Topics:

- Bias/Fairness
- Wrap-up:
  - Open directions in Deep Learning

CS 4644-DL / 7643-A
ZSOLT KIRA

- **Projects!**
  - Rubrics up: @500
  - Project due **April 29 11:59pm** (grace period **May 1st**)
  - Cannot extend due to grade deadlines!

- CIOS
  - Please make sure to fill out! Let us know about things you liked and didn’t like in comments so that we can keep or improve!
  - http://b.gatech.edu/cios

Bias & Fairness

ML and Fairness

- AI effects our lives in many ways
- Widespread algorithms with many small interactions
  – e.g. search, recommendations, social media
- Specialized algorithms with fewer but higher-stakes interactions
  – e.g. medicine, criminal justice, finance
- At this level of impact, algorithms can have unintended consequences
- Low classification error is not enough, need fairness

BUSINESS NEWS OCTOBER 10, 2018 / 3:12 AM / 6 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like

INDEPENDENT

GOOGLE'S ALGORITHM SHOWS PRESTIGIOUS JOB ADS TO MEN, BUT NOT TO WOMEN

REUTERS

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

MIT Technology Review

Intelligent Machines

How to Fix Silicon Valley's Sexist Algorithms

Computers are inheriting gender bias implanted in language data sets—and not everyone thinks we should correct it.

ProPublica

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Slide By Aaron Roth

Machine Learning and Social Norms

- Sample norms: privacy, fairness, transparency, accountability…
- Possible approaches
  - "traditional": legal, regulatory, watchdog
  - *Embed* social norms in data, algorithms, models
- Case study: privacy-preserving machine learning
  - "single", strong, definition (differential privacy)
  - almost every ML algorithm has a private version
- Fair machine learning
  - not so much…
  - impossibility results

(Un)Fairness Where?

- Data (input)
  - e.g. more arrests where there are more police
  - Label should be “committed a crime”, but is “convicted of a crime”
  - try to “correct” bias
- Models (output)
  - e.g. discriminatory treatment of subpopulations
  - build or “post-process” models with subpopulation guarantees
  - equality of false positive/negative rates; calibration
- Algorithms (process)
  - learning algorithm *generating* data through its decisions
  - e.g. don’t learn outcomes of denied mortgages
  - lack of clear train/test division
  - design (sequential) *algorithms* that are fair

When the *training data* we collect does not contain representative samples of the true distribution.

Examples:

- If we use data gathered from smart phones, we would likely be underestimating poorer and older populations.
- ImageNet (a very popular image dataset) with 1.2 million images. About 45% of these images were taken in the US and the majority of the rest in North America and Western Europe. Only about 1% and 2.1% of the images come from China and India respectively.

Data Bias

Often we are gathering data that contains (noisy) proxies of characteristics of interest. Some examples:

- Financial responsibility → Credit Score
- Crime Rate → Arrest Rate
- Intelligence → SAT Score

If these measurements are not measured equally across groups or places (or aren't relevant to the task at hand), this can be another source of bias.

Measurement Bias

Examples:

- If factory workers are monitored more often, more errors are spotted. This can result in a **feedback loop** to encourage more monitoring in the future.
  - Same principles at play with predictive policing. Minoritized communities were more heavily policed in the past, which causes more instances of documented crime, which then leads to more policing in the future.

- Women are more likely to be misdiagnosed (or not diagnosed) for conditions where self-reported pain is a symptom. In this case aspect of our data “diagnosed with X” is a biased proxy for “has condition X”.
- The feature we measure is a poor representation of the quality of interest (e.g., SAT score doesn’t actually measure intelligence)

Measurement Bias

What does it mean for a model to be fair or unfair? Can we come up with a numeric way of measuring fairness?

Lots of work in the field of ML and fairness is looking into mathematical definitions of fairness to help us spot when something might be unfair.

- There is not going to be one central definition of fairness, as each definition is a mathematical statement of which behaviors are/aren't allowed.
- Different definitions of fairness can be contradictory!

