Knowledge Augmented Deep Learning for Data Efficient, Generalizable, and Interpretable Visual Understanding

Qiang Ji
jiq@rpi.edu
Rensselaer Polytechnic Institute

Introduction

- Computer vision is about developing algorithms to automatically process images or videos to reconstruct, interpret and understand a 3D scene from its 2D images in terms of the geometric, spatial, and dynamic properties of the scene objects.

- Tasks of computer vision include object detection, object recognition, motion estimation, and 3D reconstruction.

Model-based Computer Vision

- Computer vision traditionally has been model/theory driven. Various models have been developed
  - Projection models
  - Photometric models
  - Motion models
- Domain knowledge about target objects
- These models and domain knowledge have contributed to the success of computer vision in many areas before 2012.
- These successes, however, tend to be incremental because
  - The models are either too simple to adequately model real world conditions or too complex to scale up.

Learning-based Computer Vision

- The latest developments in deep learning (since 2012) has led to significant improvements for some computer vision tasks, including object detection and object recognition.

- Object recognition with deep learning has, in fact , outperformed humans on benchmark datasets.

- Common belief: **Big data + deep learning +GPUs will solve Computer Vision problems and that no models/theories and domain knowledge are needed !**

Learning-based Computer Vision

- Data Inefficient
  - Needs big data and their annotations
    - Latest developments involve ever larger models, e.g.,the largest vision transformer has 2 billion parameters and is trained with 3 billion images.
    - These models are not feasible for many real world applications, and they are unsustainable theoretically, economically, and environmentally.
- Poor generalization
  - Do not perform well on novel data or across domains
- Lack of interpretability and performance guarantee
  - Black box and cannot accurately quantify its performance

Model-based v. Learning-based

Model-based

- Model and theory-driven
- Can exploit prior knowledge
- Explainable
- Data efficient
- Cannot scale up well
- Require strong assumptions

Learning-based

- Data-driven
- Scale up well
- Perform extremely well on within-data set
- Data inefficient
- Uninterpretable
- Cannot exploit prior knowledge

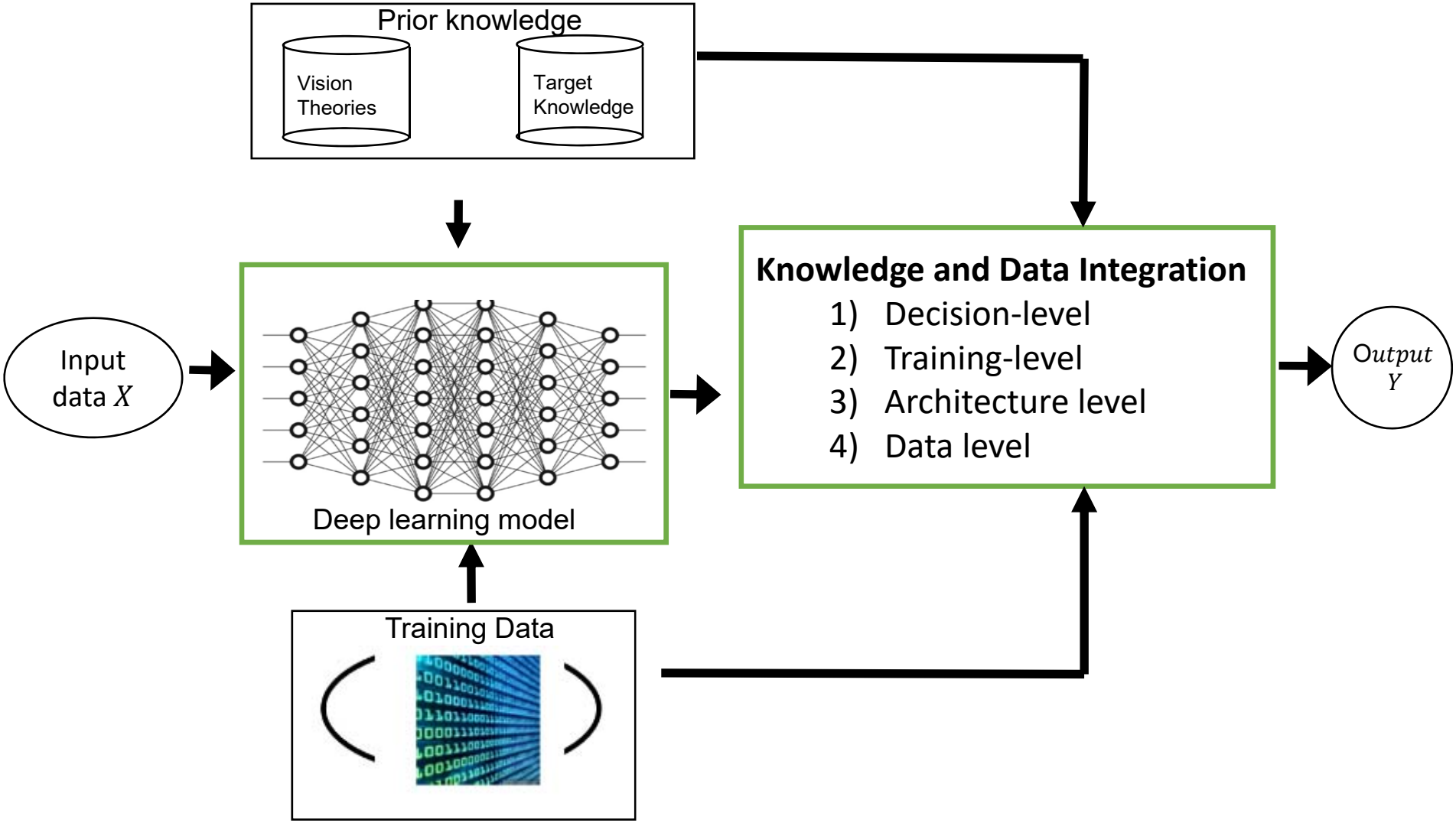
Knowledge-Augmented Computer Vision

- The long term success of computer vision requires a union of prior knowledge and data !

- Propose a hybrid vision model that combines mode-based computer vision with deep learning based computer vision to leverage their respective strengths and to ground the deep visual learning in the well-established prior knowledge

- Through the hybrid vision model, prior knowledge and data work synergistically to produce computer vision algorithms that are
  - data efficient- knowledge is transferred to the model, hence reducing dependencies on data
  - robust- prior knowledge reduces solution space and hence improve robustness
  - generalizable-prior knowledge is generic, applicable to different domains, and
  - interpretable-prior knowledge is based on human-derived theories and studies

Knowledge Augmented Computer Vision



Proposed Research

- **Knowledge identification**-identify prior knowledge from different sources

- **Knowledge integration**-integrate knowledge with image data for visual understanding

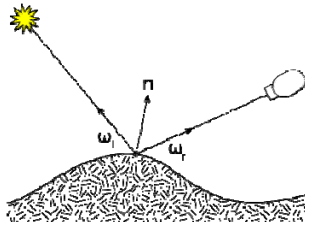
Prior Knowledge

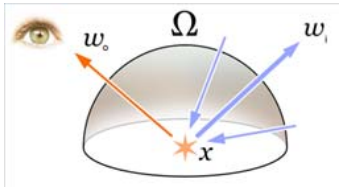
- **Computer vision models**- formal theories/principles that control the generation of the 2D images of 3D scenes.

