CS 6804: MULTIMODAL VISION

Chris Thomas

Department of Computer Science

January 18, 2023

COURSE INFORMATION

- **Course website:** https://people.cs.vt.edu/chris/cs6804_spring2023/
- **Location:** 318 Randolph Hall, Blacksburg, VA.
- **Meeting time:** Mondays and Wednesdays, 4:00 PM - 5:15 PM
- **Instructor:** Chris Thomas
- **E-mail:** chris@cs.vt.edu
  - Important note! When you e-mail me, you must put **CS6804** (no space) at the beginning of the subject line. I receive a large volume of e-mail and your e-mail might not receive a response unless you include this tag.
- **Office:** 3120C Torgersen Hall
- **Office hours:** 12:00 PM - 1:00 PM Wednesdays. My Zoom is linked on the site
- **Exam section:** 16M. Tentatively May 8, 2023, 7:45AM - 9:45AM.
  - While there are no exams in this course, we may use this time for final presentations only if absolutely necessary.
- **Canvas** will be used to submit assignments / post grades

INTRODUCTIONS

- What’s your name?

- What department are you enrolled in? Which program? Which year?

- Describe your current research to us, what are you currently working on and with which faculty? If no current research, what do you wish you were doing?

- Why are you taking this class?

- Any interesting facts we should know about you?

WHAT IS COMPUTER VISION?

- Automatic understanding of images and video
  - Computing properties of the 3D world from visual data (measurement)

- Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities (perception and interpretation)

- Algorithms to mine, search, and interact with visual data (search and organization)

Kristen Grauman and Adriana Kovashka

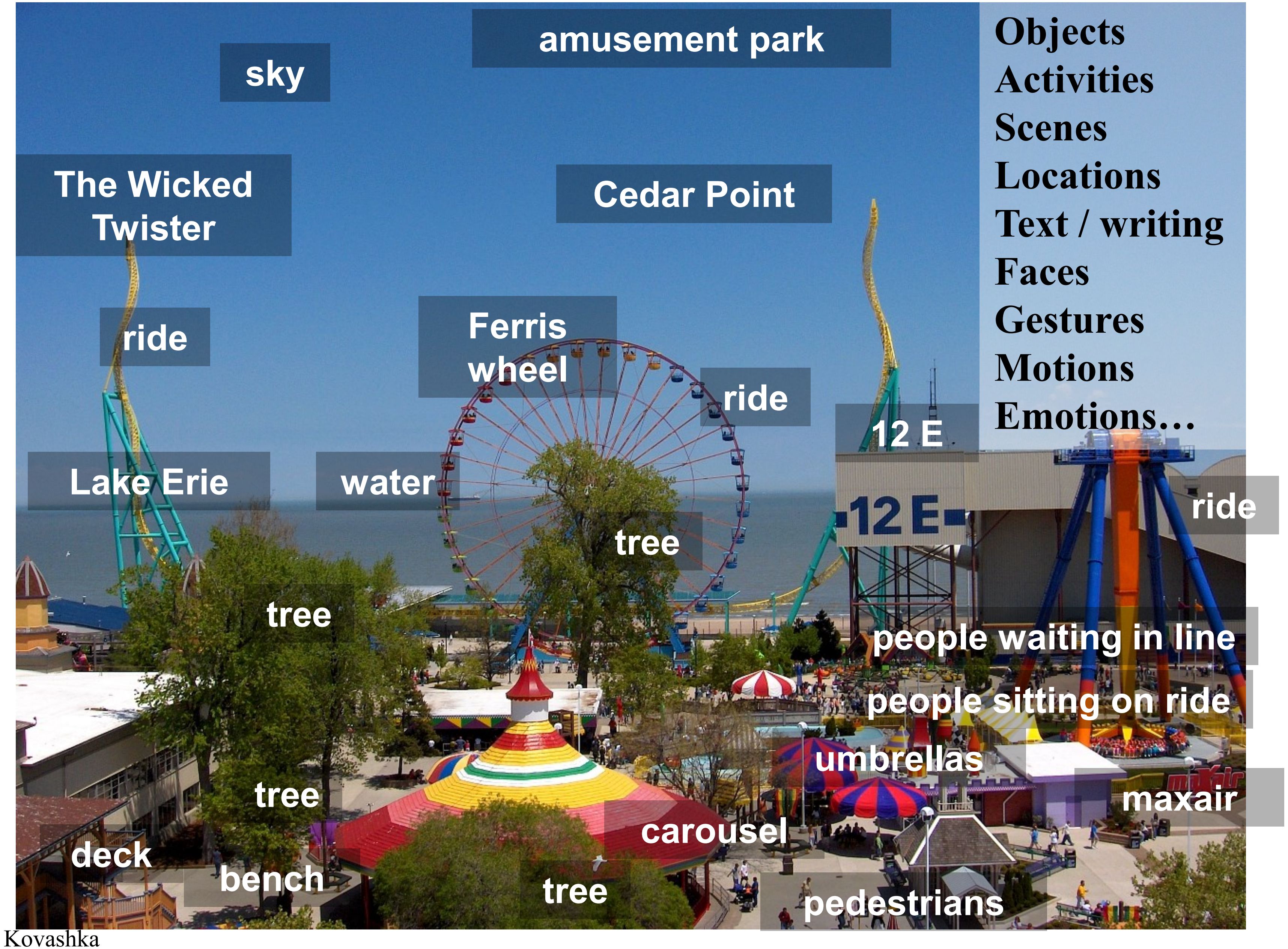
WHAT IS COMPUTER VISION?

Done?

"We see with our brains, not with our eyes“ (Oliver Sacks and others)

Kristen Grauman and Adriana Kovashka

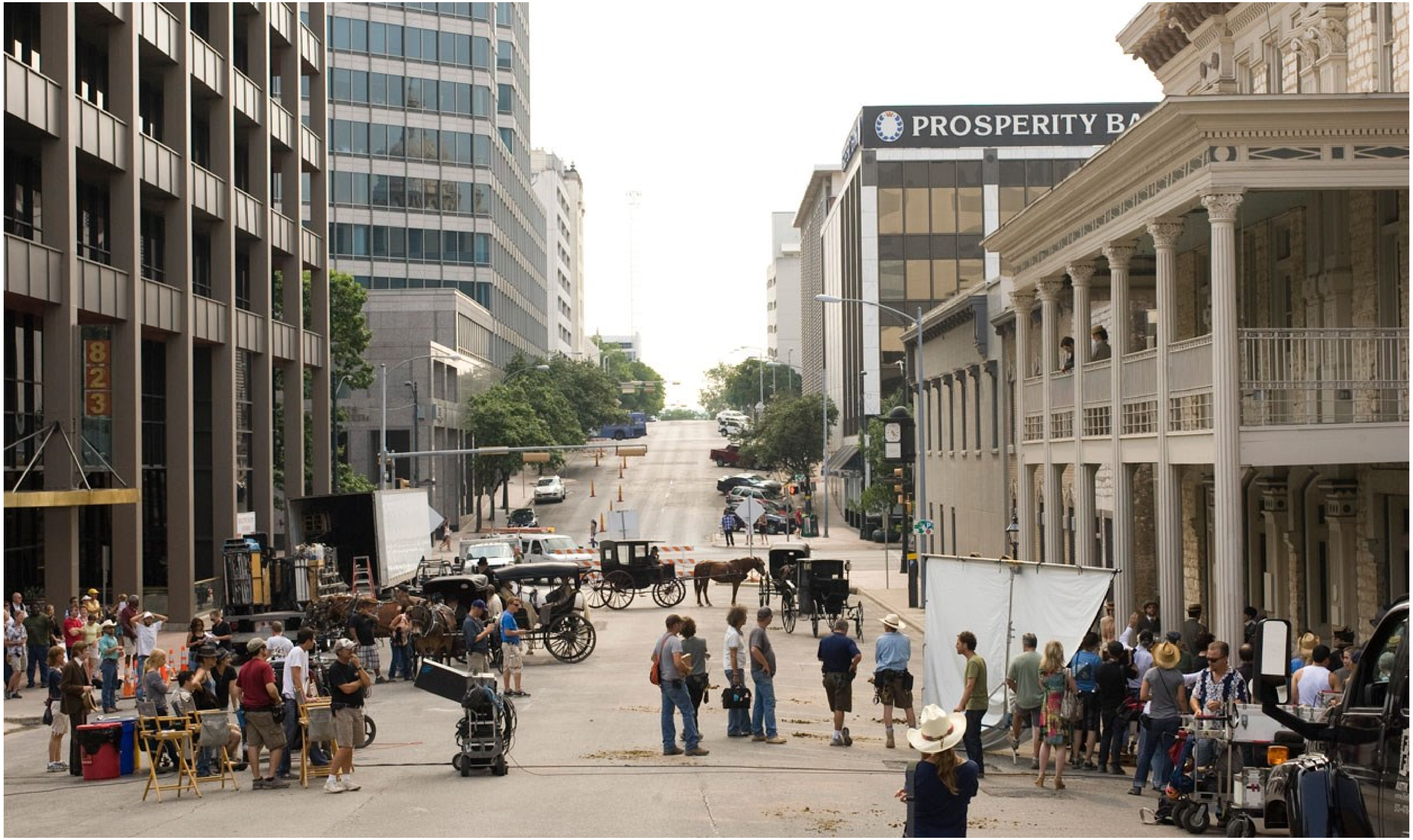
VISION FOR RECOGNITION



Objects
Activities
Scenes
Locations
Text / writing
Faces
Gestures
Motions
Emotions…

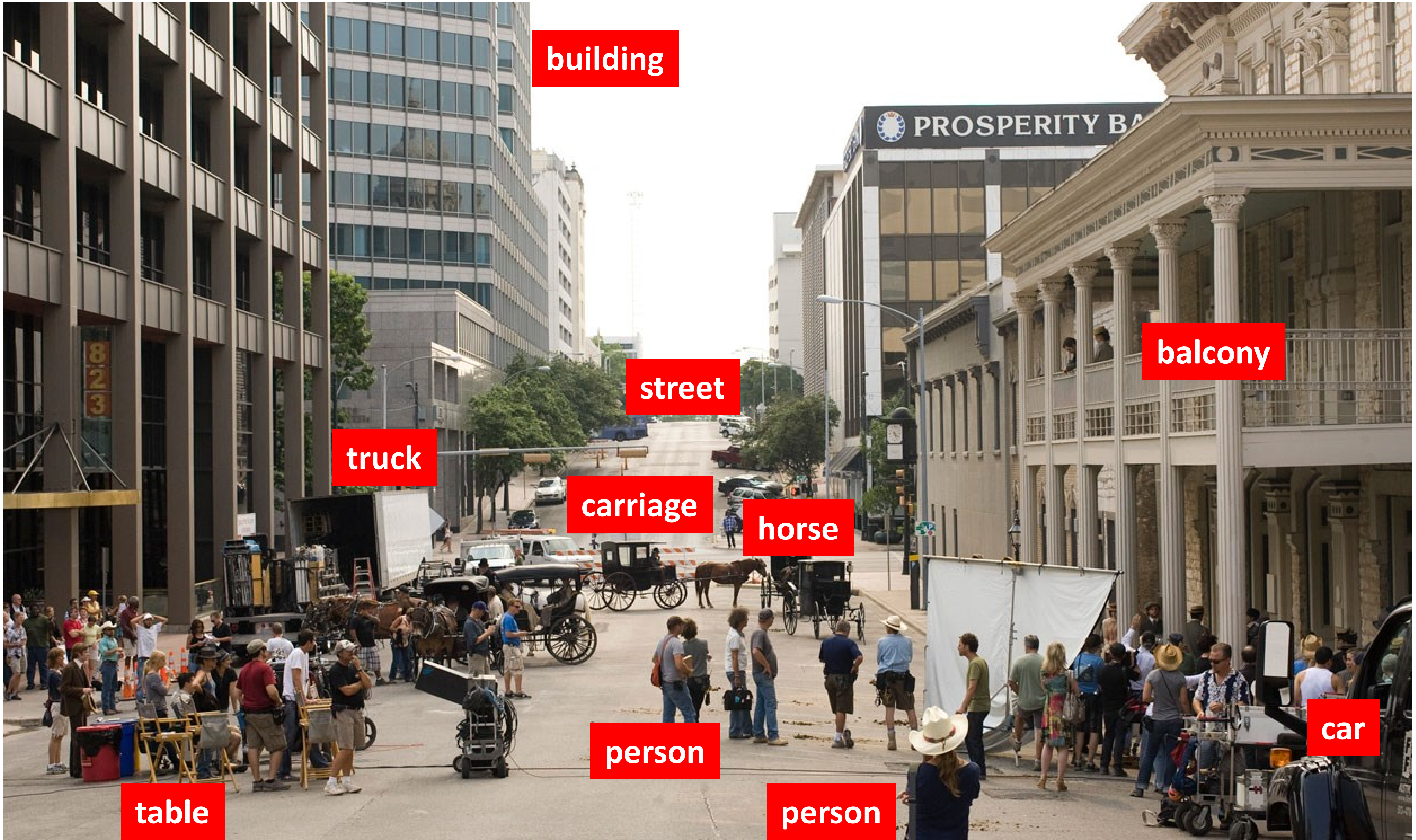
Kristen Grauman and Adriana Kovashka

SOME VISUAL RECOGNITION PROBLEMS: WHY ARE THEY CHALLENGING?