ML and Fairness

- Fairness is morally and legally motivated
- Takes many forms
- Criminal justice: recidivism algorithms (COMPAS)
  - Predicting if a defendant should receive bail
  - Unbalanced false positive rates: more likely to wrongly deny a black person bail

Table 1: ProPublica Analysis of COMPAS Algorithm

	White	Black
Wrongly Labeled High-Risk	23.5%	44.9%
Wrongly Labeled Low-Risk	47.7%	28.0%

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Why Fairness is Hard

- Suppose we are a bank trying to fairly decide who should get a loan
  - i.e. Who is most likely to pay us back?
- Suppose we have two groups, A and B (the sensitive attribute)
  - This is where discrimination could occur
- The simplest approach is to remove the sensitive attribute from the data, so that our classier doesn't know the sensitive attribute. Often called **"Fairness through unawareness"**

Table 2: To Loan or Not to Loan?

Age	Gender	Postal Code	Req Amt	A or B?	Pay
46	F	M5E	\$300	A	1
24	M	M4C	\$1000	B	1
33	M	M3H	\$250	A	1
34	F	M9C	\$2000	A	0
71	F	M3B	\$200	A	0
28	M	M5W	\$1500	B	0

Why Fairness is Hard

- However, if the sensitive attribute is correlated with the other attributes, this isn't good enough
- It is easy to predict race if you have lots of other information (e.g. home address, spending patterns)
- More advanced approaches are necessary

Table 3: To Loan or Not to Loan? (masked)

Age	Gender	Postal Code	Req Amt	A or B?	Pay
46	F	M5E	\$300	?	1
24	M	M4C	\$1000	?	1
33	M	M3H	\$250	?	1
34	F	M9C	\$2000	?	0
71	F	M3B	\$200	?	0
28	M	M5W	\$1500	?	0

Doesn't work in practice. This does not prevent historical or measurement bias. Protected attributes can be unintentionally inferred from other, related attributes (e.g., in some cities, zip code can be deeply correlated with race).

Definitions of Fairness – Group Fairness

- So we've built our classier . . . how do we know if we're being fair?
- One metric is demographic parity | requiring that the same percentage of A and B receive loans
  - What if 80% of A is likely to repay, but only 60% of B is?
  - Then demographic parity is too strong
- Could require equal false positive/negative rates
  - When we make an error, the direction of that error is equally likely for both groups

$$P(loan|no\ repay, A) = P(loan|no\ repay, B)$$
$$P(no\ loan|would\ repay, A) = P(no\ loan|would\ repay, B)$$

- These are definitions of group fairness
- Treat different groups equally"

Definitions of Fairness – Individual Fairness

- Also can talk about individual fairness | “Treat similar examples similarly"
- Learn fair representations
  - Useful for classification, not for (unfair) discrimination
  - Related to domain adaptation
  - Generative modelling/adversarial approaches

(a) Unfair representations

(b) Fair(er) representations

Figure 1: “The Variational Fair Autoencoder” (Louizos et al., 2016)

Conclusion

- This is an exciting field, quickly developing
- Central definitions still up in the air
- AI moves fast | lots of (currently unchecked) power
- Law/policy will one day catch up with technology
- Those who work with AI should be ready
  - **Think about implications of what you develop!**

Parting Thoughts

Deep Learning Fundamentals

Linear classification
Loss functions
Optimization
Optimizers
Backpropagation
Computation Graph
Multi-layer Perceptrons

Neural Network Components and Architectures

Hardware & software
Convolutions
Convolution Neural Networks
Pooling
Activation functions
Batch normalization
Transfer learning
Data augmentation
Architecture design
RNN/LSTMs
Attention & Transformers

Applications & Learning Algorithms

Semantic & instance Segmentation
Reinforcement Learning
Large-language Models
Variational Autoencoders
Diffusion Models
Generative Adversarial Nets
Self-supervised Learning
Vision-Language Models
VLM for Robotics

We Learned a Lot!

Georgia Tech

When Comparing to Humans, What’s Missing?

