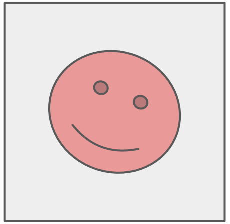
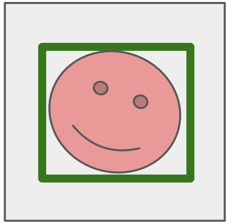
Analysis By Synthesis

Beyond Detection and Recogntion

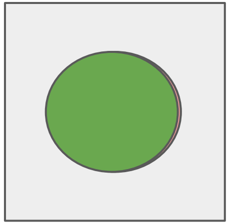
Image Analysis



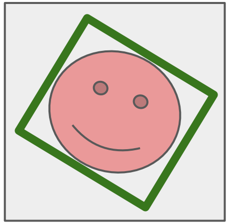
Classification: is-a-face



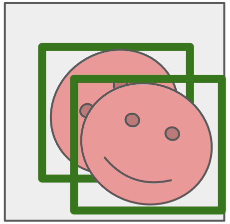
Location: find the face



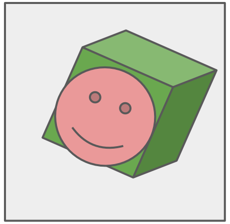
Segmentation: pixel-level location



Pose: find the face and orientation



Instance segmentation: multiple face



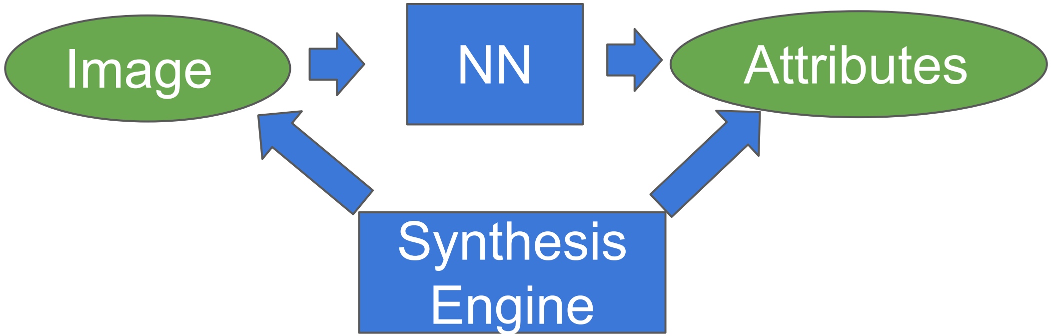
3D pose: out-of-the-plane rotation

- The curse of easy task: generalization may suffer
  - Not sure if NN has really mastered the task or just overfitted some data

Image Analysis: From Image To Attributes

- Problem
  - Labor-intensive labeling
    - E.g.: often ignore 3D pose as too hard
  - Require Balanced Training Data Coverage

Data Augmentation By Synthesizing



- Pro
  - Balanced Training Data Coverage
- Con
  - Need be Realistic (labor-intensive CG)
    - precise model
    - mostly limited to 2D