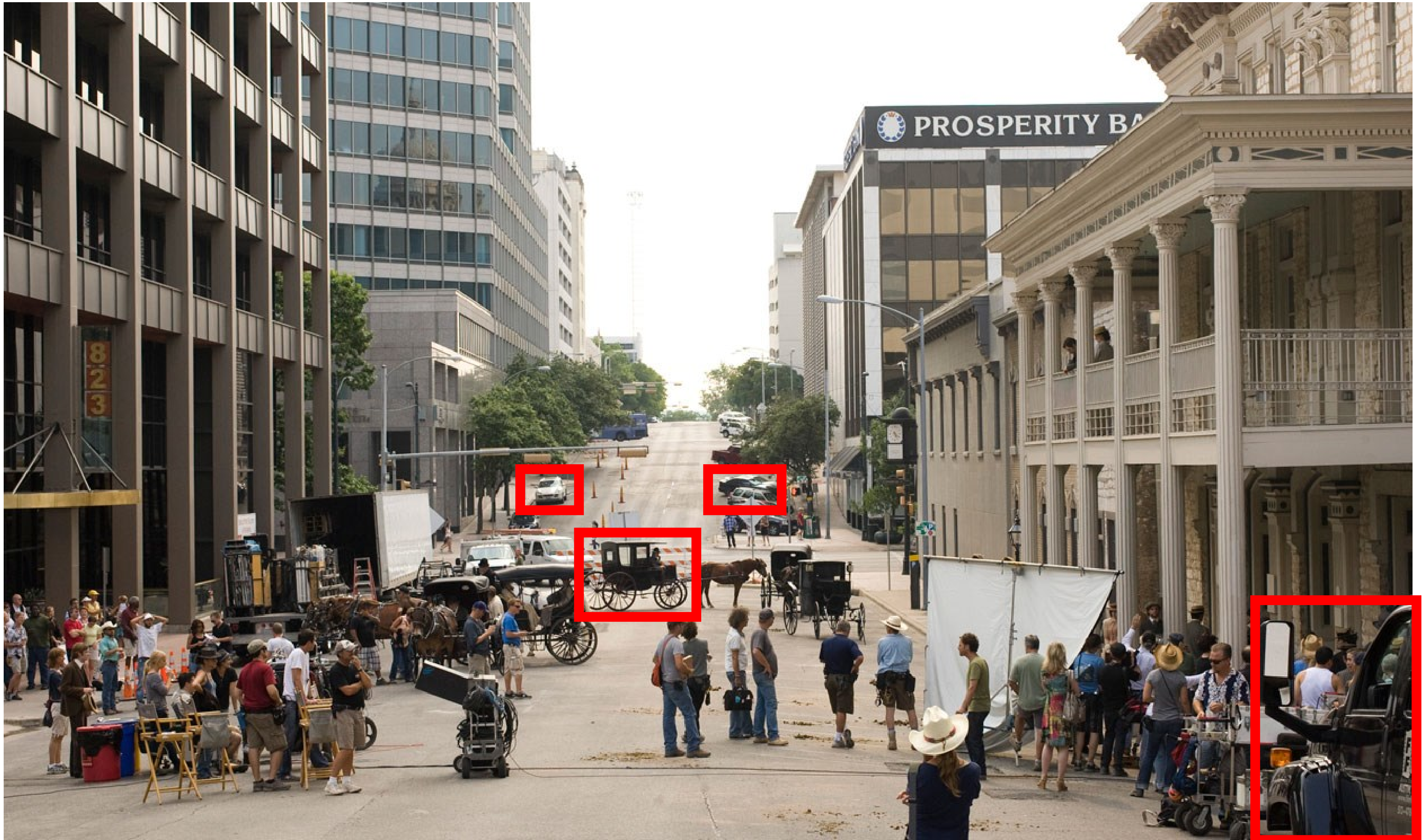
Adriana Kovashka

RECOGNITION: WHAT OBJECTS DO YOU SEE?



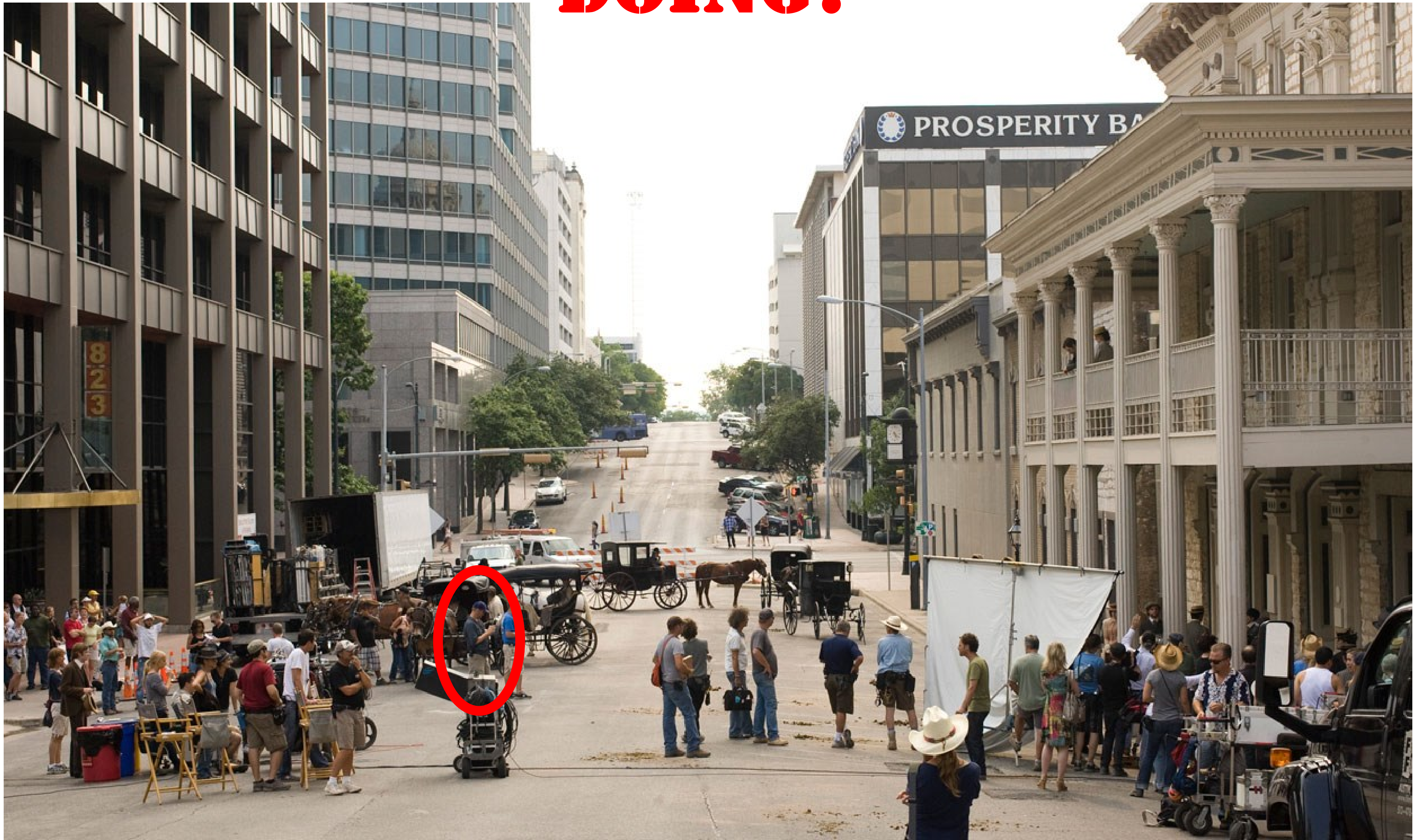
Adriana Kovashka

DETECTION: WHERE ARE THE CARS?

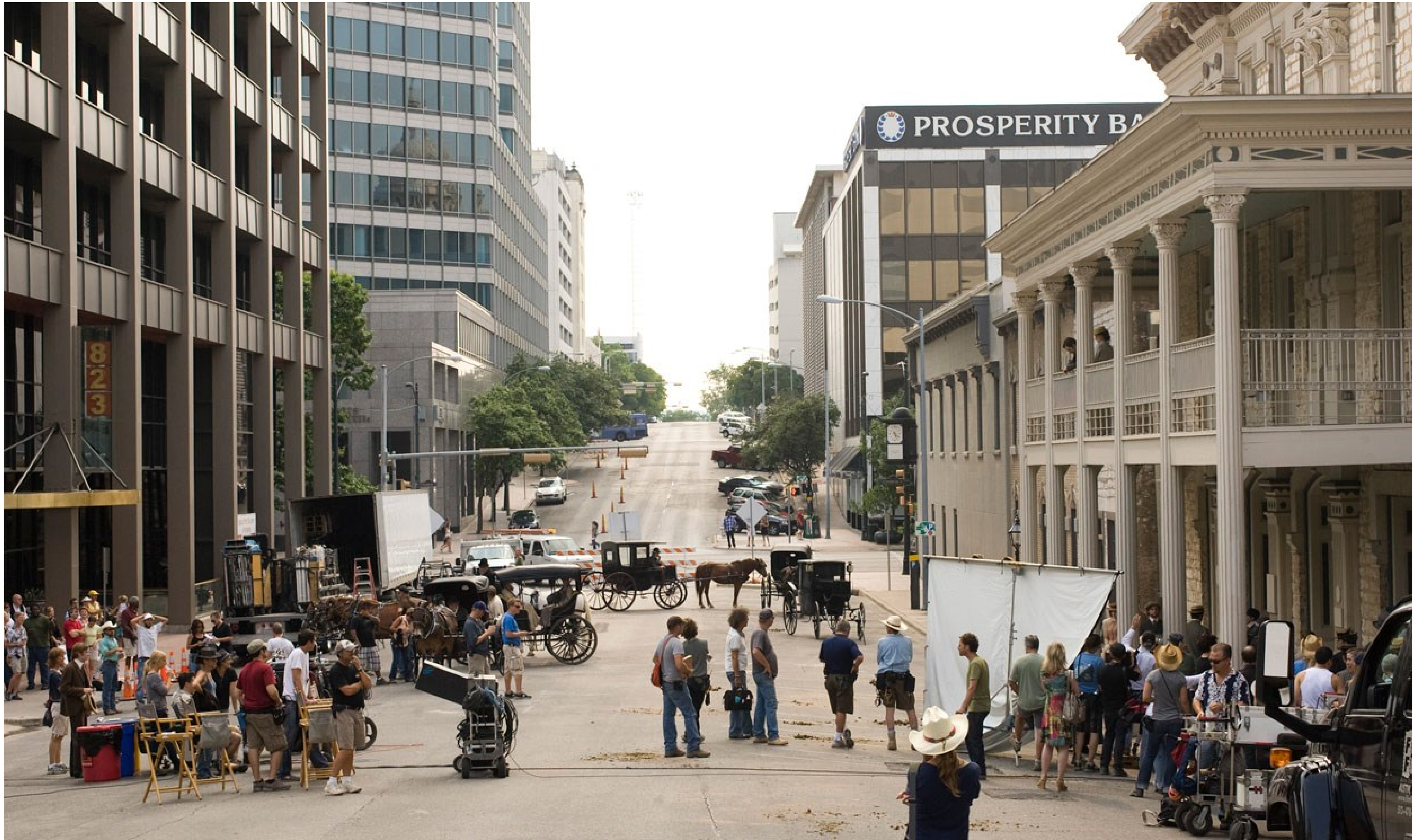


Adriana Kovashka

ACTIVITY: WHAT IS THIS PERSON DOING?



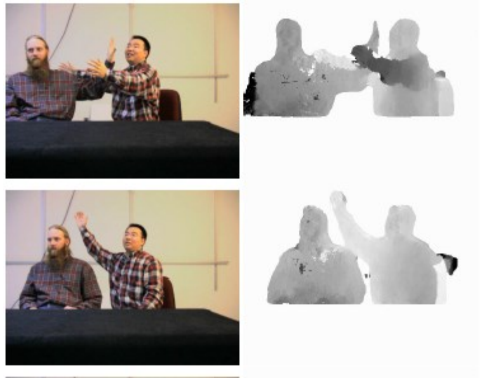
Scene: Is this an indoor scene?



Adriana Kovashka

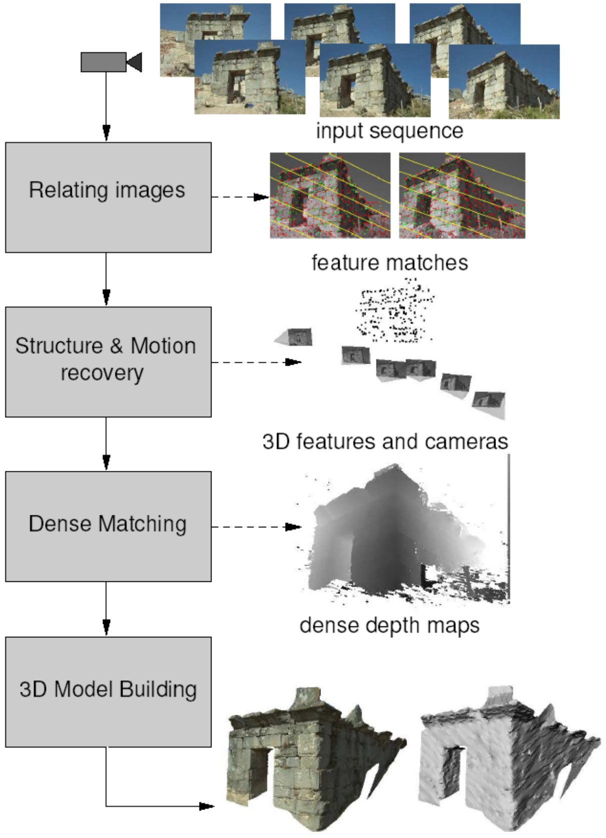
VISION FOR MEASUREMENT

Real-time stereo



Pollefeys et al.