- Memory
- Reasoning
  - What does it mean for a neural network to “think” longer?
- Planning, Search
- Deep integration of concepts and modalities

Some existing works not covered…

Current / Recent Past

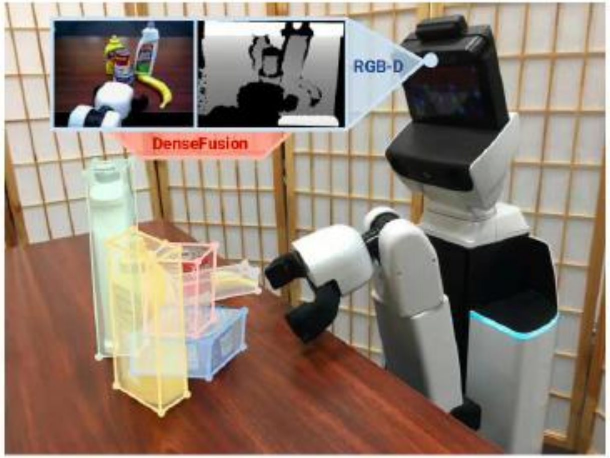
- Graph neural networks
- Meta-learning
- AutoML
- 3D perception & reconstruction / NeRFs
- Beyond supervised learning: Semi-supervised, domain adaptation, zero/one/few-shot learning
- Memory (Neural Turing Machines, etc.)
- Embodied AI & Embodied question answering
- Adversarial Learning
- Continual/lifelong learning without forgetting
- World modeling, learning intuitive/physics models
- Reasoning, Planning, Search
- Neural Theorem Proving, induction & synthesis
- Neural Radiance Fields
- MLSys and MLOps
- Evaluation…
- Alignment
- Security

Some existing works not covered…

Current / Recent Past

- Graph neural networks
- Meta-learning
- AutoML
- **3D perception & reconstruction / NeRFs**
- Beyond supervised learning: Semi-supervised, domain adaptation, zero/one/few-shot learning
- Memory (Neural Turing Machines, etc.)
- Embodied AI & Embodied question answering
- Adversarial Learning
- Continual/lifelong learning without forgetting
- World modeling, learning intuitive/physics models
- Reasoning, Planning, Search
- Neural Theorem Proving, induction & synthesis
- Neural Radiance Fields
- MLSys and MLOps
- Evaluation…
- Alignment
- Security

