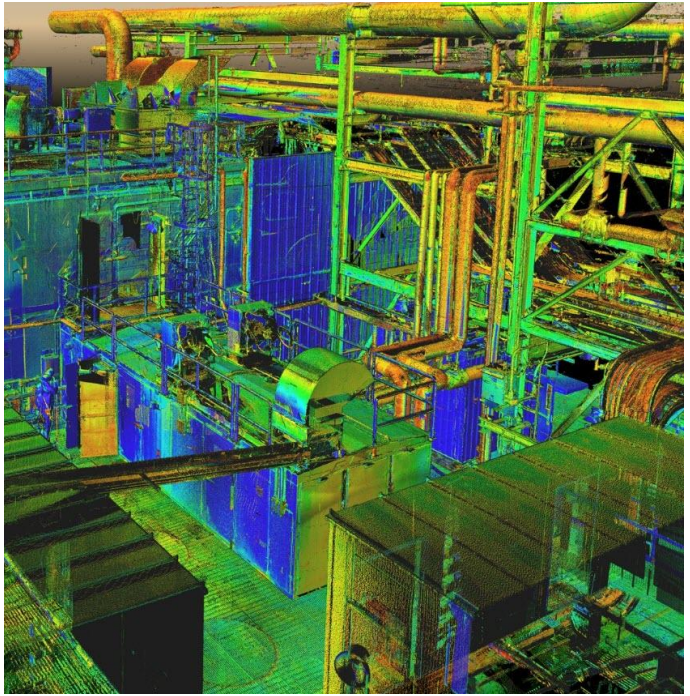
# Geometry-Aware Neural Rendering

Tobin, Joshua, Wojciech Zaremba, and Pieter Abbeel. Advances in Neural Information Processing Systems. 2019
(OpenAI, UC Berkeley)

# Motivation: Understanding 3D structure of the world



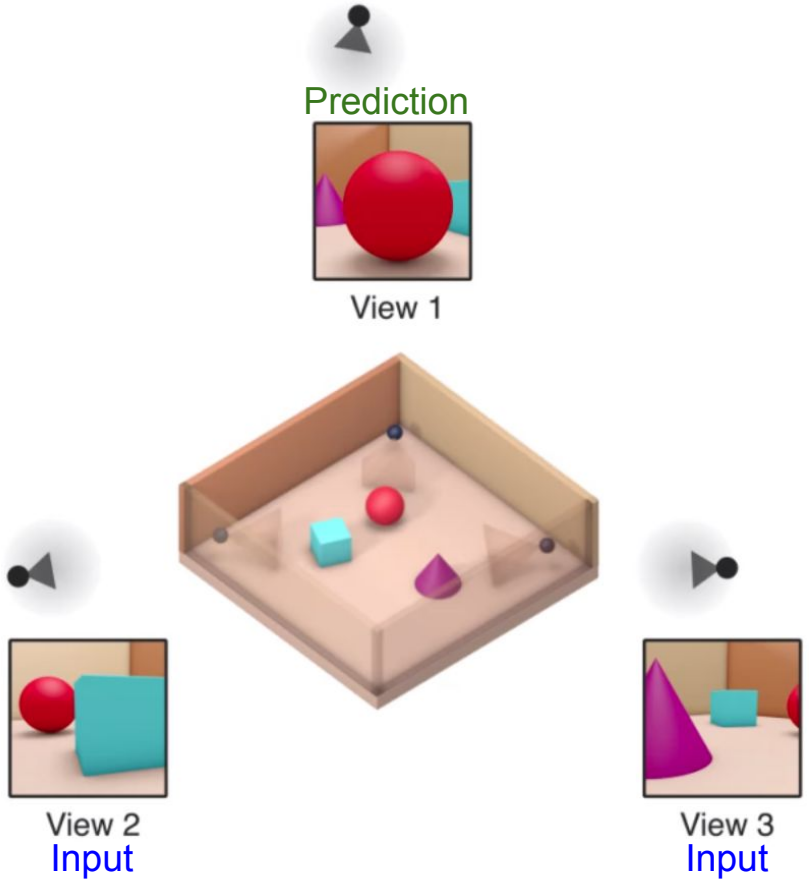
Microsoft Kinect Gen 1



3D laser scan

# Specific Goal:

- Given input views and their viewpoints
- Predict the new image for a query viewpoint



# Related and Comparative work:

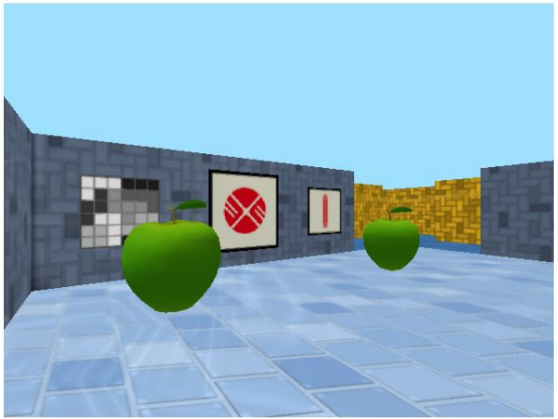
- 2D Neural Rendering
- Novel View Synthesis
- Extreme View Synthesis
- Generative Query Network

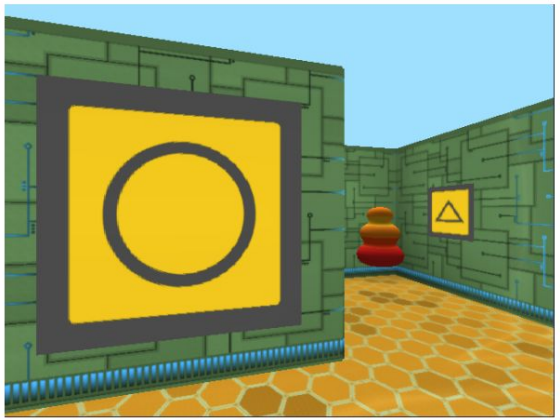
# Learning to navigate in complex environments

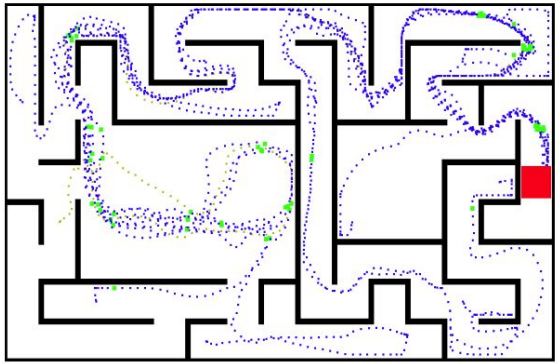
Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. arXiv preprint arXiv:1611.03673, 2016.

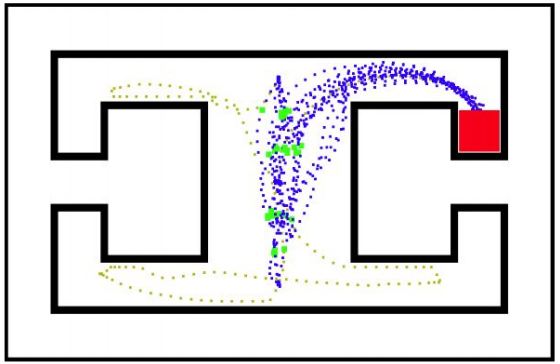
**Goal:** Learn to navigate from raw sensory input.