Structure from motion



Relating images

Structure & Motion recovery

Dense Matching

3D Model Building

input sequence

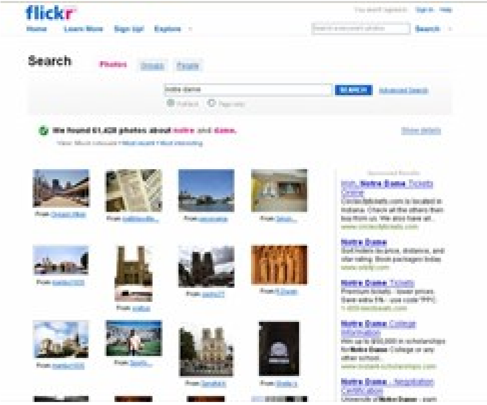
feature matches

3D features and cameras

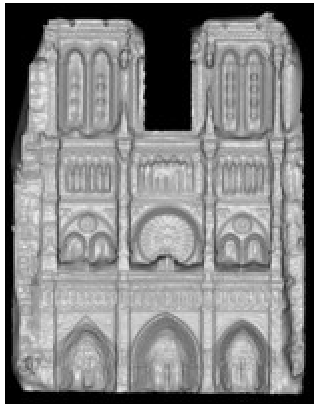
dense depth maps

3D surface model

Multi-view stereo for community photo collections



We found 61,428 photos about notre and dame.



Goesele et al.

Kristen Grauman

Slide credit: L. Lazebnik

WHAT IS COMPUTER VISION?

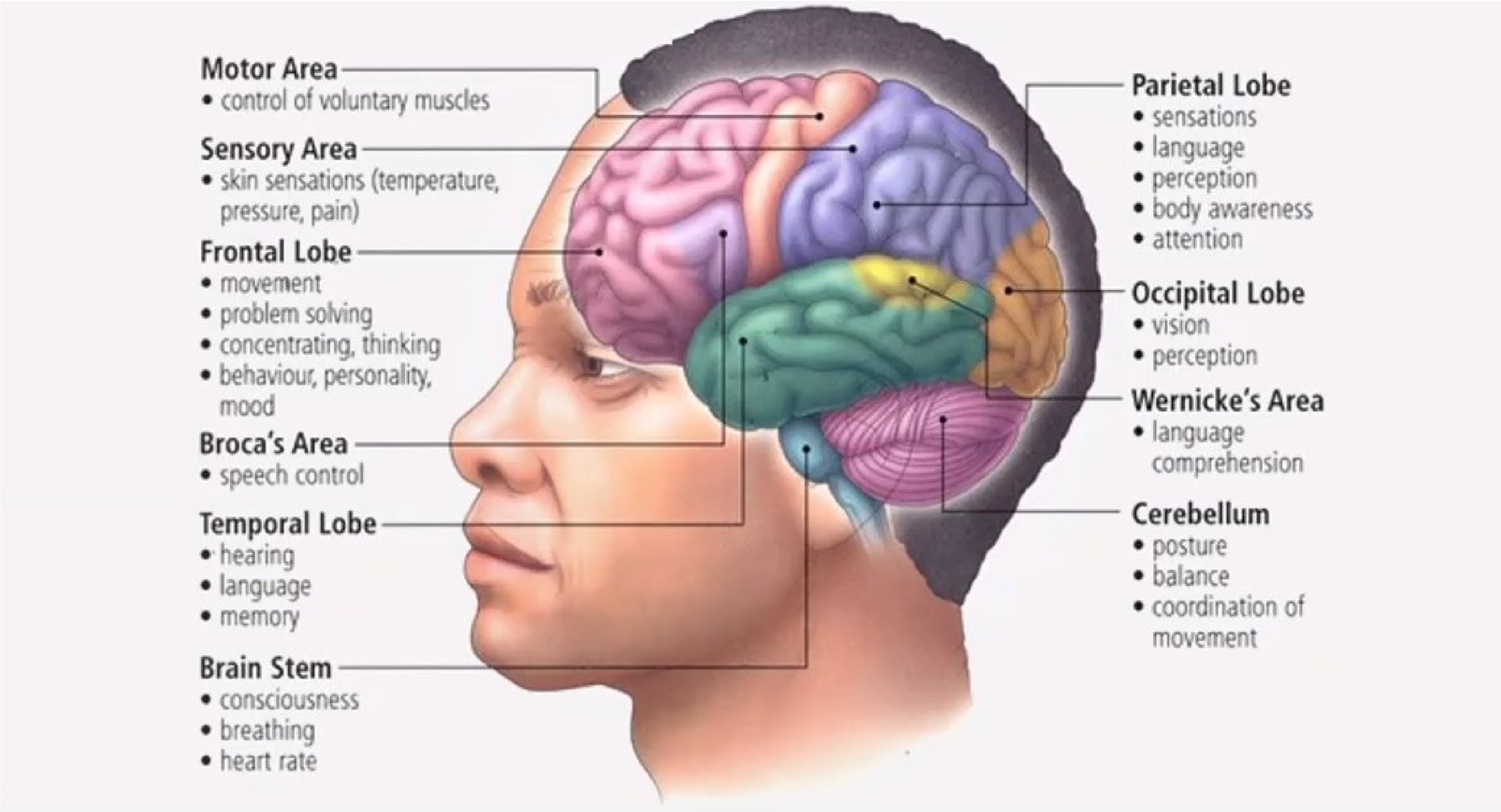
- Automatic understanding of images and video
    - Computing properties of the 3D world from visual data (measurement)
- Algorithms and representations to allow a machine to recognize objects, people, scenes, and activities (perception and interpretation)
- Algorithms to mine, search, and interact with visual data (search and organization)
- **(increasingly) Algorithms which can reason about, mine, search, learn from, and interact with *multimodal data***
    - ***Our focus***

Kristen Grauman and Adriana Kovashka

WHAT DO WE MEAN BY MULTIMODAL VISION?

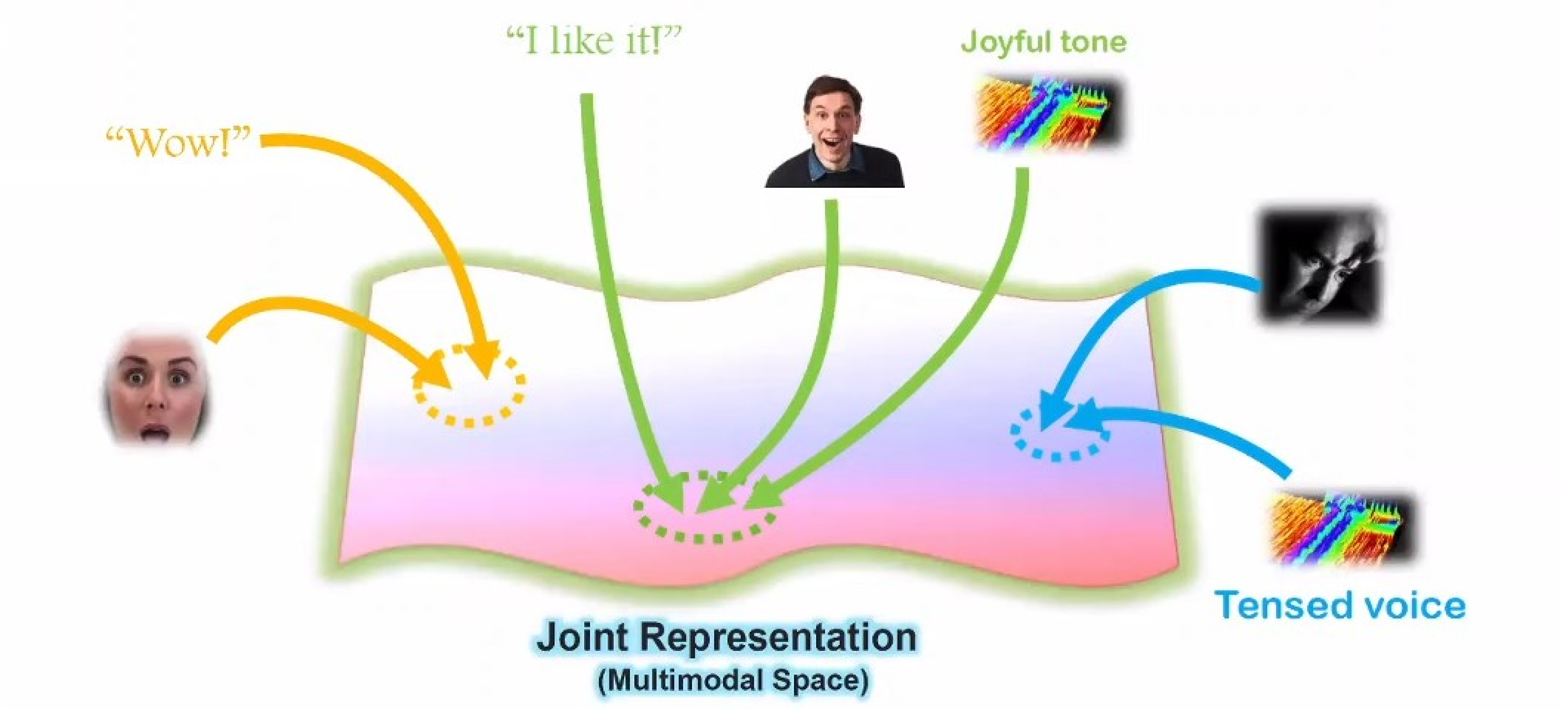
- Algorithms that can reason intelligently across different *modalities* (one or more of which is visual)
- Vision and language
  - Many (but not all) of the papers we read in this course will be focused on the intersection of vision and natural language processing
  - How can we get machines to understand the relationship between visual data and text?
- Vision and audio
  - How can we learn the association between a certain audio signal and visual inputs?
- Vision + X
  - Vision + structured knowledge
  - Vision + sensor data
  - Vision + …
- Reasoning across different types of visual data (e.g. images + 3d)

HUMAN BRAIN PROCESSES MULTIMODAL DATA



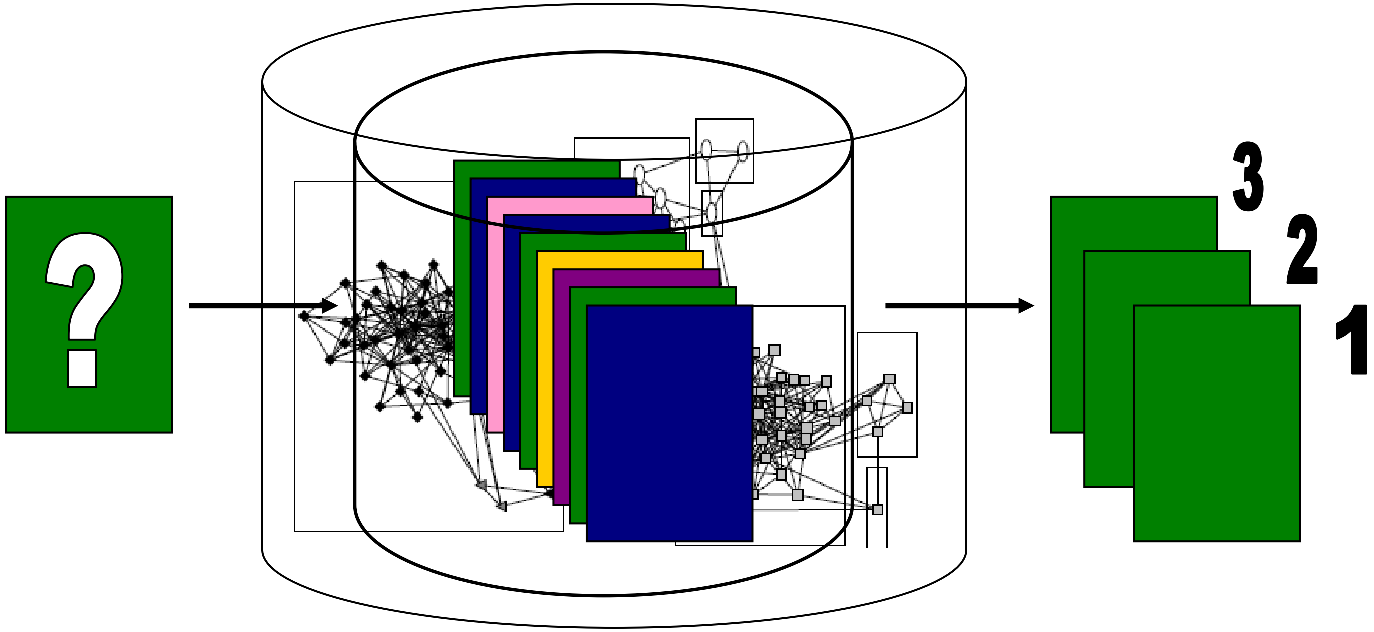
Source: https://www.humanbrainfacts.org/basic-structure-and-function-of-human-brain.php

FUNDAMENTAL CHALLENGE OF MULTIMODAL DATA



Source: Louis-Philippe (LP) Morency and Tadas Baltrusaitis, Tutorial on Multimodal ML, ACL 2017.