Perception for 3D Object Understanding: Applications



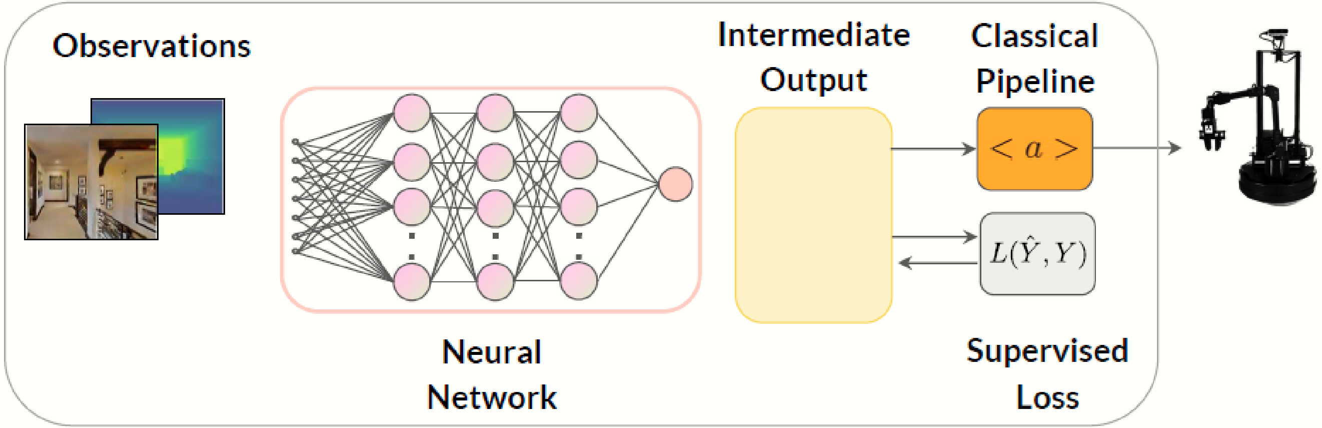
Object Grasping

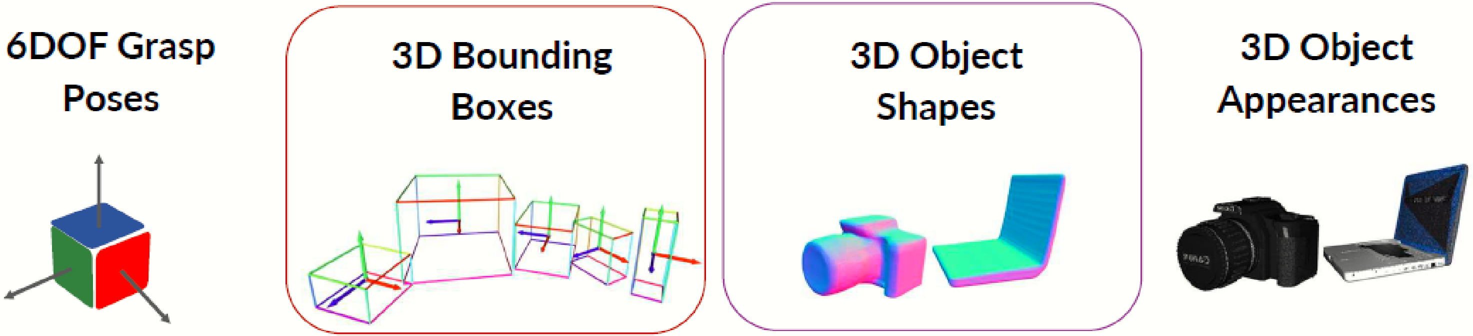
AR/VR Augmentations

[...] eggshell broken in two with an adorable chick standing next to it

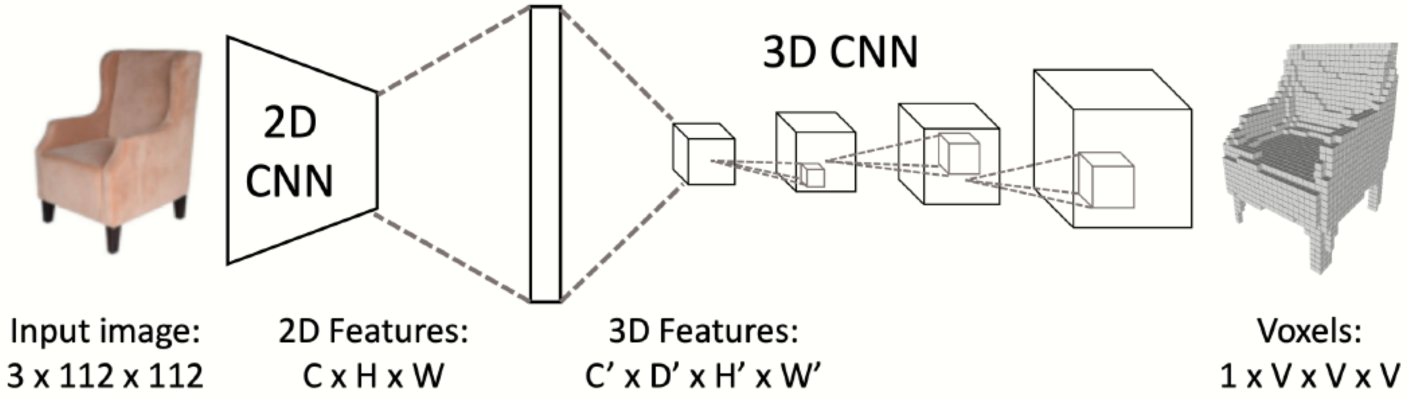
Text-to-3D