**Solution:** Actor-Critic reinforcement learning builds 2D representation in latent space.



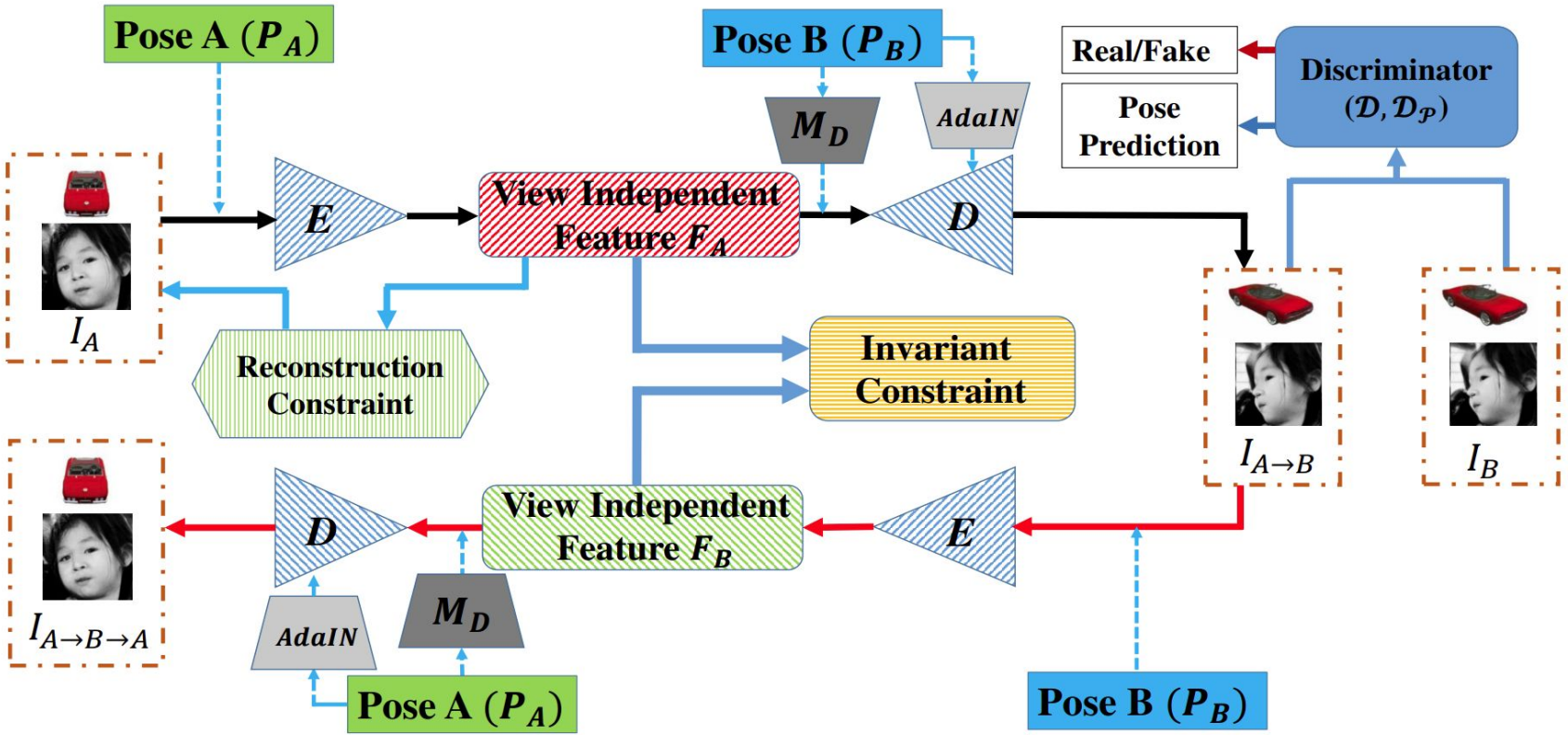






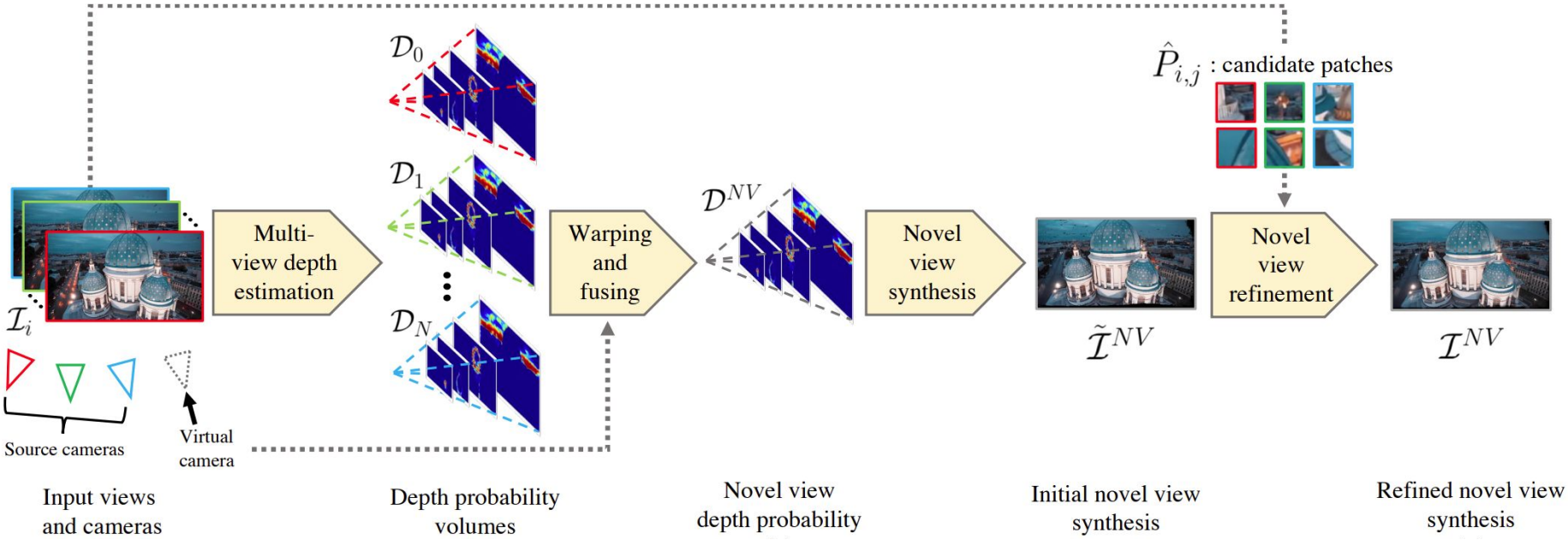
# View Independent Generative Adversarial Network for Novel View Synthesis

***Xiaogang Xu, Ying-Cong Chen, Jiaya Jia***; The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 7791-7800

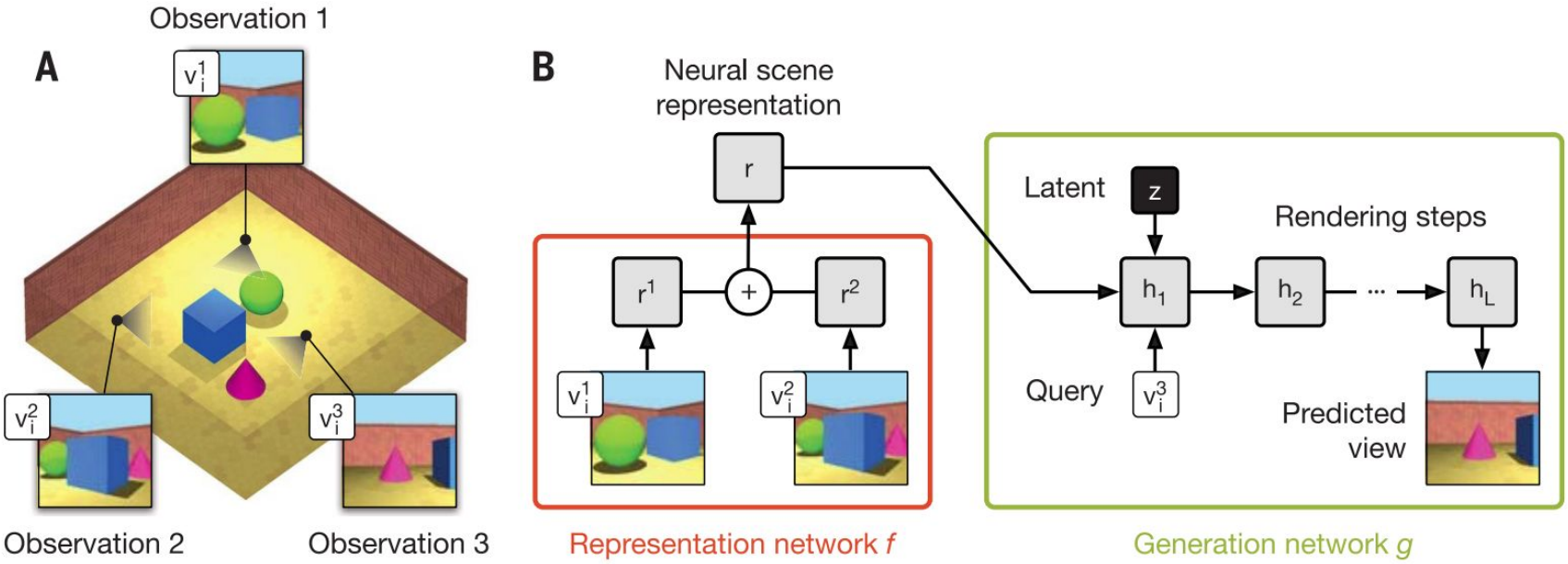


# Extreme view synthesis

Choi, I., Gallo, O., Troccoli, A., Kim, M.H. and Kautz, J., 2019. In Proceedings of the IEEE International Conference on Computer Vision (pp. 7781-7790)

