Neural scene representation and rendering

S. M. Ali Eslami, Danilo J. Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu,
Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz,
Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu and Demis Hassabis.

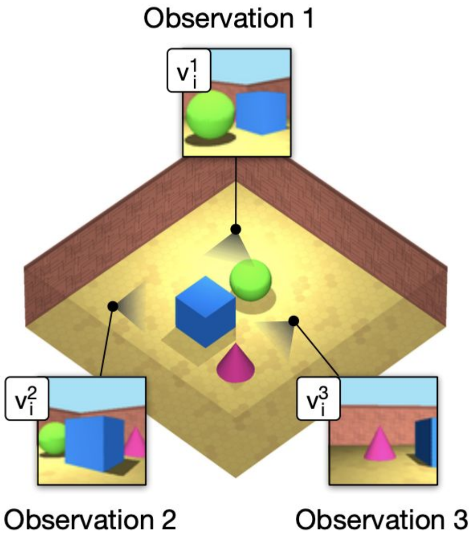
Science 15 Jun 2018: Vol. 360, Issue 6394

2020.09.22

Gyubok Lee

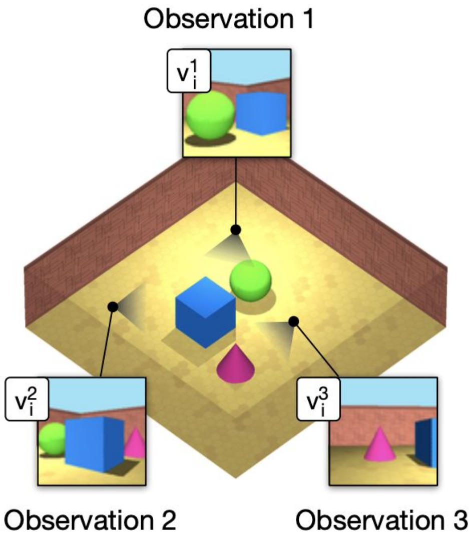
Motivation

- Humans can guess how an object looks like from perspectives that we have not seen.
  - Such visual and cognitive tasks are effortless to humans, but artificial systems are hardly capable of doing them
  - Most today's visual recognition systems are trained using large datasets (with annotated labels), which limits their capability.
  - We want machines that can automatically and fully understand the surroundings (objects, their attributes, light source, etc.) without manually giving data.