Perception for 3D Object Understanding: Current Paradigm





3D Perception

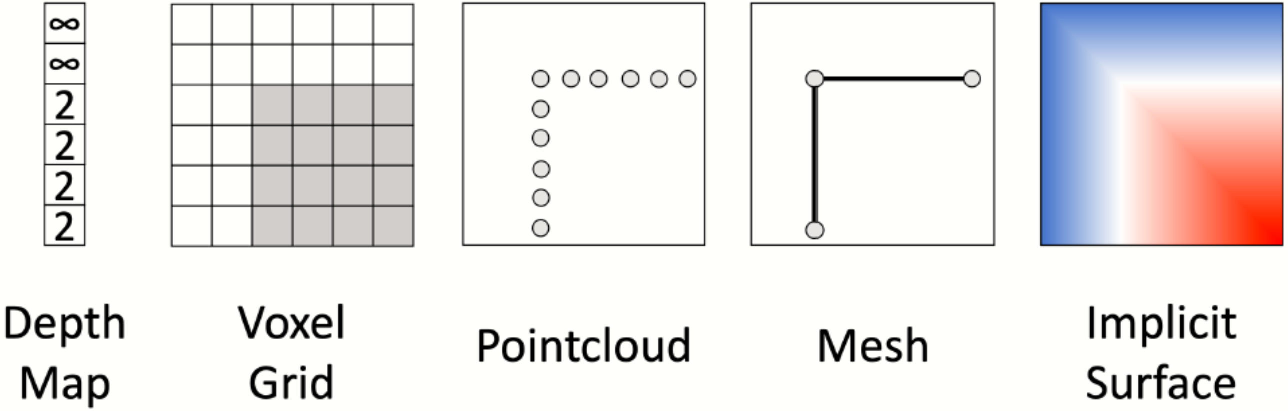


3D Convolution for Voxel-based 3D Reconstruction

Choy et al., 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction, ECCV 2016

3D Perception

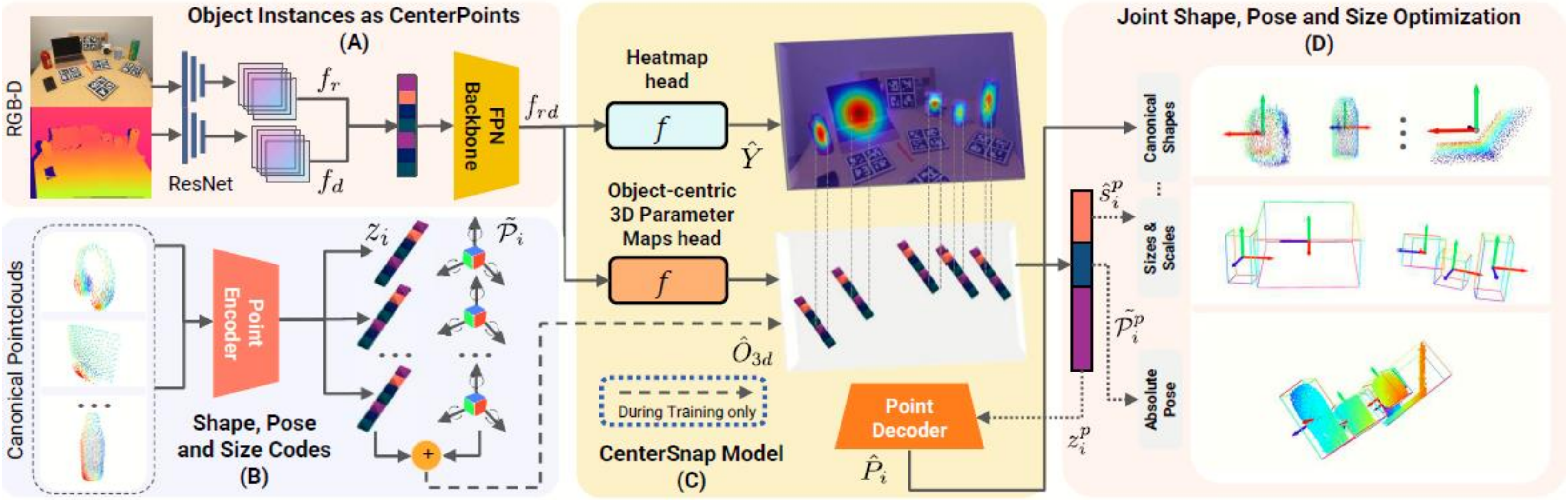
Many possible ways to represent the 3D world …