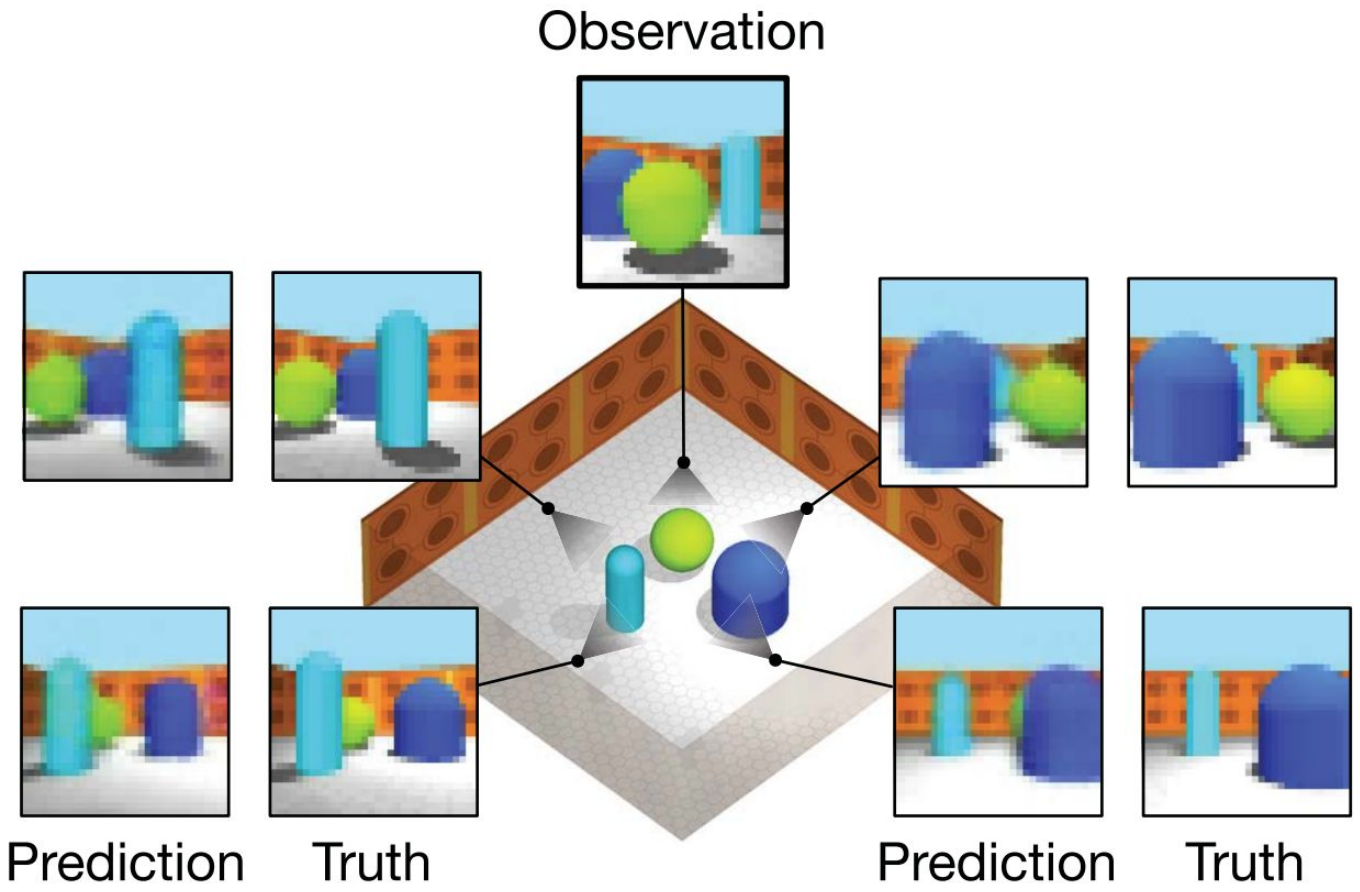


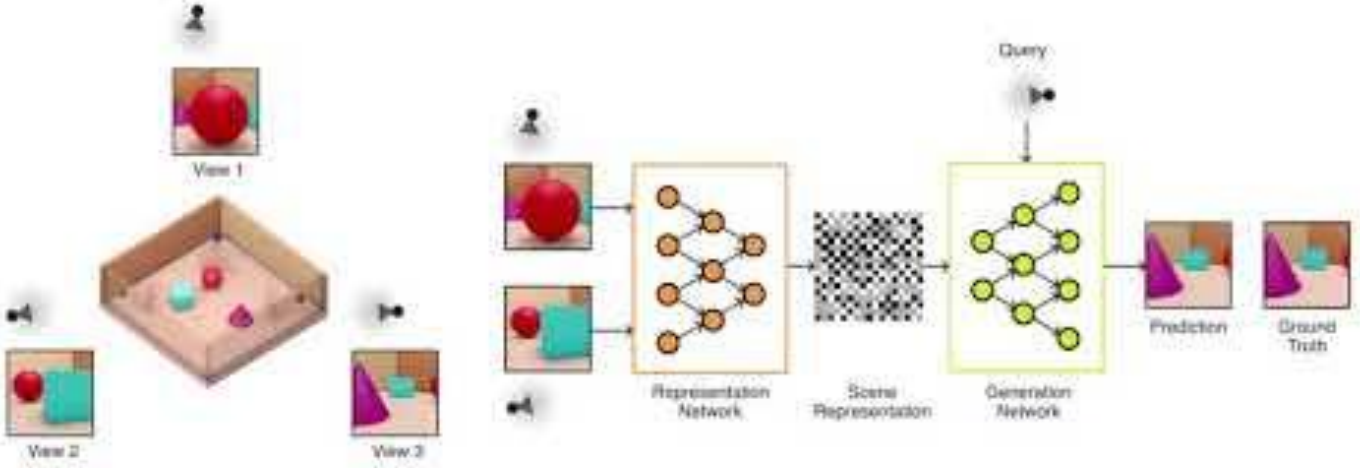
# Generative Query Network (GQN)



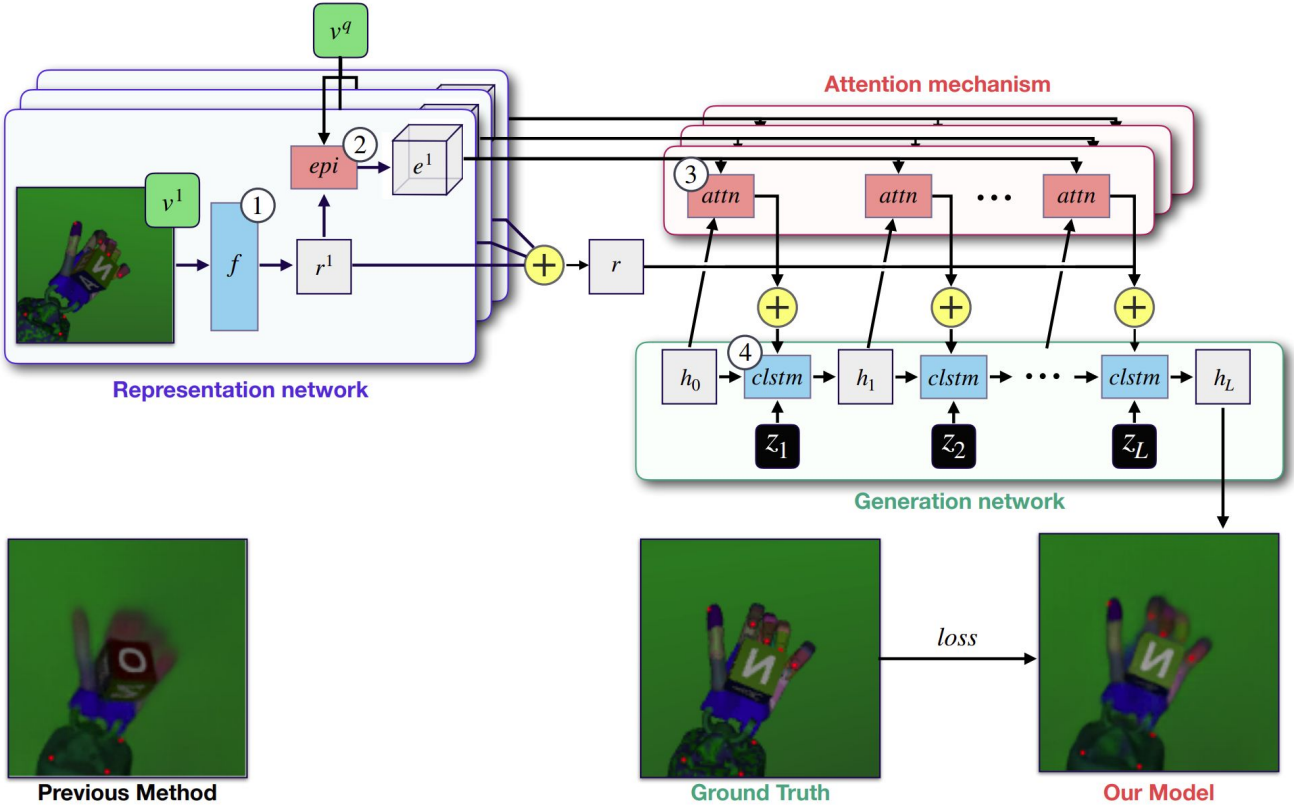

(Eslami, SM Ali, et al. "Neural scene representation and rendering." Science 360.6394 (2018): 1204-1210.)