- **Target knowledge**- theories or studies from different disciplines that govern the behaviors and properties of the target objects

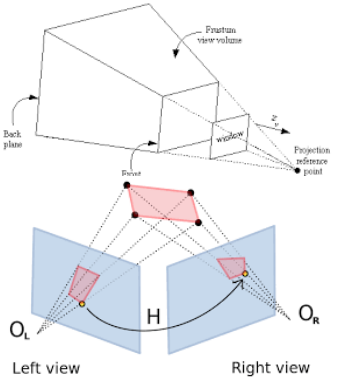
Computer Vision Models

- Photometric models
  - Illumination models: Lambertian, BDRF, and Sphere Harmonic $I(c,r) = \rho L \cdot N(\text{x,y,z})$
  - Rendering equation $L_o(x,\vec{w}) = L_e(x,\vec{w}) + \int_\Omega f_r(x,\vec{w}',\vec{w}) L_i(x,\vec{w}')(\vec{w}' \cdot \vec{n}) \mathrm{d}\vec{w}'$





- Projection models
  - Perspective, weak, affine, orthographic $\lambda \begin{bmatrix} c \\ r \\ 1 \end{bmatrix} = \begin{bmatrix} s_x f & 0 & c_0 \\ 0 & s_y f & r_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix}$
  - Multi-view geometry: fundamental matrix, epipolar constraint, homography, tri-tensor, and the ecological optics theory



- Motion models
  - Motion projection theories
  - Optical flows $\frac{\partial I}{\partial c}u + \frac{\partial I}{\partial r}v + \frac{dI}{dt} = 0$
  - Structure from motion M=RX factorization alg.

Target Knowledge

- Laws, theories, or studies from different disciplines that govern the properties and behaviors of the target objects. It varies with target domain.

- Human body is governed by theories and studies from
  - Anthropometric studies
  - Body anatomy
  - Body Biomechanics
  - Physics- kinematics and dynamics

Knowledge Integration

- Knowledge integration is to integrate the prior knowledge into the deep learning to ground the deep learning models in the well-established prior knowledge.

- Knowledge integration is challenging
  - Knowledge exists in different formats
    - Compute vision models and physics are often represented as mathematical equations (polynomial equations, ODE, integral equations, algebraic equations) ; precise and exact.

    - Target knowledge such as semantic relationships/dependencies may be expressed as constraints, graph, or logic rules; qualitative and inexact.

  - Knowledge is often incomplete, unambiguous, and uncertain

Knowledge Integration (cont'd)

Knowledge integration in different levels

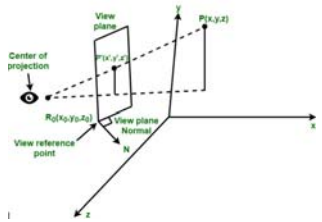
- Decision level
- Training level
- Architecture level
- Data level

Decision Level Integration

- A model-based vision algorithm and a learning based vision algorithm produce separate independent predictions.
- Their predictions are combined through a joint top-down and bottom-up inference.
- The combined prediction is expected to outperform the prediction by each algorithm alone.
- The same vision model can apply to different learning models as it is implemented independently.

