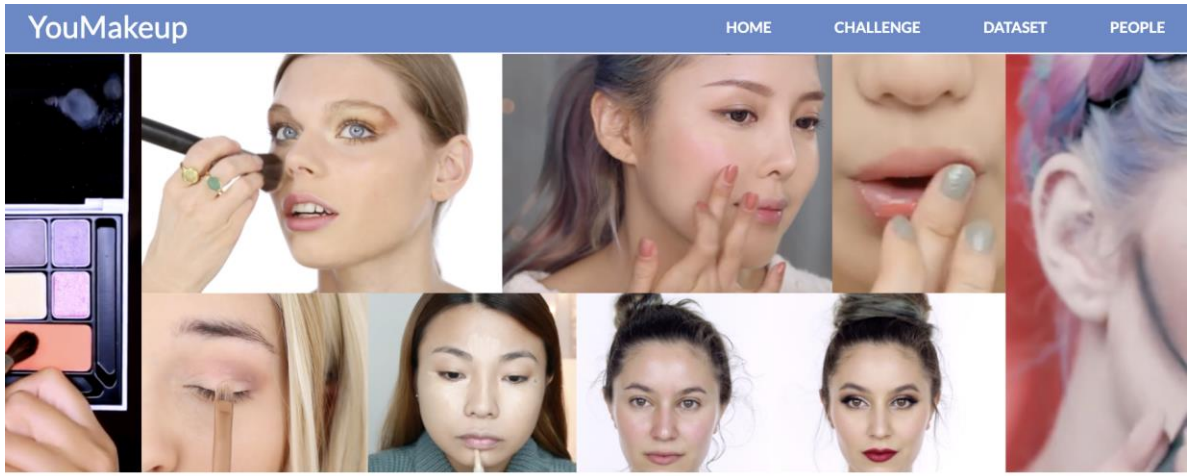
Fine-grained Understanding in Specific Domain

- FineGym
- Figure Skating
- Aist dance video
- Let's Dance
- MPII Cooking 2
- MERL Shopping

YouMakeup

Step	Time range	Area	Step Description
Step 1	00:02:10~00:02:39	eyelid	Apply eyeshadow over eyelid with brush
Step 2	00:02:40~00:02:54	brow	Draw eyebrow using the brow shadow with brush
Step 3	00:02:56~00:03:10	lashline	Draw winged eyelinear with eyelinear pen
Step 4	00:03:11~00:03:24	lip	Apply red lipgloss on lips
Step 5	00:03:25~00:03:28	cheek, hailine	Apply bronzer on the cheeks and hairline with brush
Step 6	00:03:29~00:03:34	cheekbone, forehead, nose	Apply highlighter on the cheekbones, forehead and nose with brush
Step 7	00:03:35~00:03:47	lash	Curl the lashes and apply mascara on the lashes
Step 8	00:03:54~00:04:15	lash	Apply false lashes on the lashes

Fine-grained Understanding in Specific Domain



YouMakeup Video Question Answering Challenge @CVPR LVVU Workshop 2020

Introduction

The goal of YouMakeup VQA challenge is to provide a common benchmark for fine-grained action understanding in specific domain videos. The makeup instructional videos are naturally more fine-grained than open-domain videos. Different action steps share the similar backgrounds, but contain subtle and critical differences such as actions, tools and applied facial areas, resulting in different effects on the face. Therefore, it requires fine-grained discrimination abiltilies within temporal and spatial context.

We propose two VQA sub-challenges based on YouMakeup dataset, namely Facial Image Ordering Sub-Challenge and Step Ordering Sub-Challenge. The goal of Facial Image Ordering Sub-Challenge is to understand changes of object given a certain action in flexible natural language expression, while Step Ordering Sub-Challenge aims at evaluating models' abilities in cross-modal semantic alignments between visual and texts.

YouMakeup VQA

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Conference on Computer Vision and Pattern Recognition
Language & Vision with applications to Video Understanding Workshop
https://languageandvision.github.io/youmakeup_vqa/index.html

$AI{\cdot}M^3$ *Lab*

Official Website

Challenge Task

YouMakeup VQA challenge aims to provide a common benchmark for fine-grained action understanding in specific domain videos.

Facial Image Ordering Sub-Challenge
Understand changes of object given a certain action in flexible natural language expression

Step Ordering Sub-Challenge
Evaluate models' abilities in cross-modal semantic alignments between visual and texts.

Important Date

- **Jan 15, 2020** Dataset available for download
- **April 06, 2020** Web Site and Call for Participation Ready
- **April 12, 2020** Baseline codes and models available for download
- **June 01, 2020** Results submission deadline
- **June 08, 2020** Paper submission deadline

Subjective Aspect

- Affective recognition

Future Interests

- Context encoding
- Fine-grained vs Group
  - Knowledge
- Unsupervised/self-supervised learning
- Explainable

Thank You!