Structural Causal 3D Reconstruction

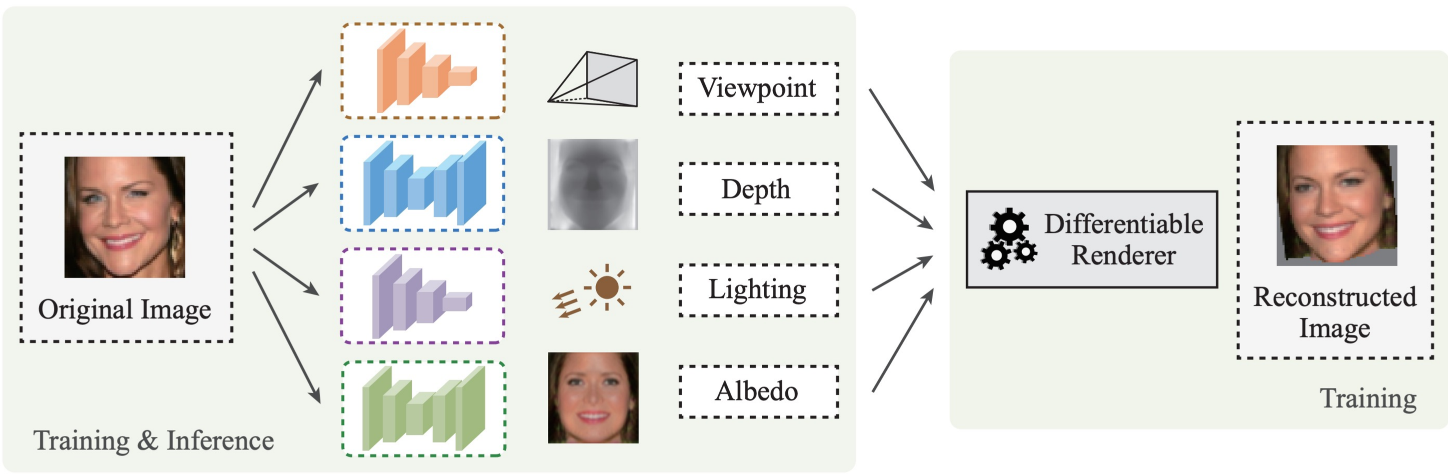
Weiyang Liu* Zhen Liu* Liam Paull Adrian Weller Bernhard Schölkopf

University of Cambridge · Université de Montréal · Max Planck Institute for Intelligent Systems · Mila

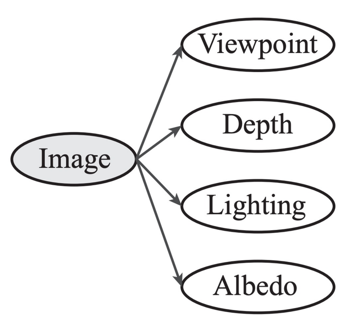
European Conference on Computer Vision, Tel Aviv 2022

Introduction

How current 3D reconstruction pipeline works

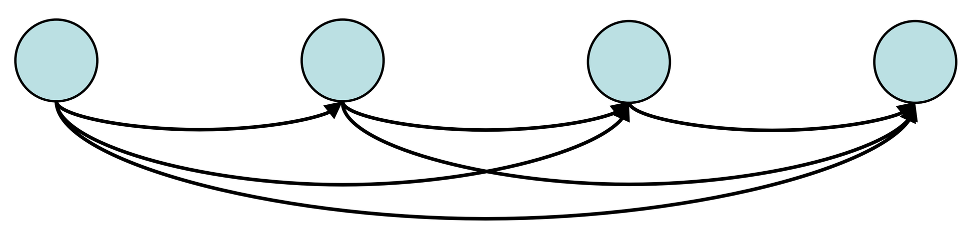


Current standard unsupervised 3D reconstruction pipeline

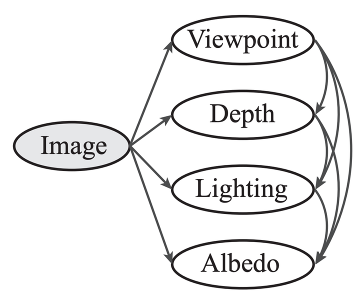


- Standard 3D reconstruction pipelines can be viewed as independently extracting four physical factors from the image.
- This is by no means an optimal design, since these four factors may have mutual dependency.

Does causal modeling of the four factors matter?