# GQN using a single context observation




View 1
View 2
View 3
Query
Representation Network
Scene Representation
Generation Network
Prediction
Ground Truth

# Epipolar Generative Query Network (E-GCN) Architecture

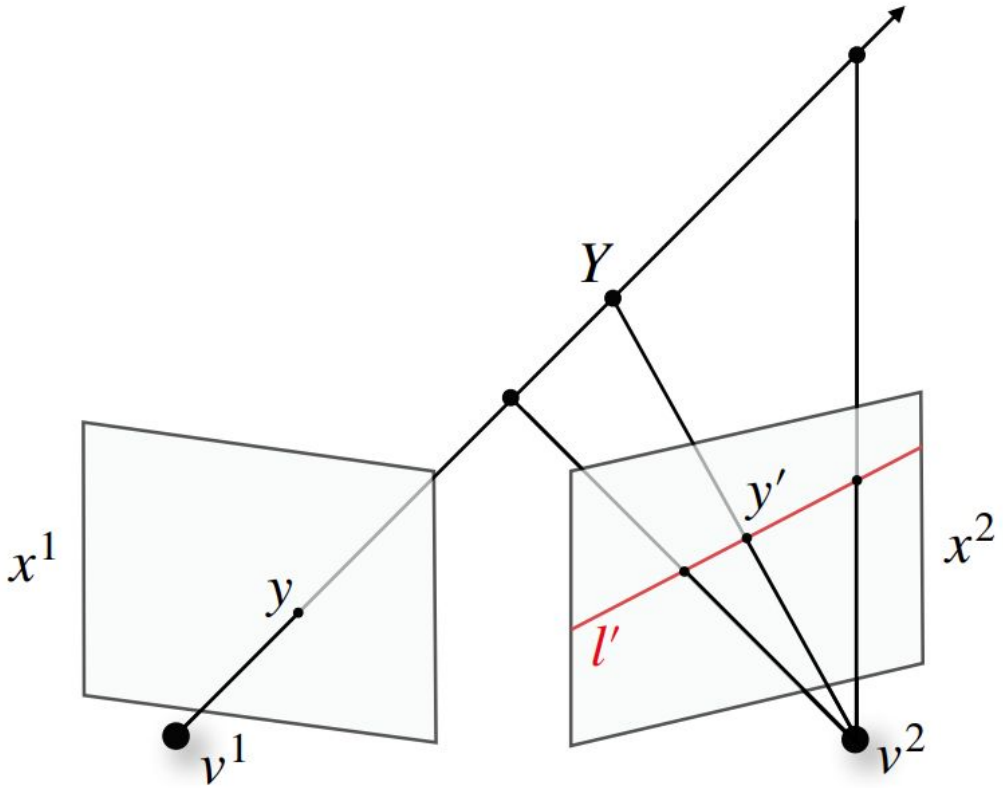


# Key Contributions:

1. Novel attention mechanism that leverages the camera geometry of the camera [Epipolar Cross-Attention (ECA)].
2. Three new datasets:
    a. Disco Humanoid
    b. OpenAI Block
    c. Room Random Objects
3. Demonstrate performance of the ECA with the Generative Query Network (E-GQN)

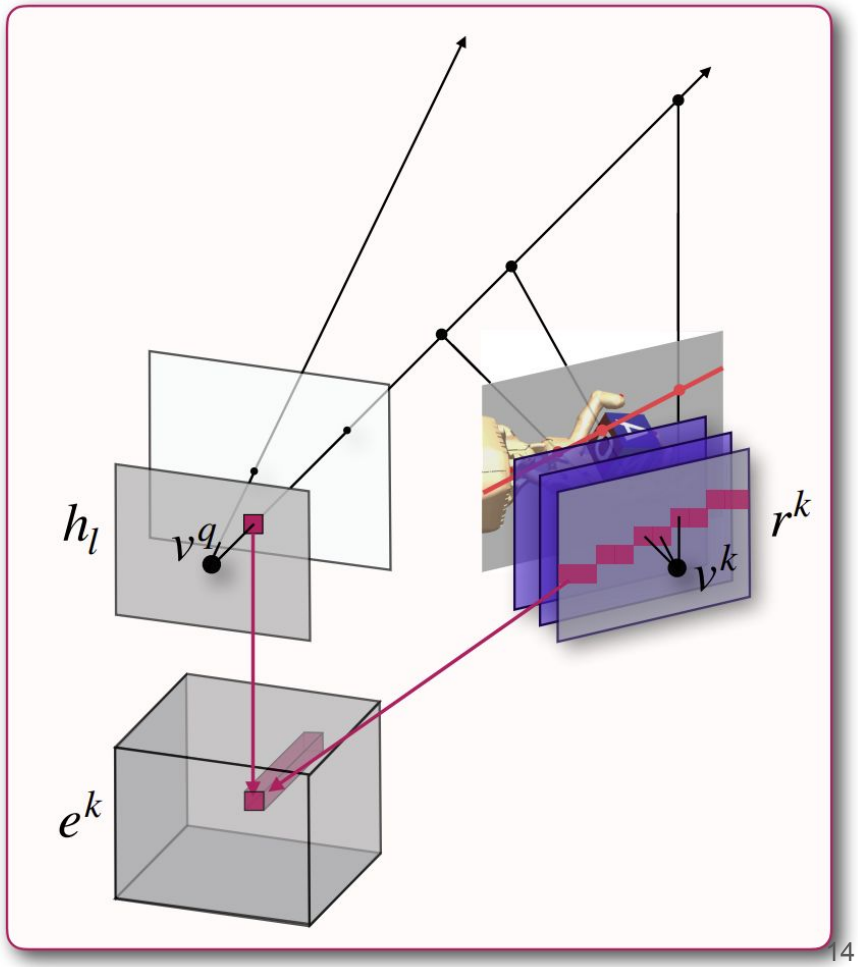
# A quick look at Epipolar Geometry (stereo vision)

- Two views ($v^1$, $v^2$) of the same 3D Point $Y$
- Ray from $v^1$ through $Y$ corresponds to $l'$
- We want to know line $l'$ in image $x^2$
- Computing $l'$ is linear
- Fundamental Matrix



# Epipolar extraction

- Given a point in our query image $h_l$
- Find the corresponding features in input image $r^k$
- Constructing new vector representation in $e^k$



# Representation Network

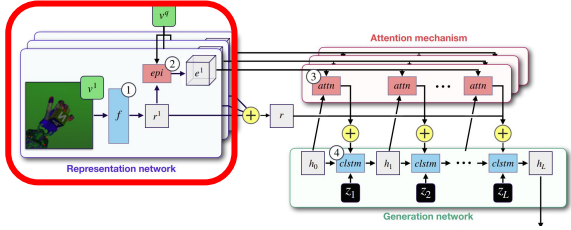
1) Images and viewpoints are passed through a CNN $f$