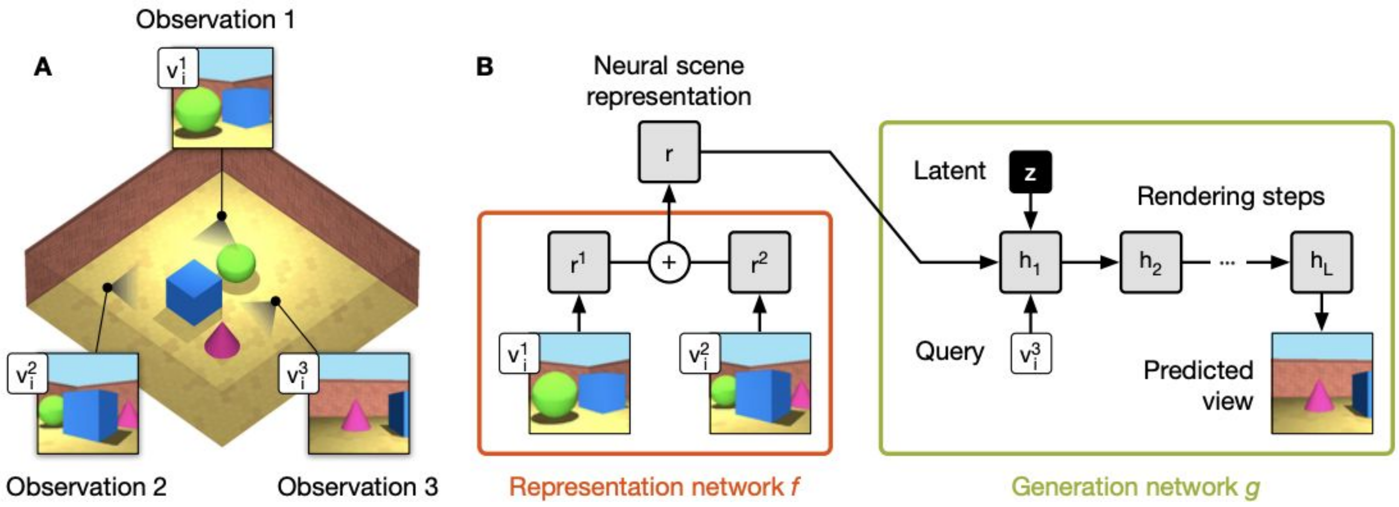
Task

- Scene Understanding
  - Many attributes in the scene (such as wall color, texture, multiple objects, their different sizes and colors)
  - If possible to imagine how different scenes looks like (answer) from a different point (query) based on previous observations (context), we can say the model understands the scene (as in a QA task in NLU)
  - We call this model Generative Query Network (GQN)



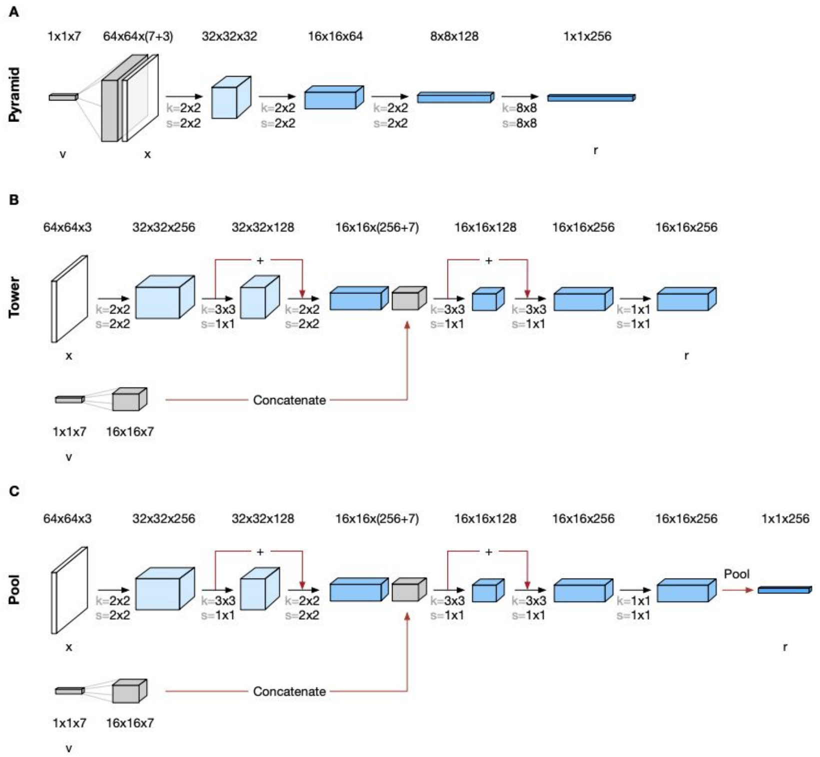
Model Architecture

- GQN has two networks: Representation network and Generative network



Model Architecture

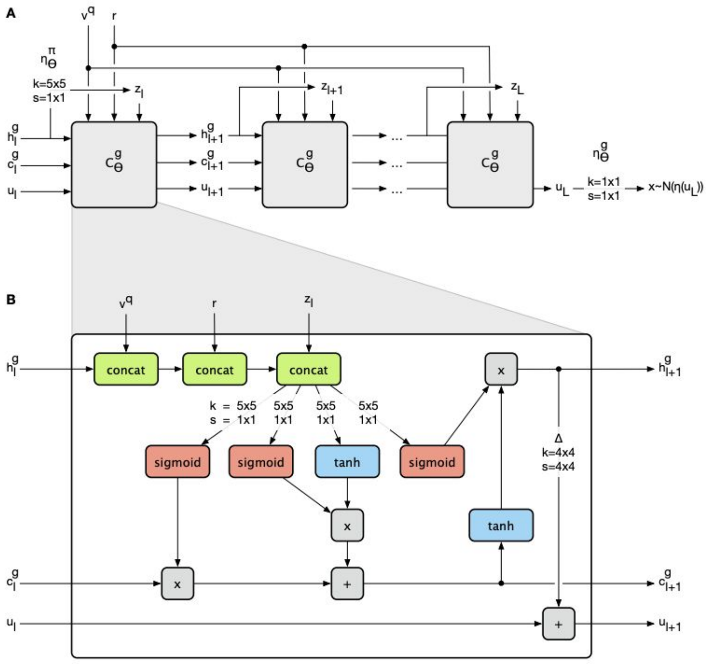
- Representation network
  - Input: $\boldsymbol{o}_i = \{(\boldsymbol{x}_i^k, \boldsymbol{v}_i^k)\}_{k=1,\ldots,K}$
  - Output: $\boldsymbol{r} = f_\theta(\boldsymbol{o}_i)$
  - If multiple r, we sum them
- Different characteristics for each network architecture:
  - Tower: fastest to learn, but less factorized
  - Pyramid & Pool: factorized across different object properties



