

A Technical Appendices and Supplementary Material

A.1 Supplementary Video Visualizations

To explore our video visualizations, please open the provided `index.html` in the supplementary materials. It presents example outputs of our model from motion video inputs. The underlying raw video files are organized in the `./video` directory.

A.2 Additional Related Work

3D Reconstruction. Many recent learning-based multi-view scene reconstruction methods for point clouds [69, 67, 68, 66] are able to generalize to single or few-view settings. These models, however, often rely heavily on supervised priors from large datasets and are designed for deterministic reconstruction. Their reconstruction focuses on depth estimation instead of recovering the intrinsic scene properties (such as diffuse color) and often have the surrounding illumination baked in the estimated point clouds. Although some view-synthesis techniques based on NeRFs [65] and Gaussian splatting [62, 60] can disentangle reflectance and lighting, they require many viewpoints and per-scene optimization. In contrast, our method, in a single feed-forward inference, leverages differential motion in short videos to successfully resolve texture–lighting ambiguity without the need for many wide-baseline images and explicit camera pose estimation.

Some recent studies [39, 63, 69] focus on generating 3D assets with text prompts or single-image conditioning, which successfully demonstrate generation of unseen parts of the target object. These models are typically trained on artist-designed, cartoon-style 3D assets with diffuse materials instead of real-world captures to increase the amount of supervision. They are also usually limited to producing mesh and texture and do not explicitly recover the reflectance properties, which for real-world objects often causes highlights and illumination to be baked into the texture.

A.3 Ablation Studies

In Figure 8a, we compare the shape estimates from our base model (U-ViT3D) with those from our full model (U-ViT3D-Mixer). Note that, as the base model uses convolutions only at high resolutions, it tends to produce over-simplified geometry estimates. On the other hand, the recovered geometry by our full U-ViT3D-Mixer model captures finer details robustly, even for shiny objects.

In Figure 8b, we show the benefits of leveraging motion cues for material estimation. When the object is static, texture estimation is highly ambiguous: the object appearance could come from painted texture, the reflected environment, or a mixture of the two. Figure 8b shows some samples of those ambiguous interpretations. Once the object moves, our model can separate out the intrinsic texture much more accurately. (See, for example, the white-boxed region where the illumination environment is removed from the estimated albedo.) A video version of these examples can be found in the supplementary webpage.

A.4 Temporal Consistency Guidance

Our model is trained on short clips (e.g., $F = 3$ frames) of objects undergoing differential motion, but it can generalize to longer sequences by additional temporal-consistency guided sampling. During inference, we apply a reconstruction loss (e.g., MSE loss) on one overlapping frame across adjacent temporal windows, encouraging the denoised predictions \hat{x} to remain consistent for the overlapping observation. By applying inference-time optimization to minimize this consistency loss as in prior work [61, 23], we can achieve temporally consistent shape and material estimation over time horizons longer than what the model is originally trained on. This lightweight guidance improves temporal coherence without the need for modification to the model or additional training on long sequences.

We visualize the results in the supplementary webpage and compare it with the StableNormal [57] model, which shows flickering and inconsistency for video-based shape estimation.

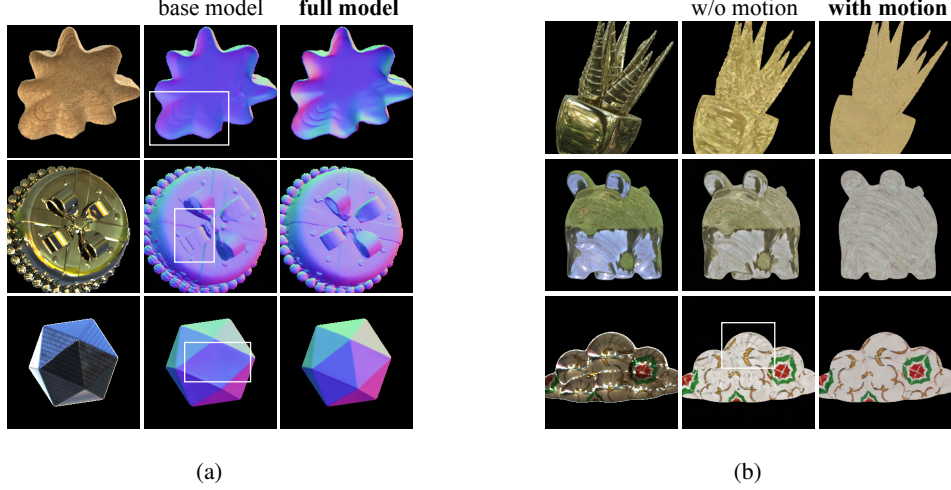


Figure 8: *Left*: Ablation of model components. The base U-ViT3D model, i.e., local attention and channel mixing layers ablated from our full model, produces over-smoothed results and struggles with specular surfaces due to complex inter-reflections. *Right*: Ablation of motion from input images. When the shiny objects are rendered as being static, the model can sometimes provide inaccurate estimates due to the inherent ambiguity between texture and illumination. Once the model observes the object undergoing motion, this ambiguity is resolved and the model produces more accurate albedo estimates.

