Computer Vision: A journey in time

Computer Vision: A journey in time

Everybody Dance Now

Caroline Chan* Shiry Ginosar Tinghui Zhou† Alexei A. Efros

UC Berkeley

Figure 1: **"Do as I Do" motion transfer:** given a YouTube clip of a ballerina (top), and a video of a graduate student performing various motions, our method transfers the ballerina's performance onto the student (bottom). Video: `https://youtu.be/mSaIrz8lM1U`

Autotuner for dancing

Everybody dance now

C Chan, S Ginosar, T Zhou... - Proceedings of the IEEE ..., 2019 - openaccess.thecvf.com

This paper presents a simple method for" do as I do" motion transfer: given a source video of a person dancing, we can transfer that performance to a novel (amateur) target after only a ...

Computer Vision: A journey in time

Everybody Dance Now

Caroline Chan[*] Shiry Ginosar Tinghui Zhou[†] Alexei A. Efros

UC Berkeley

Autotuner for dancing

Figure 1: **"Do as I Do" motion transfer:** given a YouTube clip of a ballerina (top), and a video of a graduate student performing various motions, our method transfers the ballerina's performance onto the student (bottom). Video: `https://youtu.be/mSaIrz8lM1U`

Everybody dance now

C Chan, S Ginosar, T Zhou… - Proceedings of the IEEE …, 2019 - openaccess.thecvf.com

This paper presents a simple method for" do as I do" motion transfer: given a source video of a person dancing, we can transfer that performance to a novel (amateur) target after only a …

☆ Save 99 Cite Cited by 818 Related articles All 7 versions Import into BibTeX

Computer Vision: A journey in time

Last week

Course Project: Reality Checks

For the examples shown before

- Usually resource hungry (data, compute)
- Substantial literature review (reading)
- A lot of trial and error (coding, debugging)
- Enormous time commitment

What you should consider first!

- How much compute do you have?
- How much time can you allocate as a group?
- Group size matters!
- How well does the literature fit into your background?

Course Project: Reality Checks

For the examples shown before

- Usually resource hungry (data, compute)
- Substantial literature review (reading)
- A lot of trial and error (coding, debugging)
- Enormous time commitment

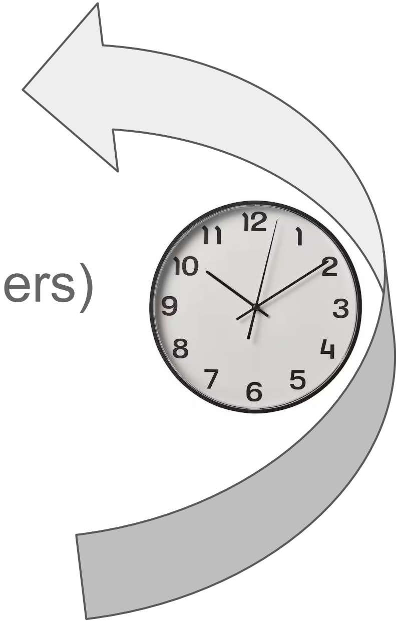
What you should consider first!

- How much compute do you have?
- How much time can you allocate as a group?
- Group size matters!
- How well does the literature fit into your background?

It is not uncommon for course projects - with more refinement at a later time; to get published