VISUAL SEARCH, ORGANIZATION



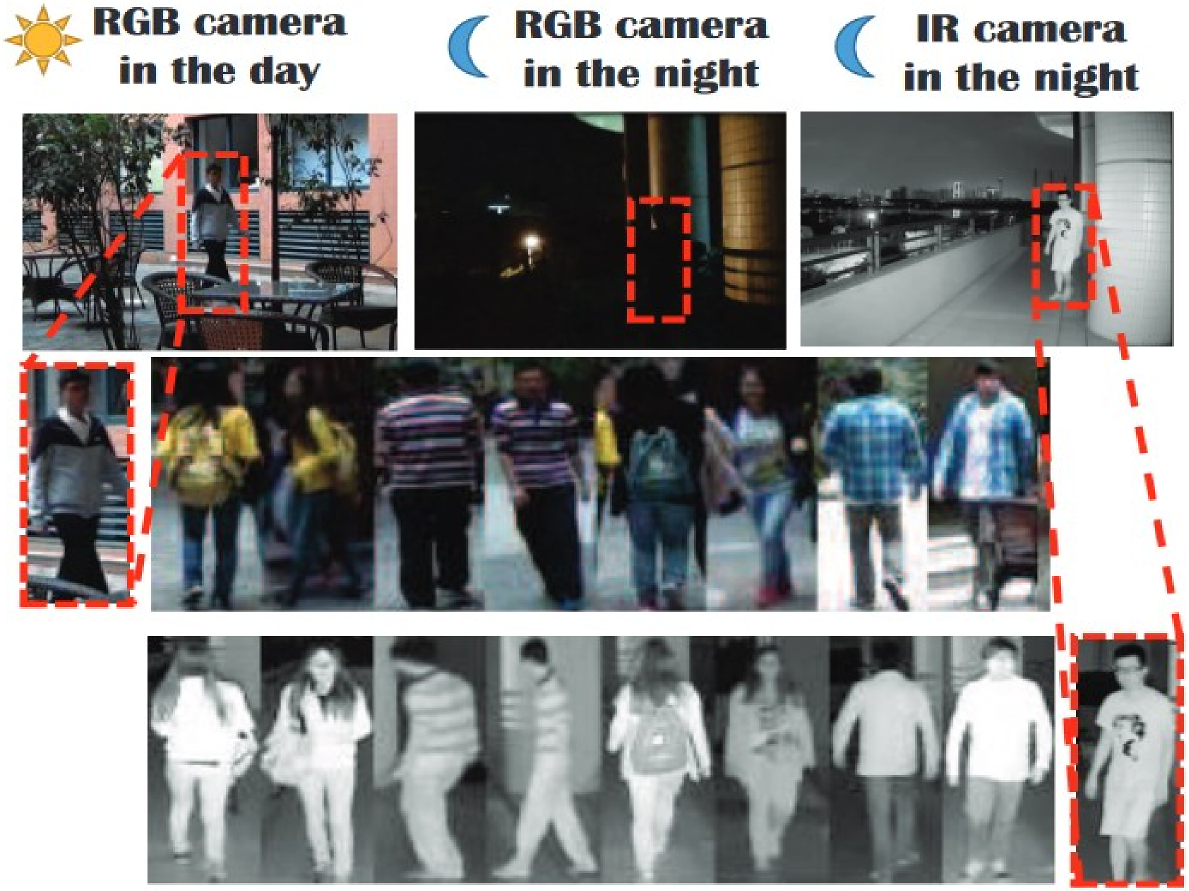
Query

Image or video archives

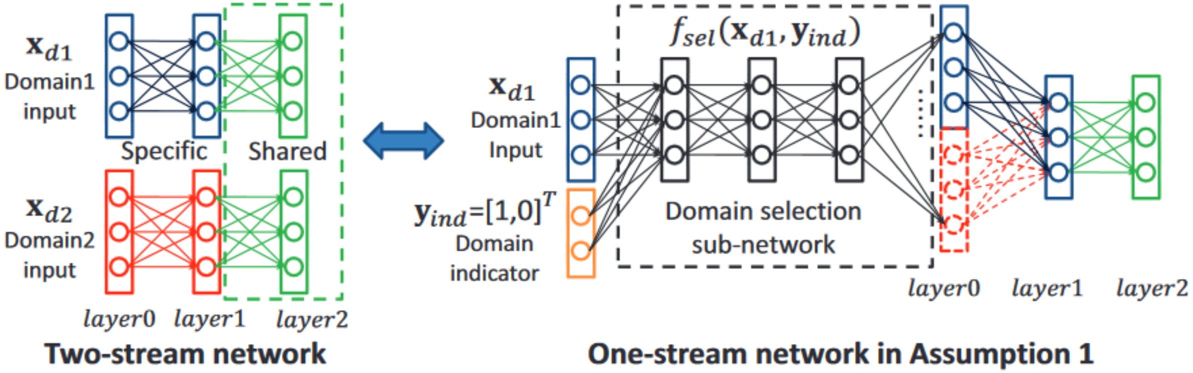
Relevant content

Adapted from Kristen Grauman / Adriana Kovashka

CROSS-MODALITY PERSON RE-ID



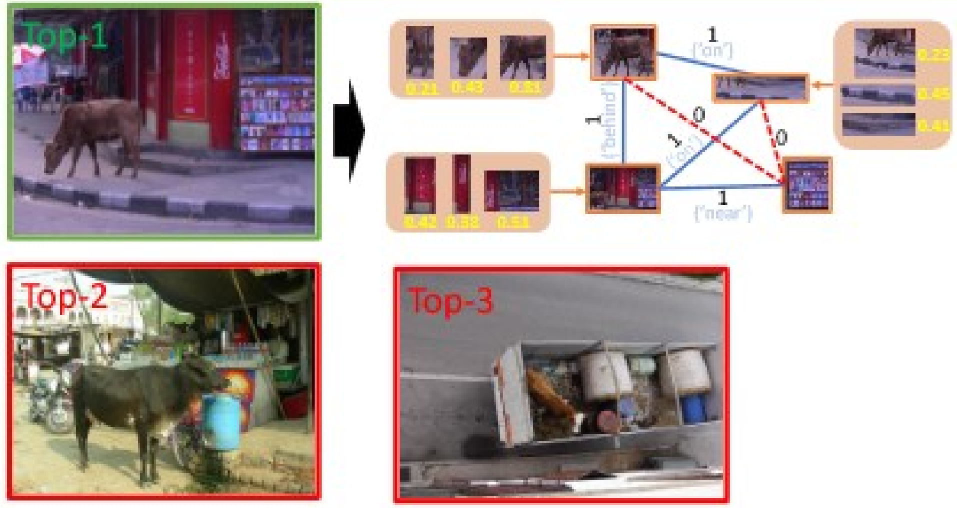
- Can you detect the same person in different visual modalities?
- How to represent visual data from these different modalities in a joint way?

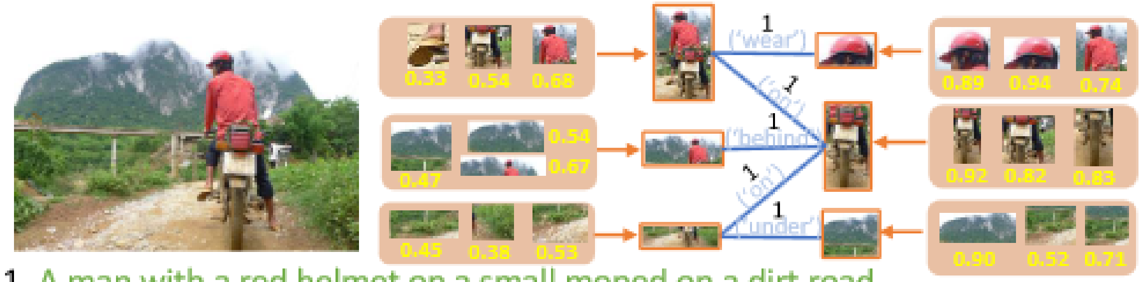


Wu et al. RGB-Infrared Cross-Modality Person Re-Identification. ICCV 2017

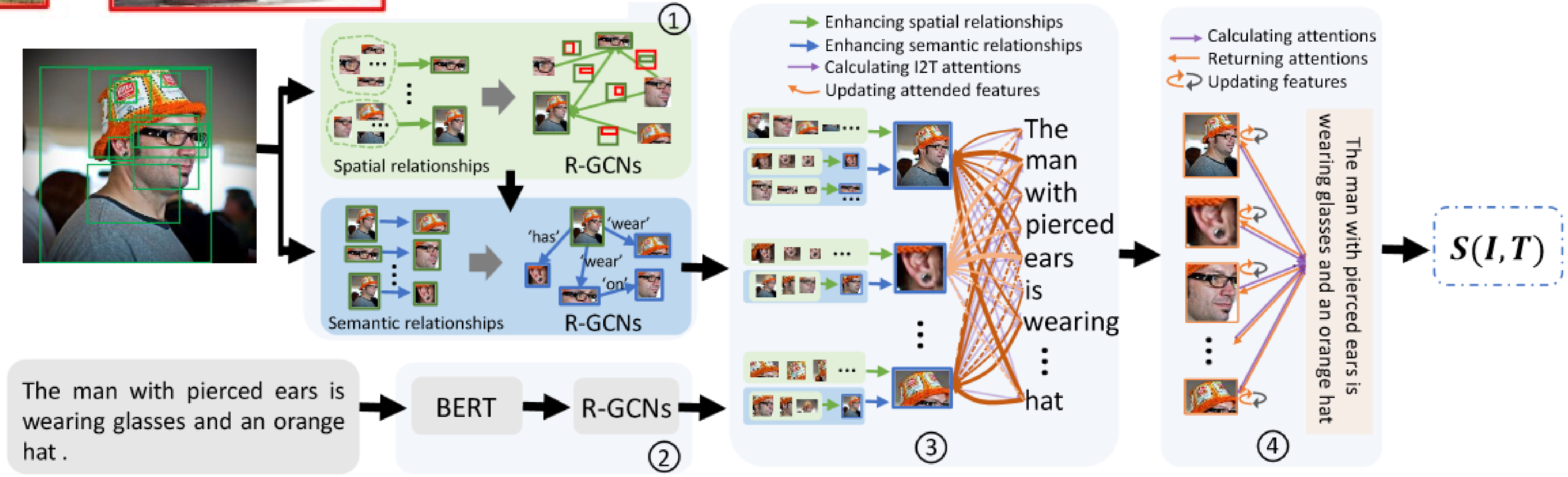
CROSS-MODAL RETRIEVAL

Query: There is a cow on the sidewalk standing in front of a door .





1. A man with a red helmet on a small moped on a dirt road .
2. A dirt path with a young person on a motor bike rests to the foreground of a verdant area with a bridge and a background of cloud-wreathed mountains .
3. A man in a red shirt and a red hat is on a motorcycle on a hill side .

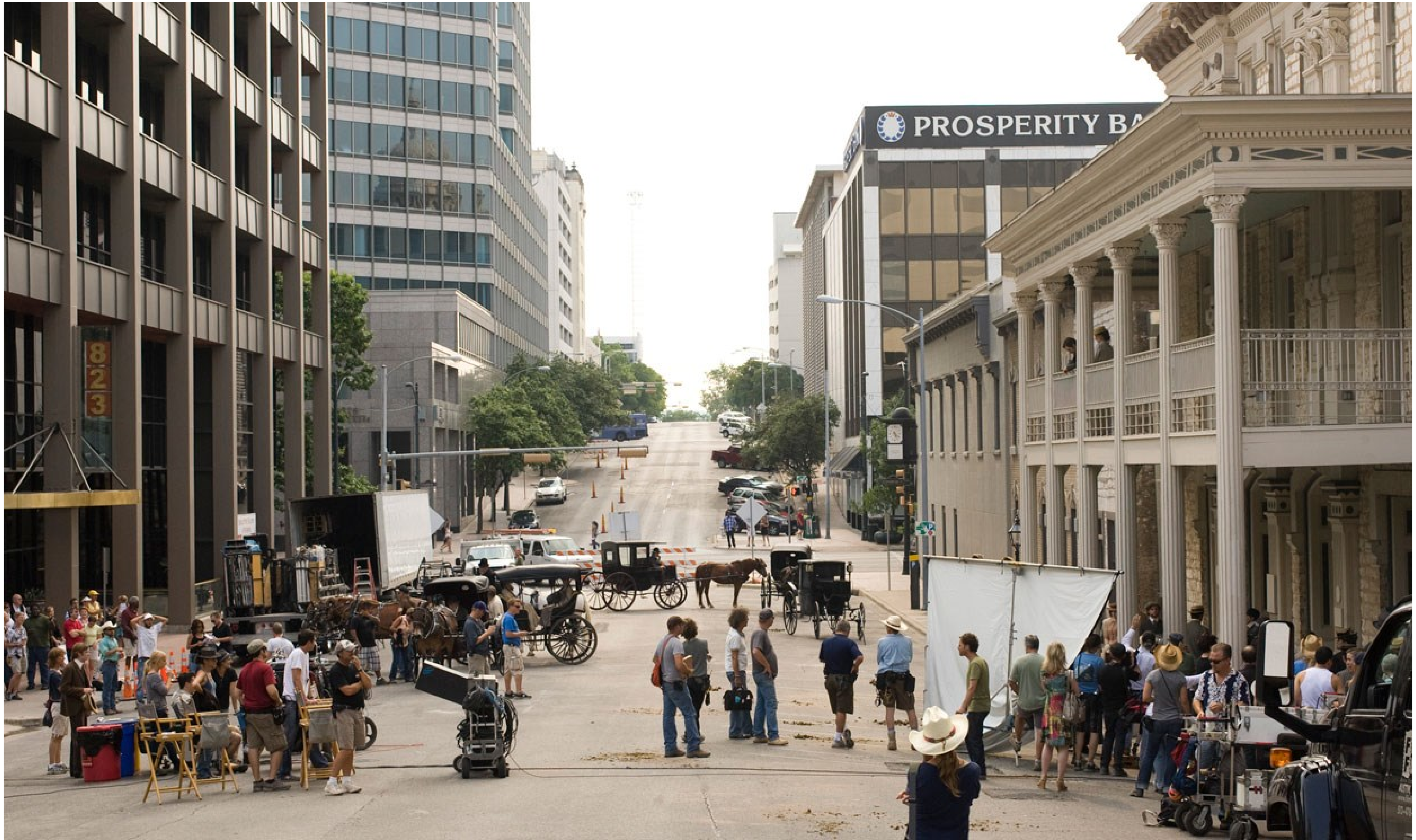


The man with pierced ears is wearing glasses and an orange hat .

Ge et al. Cross-modal Semantic Enhanced Interaction for Image-Sentence Retrieval. WACV2023.