Possible architectures

Model Architecture

- Generation network
  - Input: Query viewpoint, $\mathbf{v}^q$, and representation *r*
  - Output: Query image, $\mathbf{x}^q$
- Recurrent latent variable model (RNN + VAE)
  - Vector of latent variable *z* is split into *L* groups in an auto-regressive manner
  - Latent variable for each *l*: $\pi_\theta(\mathbf{z}_l|\mathbf{v}^q, \mathbf{r}, \mathbf{z}_{<l}) = \mathcal{N}(\mathbf{z}_l|\eta^\pi_\theta(h^g_l)),$
  - The prior: $$\pi_\theta(\mathbf{z}|\mathbf{v}^q, \mathbf{r}) = \prod_{l=1}^{L} \pi_\theta(\mathbf{z}_l|\mathbf{v}^q, \mathbf{r}, \mathbf{z}_{<l})$$



One possible architectures

Training

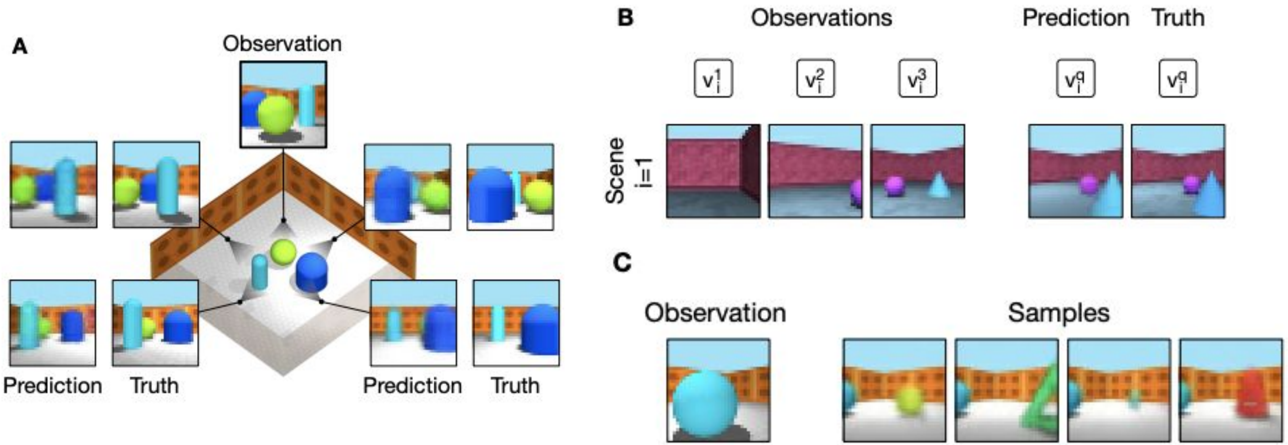
- Meta-learning style of training
  - For each iteration, we train on new different scenes, to avoid overfitting to one scene
  - Forced to learn whatever context the model gets

- Optimization
  - Variational approximation by minimizing the loss function below
  - Evidence lower bound (ELBO), here *-F(θ, φ)*, is decomposed into the reconstruction likelihood and a regularization term

$$\mathcal{F}(\theta, \phi) = \mathbb{E}_{(\mathbf{x},\mathbf{v})\sim D, \mathbf{z}\sim q_\phi} \left[-\ln \mathcal{N}\left(\mathbf{x}^q | \eta_\theta^g\left(\mathbf{u}_L\right)\right) + \sum_{l=1}^{L} \mathbf{KL}\left[\mathcal{N}\left(\cdot|\eta_\phi^q\left(\mathbf{h}_l^e\right)\right) || \mathcal{N}\left(\cdot|\eta_\theta^\pi\left(\mathbf{h}_l^g\right)\right)\right] \right]$$