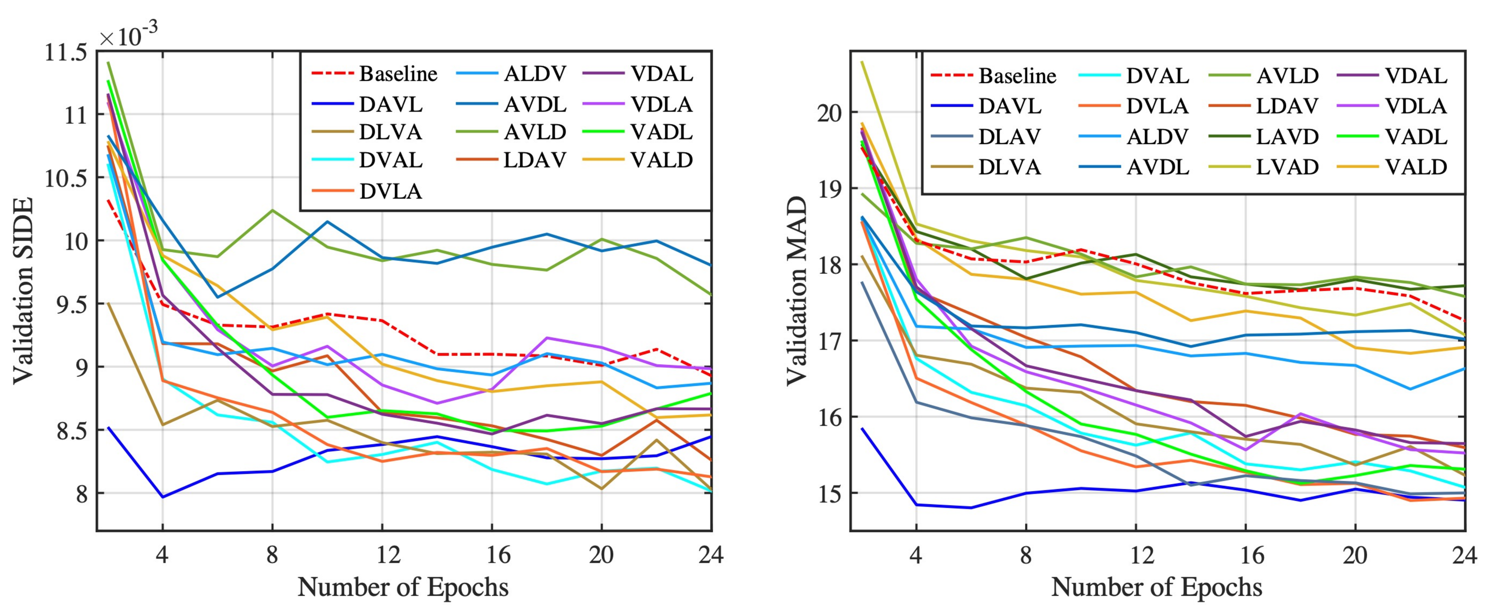


Dense ordering of the four physical factors



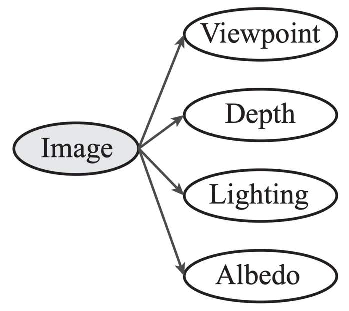
Example: VDLA

- We conduct an experiment to see the effect of different causal orderings of the four physical factors.
- We then show the empirical performance of their 3D reconstruction.
- VDLA: Viewpoint -> Depth -> Lighting -> Albedo
- SIDE: Scale Invariant Depth Error; MAD: Mean Angle Deviation;



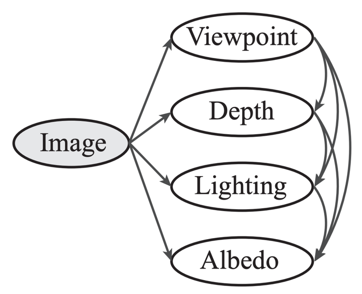
Causal ordering matters a lot!