Decision Level Integration

Model-based Vision



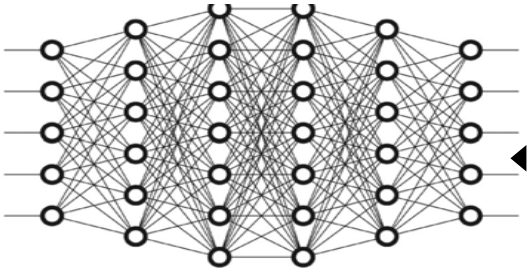
Top-down Predictions $\boldsymbol{Y}_M$

Prediction Level Integration through Bayes' Rule

Prior $\times$ Likelihood $\propto$ Posterior

Prediction **Y**

Input data $\boldsymbol{X}$



Bottom-up Estimates $\boldsymbol{Y}_N$

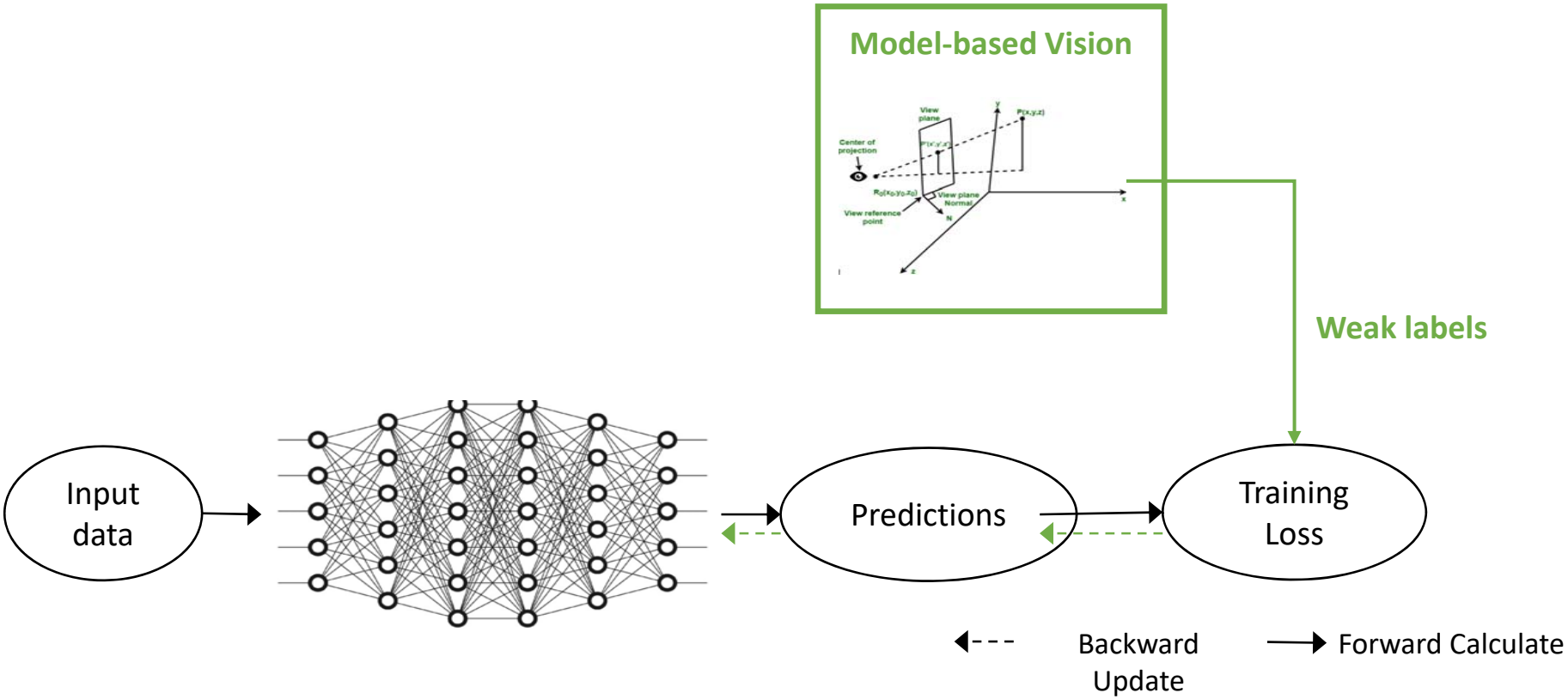
Learning-based Vision

Training Level Integration

- Weakly/self supervised learning
  - Produce weak labels by the model-based vision algorithms and pre-train the deep models using the weak labels
- Model regularization
  - Incorporate the knowledge into the loss function as regularization terms

Weakly/self-supervised Learning

1. Pre-train the model via self-supervised/weakly supervised learning



2. Fine tune the pre-trained model on target dataset using small amount of training samples .

Model Regularization

Convert target knowledge into constraints on the target variables **y**

- Variable constraints
  - Inequality or equality constraints on target variables **y**

    $a y_k + b > c$ *and* $a y_k + b = c$ or more generally, nonlinear constraints , i.e., f(y_k, a) > or = c
  - Implicit equality constraints $f(\mathbf{y}, \boldsymbol{\alpha}) = 0$
    - *f*() can be an algebraic, differential or integral equation.
- Probability constraints on p(**y**)

$$\mathrm{p}(y_k = \alpha) > \beta \quad \text{or} \quad \mathrm{p}(y_k = \alpha) > \mathrm{p}(y_j = \beta) \quad \text{or} \quad \mathrm{p}(\mathbf{y})$$

Model Regularization (cont'd)

- Construct regularization terms for the constraints
  - Inequality and equality variable constraints

$$ay_k + b \geq c \quad \Rightarrow \quad l_v(\mathbf{y}) = |\min(0,\ ay_k + b - c)| \quad \text{or } |ay_k + b + \xi - c|$$

margin variable

  - Equation constraints

$$\mathrm{f}(\mathrm{y},\alpha) = 0 \quad \Rightarrow \quad l_{\mathrm{e}}(\mathrm{y}) = |\mathrm{f}(\mathrm{y},\alpha)|$$

  - Probability constraints

$$p(y_k = \alpha) \geq \beta \quad \Rightarrow \quad l_p(\mathbf{y}) = |\min(0, p(y_k = \alpha) - \beta)| \quad \text{or } \mathrm{E}_{p(\mathbf{y})}(l(\mathbf{y}', \mathbf{y}))$$

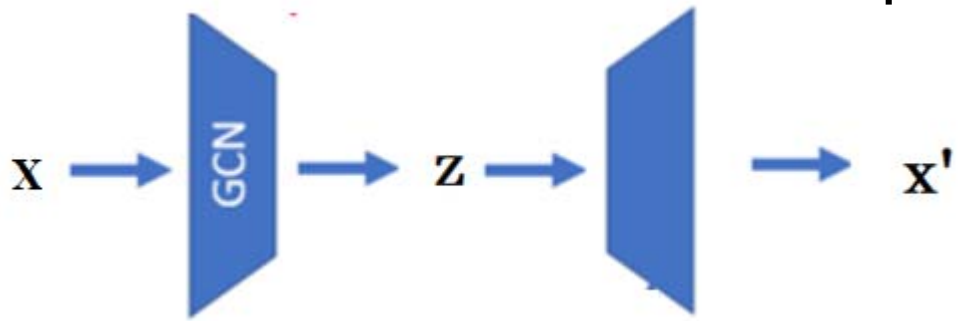
  - Total loss function

$$L(y, x, \Theta) = l(y, x, \Theta) + \lambda_v l_v(\mathbf{y}) + \lambda_e l_e(\mathbf{y}) + \lambda_p l_p(\mathbf{y})$$

Architecture Level Integration

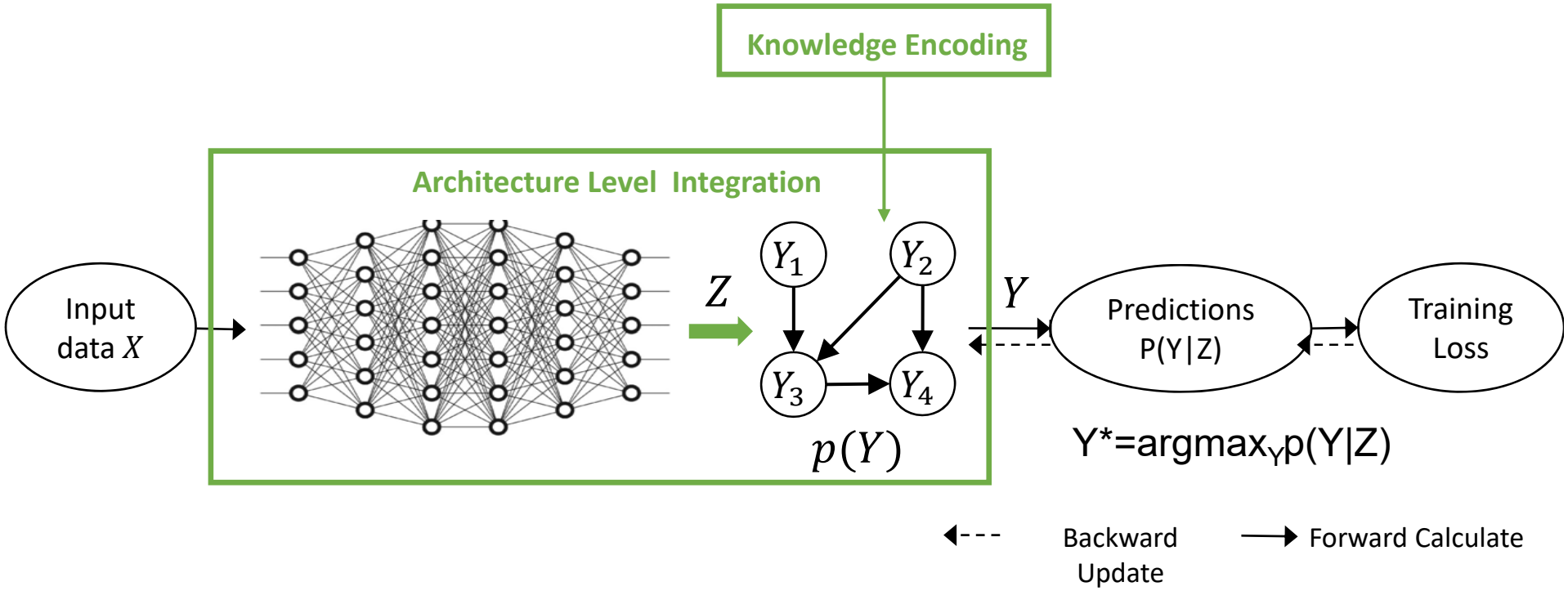
- Incorporate the prior knowledge into the architecture of the deep model by introducing additional layers, nodes, module, pathway or a new activation function.