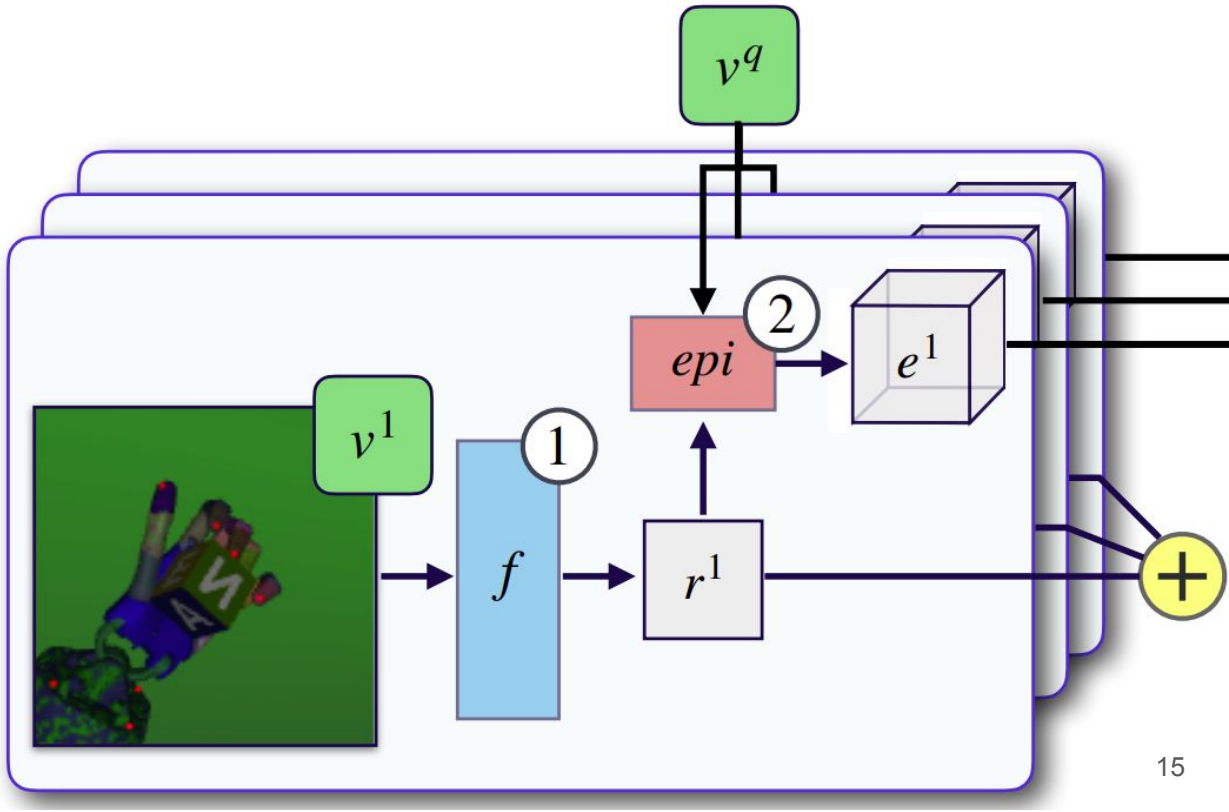
2) Produce a context representation $r$

3) Epipolar geometry (epi) used to extract the features

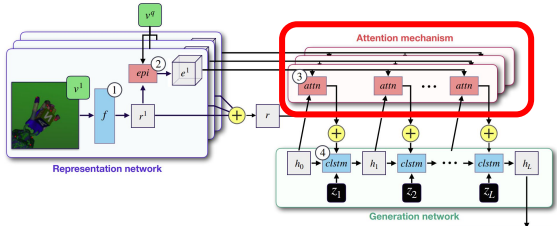
4) Extracted features stored in $e^k$

5) Epipolar representation (3D tensor) passed to Attention mechanism.

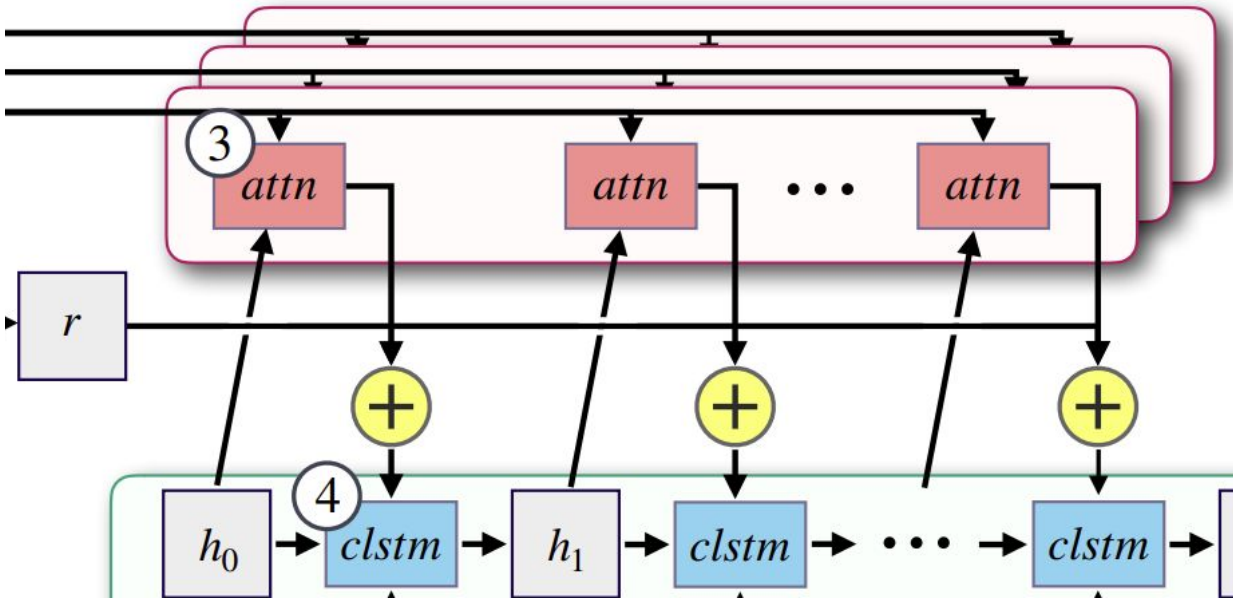


# Attention Mechanism

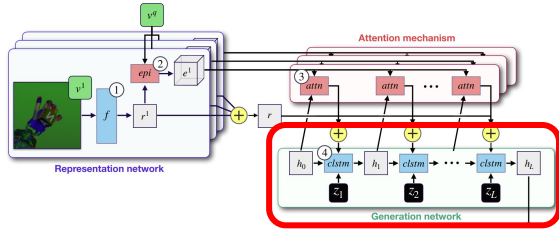


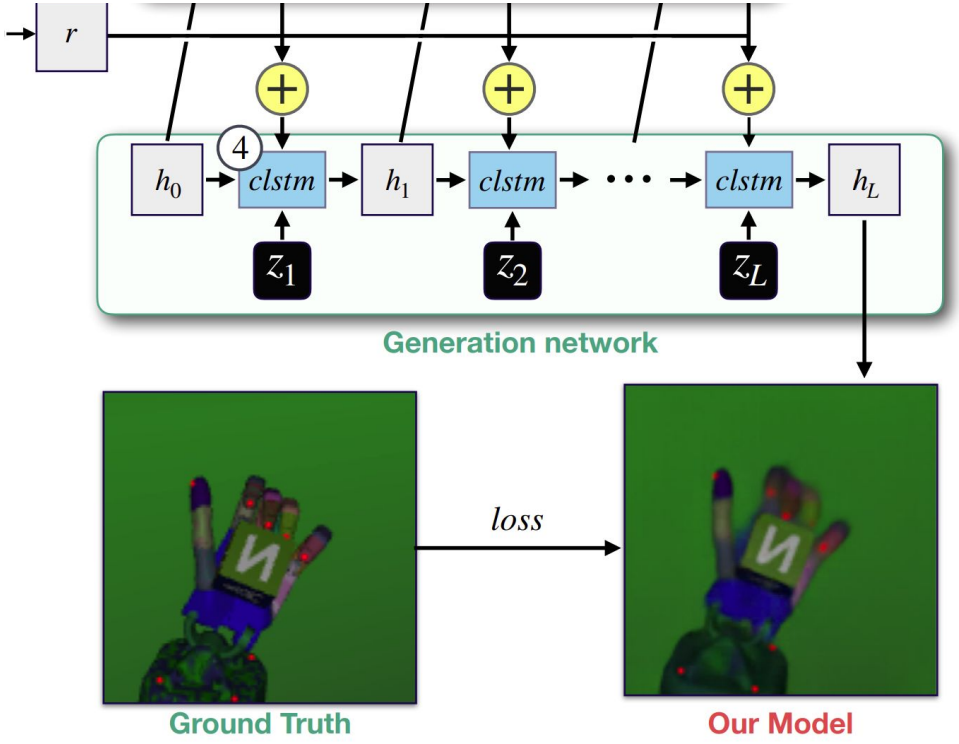

- For each step, each spatial position in decoded hidden state pass to attention mechanism
- Map captures the weighted contribution
- Attend over the epipolar representation
- At each step compute decoder input



# Generation Network

- Generator network is the skip-connection convolutional LSTM cell.
- Takes as input the attention map, previous hidden state $\boldsymbol{h}$ and $z$.
- $z$ is stochastic latent variable



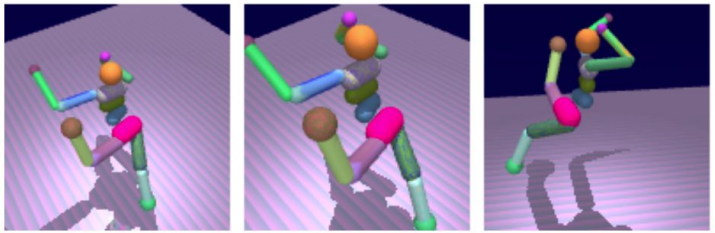



# Dataset Contributions

OpenAI Blocks



Disco Humanoid



Rooms Random Objects



# OpenAI blocks

- ShadowHand robot with a random joint configuration (20 DoF)
- Lettered cube placed in hand with random orientation
- Appearance is changed by randomizing the lighting and textures of all bodies in the scene.

