MonoPhysics: Estimating Geometry, Appearance, and Physical Parameters from Monocular Videos

Daniel Rho[1] Jun Myeong Choi[1] Matthew Thornton[1] Biswadip Dey[2] Roni Sengupta[1]

[1]UNC Chapel Hill [2]Meta

NEURAL INFORMATION PROCESSING SYSTEMS

Paper & code

Motivation

From video to a simulatable object

Recover shape, appearance and material
Then simulate how it moves in new situations



Interactive 3D



Robot planning



Accurate methods need multiple cameras

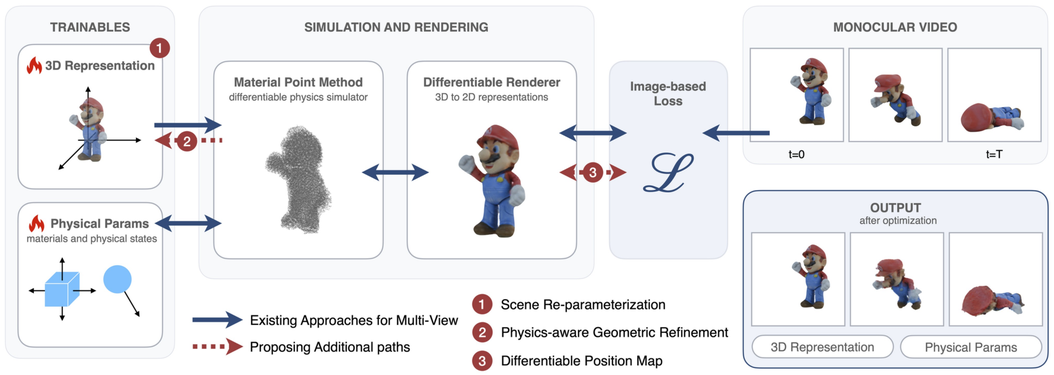
Cameras pin down shape and scale first
Most real videos come from one view

Three challenges of monocular video

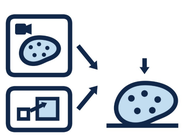
1. **Scale ambiguity**
   Size is ambiguous from one view
2. **Ambiguous geometry**
   Unseen regions stay ambiguous
3. **Weak supervision**
   Image losses only help where shapes overlap

Methods

MonoPhysics bridges the visual (3D Gaussian Splatting) and physical (MPM simulation) sides
This lets cues in a monocular video better guide both shape and material



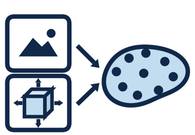
1 Scene Re-parameterization



Shape and size learned separately
One scale value for the whole object

→ **Whole object rescales coherently**

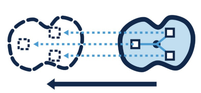
2 Physics-aware Geometry Refinement



Adaptive per-particle volume
Rendering and physics guide the shape

→ **Physics helps where vision is ambiguous**

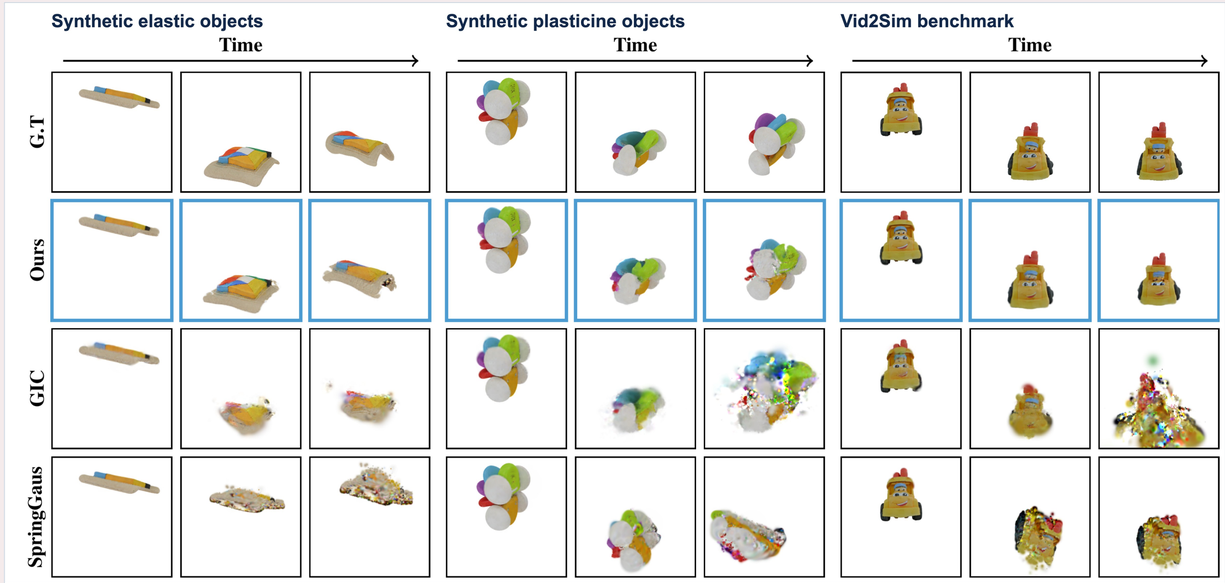
3 Differentiable Position Map



Pixel positions rendered like color
Position errors can move the shape

→ **Pulls the shape toward the target**
→ **Works even without overlap**

Evaluation



More accurate scale and shape

Lowest geometry error on both materials
Baselines often get scale or shape wrong
See: synthetic figures, table CD

Stiffness on par with multi-view

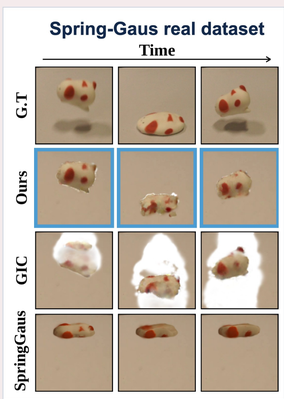
Versus the 12-camera Vid2Sim method
Lowest stiffness error on Vid2Sim and elastic
See: table log *E*

Best monocular future prediction

Best on all three benchmarks
Holds on real video, not only synthetic
See: all three figures, table PSNR

	Vid2Sim		Synthetic elastic / plasticine			Real
Method	**PSNR ↑**	**log *E* ↓**	**CD ↓**	**PSNR ↑**	**log *E* ↓**	**PSNR ↑**
Ours	**22.4**	**0.45**	**24** **19**	**21.1** **23.4**	**0.31** **0.40**	**32.3**
GIC	18.7	0.48	421 333	18.5 22.0	0.66 0.95	25.9
SpringGaus	18.9	–	693 450	16.7 17.3	–	28.9
PAC-NeRF	17.9	2.07	438 247	17.4 19.8	1.33 2.55	–
Vid2Sim† (12 cameras)	25.1	0.51	–	–	–	–

CD: shape and trajectory error, log *E*: stiffness error
PSNR: rendering quality, –: not reported
Synthetic: elastic (top), plasticine (bottom)
† From the Vid2Sim paper (all other methods use one camera)



Conclusion

Geometry, appearance and material from one camera
Three bridges link appearance and physics
More accurate future prediction than prior monocular methods

Limitations

Refinement reduces but does not remove geometry errors
Assumes known camera, ground plane and material model
Tested only on falls onto a flat plane

Acknowledgments. Supported by NIH NIBIB R21EB035832 and NSF CAREER 2543161