Ladder Network

- Model Viewpoint
- ***Implicit*** Intermediate representation, not as useful

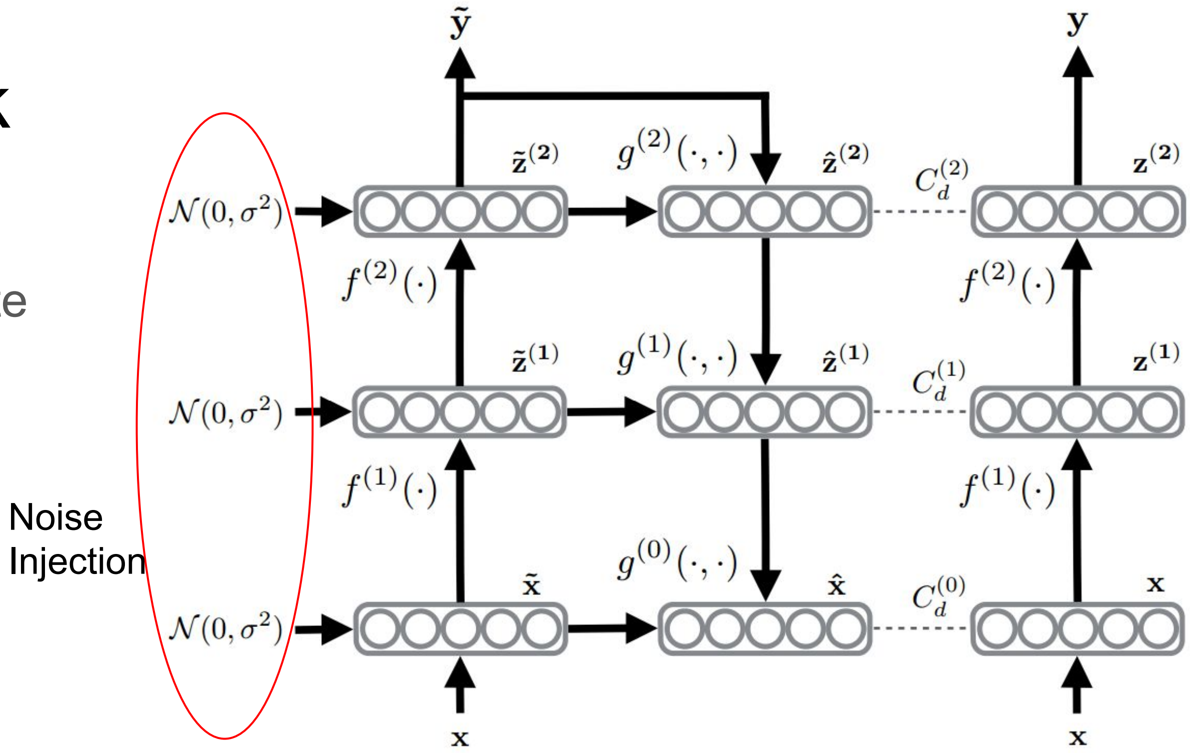
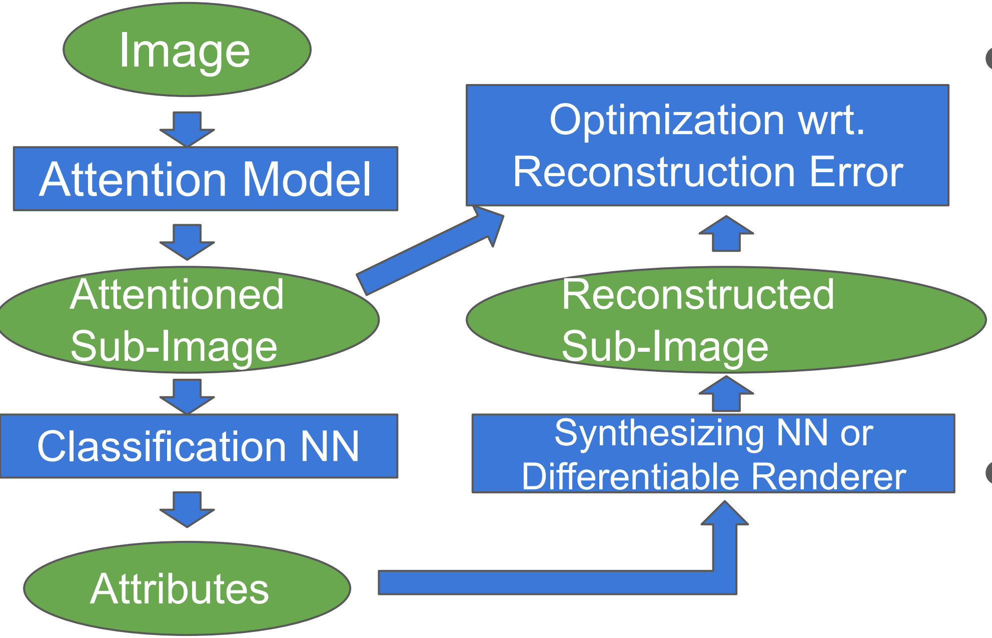


