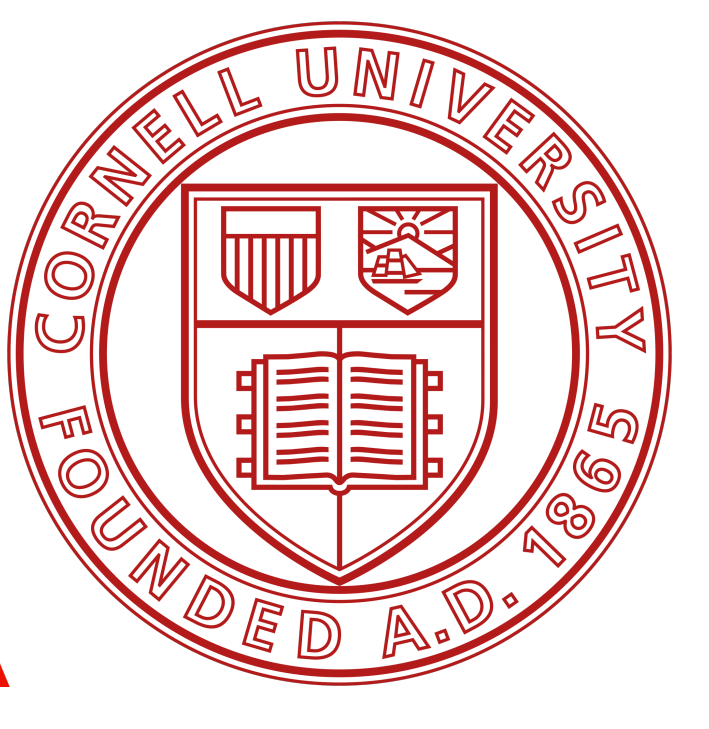
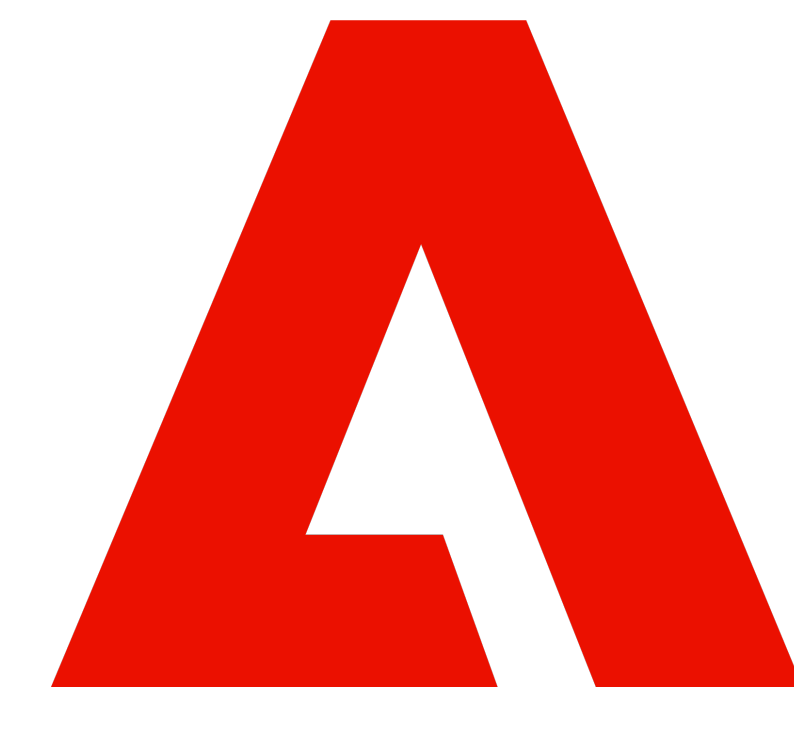


Unified Video Dense Prediction from Disjoint Data

Yihong Sun^{1,2}, Seoung Wug Oh¹, Jiahui Huang¹, Bharath Hariharan², Joon-Young Lee¹

¹Adobe Research ²Cornell University

<https://unid-video.github.io>



Project Page

Introduction

Motivation

- Scene understanding needs many properties at once — depth, materials, semantics, lighting — but existing dense-prediction annotations are fragmented across incompatible, domain-specific datasets:

Geometric Data

Depth and normals can be captured at scale via RGB-D cameras or stereo, but are confined to environments with **limited semantic diversity**.

Semantic Data

Semantic annotations require expensive labeling over diverse internet images, yielding rich category diversity but **rarely paired with geometric ground truth**.

Intrinsic Data

Intrinsic decompositions (albedo, shading) are prohibitively expensive to measure in the real world and largely **confined to synthetic domains**.