Expressiveness of Representing Conditional Distributions



$$P(\boldsymbol{V},\boldsymbol{D},\boldsymbol{L},\boldsymbol{A}|\boldsymbol{I}) = P(\boldsymbol{V}|\boldsymbol{I})\cdot P(\boldsymbol{D}|\boldsymbol{I})\cdot P(\boldsymbol{L}|\boldsymbol{I})\cdot P(\boldsymbol{A}|\boldsymbol{I})$$

Assume conditional independence

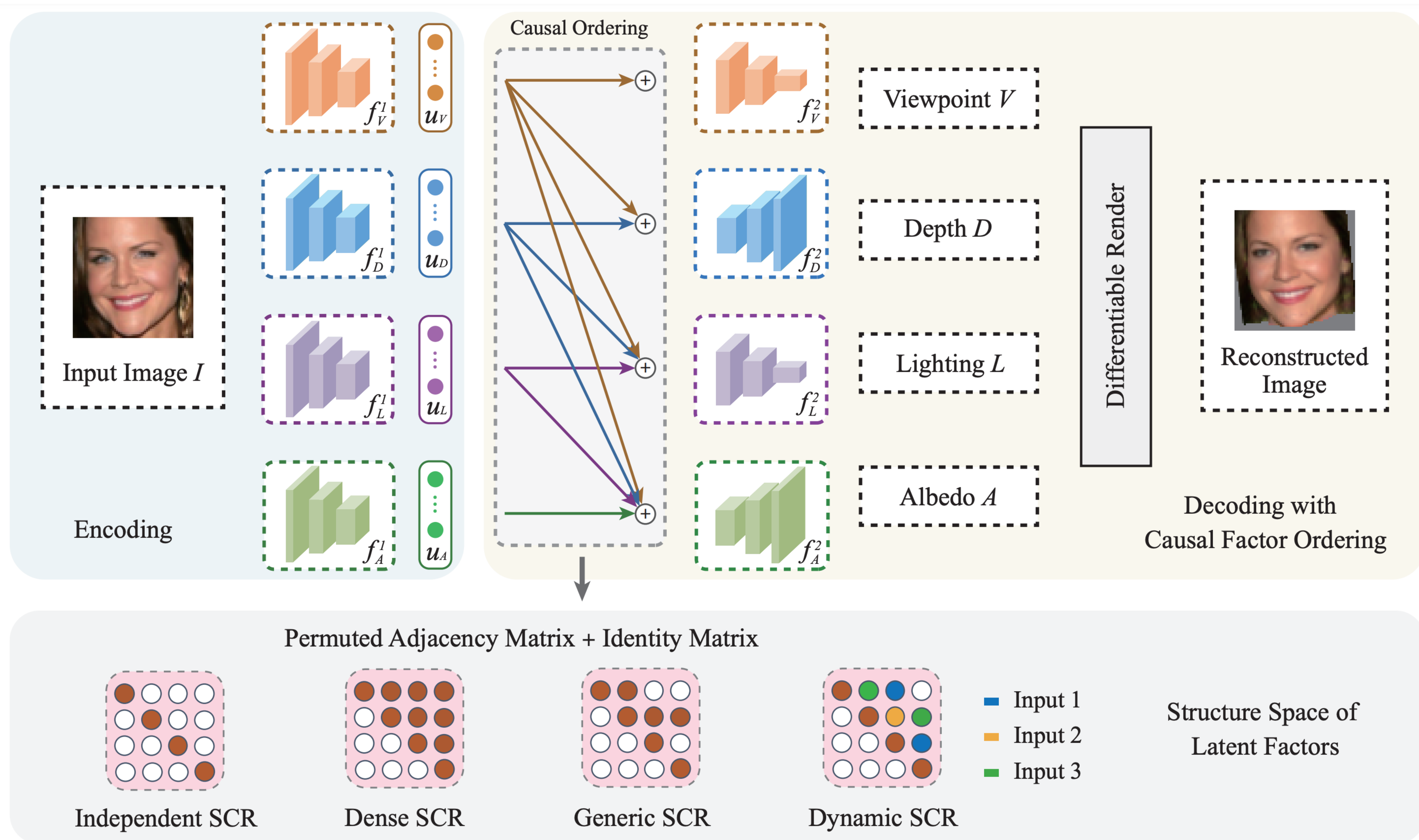


$$P(\boldsymbol{V},\boldsymbol{D},\boldsymbol{L},\boldsymbol{A}|\boldsymbol{I}) = P(\boldsymbol{V}|\boldsymbol{I})\cdot P(\boldsymbol{D}|\boldsymbol{I},\boldsymbol{V})\cdot P(\boldsymbol{L}|\boldsymbol{I},\boldsymbol{V},\boldsymbol{D})\cdot P(\boldsymbol{A}|\boldsymbol{I},\boldsymbol{V},\boldsymbol{D},\boldsymbol{L})$$

Chain factorization without any assumption

Thus more expressive!