Figure 2: A conceptual illustration of the Ladder network when $L = 2$. The feedforward path ($\mathbf{x} \rightarrow \mathbf{z}^{(1)} \rightarrow \mathbf{z}^{(2)} \rightarrow \mathbf{y}$) shares the mappings $f^{(l)}$ with the corrupted feedforward path, or encoder ($\mathbf{x} \rightarrow \tilde{\mathbf{z}}^{(1)} \rightarrow \tilde{\mathbf{z}}^{(2)} \rightarrow \tilde{\mathbf{y}}$). The decoder ($\tilde{\mathbf{z}}^{(l)} \rightarrow \hat{\mathbf{z}}^{(l)} \rightarrow \hat{\mathbf{x}}$) consists of the denoising functions $g^{(l)}$ and has cost functions $C_d^{(l)}$ on each layer trying to minimize the difference between $\hat{\mathbf{z}}^{(l)}$ and $\mathbf{z}^{(l)}$. The output $\tilde{\mathbf{y}}$ of the encoder can also be trained to match available labels $t(n)$.

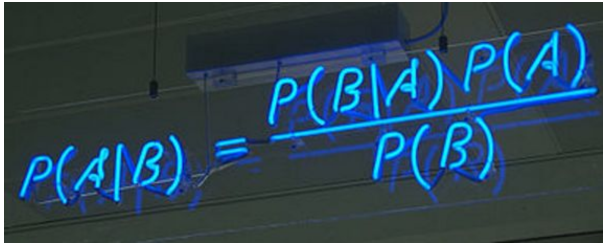
Analysis By Synthesis



- Attention + Auto-Encoder + Inference-time Optimization
- Pro
  - Can use unlabeled real data (auto-encoder)
  - Optimize reconstruction error at inference-time
  - Suffices to reconstruct part of image (attention)
- Con
  - More computation during Inference stage

Analysis By Synthesis

- Model viewpoint
    - Semi-supervised learning that can exploit unlabeled data
- Reconstruction viewpoint
    - Explain away of parts of images
    - Training time reconstruction allows auto “labeling” of data
        - can be costly
    - Test time reconstruction allows providing more reliable confidence metrics than NN classification score
- Bayesian viewpoint
    - B is a geometric shape
    - A is a character
    - Natural incorporation of
        - detection confidence P(B)
        - prior P(A), e.g. language model
        - Synthesis model P(B|A)



From Attributes to Image

- Synthesizing NN
  - Decoder part of auto-encoder
  - Relies on NN’s ability, may not be as sharp for characters
- Differentiable Renderer
  - https://github.com/mattloper/opendr/wiki

Possibility: Perfect Analysis of Document Image

- A document image is made up from
  - simple lines
  - clean background
  - non-handwriting characters
- We may synthesize the document image
  - As a consequence, can use whole image reconstruction error as metrics
  - Leading to perfect analysis

Literature: Analysis by Synthesis: 3D Object Recognition by Object Reconstruction (CVPR '14)

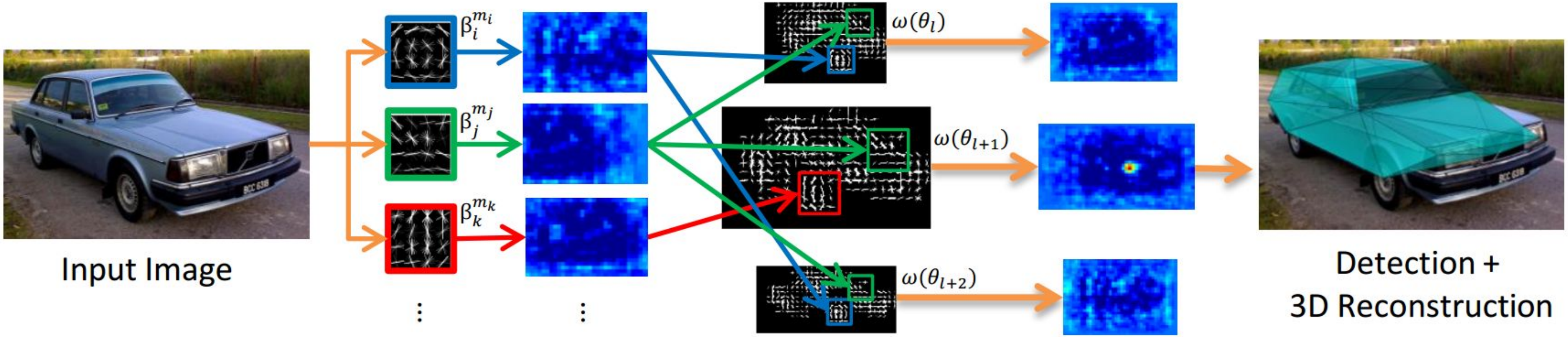


Figure 4. We search through a large collection of templates (with shared parts) by first caching part responses, and then looking up response values to score each template.