Video Annotations

Extending to video makes annotation scarcity even worse: dense semantic or intrinsic video annotation is **nearly infeasible at scale**.

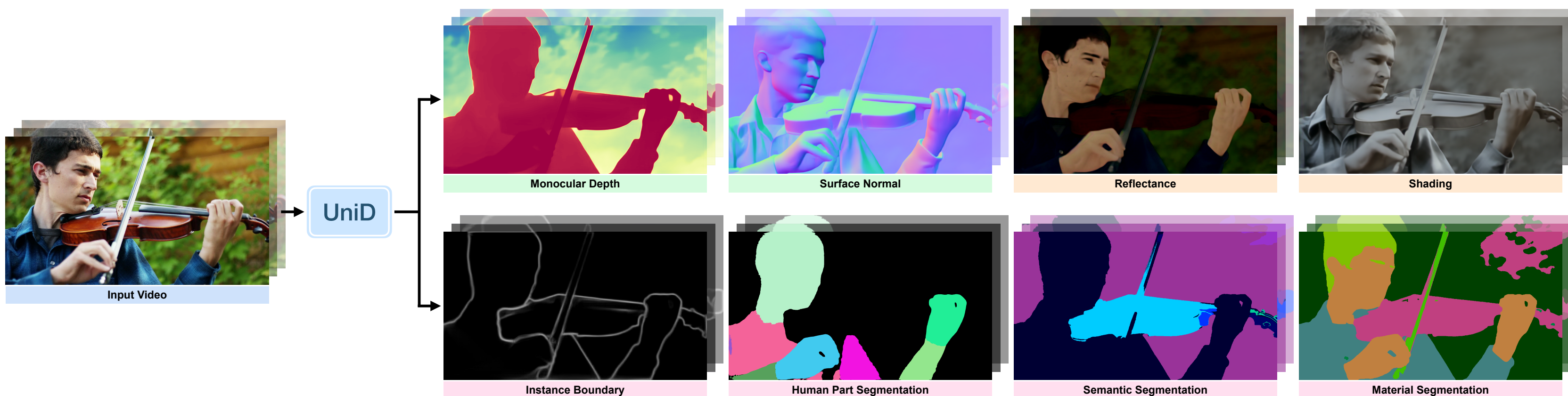
Research Question

Can we learn unified dense prediction for images and videos from **disjoint, domain-specific** datasets — **without** pseudo-labeling and **without** annotation co-occurrence?

Insights

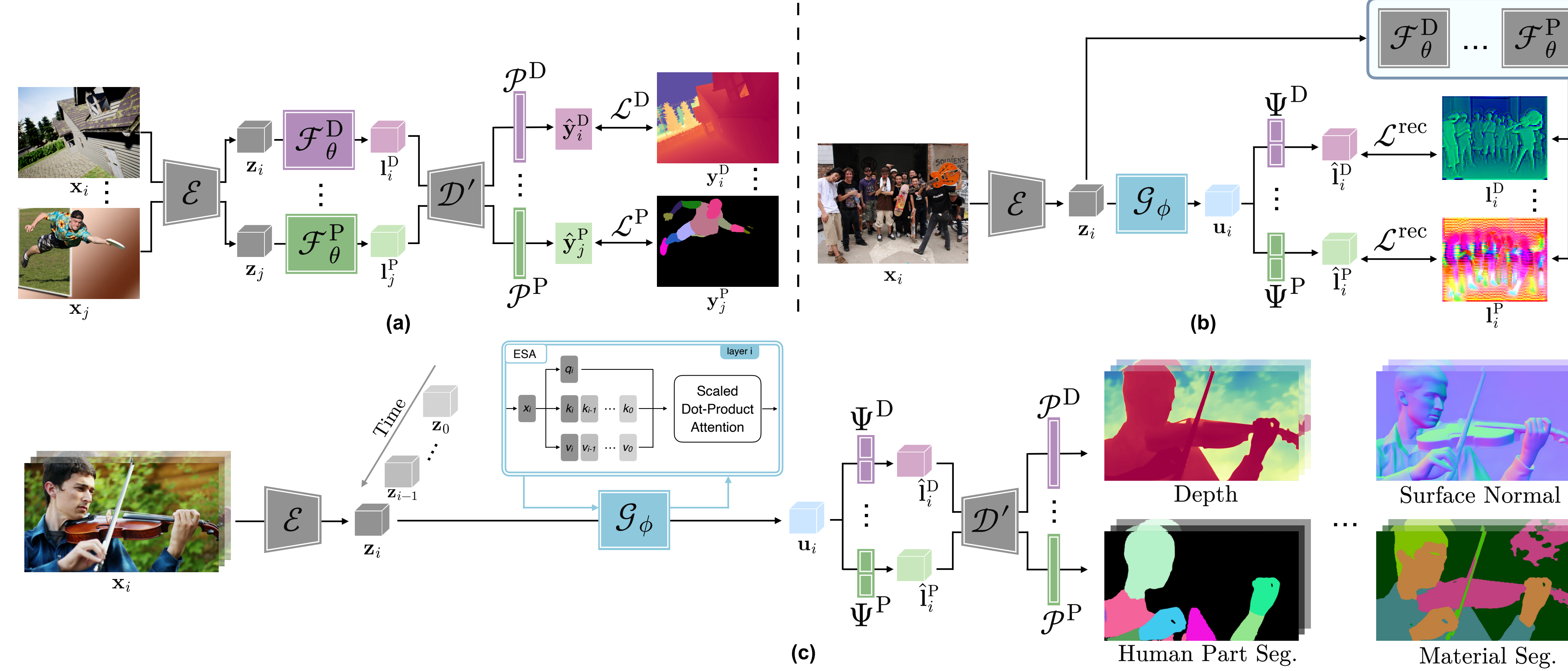
- Despite identical training data, diffusion priors **generalize** where discriminative features overfit.
- A shared feature space **improves cross-task consistency** and lets tasks without video annotations **inherit temporal stability** from those that do.
- Sharing one backbone also **improves efficiency** as temporal memory is computed **once and reused** across all eight tasks.

