IGE-Net: Inverse Graphics Energy Networks for Human Pose Estimation and Single-View Reconstruction

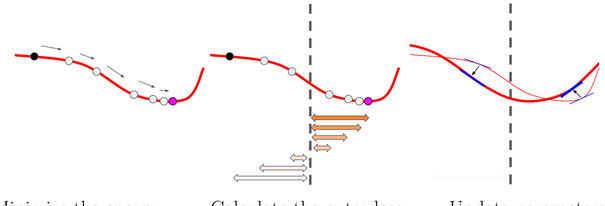
Dominic Jack [1] Frederic Maire [1] Sareh Shirazi [1] Anders Eriksson [2]

[1] School of Electrical Engineering and Computer Science, Queensland University of Technology
[2] School of Information Technology and Electrical Engineering, University of Queensland

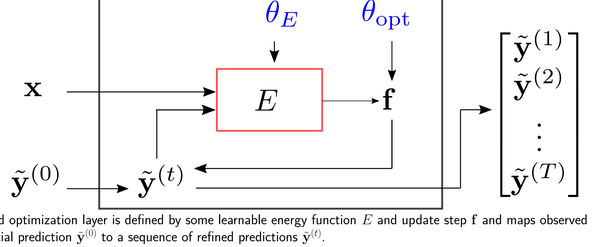
Abstract

Inferring 3D scene information from 2D observations is an open problem in computer vision. We propose using a **deep-learning based energy minimization framework** to learn a consistency measure between 2D observations and a proposed world model, and demonstrate that this framework can be trained end-to-end to produce consistent and realistic inferences. We evaluate the framework on human pose estimation and voxel-based object reconstruction benchmarks and show competitive results can be achieved with relatively **shallow networks** with **drastically fewer learned parameters** and floating point operations than conventional deep-learning approaches.

Energy Prediction Networks [1]

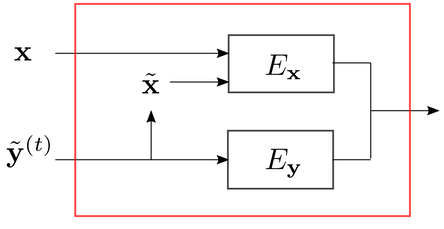


Minimize the energy (red) with respect to proposed solution. Calculate the outer loss (orange) using label information (dashed). Update parameters of the energy function (blue).



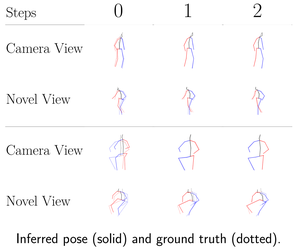
An unrolled optimization layer is defined by some learnable energy function E and update step $\mathbf{f}$ and maps observed features $\mathbf{x}$ and an initial prediction $\tilde{\mathbf{y}}^{(0)}$ to a sequence of refined predictions $\tilde{\mathbf{y}}^{(t)}$.

Human Pose Lifting

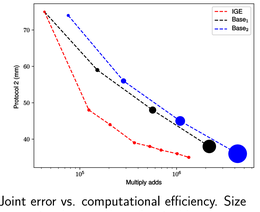


- Initial estimate $\tilde{\mathbf{y}}^{(0)}$ based on a simple MLP [2].
- $E_{\mathbf{x}}$ and $E_{\mathbf{y}}$ are both shallow MLPs with relatively few parameters.
- Pairwise-distance preprocessing enforces invariance to rotation and translation.

$$\Delta^2_{ij}(\mathbf{z}) = ||\mathbf{z}_i - \mathbf{z}_j||_2^2$$



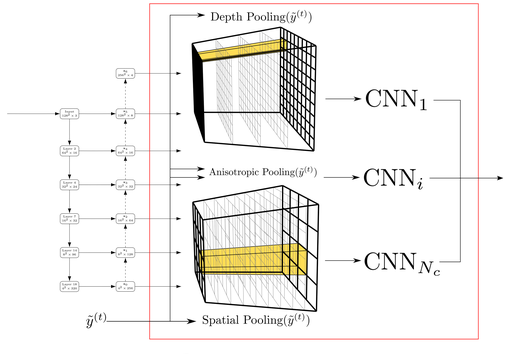
Inferred pose (solid) and ground truth (dotted).

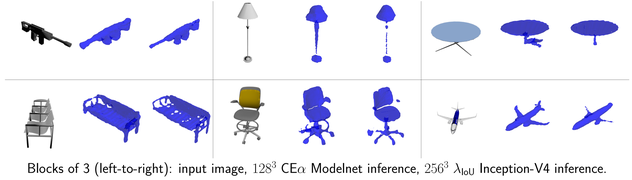


Joint error vs. computational efficiency. Size proportional to number of trainable parameters. Low, left and small are good.

+ $\sim 100\times$ fewer parameters than baseline MLP model
+ Accuracy can be traded off for computation in the same model
+ Error within 10% in 4 steps ($\sim 10\times$ fewer ops), comparable at 16 ($\sim 3\times$ fewer)
+ 2D/3D annotations need not be consistent
+ Same model can be retrofitted for different camera intrinsics
− Requires known camera intrinsics
− Inherently iterative – potentially slower on accelerators like GPUs.

3D Single View Reconstruction





Blocks of 3 (left-to-right): input image, 128^3 CEα Modelnet inference, 256^3 λ_{IoU} Inception-V4 inference.

	car				plane				table			
Resolution	32	64	128	256	32	64	128	256	32	64	128	256
OGN$_1$ [3]	64.1	77.1	78.2	76.6	-	-	-	-	-	-	-	-
MAT$_1$ [4]	68.3	78.4	79.4	**79.6**	36.7	48.8	58.0	59.6	38.6	42.3	43.5	41.3
IGE-MN$_{13}$	57.8	68.8	72.8	73.3	29.6	44.8	52.9	54.4	33.6	44.0	47.8	48.2
IGE-I4$_{13}$	57.9	70.9	74.0	75.2	30.5	47.8	57.5	57.3	34.8	46.5	52.7	50.5
IGE-MN$_1$	57.0	70.3	76.2	75.2	30.7	47.9	58.7	58.1	33.6	45.9	50.6	50.2
IGE-I4$_1$	58.4	71.2	76.5	76.5	30.1	49.2	60.5	**62.0**	35.0	46.4	52.2	**52.1**

Mean IoU (in %) trained at different resolutions and evaluated at 256^3.

[1] David Belanger, Bishan Yang, and Andrew McCallum.
End-to-end learning for structured prediction energy networks.
In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pages 429–439. JMLR.org, 2017.

[2] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little.
A simple yet effective baseline for 3d human pose estimation.
In Proceedings of the IEEE International Conference on Computer Vision, pages 2640–2649, 2017.

[3] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox.
Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs.
In Proceedings of the IEEE International Conference on Computer Vision (ICCV), volume 2, page 8, 2017.

[4] Stephan R Richter and Stefan Roth.
Matryoshka networks: Predicting 3d geometry via nested shape layers.
In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1936–1944, 2018.

tensorflow implementation
github.com/jackd/ige
d1.jack@qut.edu.au