For the examples shown before

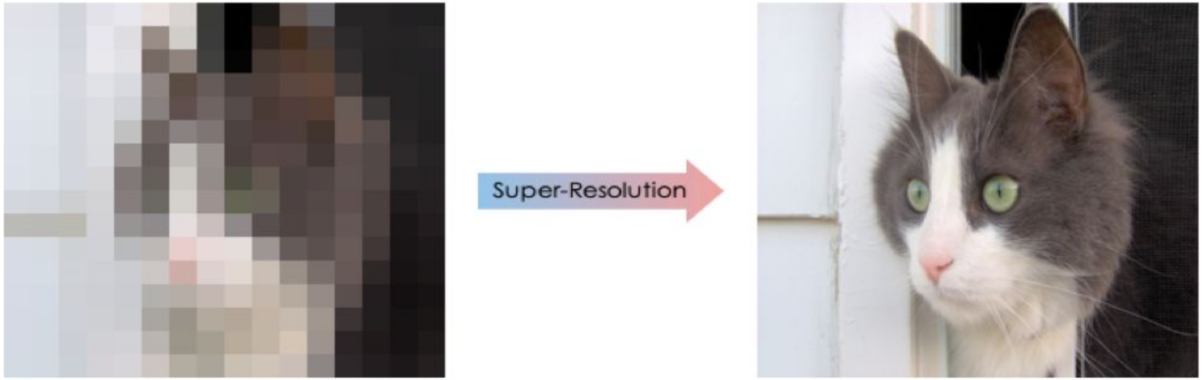
- Usually resource hungry (data, compute)
- Substantial literature review (reading)
- A lot of trial and error (coding, debugging)
- Enormous time commitment (big lab of researchers)



What you should consider first!

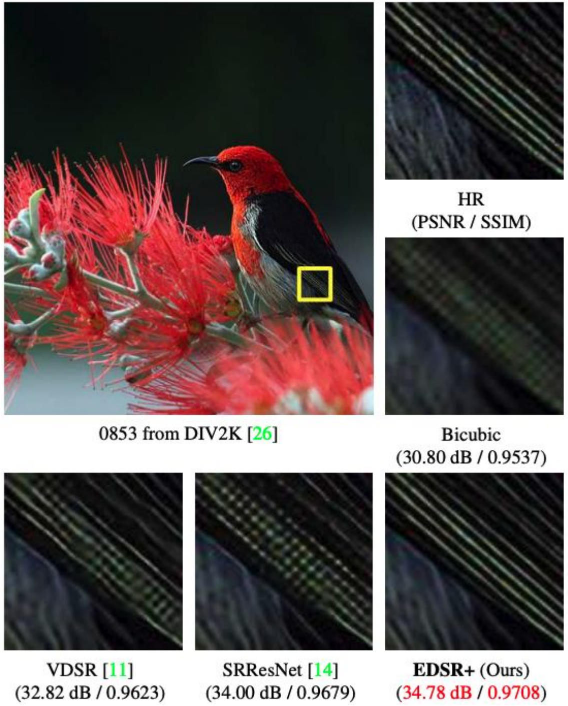
- How much compute do you have?
- How much time can you allocate as a group?
- Group size matters!
- How well does the literature fit into your background?

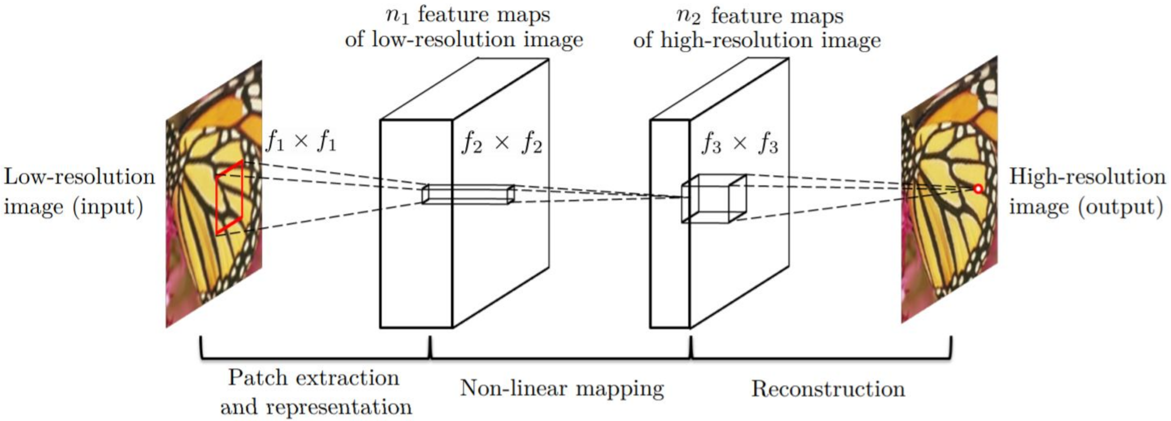
Seed Idea: Single Image Super Resolution