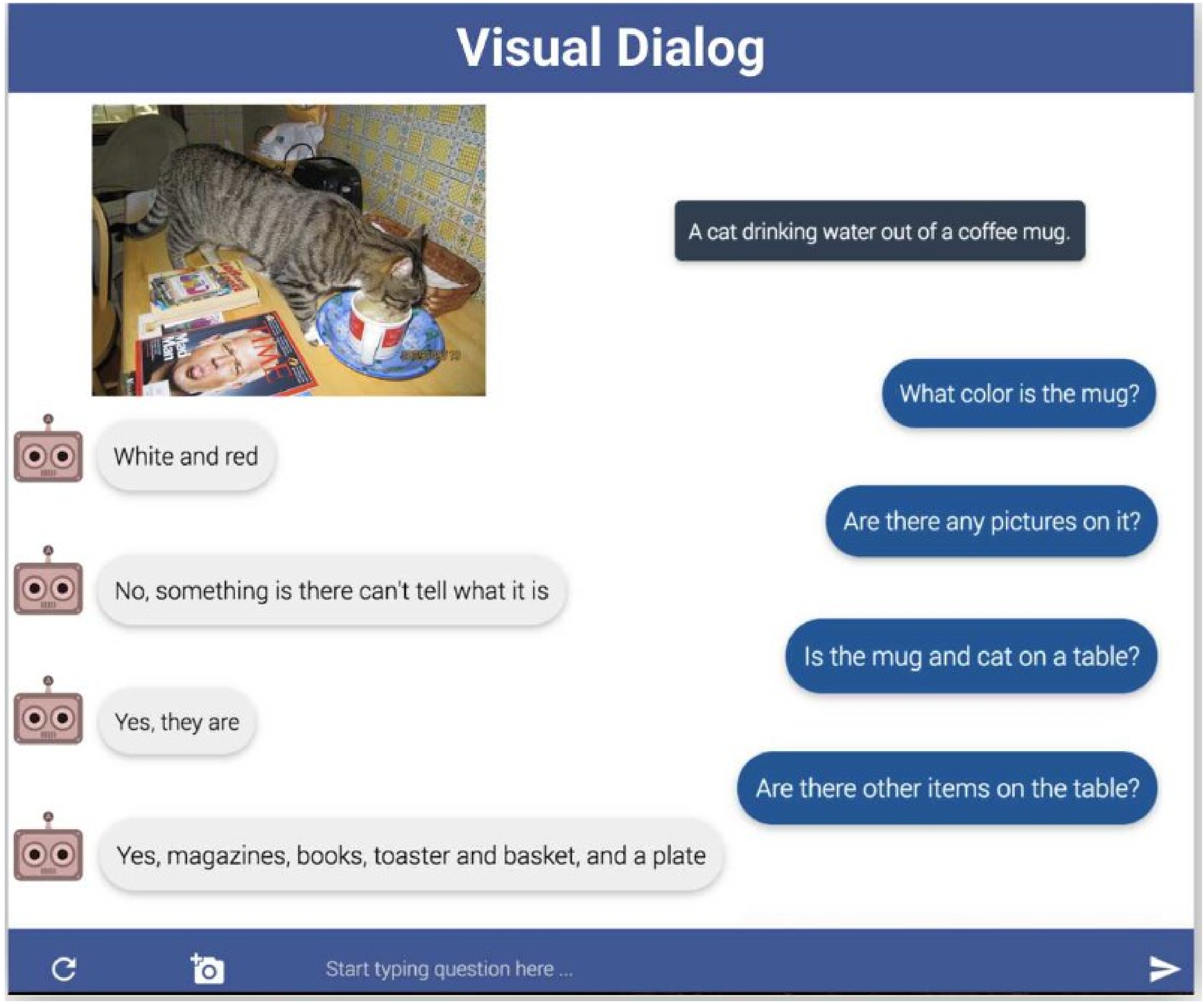
VISUAL QUESTION ANSWERING:

WHY IS THERE A CARRIAGE IN THE STREET?



Adriana Kovashka

VISUAL DIALOG



Visual Dialog

A cat drinking water out of a coffee mug.

What color is the mug?

White and red

Are there any pictures on it?

No, something is there can't tell what it is

Is the mug and cat on a table?

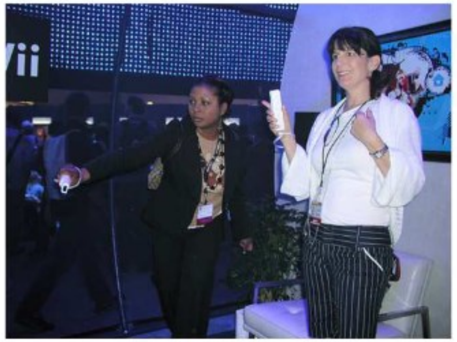
Yes, they are

Are there other items on the table?

Yes, magazines, books, toaster and basket, and a plate

Start typing question here ...

- VQA+
- Engaging in conversation about visual data with an intelligent system



Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right

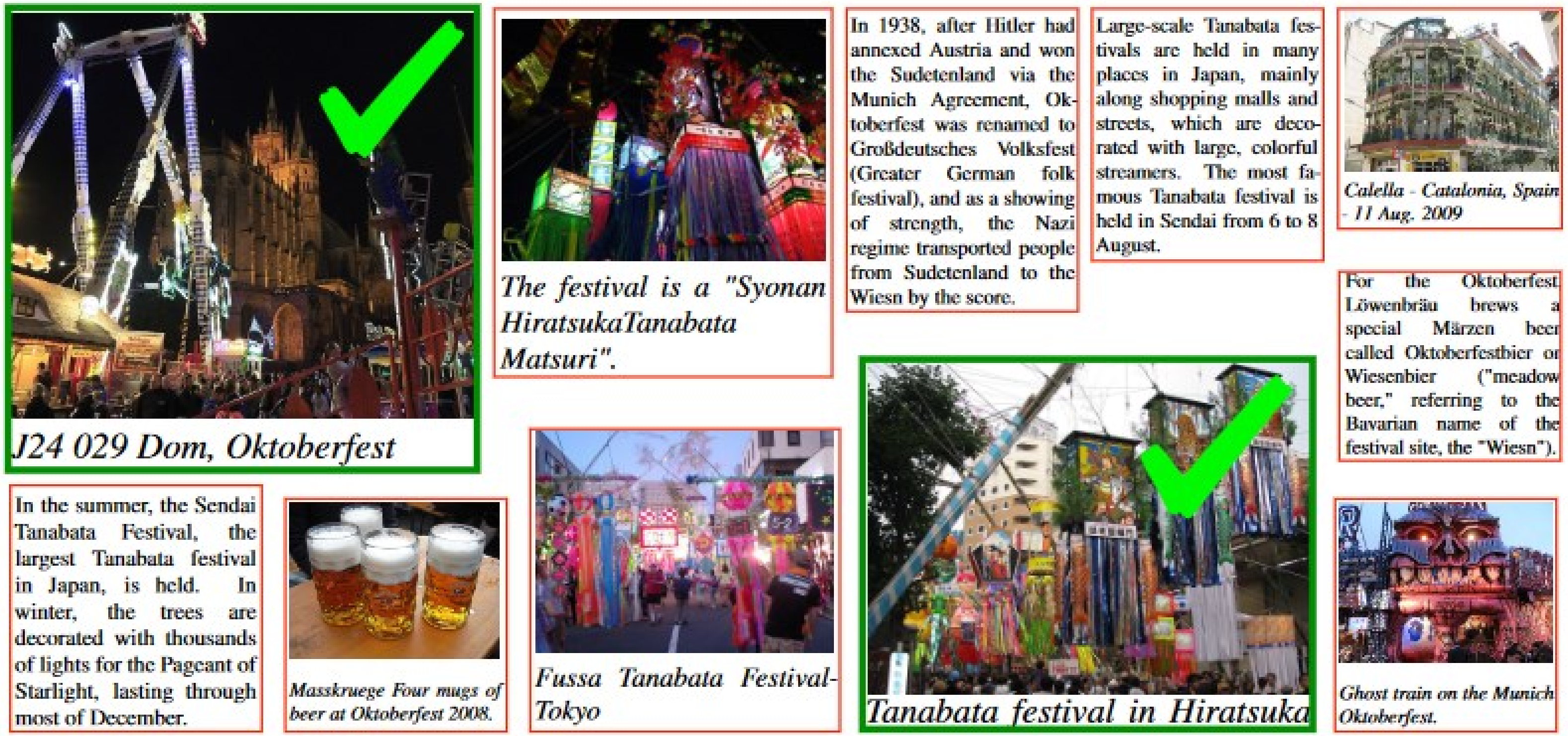
A: No, it's a woman

Das et al. CVPR 2017

MULTIMODAL QUESTION ANSWERING

- Answering textual questions which require reasoning over multimodal data

Q: At which festival can you see a castle in the background: Oktoberfest in Domplatz Austria or Tanabata festival in Hiratsuka, Japan?



J24 029 Dom, Oktoberfest

The festival is a "Syonan HiratsukaTanabata Matsuri".

In 1938, after Hitler had annexed Austria and won the Sudetenland via the Munich Agreement, Oktoberfest was renamed to Großdeutsches Volksfest (Greater German folk festival), and as a showing of strength, the Nazi regime transported people from Sudetenland to the Wiesn by the score.

Large-scale Tanabata festivals are held in many places in Japan, mainly along shopping malls and streets, which are decorated with large, colorful streamers. The most famous Tanabata festival is held in Sendai from 6 to 8 August.

Calella - Catalonia, Spain - 11 Aug. 2009

For the Oktoberfest Löwenbräu brews a special Märzen beer called Oktoberfestbier or Wiesenbier ("meadow beer," referring to the Bavarian name of the festival site, the "Wiesn").

In the summer, the Sendai Tanabata Festival, the largest Tanabata festival in Japan, is held. In winter, the trees are decorated with thousands of lights for the Pageant of Starlight, lasting through most of December.

Masskruege Four mugs of beer at Oktoberfest 2008.

Fussa Tanabata Festival-Tokyo

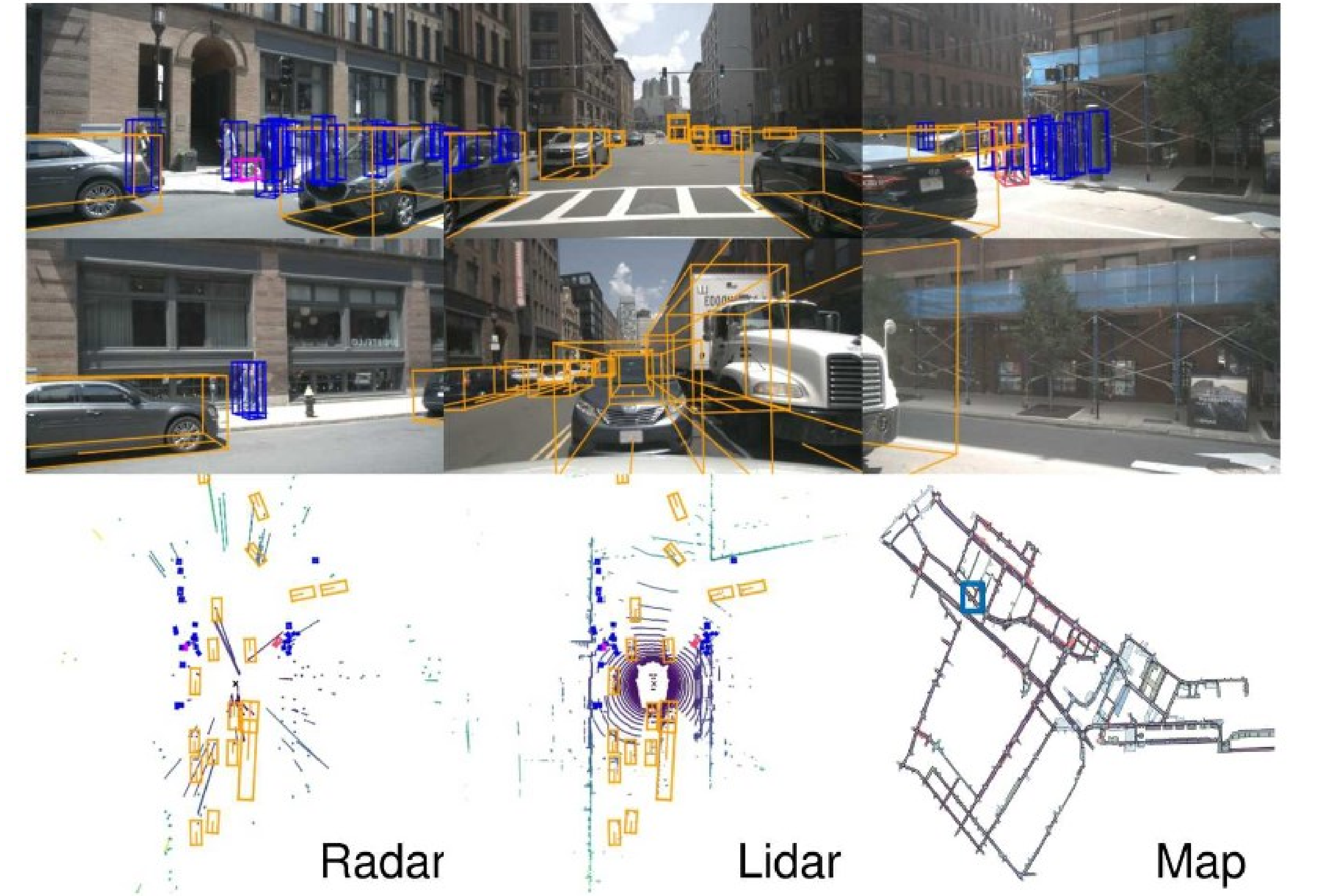
Tanabata festival in Hiratsuka

Ghost train on the Munich Oktoberfest.