An encoder and decoder deep model



Neural encoder Physics-based/numerical decoder

Architecture Level Integration



For structured prediction, replace the softmax layer with a graphical model layer that captures the dependencies among the target variables.

Data Level Integration

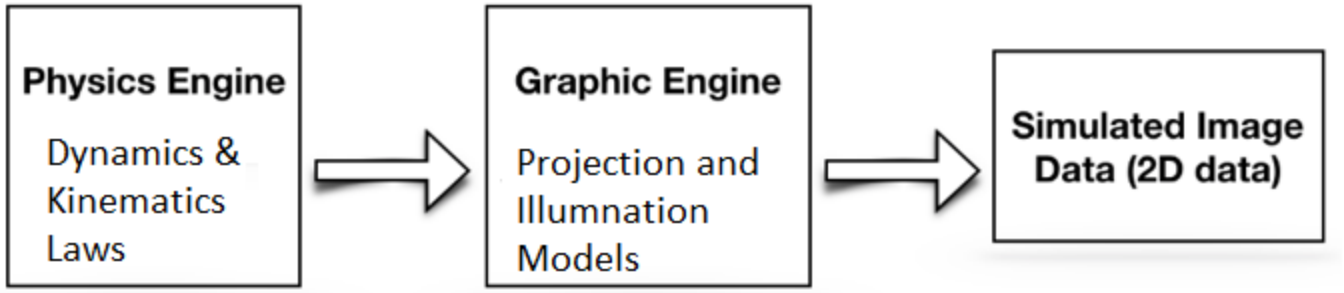
Data level integration converts the prior knowledge into synthetic data to augment real data for joint training the deep model.

- Synthetic data generation via simulators
- Synthetic data generation via knowledge datalization

Data Generation via Simulators

Encode the knowledge into the simulators and produce synthetic data by simulation

- A physics engine can be created to capture kinematic and dynamic laws to predict target's 3D physical properties, including locations and velocity.
- A graphic engine can be built to encode computer vision projection and illumination models to render images



Data Generation via Knowledge Datalization

For target knowledge represented as equations or constraints, e.g. $f(\mathbf{y},\alpha)=0$ or $a\mathbf{y}+b > c$

- Convert the knowledge into pseudo-data via sampling to represent the knowledge
- Employ Monto Carlo sampling technique (Metropolis sampling) to efficiently explore the target space to acquire samples **y** that satisfy the knowledge.
  - Uniformly sample the target space using the proposal distribution

$$p(\mathbf{y}^{(n)}|\mathbf{y}^{(n-1)},\ldots,\mathbf{y}^{(1)}) \propto \frac{1}{(2\pi\sigma^2)^{D/2}} - \frac{1}{n-1}\sum_{j=1}^{n-1}\frac{1}{(2\pi\sigma^2)^{D/2}}\exp\{-\frac{\|\mathbf{y}^{(n)}-\mathbf{y}^{(j)}\|^2}{2\sigma^2}\}$$

  - If the sample is consistent with the knowledge, accept it with a probability p. Otherwise, reject it.
  - Repeat until enough samples are collected
  - Output target samples $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N$

1) $x_1 + x_2 + x_3 < 1$
2) $x_1^2 + x_2^2 + x_3^2 < 1$
3) $x_1^2 + x_2^2 + x_3^2 < 1$ and $x_1^{1/2} + x_2^{1/2} + x_3^{1/2} > 1$
4) $x_1 \cdot x_2 \cdot x_3 < 0.1$

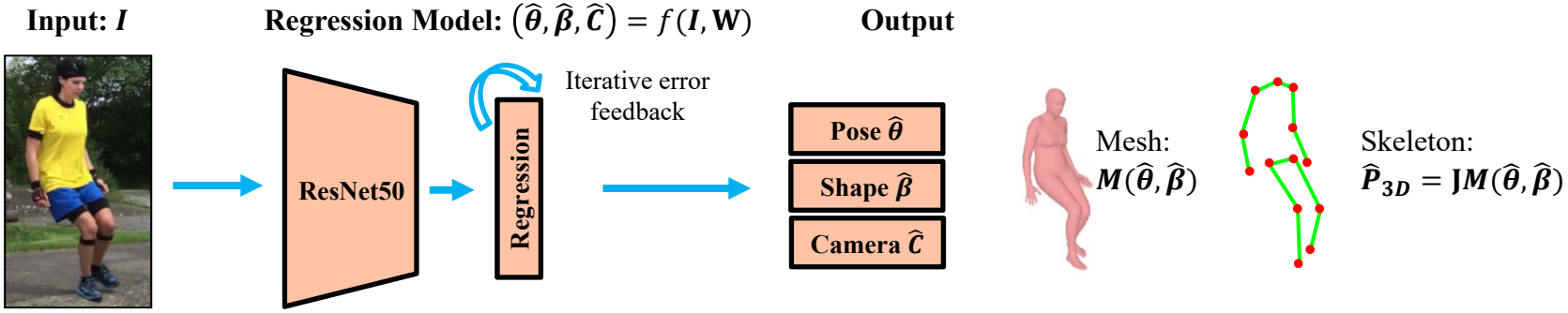
Knowledge constraints **distributioin of samples**

Computer Vision Applications

- 3D Human body shape and pose reconstruction from monocular video
  - Body knowledge

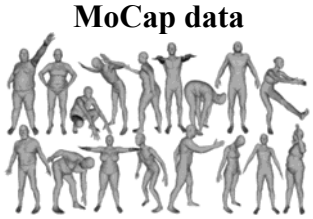
- 3D reconstruction from single image
  - Projection and illumination models

Learning-based 3D Body Reconstruction using Deformable Mesh Model



- Training requires 3D annotations $\boldsymbol{\theta}, \boldsymbol{\beta}$ and 2D landmark annotations $\boldsymbol{P}_{2D}$,

$$\mathbf{W}^* = \text{argmin } \mathcal{L}_{\text{proj}}(\hat{\mathbf{P}}_{2D}; \mathbf{P}_{2D}) + \mathcal{L}_{3D}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}; \boldsymbol{\theta}, \boldsymbol{\beta})$$



- **Issues with learning-based methods**
  - Collecting the MoCap data is expensive; MoCap data has limited diversity in demographics and poses, and they generalize poorly to unseen poses and demographics
  - Most existing methods are frame-based and hence static, and cannot exploit the body motion.