- Reconstructs a high-resolution (HR) image from a low-resolution (LR) image
- Applications: medical imaging, security, and surveillance
- One-to-many mapping relation to recover HR images from a LR image
- An ill-posed and still challenging problem in the community

Seed Idea: Single Image Super Resolution



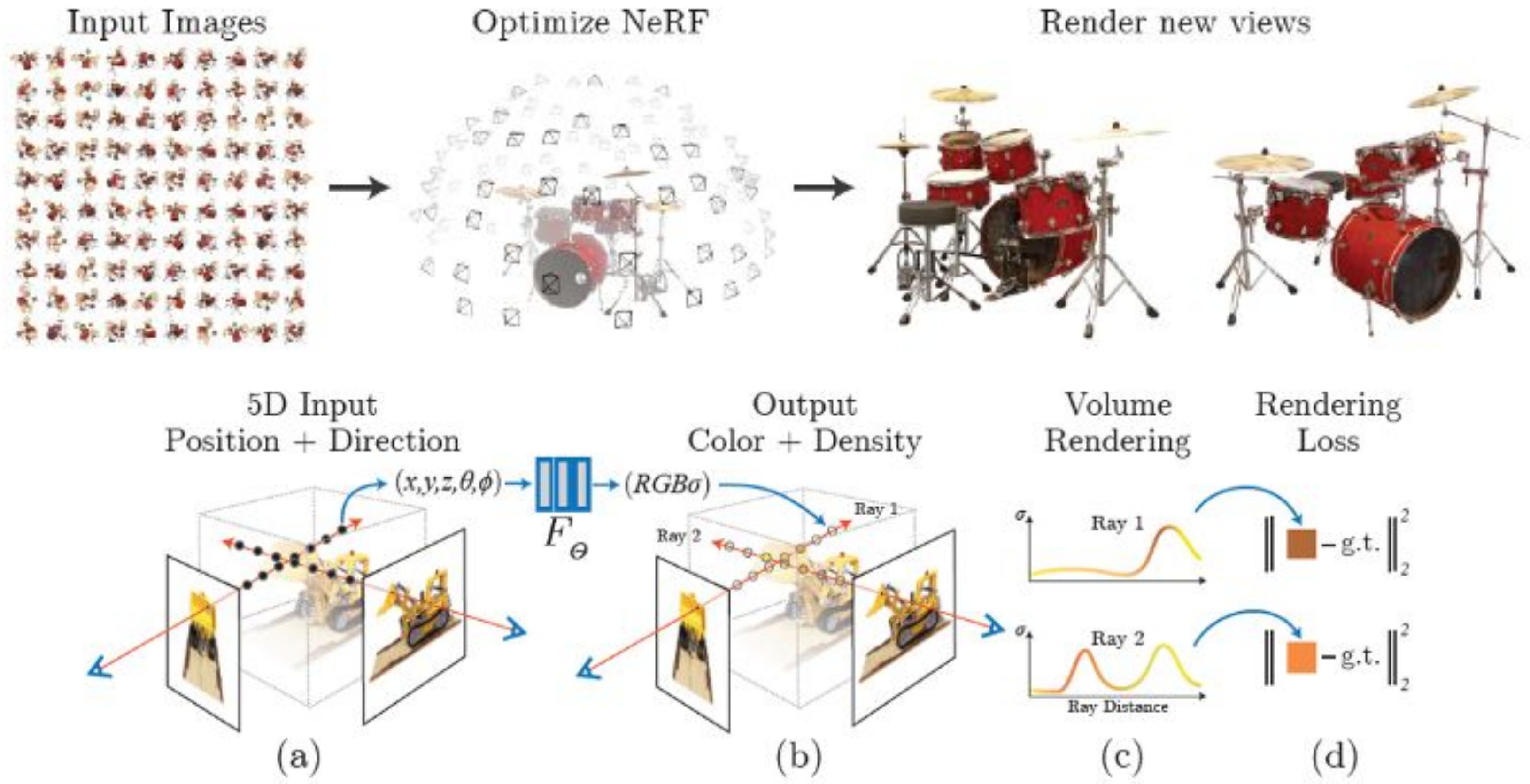


Source: https://openaccess.thecvf.com/content_cvpr_2017_workshops/w12/papers/Lim_Enhanced_Deep_Residual_CVPR_2017_paper.pdf