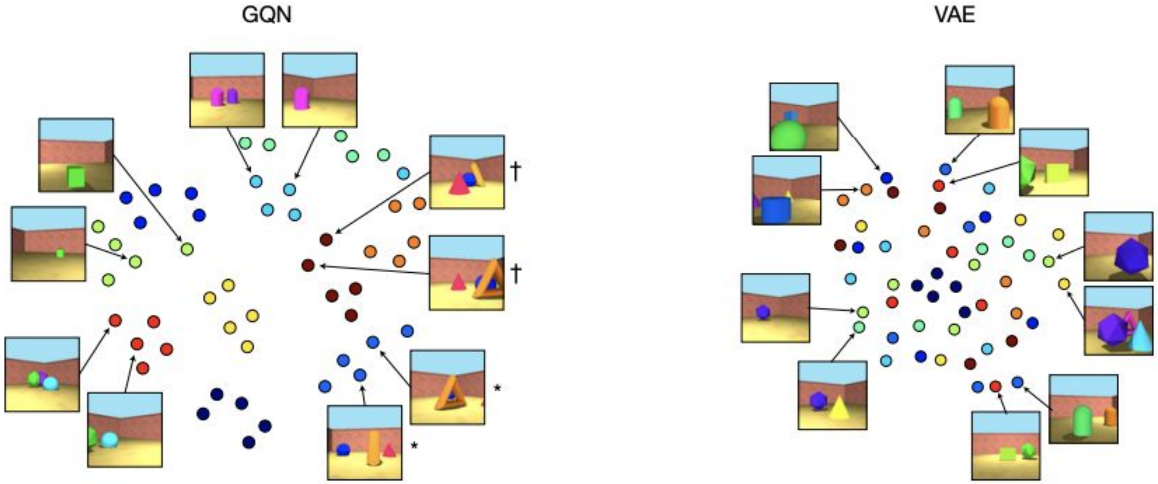
Experiments

- Neural scene representation and rendering
    - GQN's generator learns an approximate 3D renderer (a program that can generate an image when given a scene representation and camera viewpoint)
    - (A) Accurate images from arbitrary query viewpoint; (B) Consistent with laws of perspective, occlusion, and lighting; (C) Sample variability indicates uncertainty over scene contents



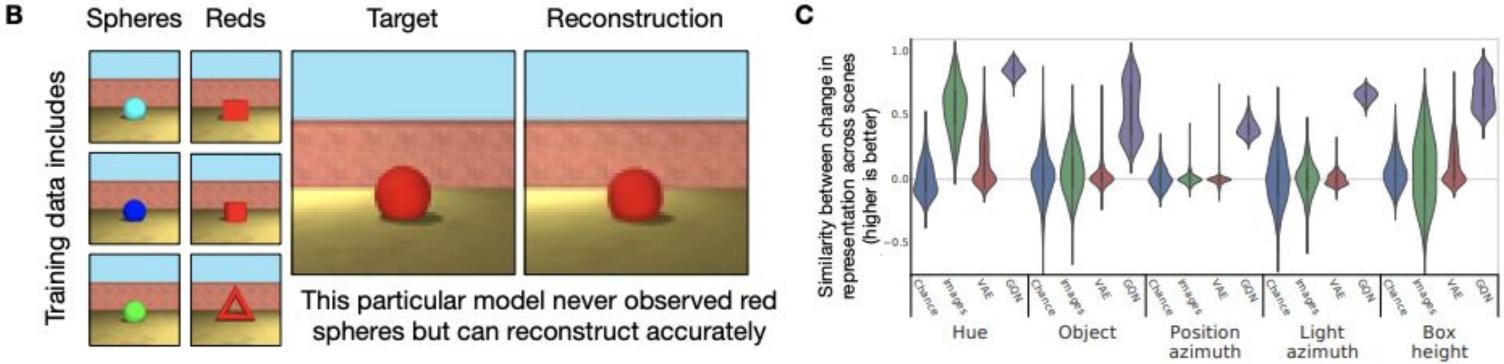
Experiments

- Viewpoint invariance
  - t-SNE embeddings visualization (GQN vs VAE)
  - VAE captures mostly wall angles; GQN can encode scene representations computed from each image individually