Literature: Enriching Object Detection with 2D-3D Registration and Continuous Viewpoint Estimation (CVPR '15)

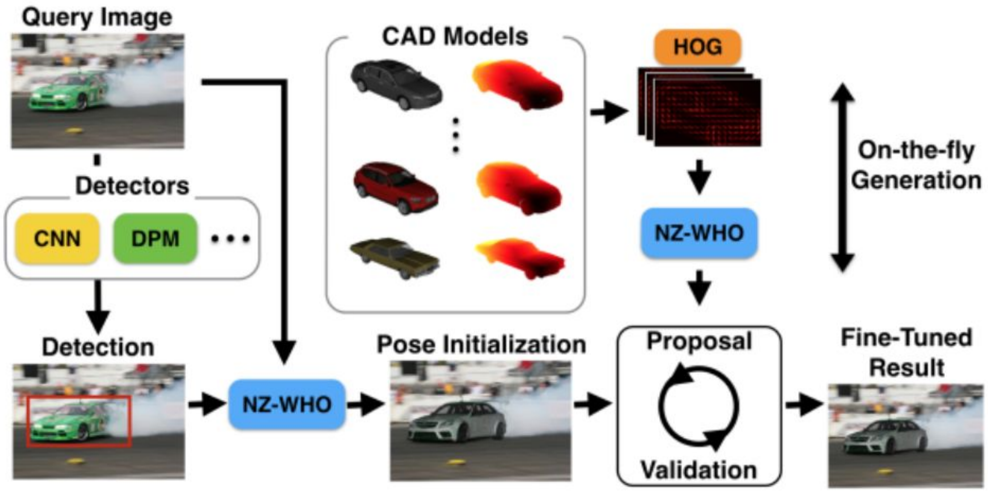
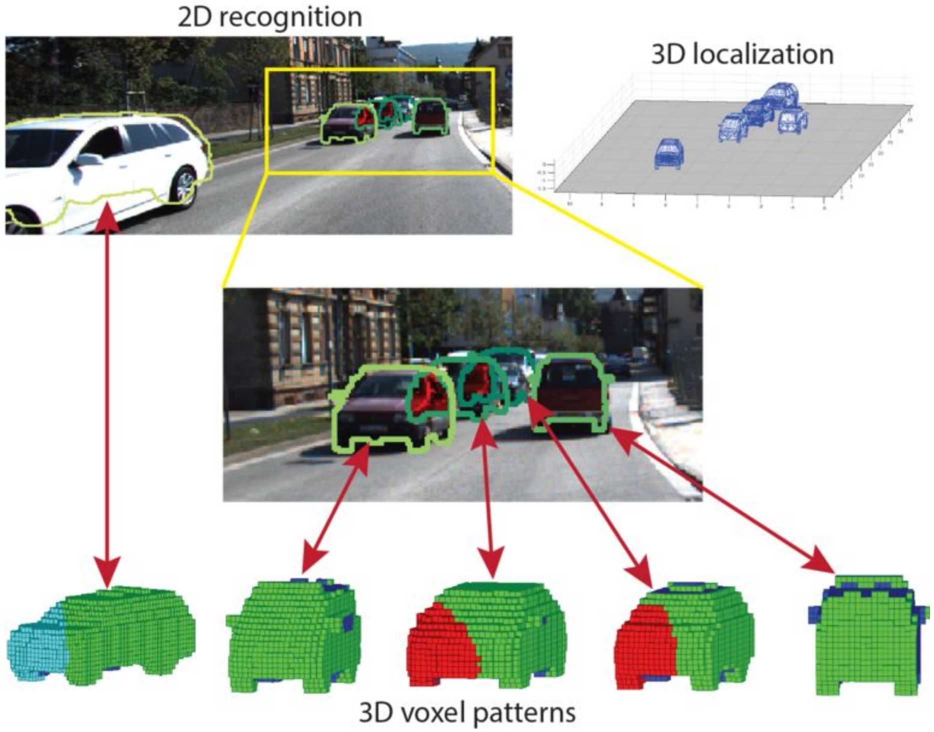
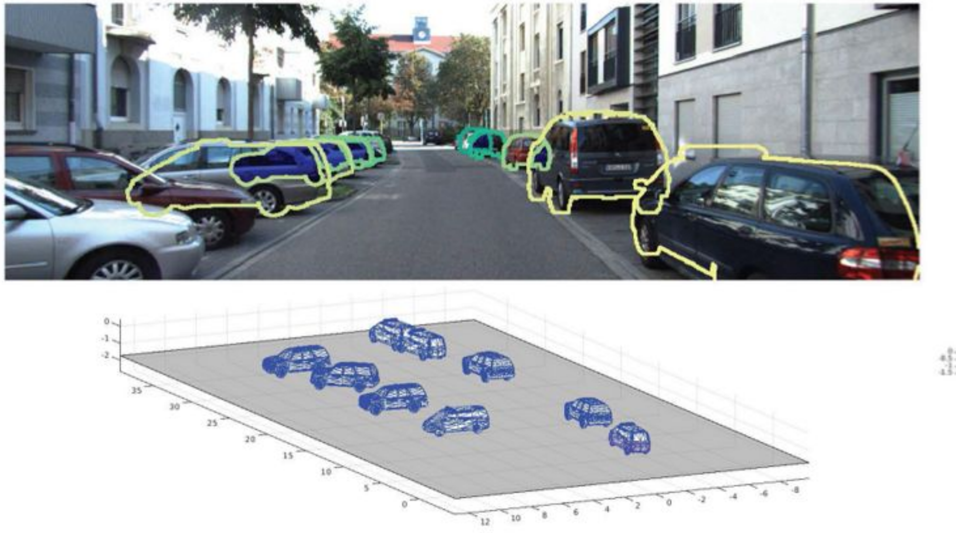