Contribution



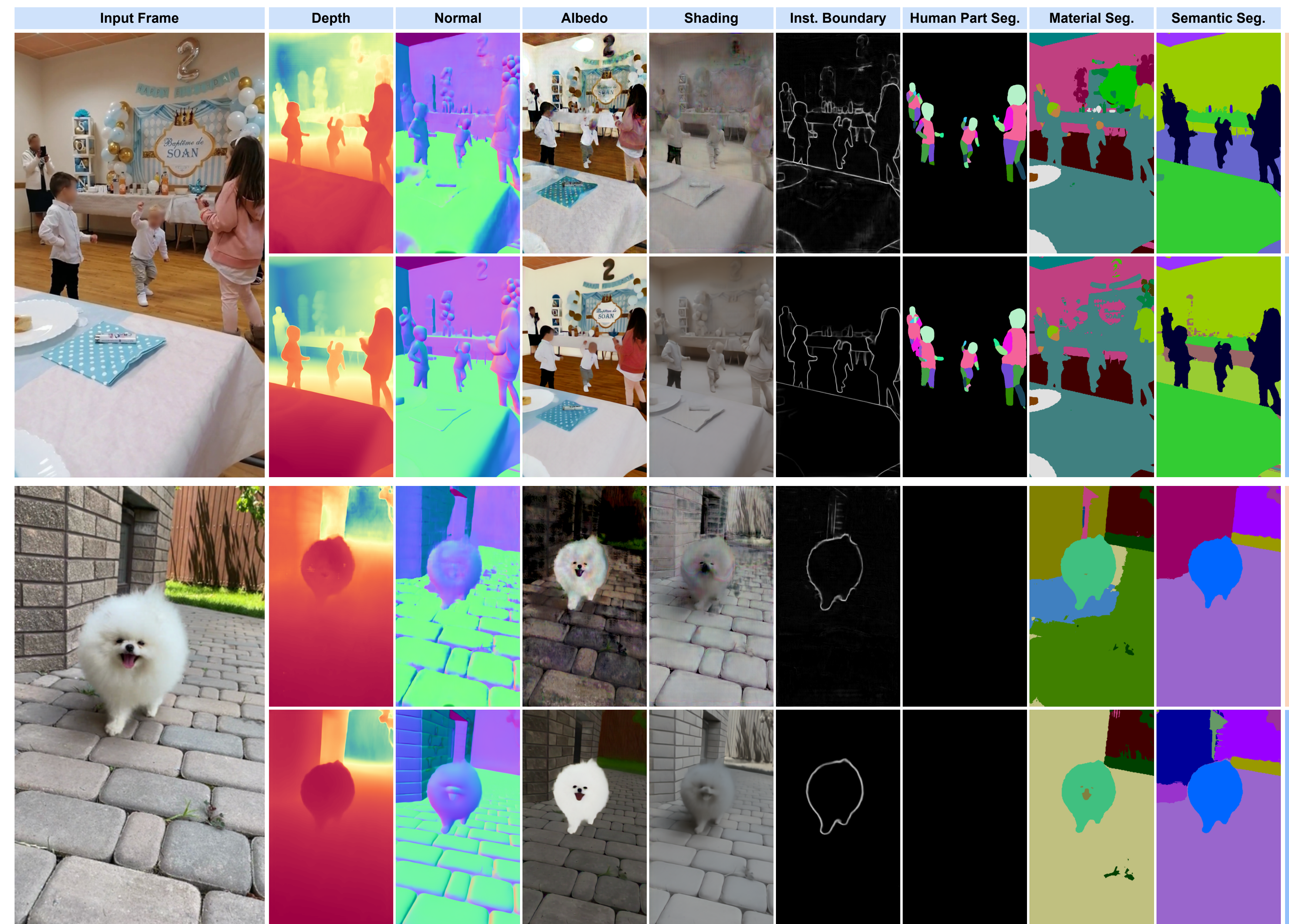
- First unified video dense-prediction model trained entirely on **disjoint, domain-specific data**: diffusion priors bridge the domain gaps, enabling **generalization** to unseen scene-task combinations without co-annotation or pseudo-labeling.
- Latent-space distillation makes **joint training tractable**, while Extended Self-Attention **propagates temporal consistency** to tasks with no video supervision.
- A single forward pass predicts **all 8 tasks** at once, **faster** than sequential multi-task or pseudo-labeling baselines.