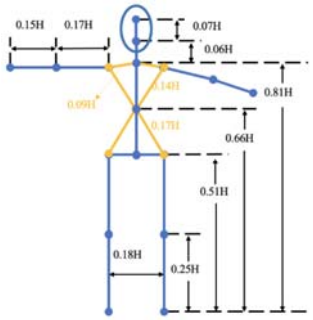
- **Usage of the generic body knowledge to replace MoCap data as supervision**
  - There exist well-established studies on human bodies from different disciplines
  - Body motion must obey the biomedical laws .
  - Exploit both static and dynamic body knowledge
  - Body knowledge is generic, applicable to different subjects under different body poses, shapes, and motions.

Generic Body Knowledge

Anthropometry:

Scientific study of human body measurements and proportions

- Human body consists of body parts connected by body joints.
- Body parts lengths follow relative body proportions.

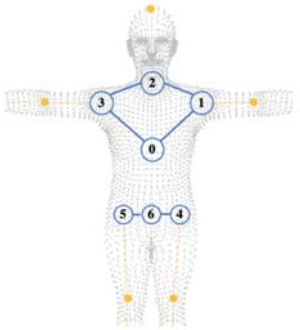


Body proportions

Body geometry:

Geometric configuration of body joints

- The shoulders, neck, and spine joints are coplanar, and the hips and pelvis joints are collinear.
- Body joints are symmetric

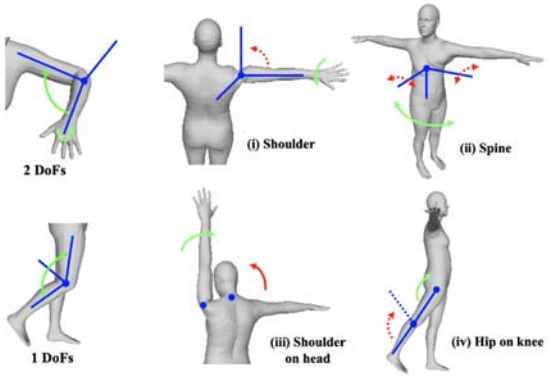


Colinear and coplanar joints

Body biomechanics:

The study of body movement mechanisms.

- Body joints have different DOFs and range of motion.
- Neighboring body joints depend on each other;.



Joint DOFs and range of motions

$0^o < \alpha_{43} < 140^o$

Inter-joint dependency

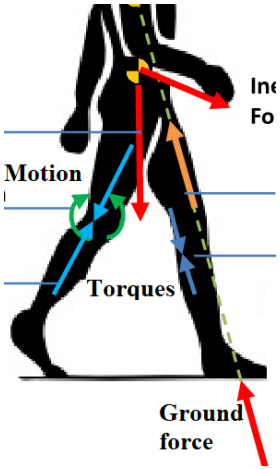
Body physics:

Study of physics in body movements.

- Non-penetrating constraint : different body parts can not penetrate each other.

Body dynamics :

Body joint torques and their motions must obey Newton second law or Euler–Lagrange equation



Generic Body Knowledge Encoding via Regularization Terms

Knowledge

Anthropometry : impose 20 body parts proportions l_i from literature .

Body geometry: impose joints 0,1,2,3 to be coplanar and joints 4,5,6 to be colinear

Body biomechanics: Restrict the joint DoF and it's range of motion ($\varphi_{min}, \varphi_{max}$) from literature .

Body non-penetrating constraint: Detect colliding mesh triangle pairs *C*; penalize the penetration through signed distance field $\psi(\cdot)$.

Regularization constraints

$$\mathcal{L}_{\text{anth}} = \frac{1}{20}\sum_{i=1}^{20}\left(\frac{\hat{l}_i - l_i}{l_i}\right)$$

$$\mathcal{L}_{\text{geometry}} = \mathcal{L}_{\text{coplanar}} + \mathcal{L}_{\text{geometry}}$$

$$\mathcal{L}_{\text{coplanar}} = \frac{|(\mathbf{P}_{01} \times \mathbf{P}_{03}) \cdot \mathbf{P}_{02}|}{\|\mathbf{P}_{01} \times \mathbf{P}_{03}\|\|\mathbf{P}_{02}\|}$$

$$\mathcal{L}_{\text{colinear}} = \frac{|\mathbf{P}_{64} \times \mathbf{P}_{65}|}{\|\mathbf{P}_{64}\|\|\mathbf{P}_{65}\|}$$

$$\mathcal{L}_{\text{biomechanic}} = \sum_{i=1}^{23}\left(\max\{\hat{\varphi}_i - \varphi_{i,max}, \varphi_{i,min} - \hat{\varphi}_i, 0\}\right)^2$$

$\mathbf{P}_{ij}$ denotes a 3D bone between joint *i* and *j*.

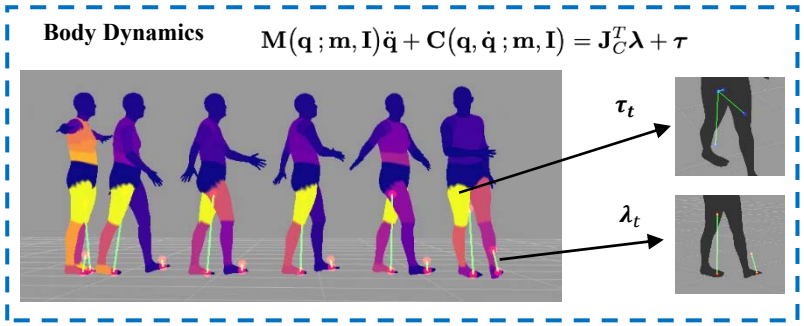
$$\mathcal{L}_{non-penetrate} = \sum_{(f_s,f_t)\in C}\left\{\sum_{\boldsymbol{v}_s \in f_s}\left\|-\boldsymbol{\Psi}_{f_t}(\boldsymbol{v}_s)\cdot\boldsymbol{n}_s\right\|^2 + \sum_{\boldsymbol{v}_t \in f_t}\left\|-\boldsymbol{\Psi}_{f_s}(\boldsymbol{v}_t)\cdot\boldsymbol{n}_t\right\|^2\right\}$$

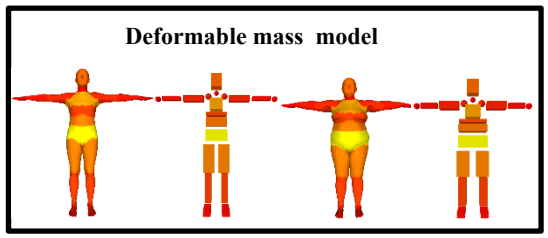
f_s, f_t are the colliding triangles in detected colliding set *C*.