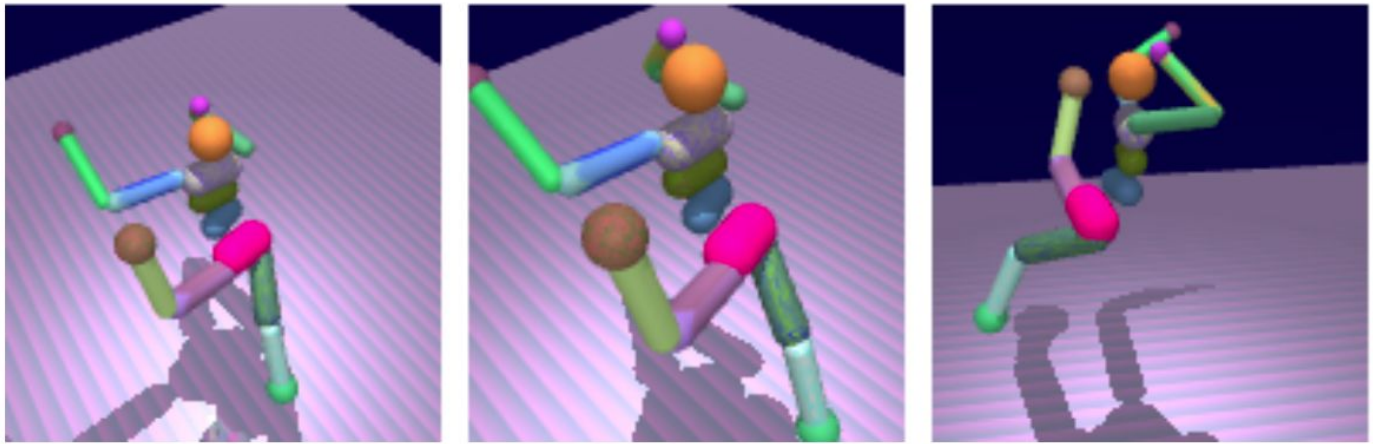
OAB



# Disco Humanoid

- Humanoid from the Humanoid-v2 gym environment (27 DoF)
- All joints are configured randomly
- Randomize lighting and textures of all bodies in the scene.

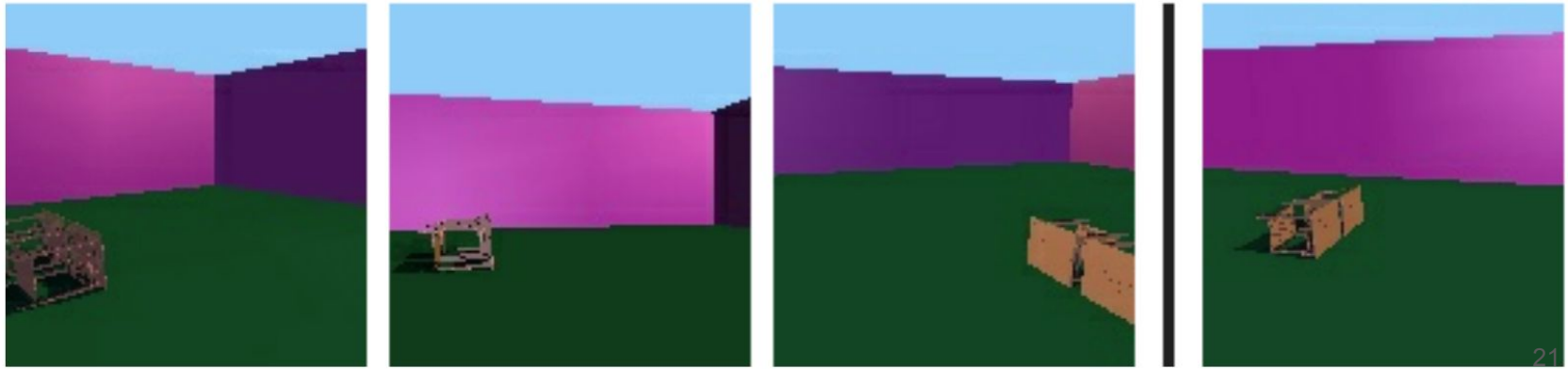
**Disco**



# Rooms Random Objects

- One to four objects from the ShapeNet dataset
- Randomly oriented and dropped into scene
- Randomize lighting and all textures