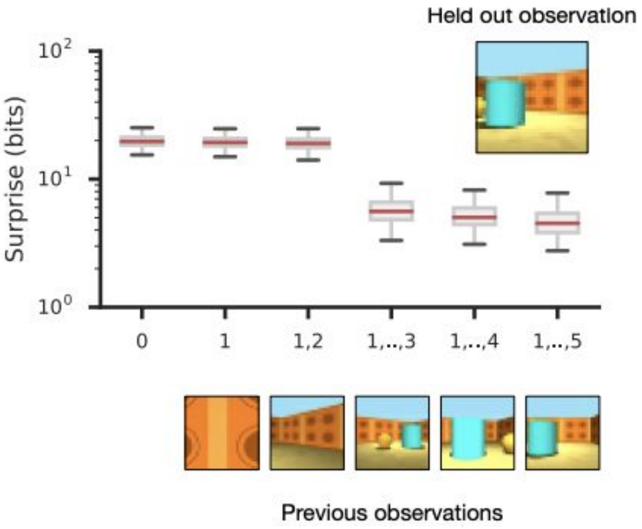
Experiments

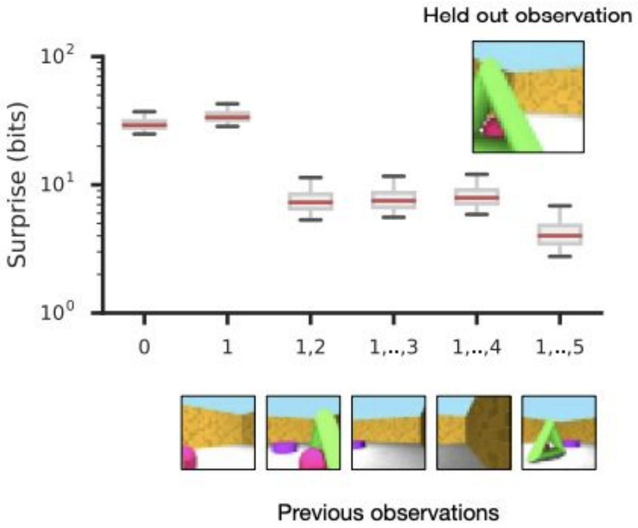
- Compositionality and factorization of the learned scene representations
  - (B) Reconstruction of holdout shape-color combinations.
  - (C) By changing one attribute, the representations are shifting (factorization of scenes representations)



Experiments

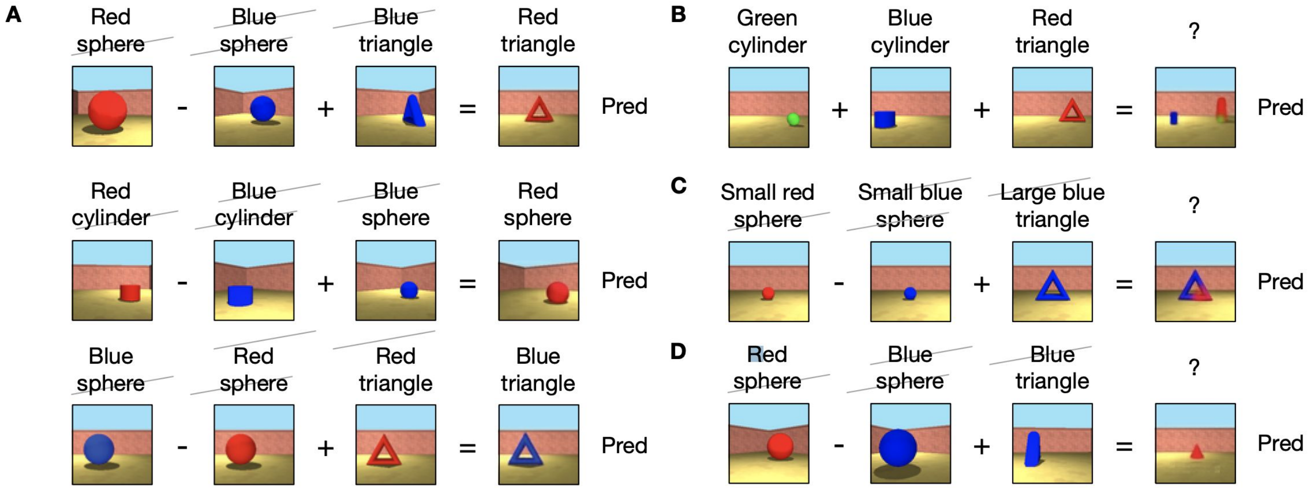
- Information gain
  - The model’s surprise of the held-out observation drops most sharply when it views the similar scenes (position, shape, color, etc.)