Body Biomechanics Encoding via Architecture Design

-- Customize the architecture of Encoder-Decoder Network

Dynamics branch





- Consider generic mass distribution [1], augment the mass: $\mathbf{m}(\boldsymbol{\beta})$
- Approximate body parts with simple geometry to compute the inertia: $\mathbf{I}(\boldsymbol{\beta})$

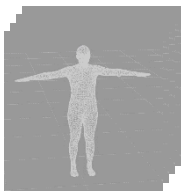
Generalized position in world frame

$\boldsymbol{q_t} = \{\boldsymbol{\theta}_t, \boldsymbol{\beta}, \boldsymbol{T}_t\}$

$\mathbf{J_{c,t}}(\mathbf{q_t}),\ \mathbf{M_t}(\mathbf{q_t};\mathbf{m},\mathbf{I}),\ \mathbf{C_t}(\mathbf{q_t},\dot{\boldsymbol{q}};\mathbf{m},\mathbf{I})$ — Jacobian, Generalized mass, Generalized bias force

$(\boldsymbol{q_i} = \boldsymbol{\theta}_t, \boldsymbol{\beta}, \boldsymbol{T}_t)$

3D mesh sequence



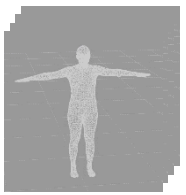
Mesh convolution → Dynamics regression → $\boldsymbol{\lambda_t}, \boldsymbol{\tau_t}$

Ground reaction force, joint actuation

Neural ODE Solver (Runge-Kutta)

$\boldsymbol{q'_{t+1}}$

Refined 3D mesh sequence



$$\ddot{\boldsymbol{q}}_t = \boldsymbol{M}_t^{-1}\boldsymbol{J}_{c,t}^T\boldsymbol{\lambda}_t + \boldsymbol{\tau}_t - \boldsymbol{C}_t$$
$$\boldsymbol{q}_{t+1} = \boldsymbol{q}_t + \Delta t\dot{\boldsymbol{q}}_t$$
$$\dot{\boldsymbol{q}}_{t+1} = \dot{\boldsymbol{q}}_t + \Delta t\ddot{\boldsymbol{q}}_t$$

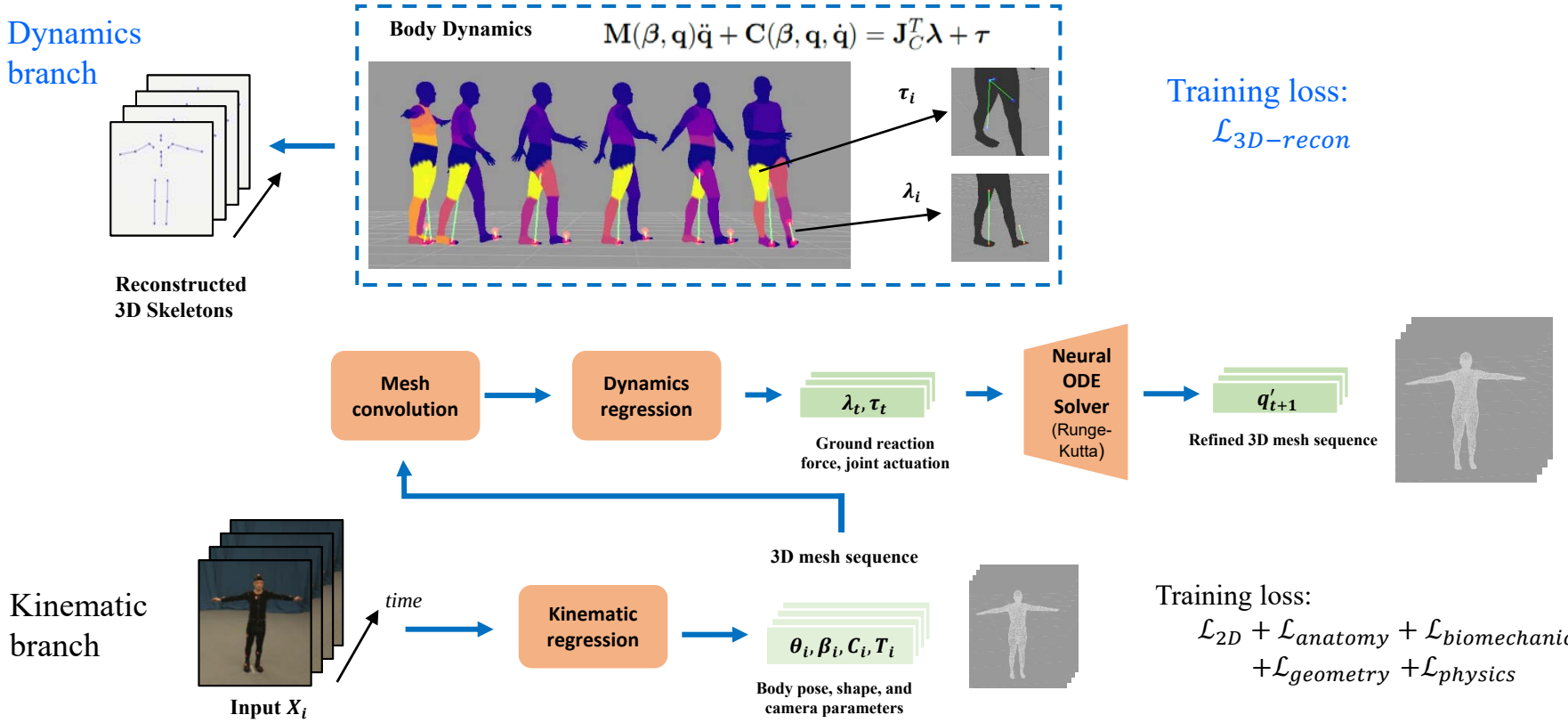
Training loss:

$$\mathcal{L}_{3D-recon} = \sum_t (q'_{t+1} - q_{t+1})^2$$

[1] Full-body musculoskeletal model for muscle-driven simulation of human gait

3D Body Reconstruction and Joint Force Estimation from Monocular Videos

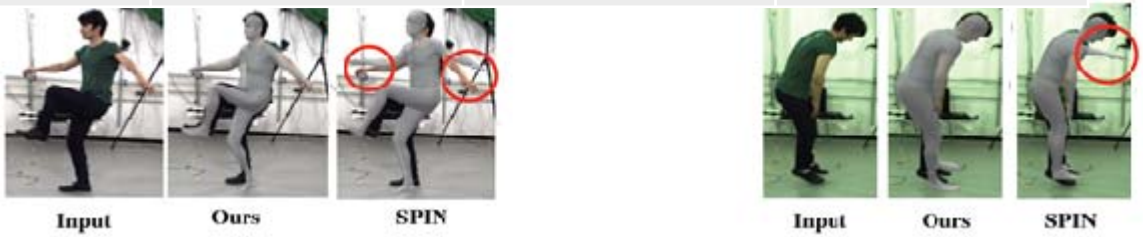
Overall framework



Evaluation on Existing Benchmarks

Quantitative evaluation

Model	Loss	Human3.6M (within-dataset)	MPI-INF-3DHP (across dataset)
HMR	2D+MoCap data	73.3	169.5
Our	2D+generic	78.5	129.3



Experiment Settings:

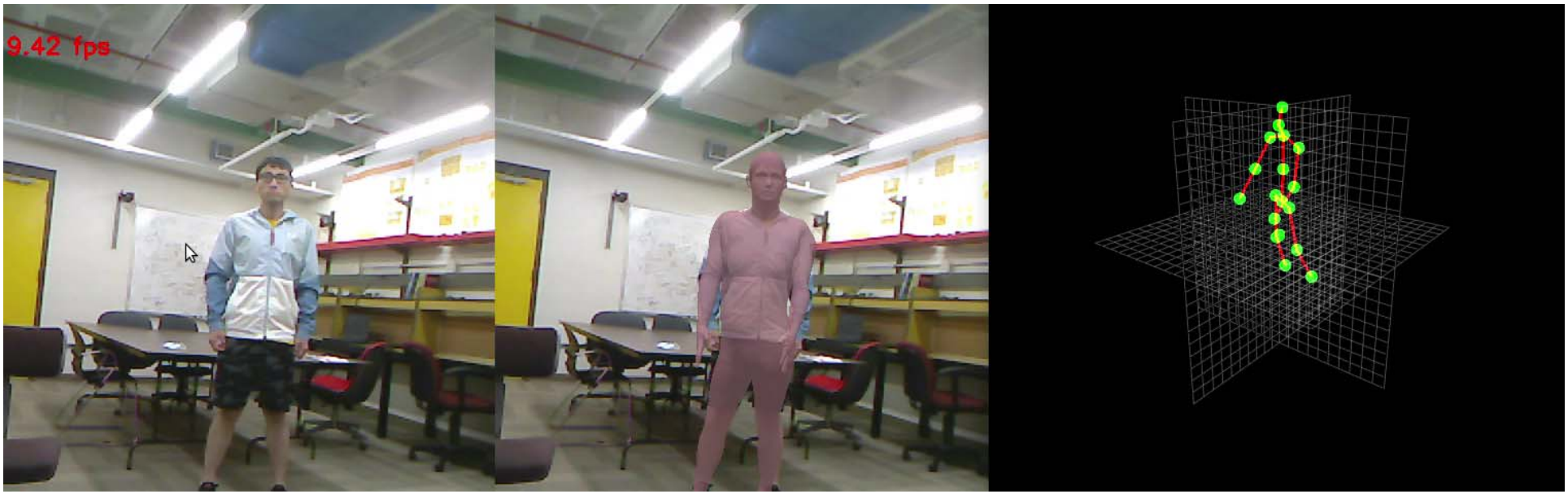
- Within dataset: Train and test on Human3.6m
- Across dataset: train on Human3.6m ar test on MPI-INF-3DHP .
- Metric: mean per-joint position error in

Comparison with SOTA data-driven method, SPIN, on unseen scenarios. Failures of SPIN are marked with red circles.

Conclusion

- Our method does not need any 3D body annotations
- Our methods achieves comparable within-dataset performance and outperforms SOTA method for across-dataset performance
- Our method generalizes better to unseen poses.

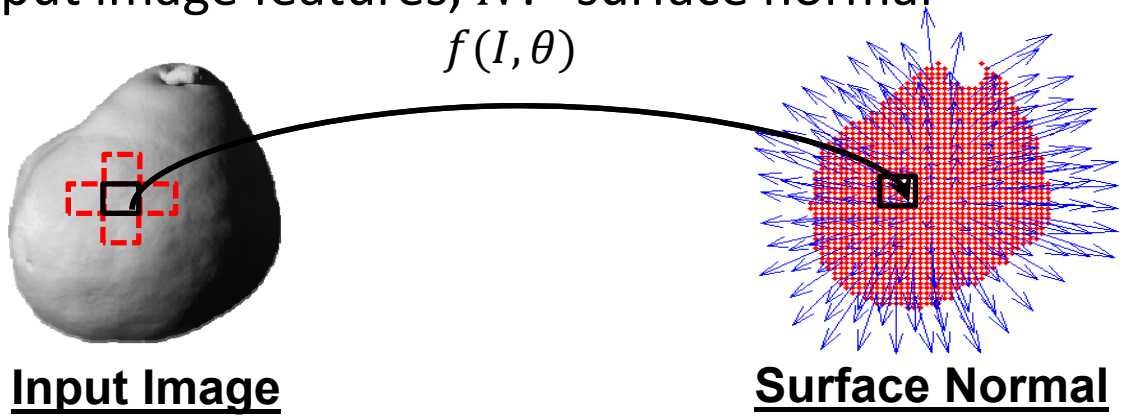
Real-time 3D Mesh and Skeleton Reconstruction



3D Shape Reconstruction from Single Image

Reconstruct 3D shape of an object from a single image

- Learning-based methods : 3D reconstruction is formulated as a learning problem, whereby training data is used to learn the mapping function.
  - Given training data $(I_1, N_1), \cdots, (I_n, N_n)$
  - I: input image features, N: surface normal



  - Learn the mapping function $f(I, \theta)$ to predict N from I

$$\theta^* = \arg min_{\theta} \sum_{i=1}^{m} \sum_{j \in N(i)} \left\| N_i - f(x_j, \theta) \right\|_2^2$$

Does not generalize well to different objects or shapes.