SCR: Structural Causal 3D Reconstruction



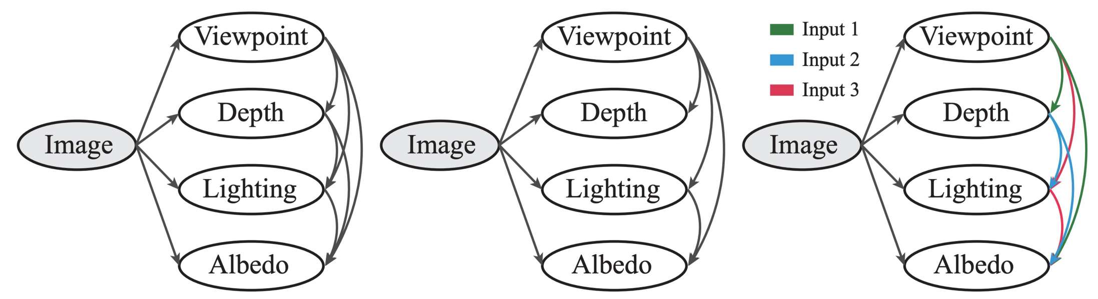
Our SCR framework for learning causal ordering of factors

- Encoding of viewpoint, depth, lighting and albedo:

$$\boldsymbol{V} = f_V^2\left(f_V^1(\boldsymbol{I}), \boldsymbol{M}_V^\top \boldsymbol{u}\right) \qquad \boldsymbol{D} = f_D^2\left(f_D^1(\boldsymbol{I}), \boldsymbol{M}_D^\top \boldsymbol{u}\right)$$

$$\boldsymbol{L} = f_L^2\left(f_L^1(\boldsymbol{I}), \boldsymbol{M}_L^\top \boldsymbol{u}\right) \qquad \boldsymbol{A} = f_A^2\left(f_A^1(\boldsymbol{I}), \boldsymbol{M}_A^\top \boldsymbol{u}\right)$$

- Three types of causal orderings:



Learning dense SCR via Bayesian optimization

- Dense SCR: order permutation of viewpoint, lighting, depth and albedo
- Evaluation on a half-out validation set

Learning Generic SCR via Optimization Unrolling

- Method I: Training with differentiable DAG regularization:

$$\min_{\boldsymbol{M}} \mathcal{L}_{\text{val}}(\boldsymbol{W} - \eta\nabla_{\boldsymbol{W}}\mathcal{L}_{\text{train}}(\boldsymbol{W},\boldsymbol{M}),\boldsymbol{M}) + \lambda_{\text{DAG}}\mathcal{H}(\boldsymbol{M})$$

Unrolling the training gradients to the validation loss

- $\boldsymbol{M}$: adjacency matrix of causal ordering; $\boldsymbol{W}$: weights of encoders
- Method II: Training on top of the adjacency matrix in dense SCR

$$\min_{\boldsymbol{M}} \mathcal{L}_{\text{val}}(\boldsymbol{W} - \eta\nabla_{\boldsymbol{W}}\mathcal{L}_{\text{train}}(\boldsymbol{W},\boldsymbol{M}\circ\boldsymbol{M}^*_{\text{dense}}),\boldsymbol{M}\circ\boldsymbol{M}^*_{\text{dense}})$$

- $\boldsymbol{M}$*: The learned dense causal ordering from Bayesian optimization

Learning Dynamic SCR via Masked Self-Attention

- Similar idea to Method II in learning generic SCR
- Making the predicted $\boldsymbol{M}$ a function of the input image:

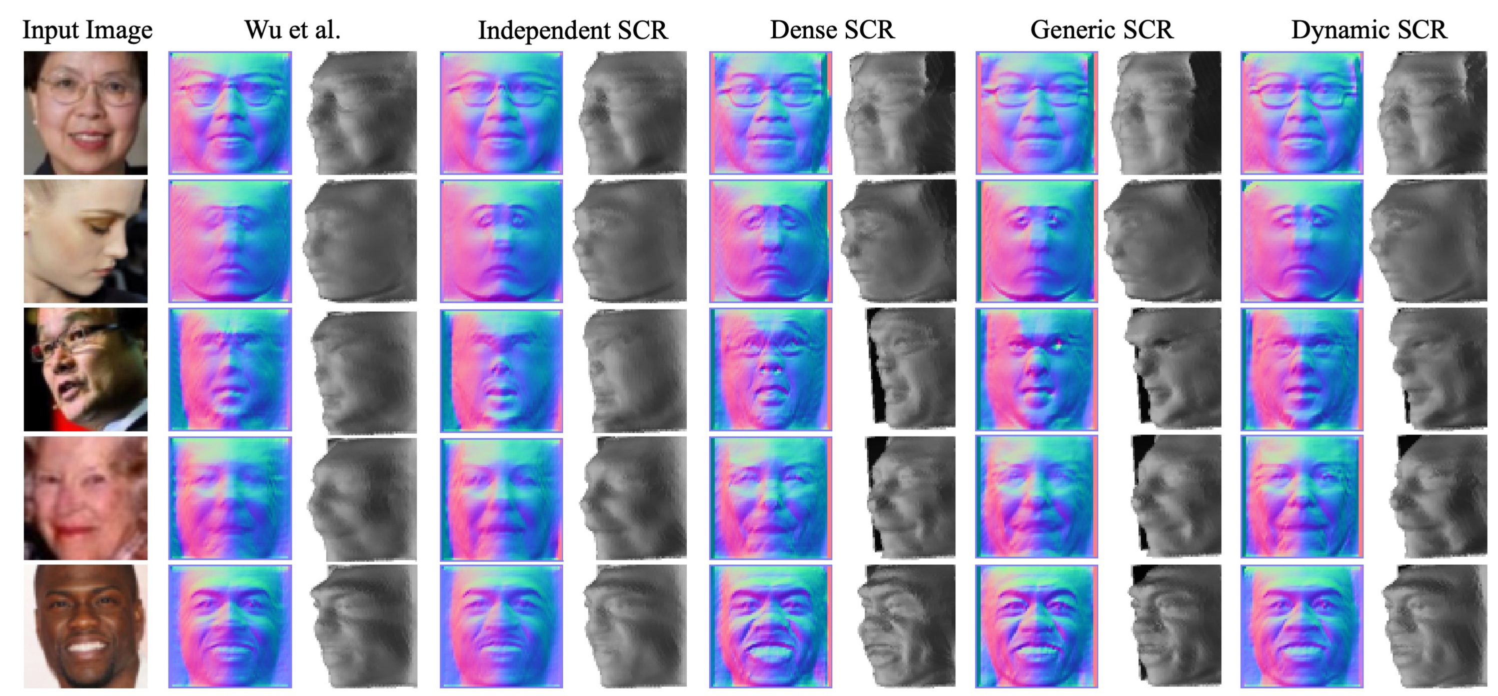
$$\min_{\boldsymbol{W}} \mathcal{L}_{\text{train}}(\boldsymbol{W}, q(\boldsymbol{u})\circ\boldsymbol{M}^*_{\text{dense}})$$

- q(u) is a parameter-free function, realized by self attention among V,D,L,A.

Experiments and Results

Method	SIDE ($\times 10^{-2}$) ↓	MAD (deg.) ↓
Supervised	**0.410** ±0.103	**10.78** ±1.01
Constant Null Depth	2.723 ±0.371	43.34 ±2.25
Average GT Depth	1.990 ±0.556	23.26 ±2.85
Wu et al. [75] (reported)	**0.793** ±0.140	16.51 ±1.56
Ho et al. [27] (reported)	0.834 ±0.169	**15.49** ±1.50
Wu et al. [75] (our run)	0.901 ±0.190	17.53 ±1.84
Independent SCR	0.895 ±0.183	17.36 ±1.78
Dense SCR (random)	1.000 ±0.275	17.66 ±2.09
Dense SCR (BO)	**0.830** ±0.205	**14.88** ±1.94
Generic SCR (Eq. 5)	0.859 ±0.215	15.17 ±1.92
Generic SCR (Eq. 6)	**0.820** ±0.190	**14.79** ±1.96
Dynamic SCR (Sigmoid)	0.827 ±0.220	14.86 ±2.02
Dynamic SCR (Cosine)	**0.815** ±0.232	**14.80** ±1.95

Depth reconstruction results on BFM



Textureless view and canonical normal map on CelebA