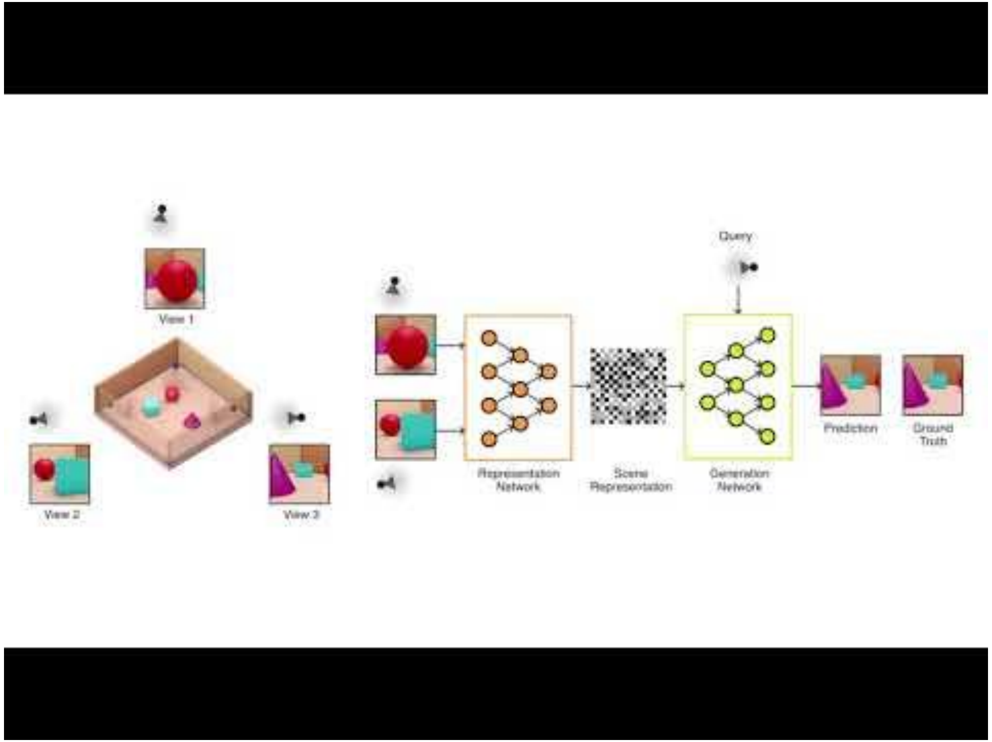


Experiments

- Scene algebra
  - The model can correctly modify/recombine scenes in a variety of settings
  - But fail to combine different objects in (B) and objects with different sizes in (C)



Summary


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

Related Work

- Traditional structure-from-motion, structure-from-depth and multi view geometry techniques
    - Requires 3D structure of the environment
- Classical neural generative models (e.g. auto-encoding, density models)
    - Capturing only the distribution of observed images
- Viewpoint transformation networks
    - Requires explicit relationships; non-probabilistic and limited in scale

Contribution

- GQN learns representations that adapt to and compactly capture the important details of its environment (positions, color, objects, textures, lights, etc.) without any human labelling of the scenes
- GQN learns disentangled semantics (though not interpretable by humans) by itself and in a generally applicable manner
- GQN learns a powerful neural renderer that is capable of producing accurate and consistent images of scenes from new query viewpoints
- A step towards fully unsupervised scene understanding