Methods



Method Overview. (a) **Task specialist training**: a diffusion backbone and pixel projector finetuned per task on disjoint data, with VAE encoder/decoder frozen. (b) **Unified model training**: one shared backbone and K lightweight latent projectors distill every specialist's frozen latents. (c) **Streaming video inference**: Extended Self-Attention over a frame memory bank gives temporally consistent, single-pass prediction of all 8 tasks..

Qualitative Results



Qualitative Results. The proposed UniD produces sharper, more temporally- and cross-task-consistent predictions than DINOv3-H, especially on out-of-distribution scenes.

Quantitative Results

Method	NYUv2		KITTI		ETH3D		ScanNet		DIODE	
	AbsRel ↓	$\delta_1 \uparrow$	AbsRel ↓	$\delta_1 \uparrow$	AbsRel ↓	$\delta_1 \uparrow$	AbsRel ↓	$\delta_1 \uparrow$	AbsRel ↓	$\delta_1 \uparrow$
Spec.										
DAv2	0.043	98.1	0.076	94.7	0.127	88.2	0.043	98.1	0.260	75.9
Marigold	0.055	96.4	0.099	91.6	0.065	95.9	0.064	95.2	0.308	77.3
Lotus	0.051	97.2	0.081	93.1	0.061	97.0	0.055	96.5	0.228	73.8
UniD ($\mathcal{F}_\theta^D$)	0.059	96.0	0.102	90.9	0.070	95.4	0.069	95.0	0.299	77.4
Unified										
4M-21-XL	0.068	95.1	0.105	89.6	0.070	95.3	0.065	95.5	0.331	73.4
DICEPTION	0.060	96.3	0.085	91.8	0.070	96.6	0.071	94.8	0.304	72.2
StableMTL	0.090	92.4	0.150	79.8	0.116	87.9	0.105	89.5	0.330	73.8
DINOv3-H	0.092	92.9	0.121	86.5	0.087	93.7	0.089	93.6	0.314	77.5
UniD ($\mathcal{G}_\phi$)	0.059	96.3	0.110	89.5	0.073	95.5	0.063	96.2	0.301	77.3

Tab. 1: Monocular Depth Estimation

Method	NYUv2		ScanNet		iBims-1		Sintel	
	mean ↓	11.25% ↑	mean ↓	11.25% ↑	mean ↓	11.25% ↑	mean ↓	11.25% ↑
Spec.								
DSINE	16.4	59.6	16.2	61.0	17.1	67.4	34.9	21.5
StableNormal	18.6	53.5	17.1	57.4	18.2	65.0	36.7	14.1
Lotus	16.2	59.8	14.7	64.0	17.1	66.4	32.3	22.4
UniD ($\mathcal{F}_\theta^N$)	16.7	59.9	15.4	63.7	16.3	69.0	34.3	22.2
Unified								
4M-21-XL	18.4	49.8	17.4	48.9	18.6	61.4	41.6	12.4
DICEPTION	19.7	51.4	19.3	54.5	17.8	66.2	37.2	19.6
StableMTL	19.6	48.9	20.0	45.8	19.6	58.1	40.5	11.0
DINOv3-H	16.8	57.0	16.0	58.7	16.7	66.8	32.3	18.9
UniD ($\mathcal{G}_\phi$)	16.2	59.9	14.6	65.1	16.6	67.4	33.0	21.4