A: You can see a castle in the background at Oktoberfest in Domplatz, Austria

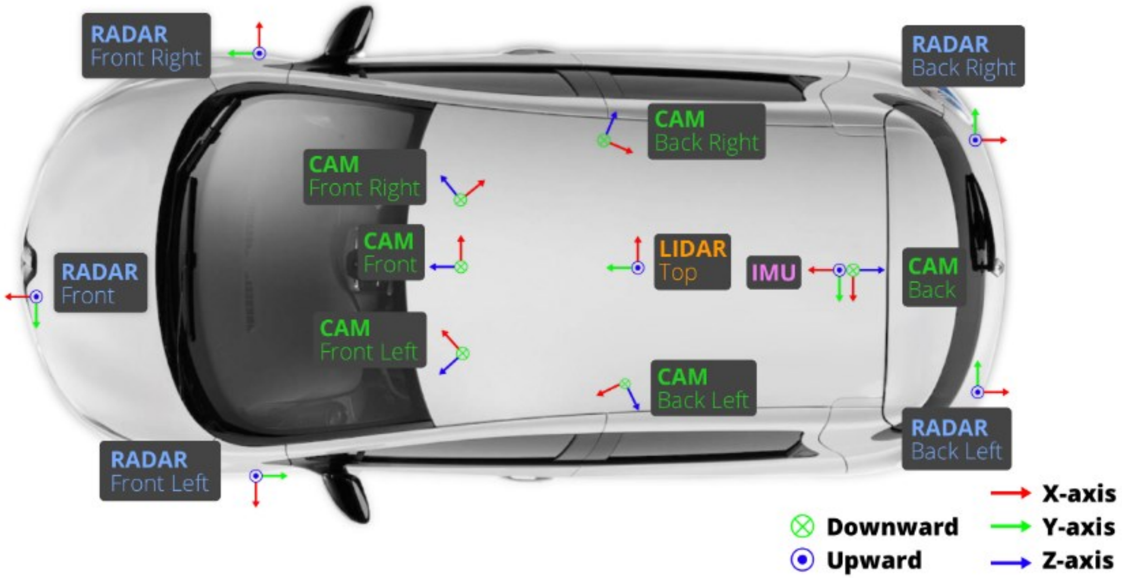
Chang et al. WebQA: Multihop and Multimodal QA. CVPR 2022.

VISION AND SENSOR DATA



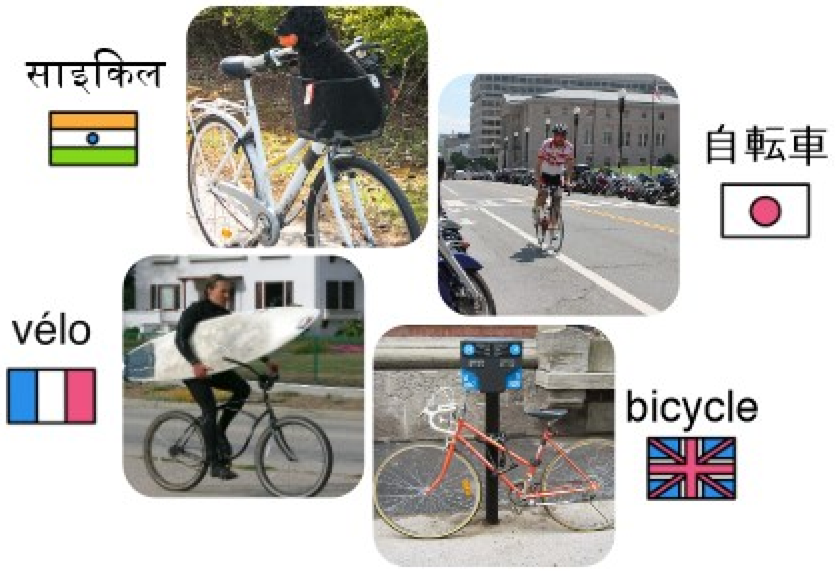
"Ped with pet, bicycle, car makes a u-turn, lane change, peds crossing crosswalk"

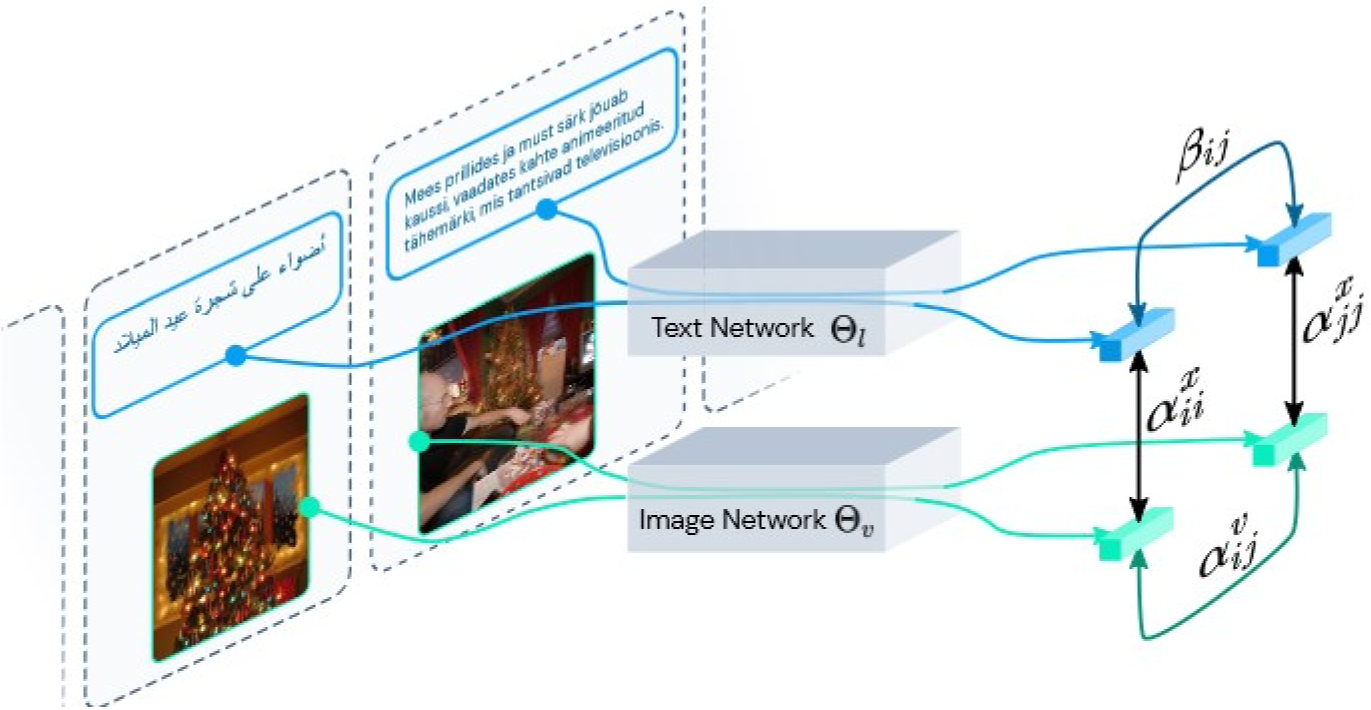
- How can a machine learning system reason across visual data (video feed) and additional sensor data (e.g. radar, Lidar, etc.)



nuScenes: A multimodal dataset for autonomous driving. Caesar et al. CVPR 2020.

IMPROVING NLP TASKS WITH VISION

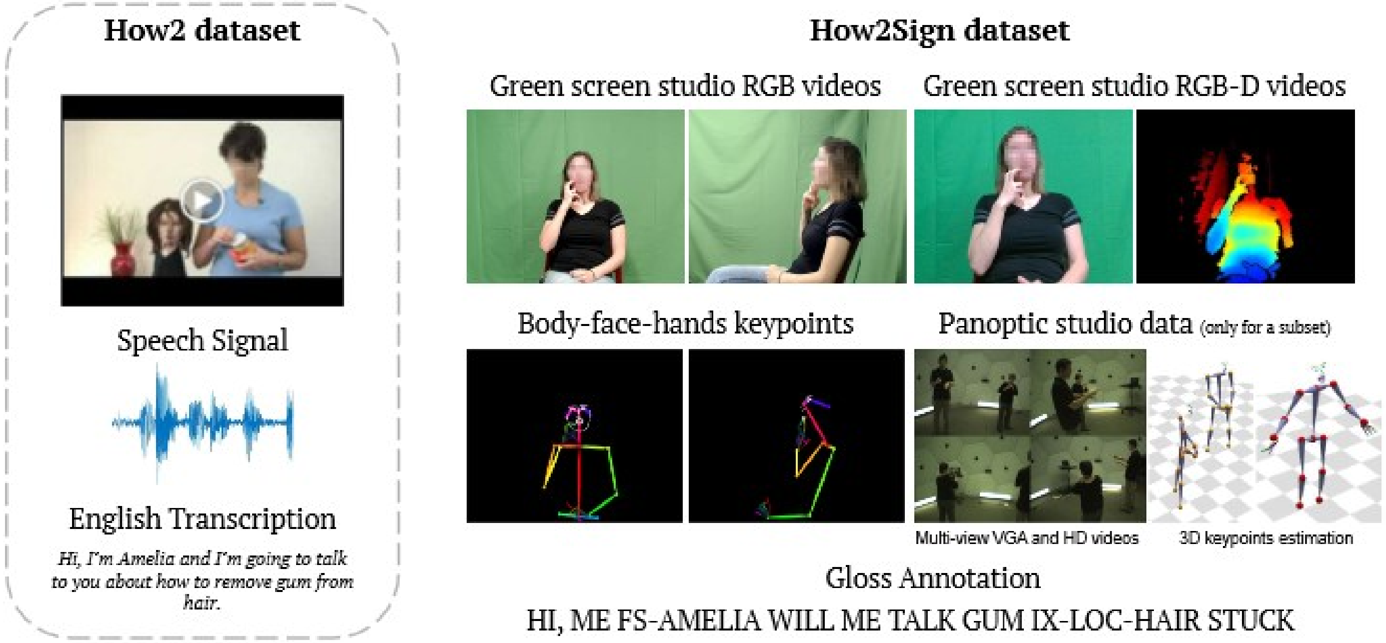


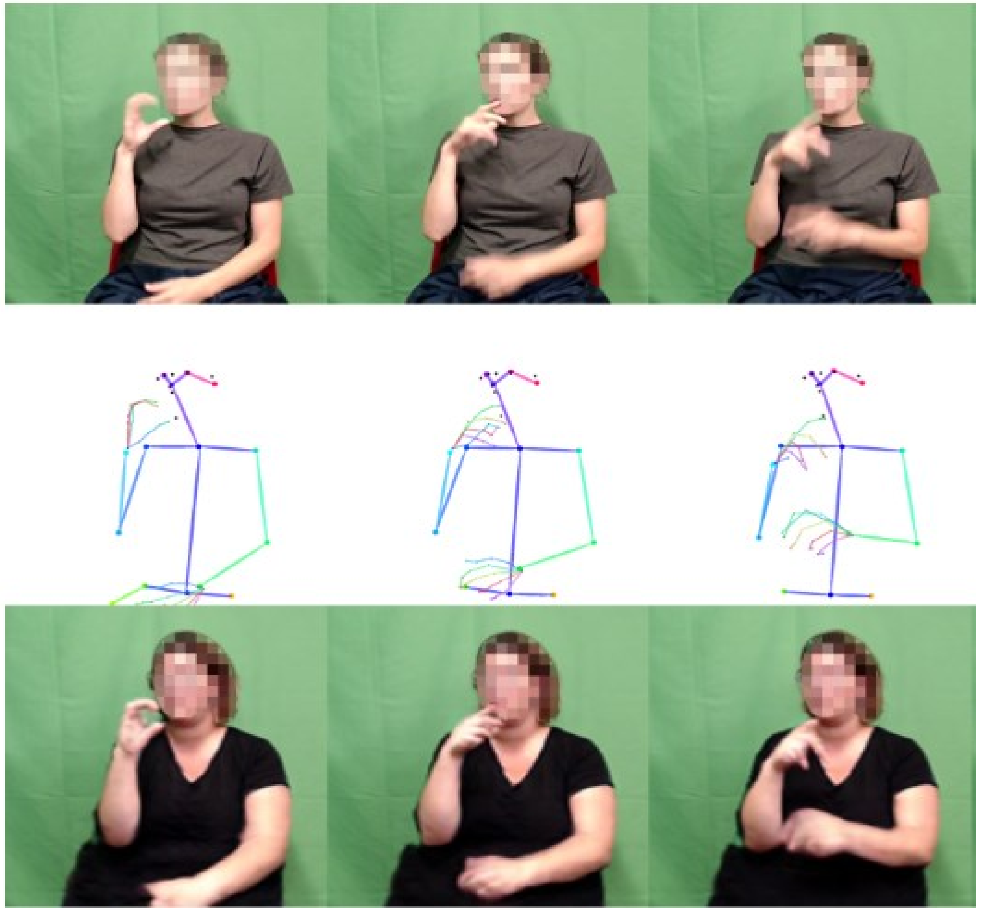

Mees prillides ja must särk jõuab kaussi, vaadates kahte animeeritud tähemärki, mis tantsivad televisioonis.
أضواء على شجرة عيد الميلاد
Text Network Θ_l
Image Network Θ_v
β_{ij}
α^x_{ii}
α^x_{jj}
α^v_{ij}

- Key observation: Bicycle is the same no matter what you call it! Use it to align different languages!

Suris et al. Globetrotter: Connecting Languages by Connecting Images. CVPR 2022 (oral).