Seed Idea: NEural Radiance Fields (NeRF)

- Creating a 3D view from a series of 2D images. (View Synthesis)
- Creating realistic visual effects, simulations, and scenes
- Volume rendering enables you to create a 2D projection of a 3D discretely sampled dataset.

Seed Idea: Neural Radiance Fields (NeRF)



Seed Idea: Improving Spatial Reasoning in
- Stable Diffusion / T2I models / GenAI for vision

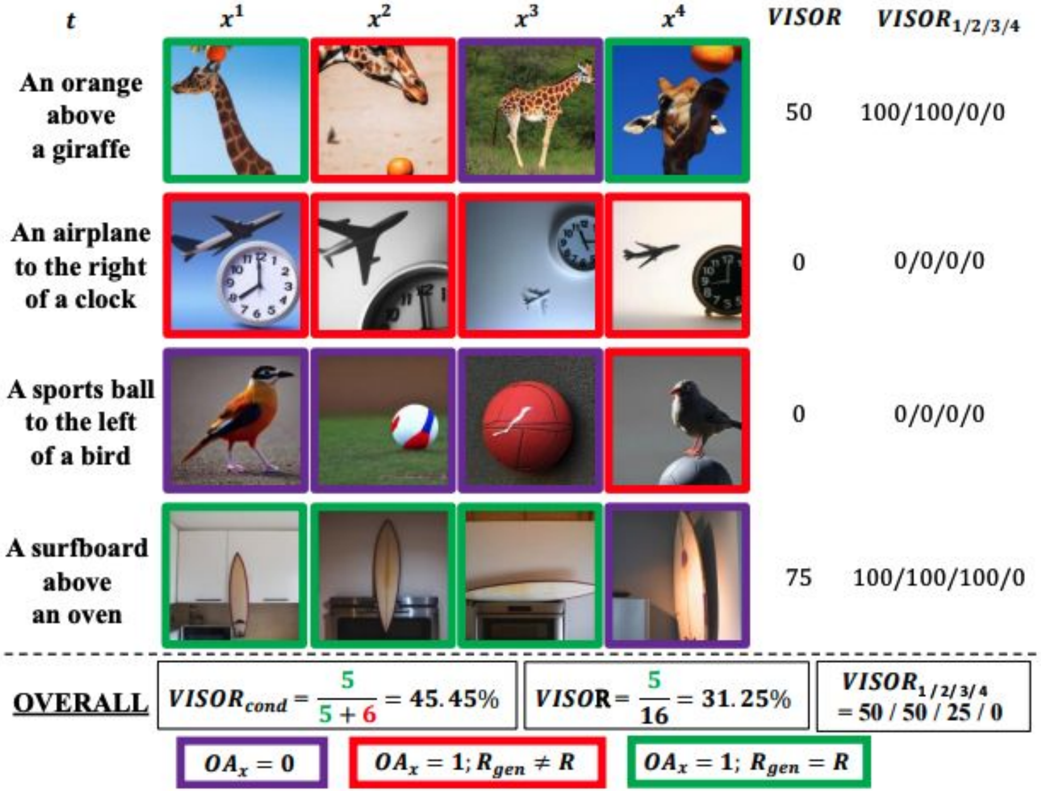
Benchmarking Spatial Reasoning Abilities of Text-to-Image Generative Models