Figure 1: Using a database of 3D CAD models, we generate NZ-WHO templates which can be used to either detect objects directly or enrich the output of an existing detector with high-quality, continuous pose and 3D CAD model exemplar.

Figure 8: Effect of fine tuning. (left) original image, (middle) initial detection, (right) continuous fine tuning using Single-Component Metropolis Hastings

Literature: Data-Driven 3D Voxel Patterns for Object Category Recognition (CVPR ‘15)





Literature: Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations (COGSCI ‘15)

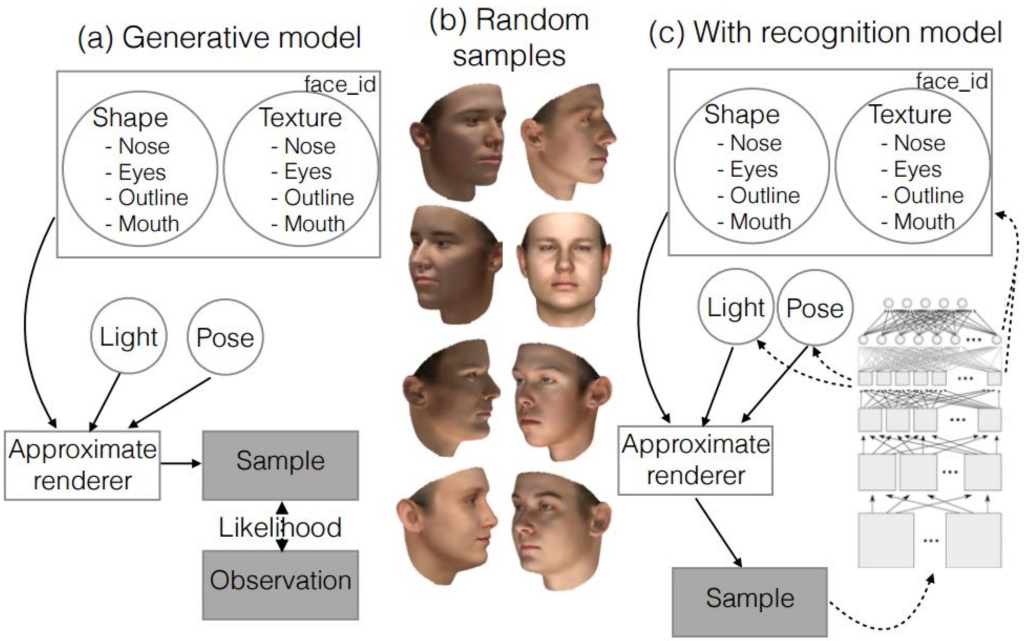
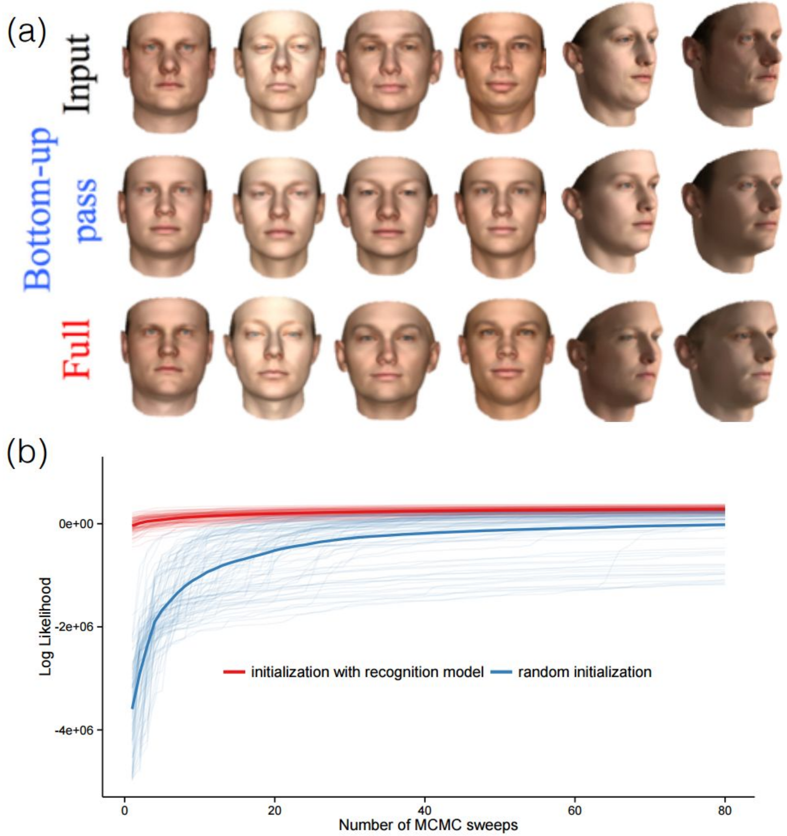