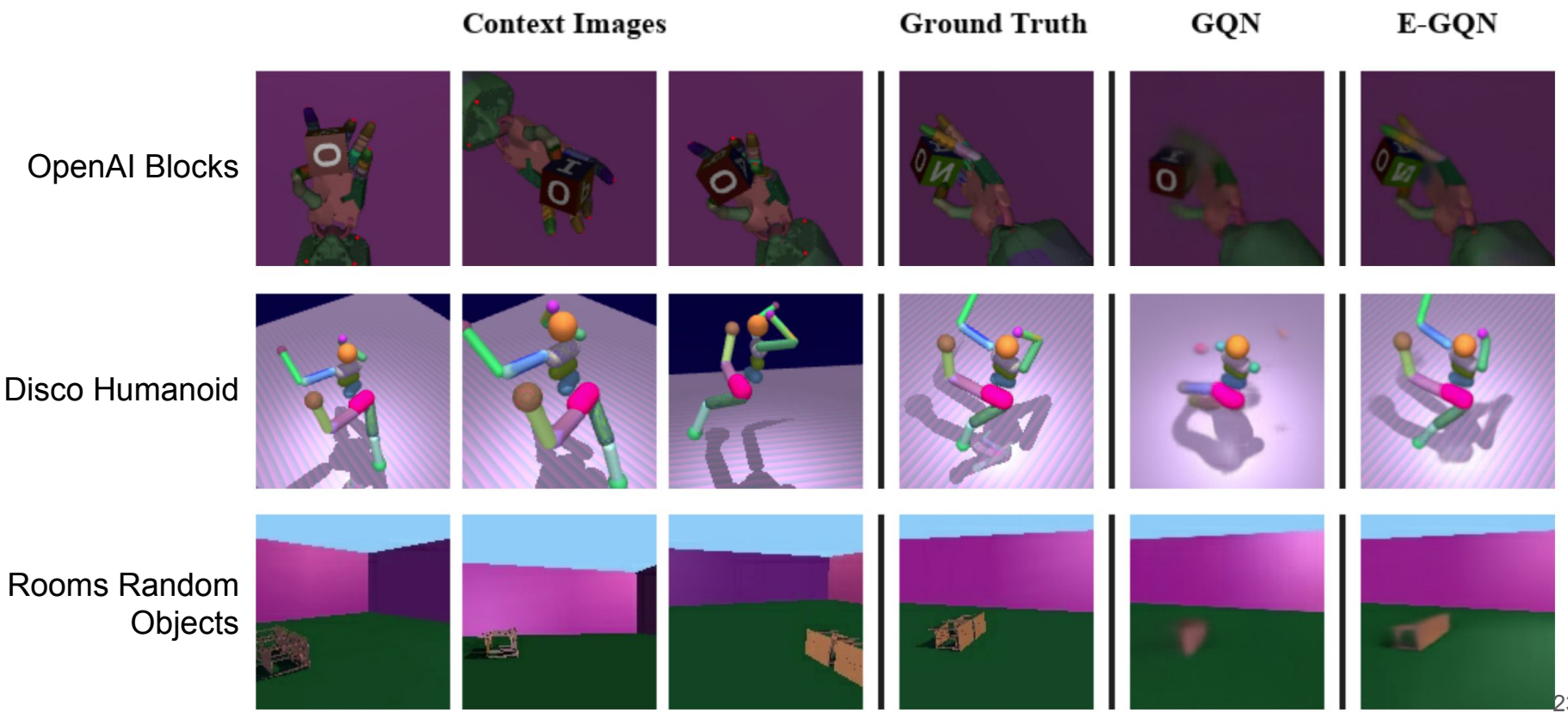
RRO

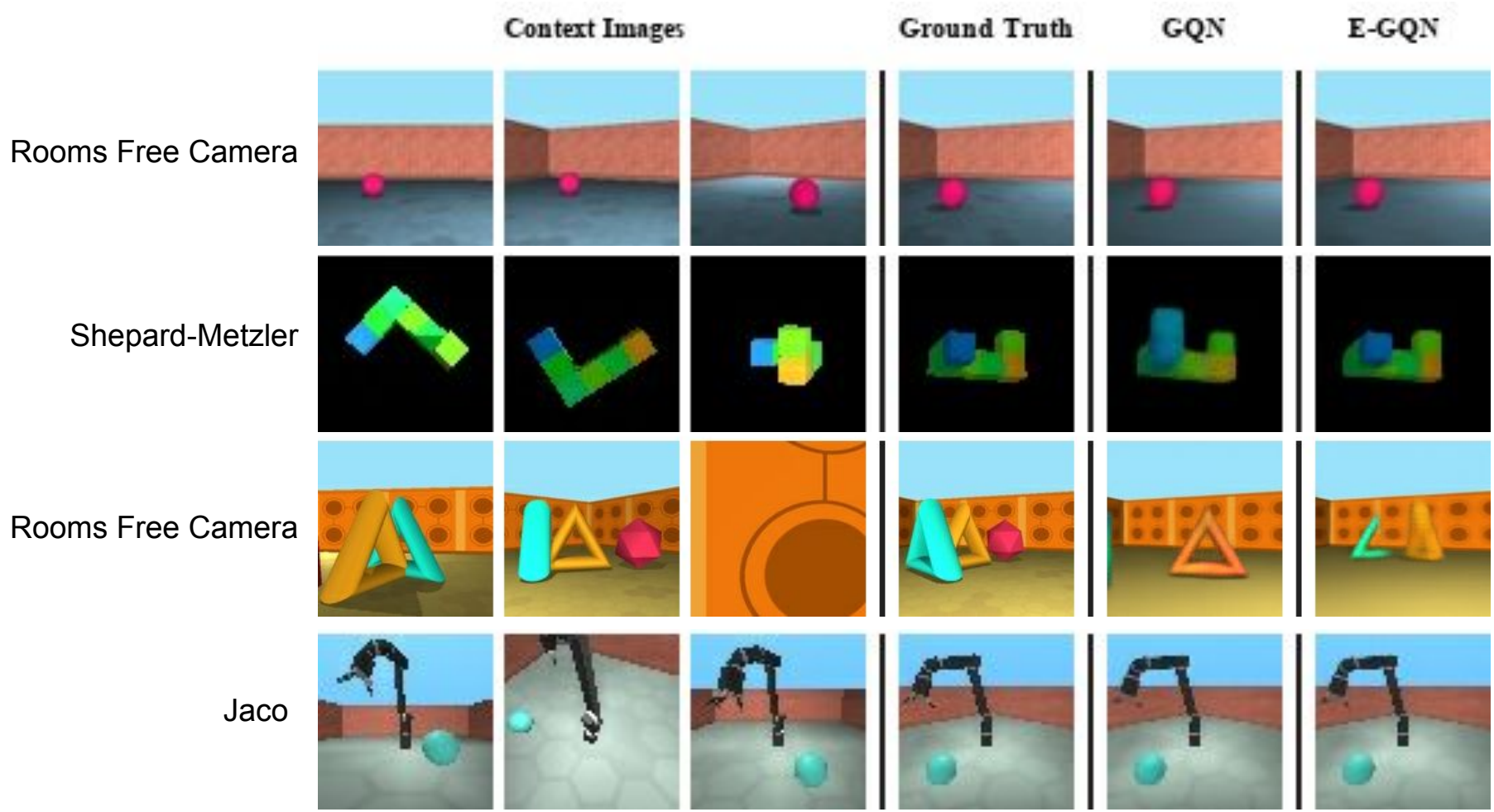


# Training

- 64 x 64 images for GQN
- 128 x 128 new datasets
- GQN datasets batch size of 36
- New datasets batch size of 32
- Trained models on 25M examples

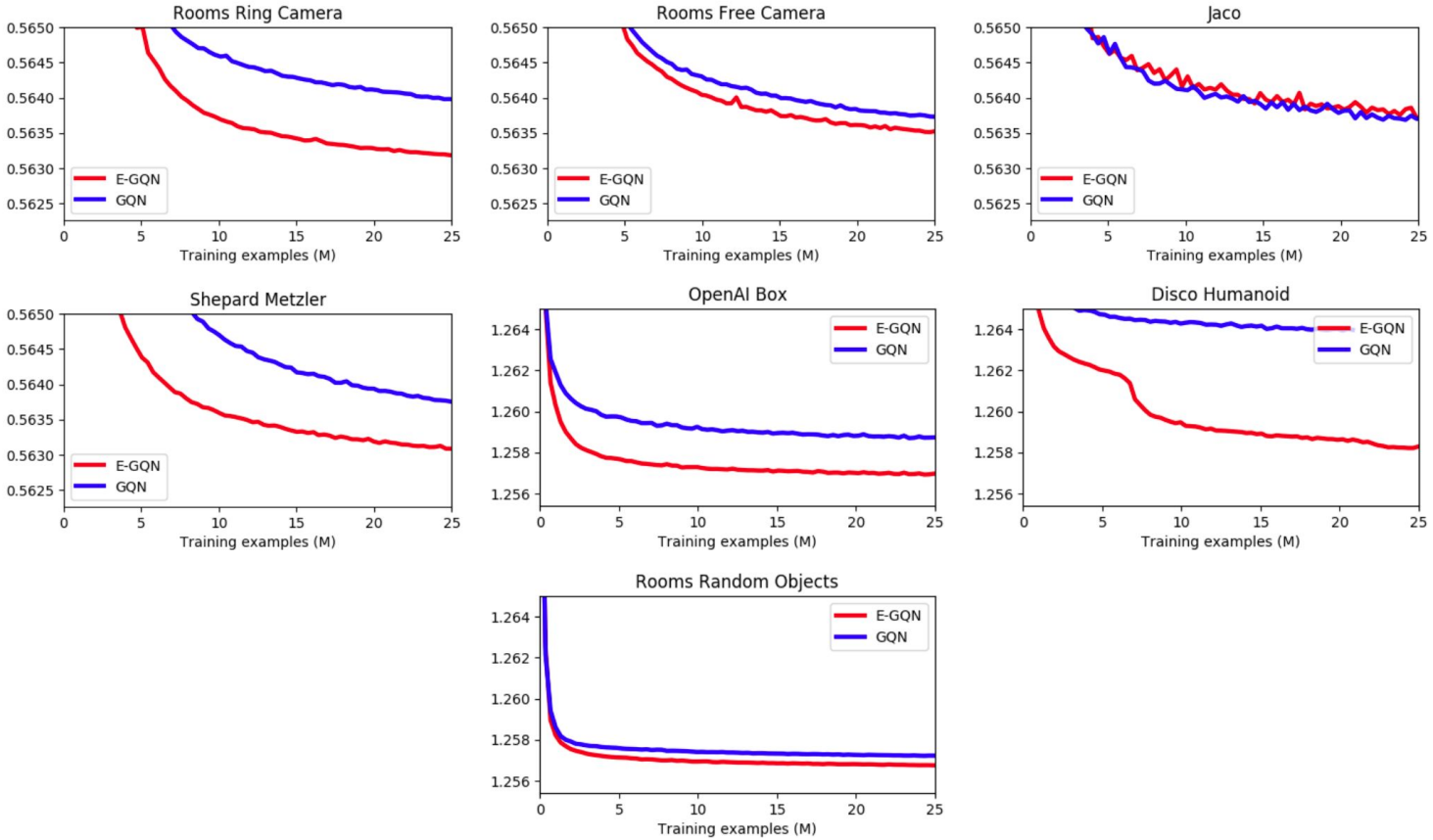
# Qualitative Results:




Context Images
Ground Truth
GQN
E-GQN
Rooms Free Camera
Shepard-Metzler
Rooms Free Camera
Jaco

# Quantitative Results

Maximizing the Evidence Lower Bound (ELBO)



# Quantitative Results

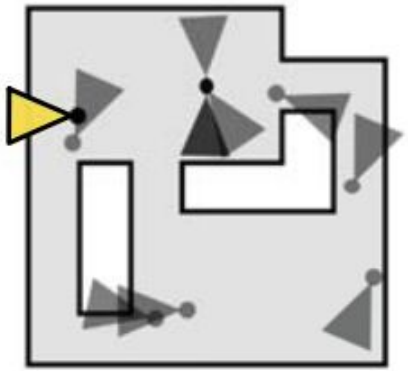
| Dataset | Mean Absolute Error (pixels) | | Root Mean Squared Error (pixels) | | ELBO (nats / dim) | |
|---|---|---|---|---|---|---|
| | GQN | E-GQN | GQN | E-GQN | GQN | E-GQN |
| rrc | $7.40 \pm 6.22$ | $\mathbf{3.59 \pm 2.10}$ | $14.62 \pm 12.77$ | $\mathbf{6.80 \pm 5.23}$ | $0.5637 \pm 0.0013$ | $\mathbf{0.5629 \pm 0.0008}$ |
| rfc | $12.44 \pm 12.89$ | $\mathbf{12.05 \pm 12.79}$ | $\mathbf{26.80 \pm 21.35}$ | $27.65 \pm 20.72$ | $\mathbf{0.5637 \pm 0.0011}$ | $0.5639 \pm 0.0012$ |
| jaco | $4.30 \pm 1.12$ | $\mathbf{4.00 \pm 0.90}$ | $8.58 \pm 2.94$ | $\mathbf{7.43 \pm 2.32}$ | $0.5634 \pm 0.0007$ | $\mathbf{0.5631 \pm 0.0005}$ |
| sm7 | $3.13 \pm 1.30$ | $\mathbf{2.14 \pm 0.53}$ | $9.97 \pm 4.34$ | $\mathbf{5.63 \pm 2.21}$ | $0.5637 \pm 0.0009$ | $\mathbf{0.5628 \pm 0.0004}$ |
| oab | $10.99 \pm 5.13$ | $\mathbf{5.47 \pm 2.54}$ | $22.11 \pm 8.00$ | $\mathbf{10.39 \pm 4.55}$ | $1.2587 \pm 0.0018$ | $\mathbf{1.2569 \pm 0.0011}$ |
| disco | $18.86 \pm 7.16$ | $\mathbf{12.46 \pm 9.27}$ | $32.72 \pm 6.32$ | $\mathbf{22.04 \pm 11.08}$ | $1.2635 \pm 0.0055$ | $\mathbf{1.2574 \pm 0.0007}$ |
| rro | $10.12 \pm 5.15$ | $\mathbf{6.59 \pm 3.23}$ | $19.63 \pm 9.14$ | $\mathbf{12.08 \pm 6.52}$ | $1.2573 \pm 0.0011$ | $\mathbf{1.2566 \pm 0.0009}$ |