AUTOMATIC SIGN LANGUAGE

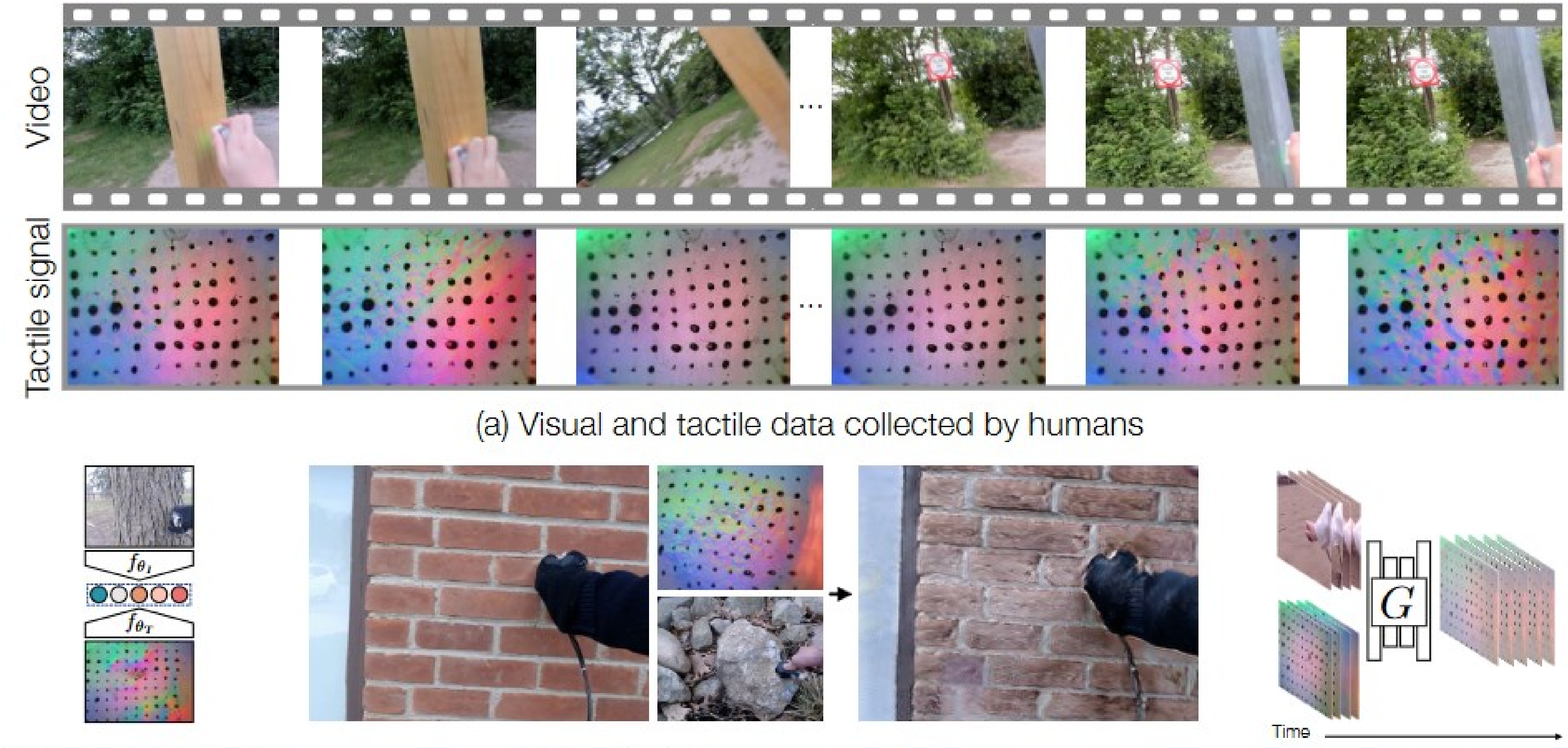

How2 dataset
Speech Signal
English Transcription
Hi, I'm Amelia and I'm going to talk to you about how to remove gum from hair.
How2Sign dataset
Green screen studio RGB videos
Green screen studio RGB-D videos
Body-face-hands keypoints
Panoptic studio data (only for a subset)
Multi-view VGA and HD videos
3D keypoints estimation
Gloss Annotation
HI, ME FS-AMELIA WILL ME TALK GUM IX-LOC-HAIR STUCK



- Multiple modalities of data! RGB videos, RGB-D videos, body keypoints, panoptic studio data, etc.

Duarte et al. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. CVPR 2021

EMBODIED VISION: VISION AND TOUCH



(a) Visual and tactile data collected by humans

(b) Self-supervision

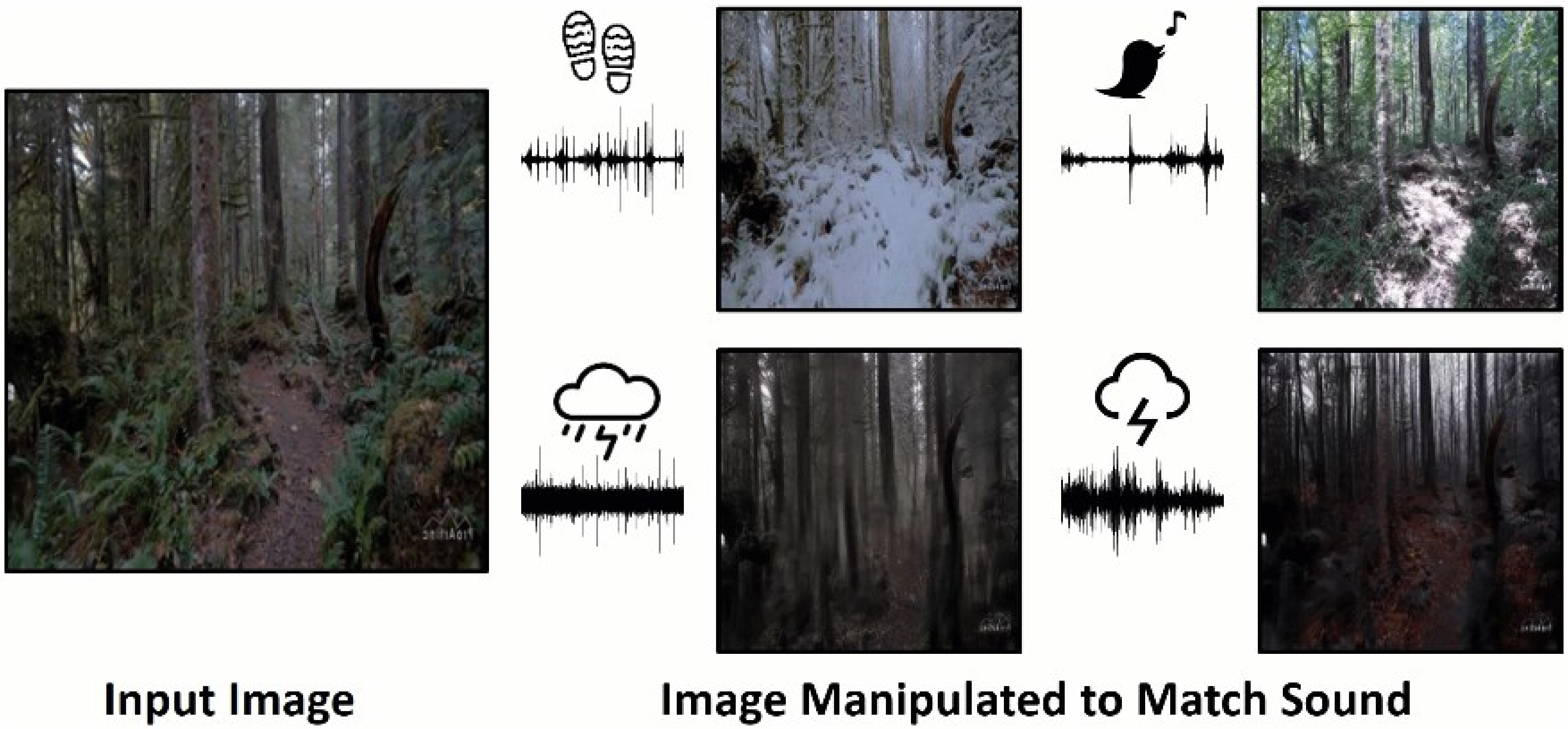
(c) Tactile-driven image stylization

(d) Future prediction

- Can we use knowledge of how objects feel to improve visual representations?

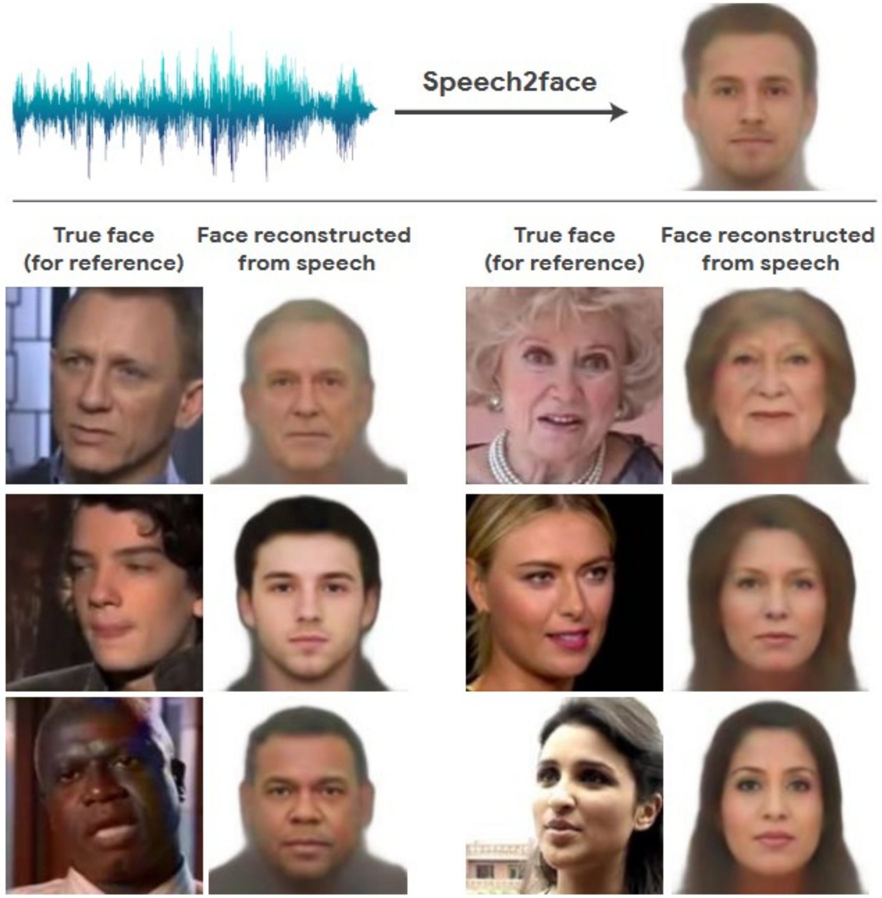
Yang et al. Touch and Go: Learning from Human-Collected Vision and Touch. NeurIPS 2022.

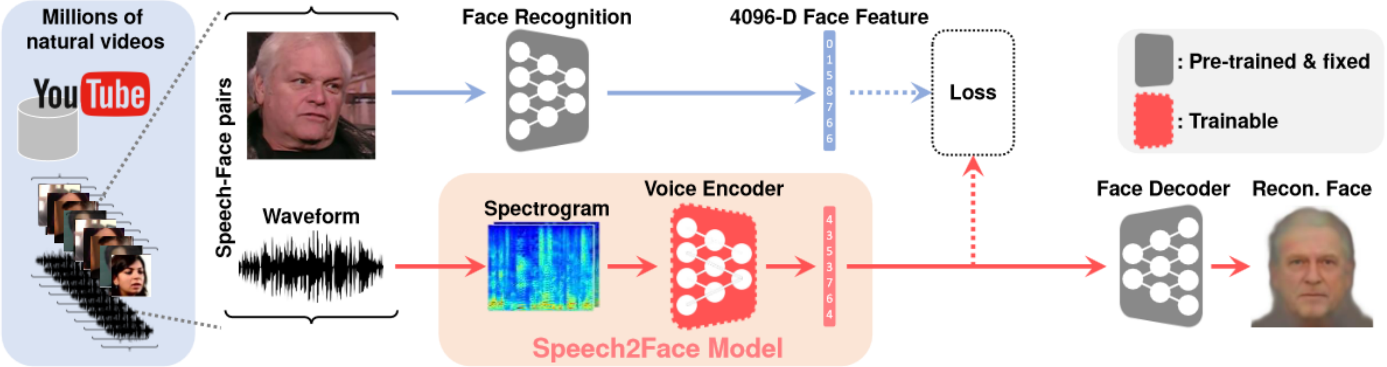
AUDIO-DRIVEN IMAGE STYLIZATION

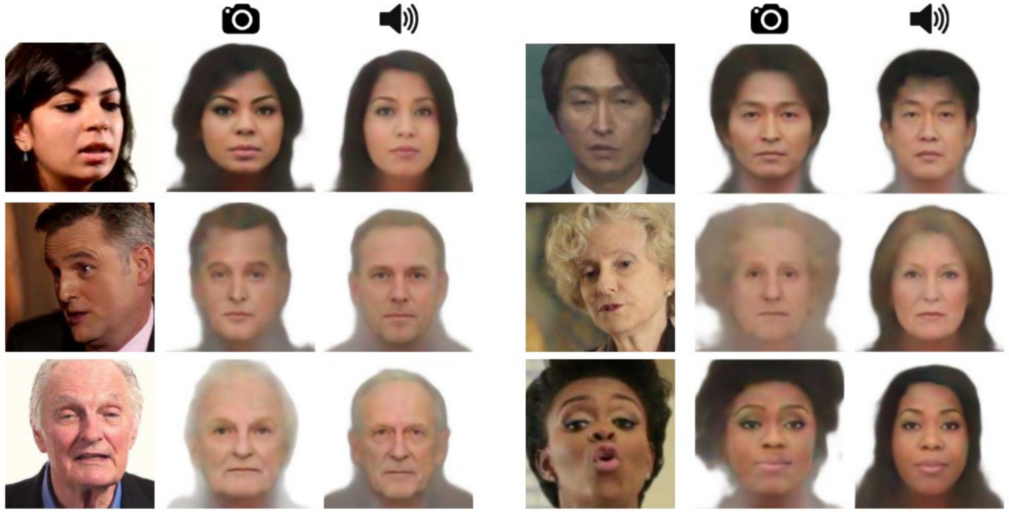


- Given a sound, how can we change an image to match it?
- Basic idea: Learn the visual patterns associated with different sounds and then transfer to image

IMAGE GENERATION FROM SPEECH


Speech2face
True face (for reference)
Face reconstructed from speech
True face (for reference)
Face reconstructed from speech

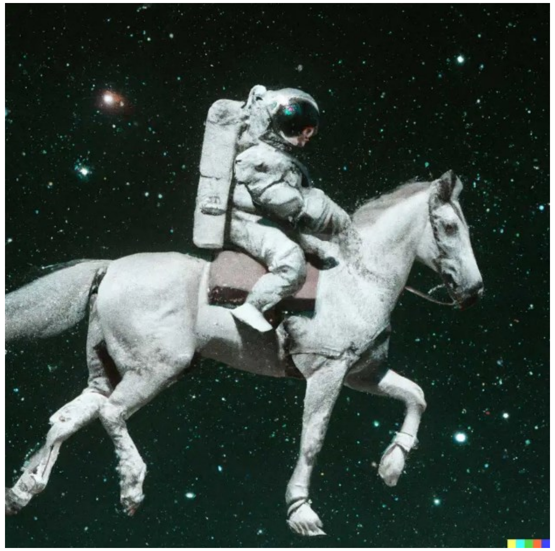




- Train models to reconstruct a person’s face from their voice

Oh et al. Speech2Face: Learning the Face Behind a Voice. CVPR 2019

IMAGE GENERATION FROM TEXT



A photo of an astronaut riding a horse.