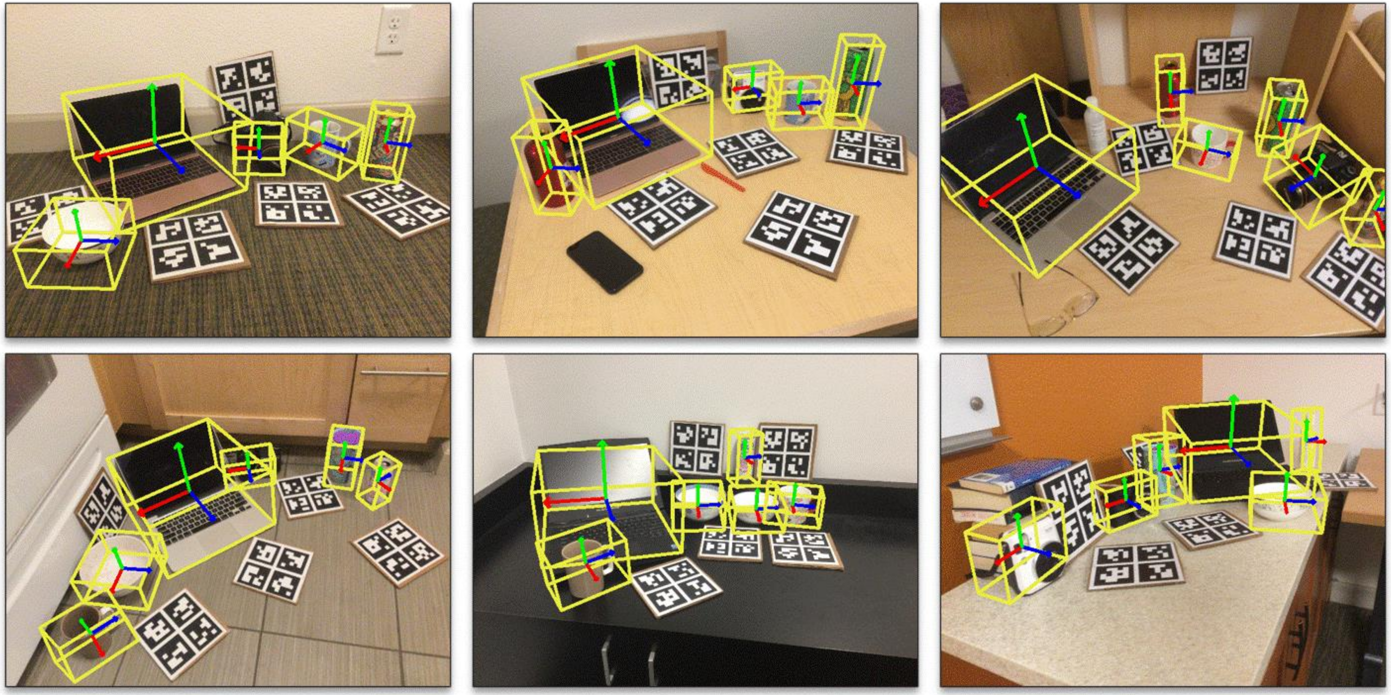
Each representation requires different neural network architectures!

Figure credit: Justin Johnson

CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

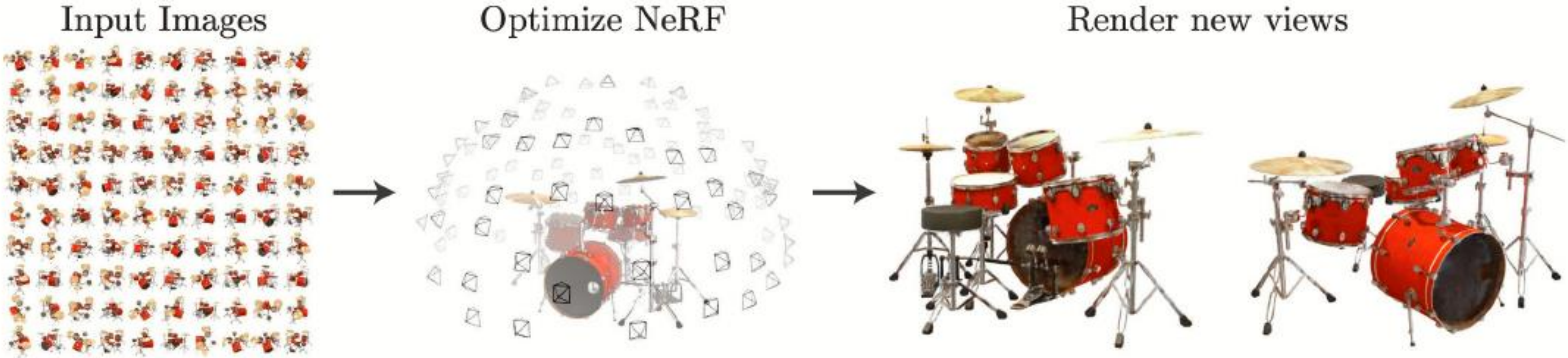


[Ref] M.Z.Irshad, T.Kollar, M.Laskey, K.Stone, Z.Kira, " CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and Categorical 6D Pose and Size Estimation, **ICRA 2022**