Figure 2: (a) Overview of the inverse graphics model. (b) Random draws from the model. (c) Training and the use of the recognition model.



Literature: Scene Text Segmentation via Inverse Rendering (ICDAR'13)

(a) input (b) segmentation (c) rendered

Literature: See the Difference: Direct Pre-Image Reconstruction and Pose Estimation by Differentiating HOG

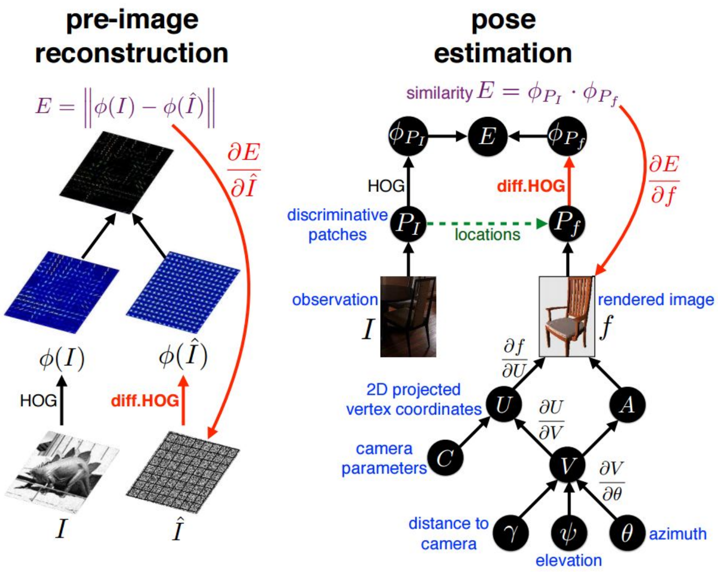


Figure 1: We exploit the piecewise differentiability of the popular HOG descriptor for end-to-end optimization. The figure shows applications on the pre-image reconstruction given HOG features as well as the pose estimation task based on the same idea.