Tab. 2: Surface Normal Estimation

Method	Albedo IIW	Shading SAW	Boundary Mapillary	Semantic ADE20K	Material ADE20K	Human Part Pascal-Human
	WHDRI5% ↓	P@0.7/AP% ↑	odsF ↑	mIoU / fwIoU	mIoU / fwIoU	mIoU / fwIoU
Spec.						
ColorfulShading	0.175	81.3 / 79.4	-	-	-	-
Mask2Former	-	-	35.6	-	-	-
UniD ($\mathcal{F}_\theta^B$)	0.201	93.5 / 91.5	56.3	47.6 / 71.6	67.2 / 76.2	66.9 / 88.4
Unified						
DINOv3-H	0.289	87.8 / 88.0	55.0	62.4 / 74.8	69.1 / 75.7	69.0 / 89.1
StableMTL	0.190	82.7 / 63.7	-	-	-	-
UniD ($\mathcal{G}_\phi$)	0.207	94.6 / 92.6	57.2	35.7 / 66.2	44.1 / 81.4	65.9 / 88.3
UniD ($\mathcal{G}_\phi$) + FT	-	-	-	44.8 / 68.9	66.0 / 76.0	65.4 / 88.0

Tab. 3: Albedo/Shading/Boundary/Semantic/Material/Part Segmentation

Method	Depth ScanNet	Normal InteriorNet	Albedo InteriorNet	Shading InteriorNet	Boundary YT-VIS	Semantic VIPSeg
	AbsRel ↓	$\delta_1 \uparrow$	mVC-4/16 ↑	mVC-4/16 ↑	mVC-4 ↑	mVC-4 ↑
Spec.						
Video Depth Anything	0.063	96.8	-/-	-/-	-	-
NormalCrafter	-	-	84.6 / 70.9	-/-	-	-
ColorfulShading	-	-	-	78.9 / 55.1	73.5 / 41.5	-
UniD ($\mathcal{F}_\theta^B$)	0.101	91.2	85.9 / 76.4	86.9 / 70.1	92.5 / 81.9	93.6
Unified						
DICEPTION	0.123	85.7	68.0 / 47.1	-/-	-	-
DINOv3-H	0.120	86.9	78.2 / 62.2	77.8 / 53.4	90.2 / 77.6	92.0
DICEPTION + ESA[16,1]	0.118	85.9	66.0 / 48.6	-/-	-	-
DINOv3-H + ESA[16,1]	0.150	79.8	79.0 / 63.6	76.4 / 54.1	92.4 / 82.7	92.4
UniD ($\mathcal{G}_\phi$)	0.103	91.0	84.7 / 75.8	88.8 / 73.7	94.5 / 85.2	93.7

Tab. 4: Temporal Consistency on Video Prediction Tasks

Method	M	$\mathcal{E}$ (ms)	U-Net (ms)	Ψ^k (ms)	$\mathcal{D}$ (ms)	$\mathcal{P}^k$ (ms)	Total (s)	FPS
StableMTL	0	-	-	-	-	-	0.03 + 0.58 × K	0.21
4M-21-XL	0	-	-	-	-	-	0.29 + 0.43 × K	0.27
DICEPTION	0	-	-	-	-	-	0.07 + 4.15 × K	0.03
DINOv3-H	0	-	-	-	-	-	0.21 + 0.02 × K	2.67
UniD ($\mathcal{F}_\theta^B$)	0	35.9	58.6 × K	-	42.3 × K	1.1 × K	0.04 + 0.10 × K	1.17
UniD ($\mathcal{G}_\phi$)	0	35.9	58.6	3.1 × K	42.3 × K	1.1 × K	0.09 + 0.05 × K	2.14
DICEPTION + ESA	16	-	-	-	-	-	0.07 + 5.05 × K	0.02
DINOv3-H + ESA	16	-	-	-	-	-	1.30 + 0.02 × K	0.69
UniD ($\mathcal{F}_\theta^B$)	16	36.5	235.0 × K	-	42.2 × K	1.1 × K	0.04 + 0.28 × K	0.44
UniD ($\mathcal{G}_\phi$)	16	36.5	235.0	3.0 × K	42.2 × K	1.1 × K	0.27 + 0.05 × K	1.56

Tab. 5: Per-component inference analysis with varying memory size (M)

Conclusion

- Diffusion priors alone are enough to bridge domain gaps across disjoint, incompatible datasets — no pseudo-labeling, no co-annotation.

Acknowledgements

- This work was conducted in part during an internship at Adobe Research, supported in part by NSF (IIS-2144117); Yihong Sun is supported by an NSF GRFP.