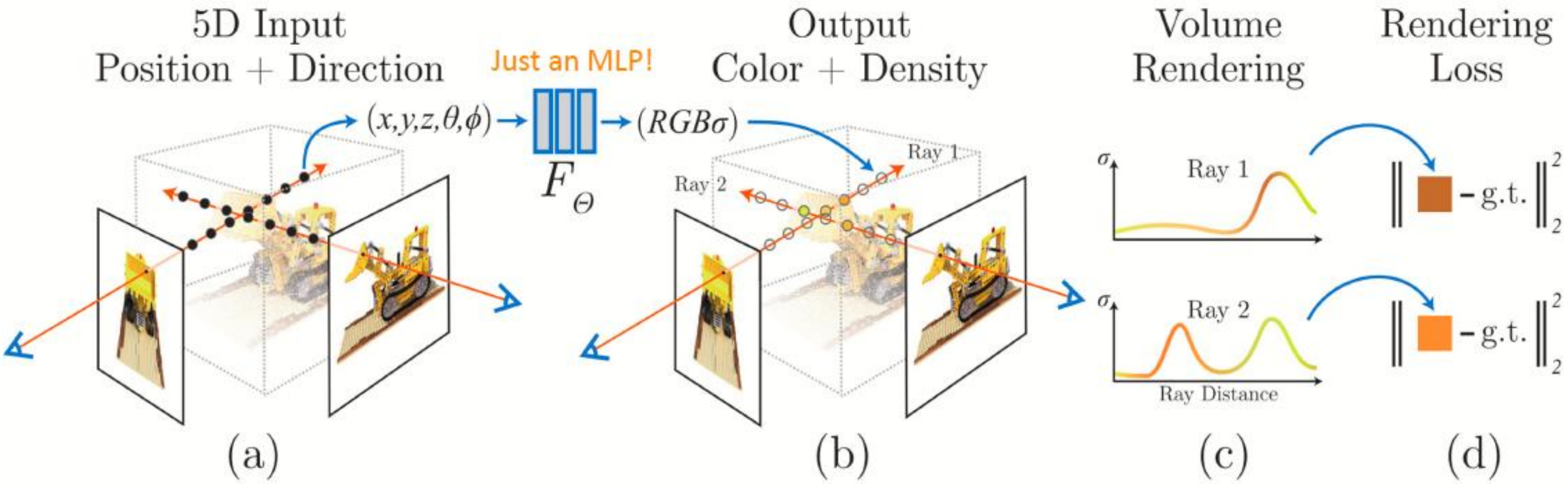


Neural Radiance Field

Neural Radiance Field: View Synthesis



Neural Radiance Field



Very slow to train & render!

Requires many tricks to render high-quality images

One model per scene

Some existing works not covered…

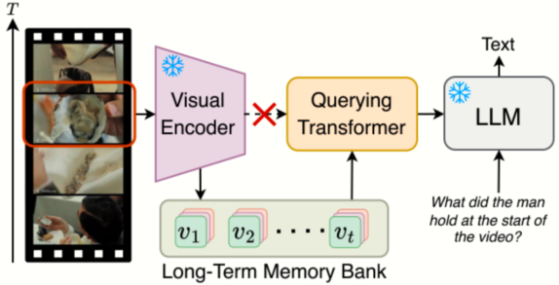
Current / Recent Past

- Graph neural networks
- Meta-learning
- AutoML
- 3D perception & reconstruction / NeRFs
- Beyond supervised learning: Semi-supervised, domain adaptation, zero/one/few-shot learning
- **Memory (Neural Turing Machines, etc.)**
- Embodied AI & Embodied question answering
- Adversarial Learning
- Continual/lifelong learning without forgetting
- World modeling, learning intuitive/physics models
- Reasoning, Planning, Search
- Neural Theorem Proving, induction & synthesis
- Neural Radiance Fields
- MLSys and MLOps
- Evaluation…
- Alignment
- Security