RRC: Rooms Ring Camera
RFC: Rooms Free Camera
Jaco: robotic Jaco arm
SM7: Shapard-Metzler-7 (lego cubs)
OAB: Open-AI blocks
Disco: Disco-humaniod!
RRO: rooms-random-objects

# Limitations:

- E-GQN limited to epipolar correspondence in inputs
- GQN predicts unseen maze wall (yellow)
- GQN does Scene Algebra

Observations



Prediction



Truth

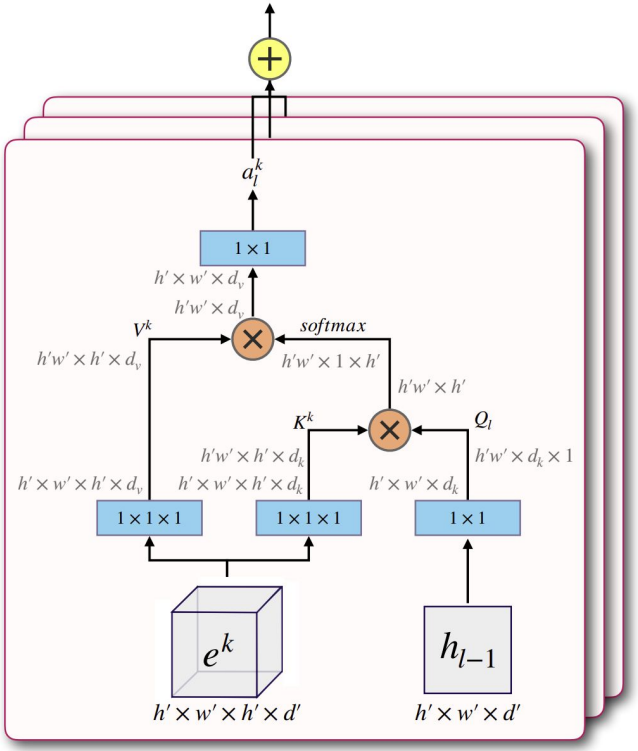


# Questions

# Supplemental slides

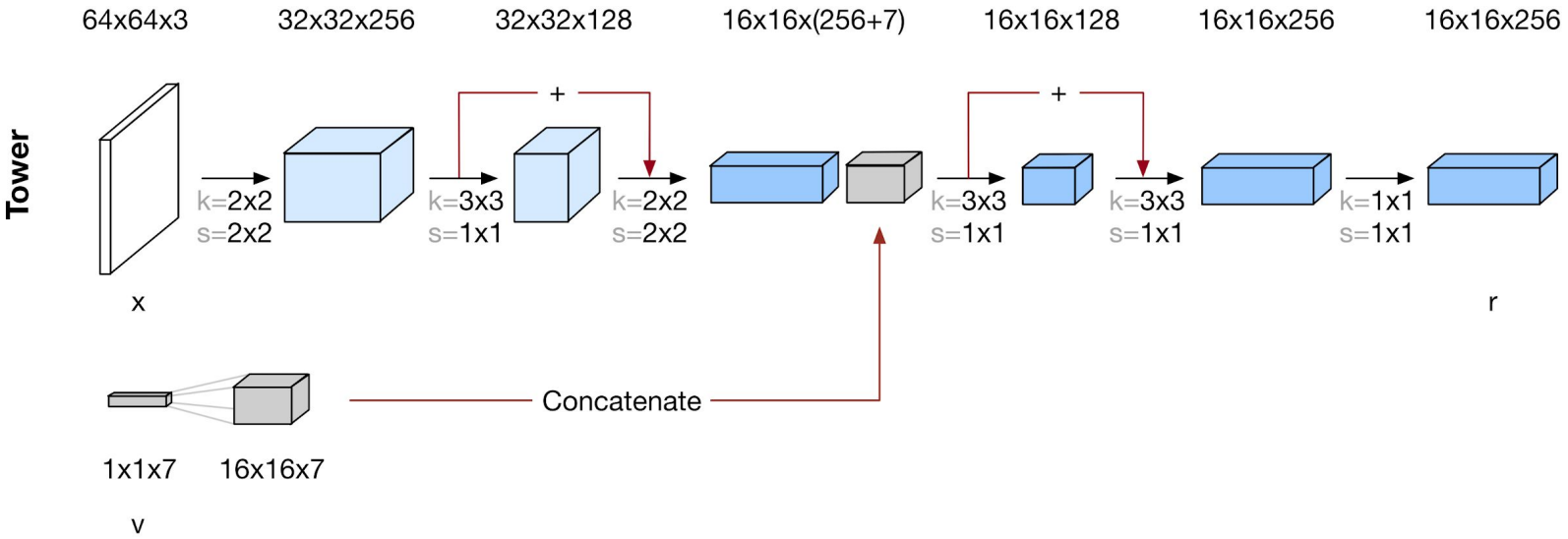
# Attention Mechanism

Our attention mechanism. Blue rectangles denote convolutional layers with given kernel size. “×" denotes batch-wise matrix multiplication, and “+" element-wise summation. The previous decoder hidden state $h_{l-1}$ is used to compute a query tensor $Q_l$ by linear projection. The epipolar representation $e^k$ is also linearly projected to compute a key tensor $K^k$ and value tensor $V^k$ . $K^k$ and $Q_l$ are matrix-multiplied to form unnormalized attention weights, which are scaled by 1/ $\sqrt{d_k}$. A softmax is computed along the final dimension, and the result is multiplied by $V^k$ to get an attention score.




(b) Attention mechanism

# GQN “Tower”


Tower
64x64x3
32x32x256
32x32x128
16x16x(256+7)
16x16x128
16x16x256
16x16x256
+
+
k=2x2
s=2x2
k=3x3
s=1x1
k=2x2
s=2x2
k=3x3
s=1x1
k=3x3
s=1x1
k=1x1
s=1x1
x
r
Concatenate
1x1x7
16x16x7
v

# Disadvantages

- Epipolar representation
- E-GQN 30% slower then GQN

There are two performance disadvantages of our implementation of E-GQN. First, E-GQN requires computing the epipolar representation e k for each context viewpoint. Each e k is a h 0 × w 0 × h 0 × d 0 tensor, which can could cause issues fitting e k into GPU memory for larger image sizes. Second, due to extra computation of e k and the attention maps al , E-GQN processes around 30% fewer samples per second than GQN in our experiments. In practice, E-GQN reaches a given loss value significantly faster in wall clock time on most dastasets due to better data efficiency

# Problem Statement

Given K images x k ∈ X and corresponding camera viewpoints v k ∈ V of a scene s, the goal of neural rendering is to learn a model that can accurately predict the image x q the camera wouldsee from a query viewpoint v q. More formally, for distributions of scenes p(S) and images with corresponding viewpoints p(V, X | s), the goal of neural rendering is to learn a model that maximizes

$$\mathbb{E}_{s\sim p(S)}\mathbb{E}_{v^q,x^q\sim p(V,X|s)}\mathbb{E}_{v^k,x^k\sim p(V,X|s)}\log p\left(x^q|(x^k,v^k)_{k=\{1,\cdots,K\}},v^q\right)$$

# Epipolar Cross Attention (ECA)

- ECA computes a response at a given spatial position as a weighted sum at all relevant positions of feature maps from the context images.
- Relevant features are those on the epipolar line in the context viewpoint.

# Attention Mechanism:

- Map the previous decoder hidden state $h_{l-1}$ and the epipolar representations $e^k$ to an attention score $a^k_l$
- Lambda is a learnable parameter
- The attention maps are provided a skip connection to

$$a_l = \lambda \sum_k a_l^k + \sum_k r^k$$