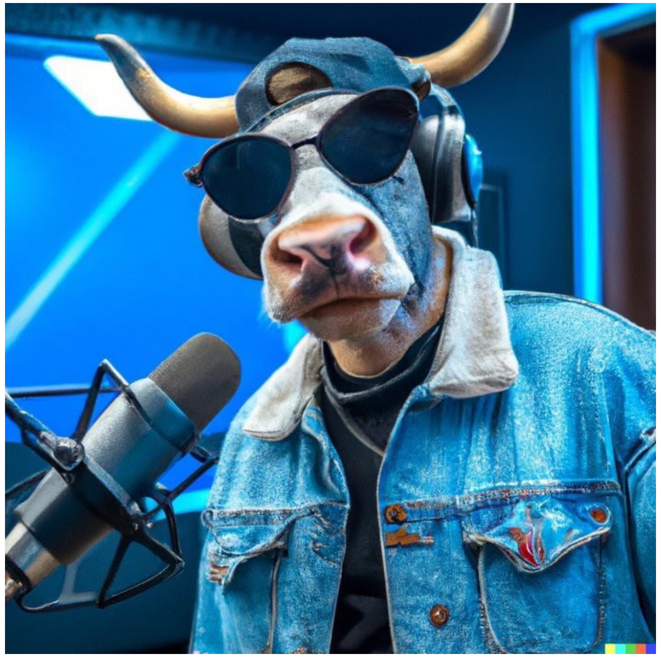
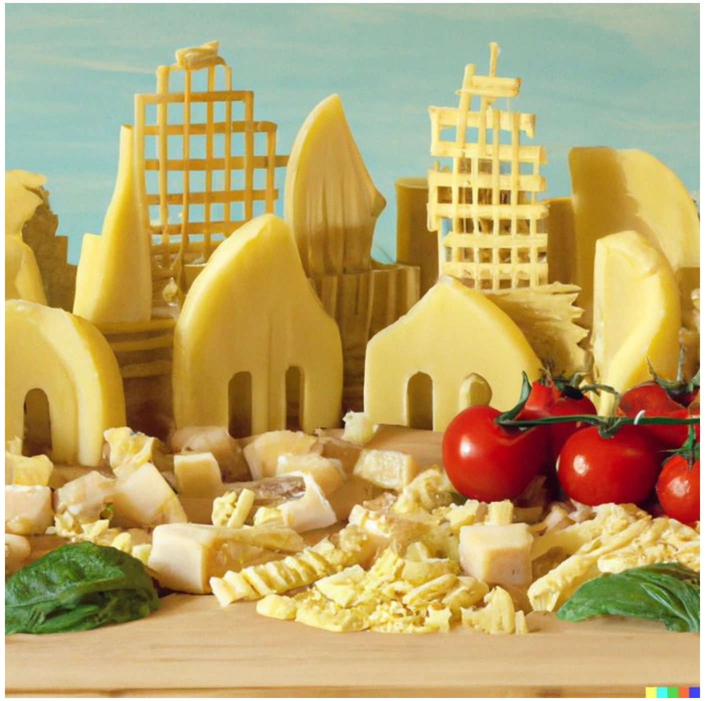


Photo of hip hop cow in a denim jacket recording a hit single in the studio



An italian town made of pasta, tomatoes, basil and parmesan

Images generated using Dall-E 2

https://qz.com/2176389/the-best-examples-of-dall-e-2s-strange-beautiful-ai-art

COURSE STRUCTURE

COURSE OBJECTIVES

- Learning about state-of-the-art methods in multimodal computer vision
- Learning to think critically about research
  - Applies beyond this class!
  - This involves developing the ability to critically assess research papers you encounter and to understand how different works are connected.
- Learning how to conceive novel ideas and extensions of existing research methods and implement your ideas
- Learning how to clearly write up and present research
- Learning how to work and collaborate with others in a research group

AM I PREPARED FOR THIS COURSE?

- This course will involve reading, understanding, writing about, and implementing ideas from recent research papers
- Ideally, you should have experience in deep learning and should have worked with a deep net before
- If you have taken machine learning courses and understand the basics of how models are trained, used, and evaluated, you should be ok
- If you are able to understand the initial reading reasonably well, you should be OK for the rest of the course
- However, if you are concerned about your ability to complete the course, talk to me sooner than later
    - Course schedule still needs to be finalized depending on number of class presentations and topics

COMPONENTS OF YOUR FINAL GRADE

- Final project – 45% of your final grade
  - Project proposal (5% of final grade)
  - Project status report (5% of final grade)
  - Project presentations (15% of final grade)
  - Project final report (20% of final grade)

- In-class participation and discussion – 15%

- Paper presentation – 20% of final grade

- Paper reviews – 20% of final grade

PAPER REVIEWS

- Students will be required to critically assess and review multimodal computer vision research papers as part of this course
- Required to write one paper review for each class presentation
  - First presentation is January 25th
  - Presenting students don't have to submit
- Must review the primary paper
- Paper reviews should be 1-2 pages long, single-spaced
- Refer to the CVPR reviewer tutorial slides (linked on website) for details on what makes a good review
- Upload paper reviews to Canvas by 10:00 PM the day before the class it is being presented in.
- Late submissions are accepted until 12:00 PM the day of class, with penalty.
  - You get three free late days (submitting late without penalty)

PAPER REVIEWS - CONTENTS

- **Summary**
  - Explaining what the paper is trying to do and how the paper proposes to do it
  - Primary novelty and contributions of the paper, rather than unimportant details.
    - ➢ E.g. summary of a new model architecture or loss function
  - How the method is experimentally evaluated and any significant findings or results
- **Relation to prior work**
  - Next summarize the paper's relation to prior work and why its contributions are (or are not) significant
- **Strengths**
  - The review should mention **at least three** strengths of the approach.
- **Weaknesses**
  - What do you feel detracts from the paper's contributions? Your review should mention **at least three** weaknesses.
- **Future work**
  - Propose at least one possible extension of the paper. This might be a fix to a weakness you identified (e.g. a modified model or loss function) or you might propose how the techniques developed in the paper could be applied in some novel way for a different task. You should not, however, simply rephrase or repeat the future work suggested by the paper itself.
- An example review by me (will be uploaded to Canvas)

PARTICIPATION AND ATTENDANCE

- You are expected to come to each class having carefully read the assigned papers and prepared with any questions
- You should actively participate in this class by asking meaningful questions of the speaker
- Use your paper reviews as a guide to contribute to discussions and make comments about the paper's strengths and weaknesses
- Questions asked by others in the class are addressed to the class and other class members should jump in and contribute if they know the answer to a question
- The paper presenter is expected to lead and moderate the discussion

PAPER PRESENTATIONS

- Each student in the class will give 1-2 presentations (depending on class schedule)

- Each presentation will cover one primary paper and at least one background paper

- The paper presentation should be 45 minutes long and should be highly polished and well rehearsed
  - **Practice your presentation!!!!**

- After the presentation, you will moderate a 20-25 minute discussion session
  - You should prepare topics for discussion and drive the discussion!

PAPER PRESENTATIONS

- Paper presentations should, at a minimum,
  - Clearly define what problem the paper is addressing.
  - Provide motivation for why the problem is important, interesting, and/or challenging.
  - Address prior related work that has attempted to address this problem (or a related problem).
  - Describe, in detail, the proposed approach for the problem. Explain how the paper is evaluated.
  - Discuss key strengths and weaknesses of the paper.
  - Propose ideas for future work and identify any open research questions.

PAPER PRESENTATIONS

- Key point: You are free to use slides and resources from the internet
  - Hint: Authors often publish their slides or you might find them on the conference site
- **You must clearly cite all your sources and give slide credits!**
  - Make sure to use your own words
- Slides shouldn’t just contain walls of text (this presentation isn’t a good example!)
- Slides will be made available on the class web page
- You must upload your presentation slides to Canvas by 10:00 PM on the day before the intended class presentation.
- Students will complete *peer reviews* at the end of every class session

FINAL PROJECT

- This course has no exams or programming homework assignments
- Instead, you will complete a student-driven group project
  - This is *your* project. Have fun with it! I encourage you to do something creative.
- A report will be due at the end of the course
- You must work in a group of 3-4 students.
  - More is expected of larger groups!
- The final report **should be highly polished** and resemble a conference paper like those you have read in this class.

FINAL PROJECT

- The final project is open-ended, as long as it leverages multimodal computer vision.
- Final projects should fall into one of these categories:
  - Extend one of the papers we covered in class in a significant way, complete with a thorough experimental evaluation;
  - Propose a novel method or approach for solving a multimodal vision problem we discussed in class or that is already known in the literature and thoroughly evaluate it; or
  - Propose a completely new multimodal vision problem and explain why it is significant and needs solving, implement an approach to solve the problem, and evaluate the approach
- All projects must be thoroughly experimentally validated

FINAL PROJECT COMPONENTS

- Project proposal (5% of final grade) - due March 3rd, 10:00 PM
  - You should form your groups as soon as possible, noting that there may be a few add/drops
  - You should begin thinking about your final project ***today!***
    - ➢ Even from this short introduction talk, you may start forming ideas about topics that interest you
    - ➢ More technical details will come later on during paper presentations
- Project status report (5% of final grade) - due April 7, 10:00 PM
- Project presentations (15% of final grade) - *to be determined* (last classes of term)
- Project final report (20% of final grade) - due 9:45 AM, May 8
- Everything submitted on Canvas

PROJECT PROPOSAL – MARCH 3, 10:00 PM

- **Advice**: Don’t underestimate the project proposal!
  - Your proposal needs to be 3-5 pages (excl. references) long in CVPR format
- It should have:
  - A clear problem statement which describes the goal of the project.
  - A thorough literature review which shows how yours differs.
  - A detailed description of the proposed approach. The authors should describe new loss functions they plan to use, changes to existing models, etc.
  - The proposed experimental evaluation protocol and expected results.
- You are encouraged to discuss your plans before proposals are due during office hours

PROJECT STATUS REPORT –APRIL 7, 10:00 PM

- The status report iterates on your proposal to describe what you have completed already.
- It should bring your proposal draft text closer in line with your final report
  - You can re-use text from your proposal
- 3-5 pages (excl. references)
- Must include:
  - Introduction
  - Related Work
  - Approach
  - Results
- Your status report should start to resemble a conference paper, except should describe your group's progress

FINAL PROJECT PRESENTATIONS - TBD

- Final project presentations will be conducted at the end of the term
    - Projects will be presented as groups, with group members deciding how to divide up the presentation
- Should address the same points as paper presentations
- Length is TBD (determined based on number of groups)
- Should be well-rehearsed by your group, clearly explained, and polished

FINAL REPORT

- The final report should resemble a CVPR conference paper (8 pages of content, excluding references)
- That means having polished figures, tables, qualitative results, etc.
- **Your final report should have the same quality of presentation as other papers you have read in this class**
- Must follow traditional CVPR outline:
  - Abstract
  - Introduction
  - Related work
  - Approach
  - Results
  - Conclusion
- After the report, write a page where each student in the group documents everything they contributed to the project and how work was divided.

(UNGRADED) HOMEWORK

- By 10:00 PM tomorrow (1/19):
  - Go through the list of topics on the course webpage
  - Rank each topic based on most interesting to you to least interesting
  - If there are any topics you would like to present, feel free to list those and any papers
- Conference papers you suggest should come from CVPR, ICCV, ECCV, NeurIPS, or ACL
- No guarantees whether we will be able to cover your paper suggestions or topics
- I will try to match students with the topics that most interest them
- Also, need a volunteer for the first paper presentation: Jan. 25th
  - Will be on multimodal architectures
- Will finalize the presentation schedule ASAP

QUIZ – TO ASSESS EVERYONE'S STRENGTHS

- 1. Describe two loss functions used for aligning multimodal representations and explain how they works (1-2 sentences each)
- 2. Imagine I wish to detect objects in an image my model has never seen training examples of.
  - What is this type of object detection called?
  - How might I perform it?
- 3. Name two popular computer vision architectures (transformer-based or CNN).
- 4. Name a multimodal vision model architecture.
- 5. Name a popular text-only transformer-based architecture.
- 6. What's the difference between self-attention, co-attention, and cross-attention?
- 7. What do you usually have to do to an image before running it through a CNN?
- 8. What is overfitting and what are some strategies to overcome it?