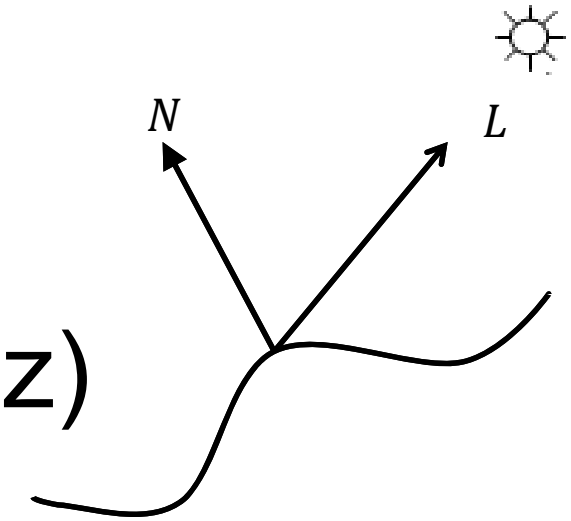
3D Shape Reconstruction from Single Image

- Model based
  - 3D shape reconstruction from a singe image (shape from X) using computer vision models, including illumination and projection models.
  - Problems are ill-posed and need illumination direction

- Hybrid approach that augments the learning-based approach with computer vision models.

Illumination model

- Lambertian model
  - Surface reflects light equally in all directions
  - I: intensity of reflected light
  - ρ: surface albedo
  - L: incident light
  - N: surface normal



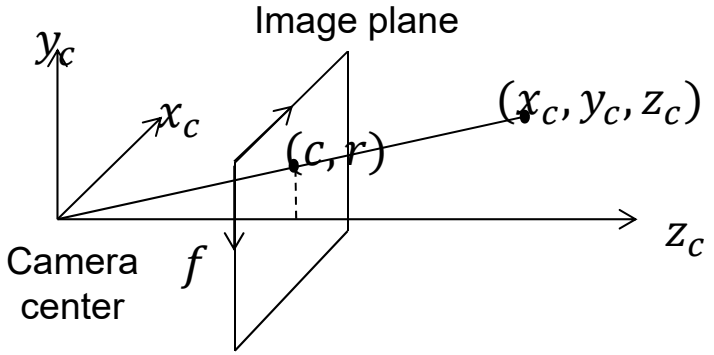
$$I(c,r) = \rho L \cdot N(\mathrm{x,y,z})$$

Shape from shading method

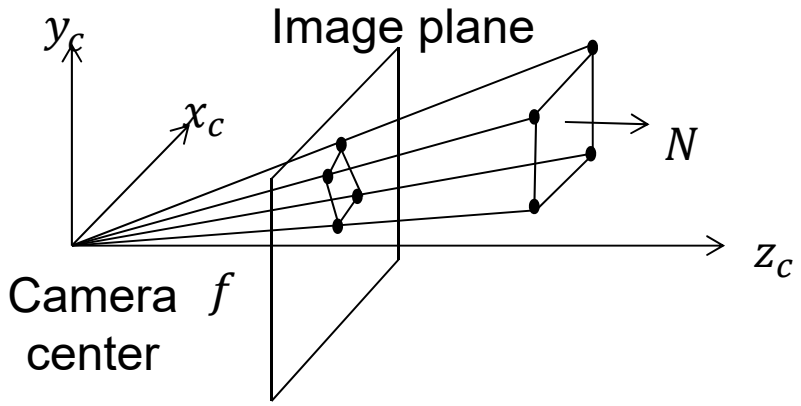
Projection Model

- Full perspective projection

$$\lambda \begin{bmatrix} c \\ r \\ 1 \end{bmatrix} = \begin{bmatrix} s_x f & 0 & c_0 \\ 0 & s_y f & r_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix}$$



- Assuming the object surface can be locally approximated by a plane and the plane normal is orthogonal to any line segments on the plane



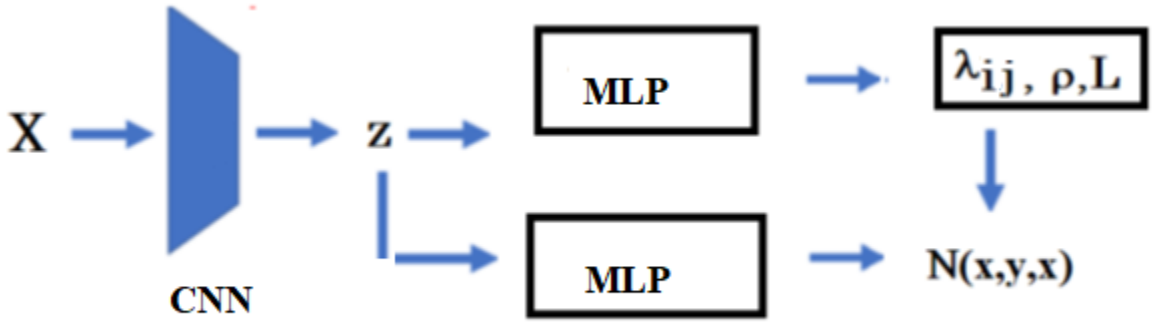
$$\left(\lambda_{ij} \begin{bmatrix} c_i \\ r_i \\ 1 \end{bmatrix} - \begin{bmatrix} c_j \\ r_j \\ 1 \end{bmatrix} \right) \cdot N = 0$$

Combining Data and Vision Models

- Basic idea
  - Augment the learning-based approach with illumination and projection models through regularization

$$N_1^*, \cdots, N_m^*, L^*, \rho^* = \arg\min \underbrace{\sum_{i=1}^{m} \sum_{j \in N(i)} \left\| N_i - f(x_j, \theta) \right\|_2^2}_{Data\ term} + \underbrace{\lambda \sum_{i=1}^{m} \| I_i - \rho L^T N_i \|^2}_{illumination\ model} + \underbrace{\beta \sum_{i=1}^{m} \sum_{j \in N(i)} \left\| \left(\frac{\lambda_i}{\lambda_j} \begin{bmatrix} c_i \\ r_i \\ 1 \end{bmatrix} - \begin{bmatrix} c_j \\ r_j \\ 1 \end{bmatrix} \right) \cdot N_i \right\|_2^2}_{Projection\ model}$$

$$+ \underbrace{N_x^2 + N_y^2 + N_z^2}_{smoothness\ term}$$



Experimental Results

Performance on benchmark datasets

Method	Error	
	MIT/Berkley	Harvard
CNN	34.95	28.54
Our model	20.75	21.28

Comparison with related work

Method	Error	
	MIT/Berkley	Harvard
GVA [182]	32.09	33.23
SIFS [6]	22.92	37.24
Our model	**20.75**	**21.28**

Conclusions

- Introduce a hybrid vision model based on combining model-based vision with learning based vision through systematic integration of prior knowledge with data
- Propose two sources of prior knowledge: target knowledge and computer vision models
- To handle the challenges with knowledge and data integration, introduce four levels of data and knowledge integration: decision-level, training-level, architecture level, and data level.
- Experiments demonstrate the hybrid model can significantly improve deep model data efficiency (even zero-shot), generalization performance, robustness, and model interpretability.


Any
Question