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, Yezhou Yang

https://arxiv.org/pdf/2212.10015.pdf

- Large scale spatial reasoning Dataset SR2D: image, text pair describing two or more objects and the spatial relationships between them with linguistic variations.
- Metric to quantify visual reasoning performance: VISOR

Seed Idea: Improving Spatial Reasoning in
- Stable Diffusion / T2I models / GenAI for vision



Seed Idea: Improving Spatial Reasoning in
- Stable Diffusion / T2I models / GenAI for vision

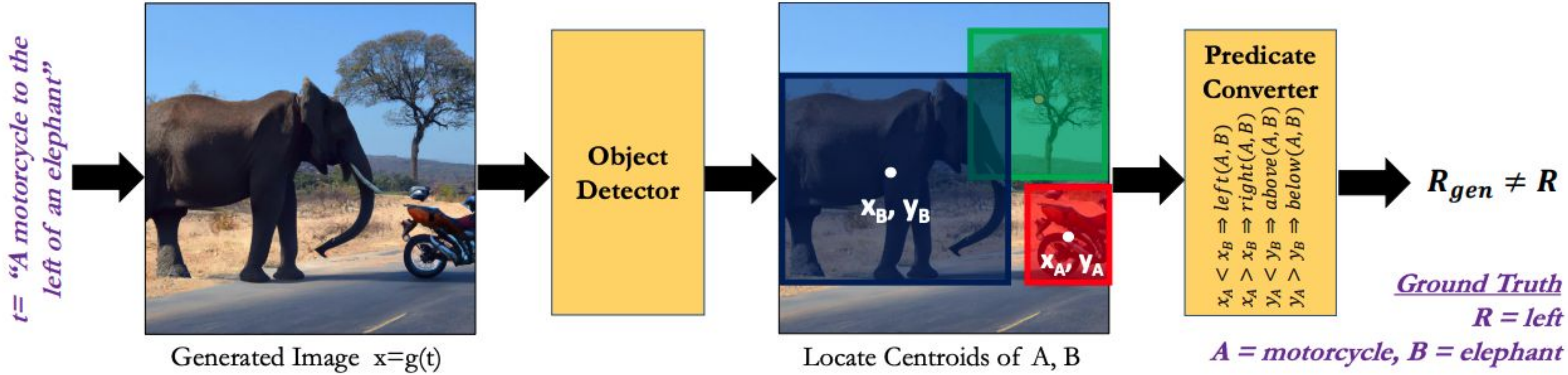


Fig. 3: For text t and corresponding generated image $x = g(t)$, object centroids are located and converted into predicates indicating the spatial relationship between them. These predicates are compared with the ground truth relationship R to obtain the VISOR score.

Seed Idea: Improving Spatial Reasoning in
- Stable Diffusion / T2I models / GenAI for vision

(1) Rate the **quality** of the image. ○1 ○2 ○3 ◉4 ○5

("1" being artificial (e.g. a sketch or cartoon) and "5" being natural (a real photograph))

(2) How likely is the scene to occur in **real life ?** ○1 ○2 ○3 ○4 ◉5

(Rate from "1" (least likely) to "5" (most likely)

(3) How many objects are in this image? 3

(4) Object A: The image contains a wine glass ◉ True ○ False

(5) Object B: The image contains a sandwich ◉ True ○ False

(6) Choose the spatial relationship between the wine glass and the sandwich.