MA-LMM: Memory-Au
Model for Long-Ter

Bo He[1,2], Hengduo Li[2], Young
Ashish Shah[2], Abhina
[1]University of Maryland, College Park

arX



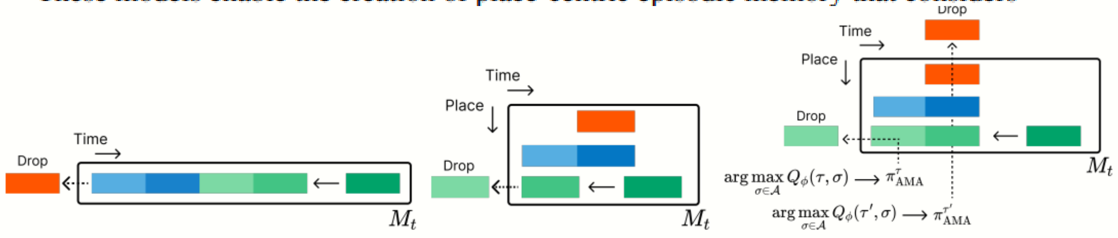
MA-LMM Long-term memory bank auto-regressively stores and accumulates past video information.

SPATIALLY-AWARE TRANSFORMER FOR EMBODIED AGENTS

Junmo Cho[1*], Jaesik Yoon[1,2*], Sungjin Ahn[1]
[1]KAIST & [2]SAP

ABSTRACT

Episodic memory plays a crucial role in various cognitive processes, such as the ability to mentally recall past events. While cognitive science emphasizes the significance of spatial context in the formation and retrieval of episodic memory, the current primary approach to implementing episodic memory in AI systems is through transformers that store temporally ordered experiences, which overlooks the spatial dimension. As a result, it is unclear how the underlying structure could be extended to incorporate the spatial axis beyond temporal order alone and thereby what benefits can be obtained. To address this, this paper explores the use of Spatially-Aware Transformer models that incorporate spatial information. These models enable the creation of place-centric episodic memory that considers



(a) FIFO Memory **(b)** Place Memory **(c)** Adaptive Memory Allocator (AMA)

Graves et. al, Neural Turing Machines

Beyond A^*: Better Planning with Transformers via Search Dynamics Bootstrapping

Lucas Lehnert[1], Sainbayar Sukhbaatar[1], Paul Mcvay[1], Michael Rabbat[1], Yuandong Tian[1]

[1]FAIR at Meta

While Transformers have enabled tremendous progress in various application settings, such architectures still lag behind traditional symbolic planners for solving complex decision making tasks. In this work, we demonstrate how to train Transformers to solve complex planning tasks and present **Searchformer**, a Transformer model that optimally solves previously unseen Sokoban puzzles 93.7% of the time, while using up to 26.8% fewer search steps than standard A^* search. Searchformer is an encoder-decoder Transformer model trained to predict the *search dynamics* of A^*. This model is then fine-tuned via expert iterations to perform fewer search steps than A^* search while still generating an optimal plan. In our training method, A^*'s search dynamics are expressed as a token sequence outlining when task states are added and removed into the search tree during symbolic planning. In our ablation studies on maze navigation, we find that Searchformer significantly outperforms baselines that predict the optimal plan directly with a 5–10× smaller model size and a 10× smaller training dataset. We also demonstrate how Searchformer scales to larger and more complex decision making tasks like Sokoban with improved percentage of solved tasks and shortened search dynamics.

Correspondence: {lucaslehnert, yuandong}@meta.com

Things to Watch out For

- Research is cyclical
  - SVMs, boosting, probabilistic graphical models & Bayes Nets, Structural Learning, Sparse Coding, Deep Learning
  - Deep learning is unique in its depth and breadth, but...
  - Deep learning may be improved, reinvented, combined, overtaken

- Learn fundamentals for techniques across the field:
  - Know the span of ML techniques and choose the ones that fit your problem!
  - **Be responsible** in 1) how you use it, 2) promises you make and how you convey it

- Try to understand landscape of the field
  - Look out for what is coming up next, not where we are

- Have fun!

Open Discussion

Thank you!