Multiple Options may be possible. If there are more instances of the same type (example: two dogs and one cat) then select all possible relationships between each dog and the cat. ***[IMPORTANT]*** *Choose "N/A" if you answered "False" for either question (2) or (3)*

- ☑ wine glass to the left of sandwich
- ☐ wine glass to the right of sandwich
- ☐ wine glass above sandwich
- ☐ wine glass below sandwich
- ☐ N/A

(7) If you answered **True** for both (4) and (5), are the two objects merged or distinct ○ Merged ◉ Distinct ○ N/A

[IMPORTANT] *Choose "N/A" if you answered "False" for either question (2) or (3*

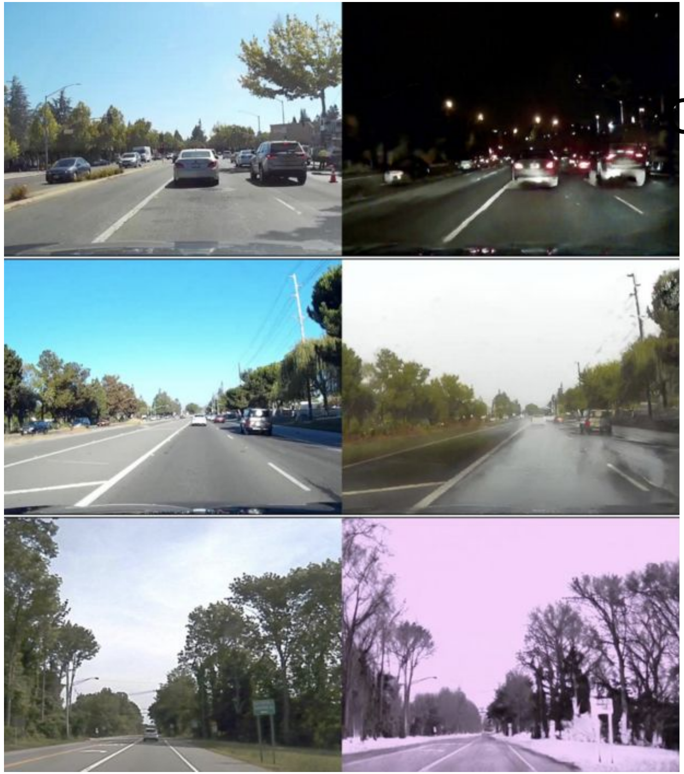
Fig. 5: The human study interface with an image on the left and seven multiple choice questions about it.

Metric	CogView2	DALLE-v2	SD	SD-CDM
OA	73.07	73.87	79.25	80.21
$\text{VISOR}_{\text{uncond}}$	88.48	77.41	88.43	88.80
$\text{VISOR}_{\text{cond}}$	75.02	75.62	76.95	74.69

TABLE 5: Agreement(%) of human responses with automated metrics

Seed Idea: Mitigation Domain Gap for CV systems in Urban Driving Scenes

- Internalize data to learn representation
- Great for downstream tasks
- Are these learned representations “general” enough?
- DG = accurate prediction on previously unseen domain



Can a model trained on the images in left

Generalize to these images in right?

Seed Idea: Mitigation Domain Gap for CV systems in Urban Driving Scenes

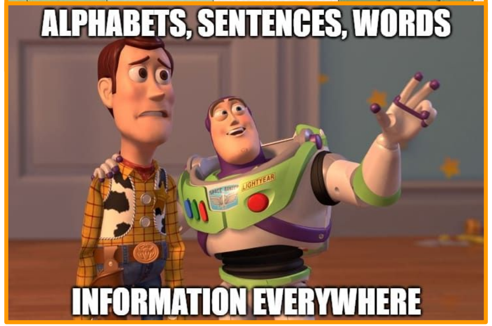
How does concurrent research deal with DG?

- SSDG vs MSDG
- Data Augmentations (RL, Adv)
- Learning domain invariant features
- Special focus on generalizing to the statistics of the unseen domain

Seed Idea: Mitigation Domain Gap for CV systems in Urban Driving Scenes

Leveraging Large Language Models to understand domains

- Language is the most active medium of communicating intelligence and has been called (Pinker 19941) “the jewel in the crown of cognition.”
- How well can LLMs describe domain informations?
- Can encode(description) generate domain invariant data?
- Can encode(description) facilitate domain invariant learning?




```
prompt = ["a photograph of an astronaut riding a horse"]
```

Bringing in Your Own Ideas !!!