506 A.5 Reflectance Model

507 We describe the full reflectance model based on the Disney principled BSDF model [9]. Recall the
508 reflectance equation

$$f_r(\omega_i, \omega_o; M) = (1 - \gamma) \frac{\rho_d}{\pi} [f_{\text{diff}}(\omega_i, \omega_o) + f_{\text{retro}}(\omega_i, \omega_o; r)] + f_{\text{spec}}(\omega_i, \omega_o; \rho_d, \rho_s, r, \gamma), \quad (6)$$

509 where $M = \{\rho_d, \rho_s, r, \gamma\}$ consists of the material parameters: diffuse albedo ρ_d , specular term, ρ_s
510 roughness r , metallic-ness coefficient γ and n denotes the spatially varying surface normal at a point.

511 The diffuse term depends on the angle of incident light θ_i and outgoing direction θ_o :

$$f_{\text{diff}} = (1 - F_i/2) (1 - F_o/2), \quad \text{where } F_i = (1 - \cos \theta_i)^5, \quad F_o = (1 - \cos \theta_o)^5. \quad (7)$$

512 To capture retro-reflective highlights, we add

$$f_{\text{retro}} = R_R (F_i + F_o + F_i F_o (R_R - 1)), \quad \text{where } R_R = 2\gamma \cos^2 \theta_d, \quad \cos \theta_d = h \cdot \omega_i, \quad (8)$$

513 with h denoting the half-vector between the incident and outgoing directions.

514 The specular reflection is based on the microfacet model with GGX distribution:

$$f_{\text{spec}} = \frac{F D G}{4 (\omega_i \cdot n) (\omega_o \cdot n)} = \frac{F D G}{4 \cos \theta_i \cos \theta_o}. \quad (9)$$

515 Here, D is a microfacet distribution function defined by the roughness parameter r ,

$$D(h) = \frac{r^4}{\pi ((n \cdot h)^2 (r^4 - 1) + 1)^2}, \quad (10)$$

516 and G denotes the masking-shadowing function

$$G = G_1(\omega_i) G_1(\omega_o), \quad \text{where } G_1(\omega) = \frac{2}{1 + \sqrt{1 + r^4 (1 - n \cdot \omega)^2 / (n \cdot \omega)^2}}. \quad (11)$$

517 The Fresnel term F blends dielectric and metallic responses:

$$F = (1 - \gamma) F_{\text{dielectric}} + \gamma F_{\text{Schlick}}, \quad (12)$$

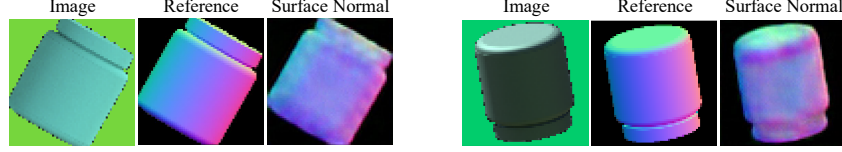


Figure 9: A model trained to predict surface albedo learns representation useful for estimating surface shapes. We show two examples with the ground truth shape (middle) and the readout shape from simple convolutional probes (right).

518 with

$$F_{\text{dielectric}} = \frac{1}{2} \left(\left(\frac{\cos \theta_i - \eta \cos \theta_t}{\cos \theta_i + \eta \cos \theta_t} \right)^2 + \left(\frac{\cos \theta_t - \eta \cos \theta_i}{\cos \theta_t + \eta \cos \theta_i} \right)^2 \right), \text{ where } \eta = \frac{2}{1 - \sqrt{0.08 \rho_s}} - 1, \quad (13)$$

519 where θ_t is the angle between the normal and the transmitted ray computed using Snell’s Law and

$$F_{\text{Schlick}} = \rho_d + (1 - \rho_d) (1 - \cos \theta_d)^5. \quad (14)$$

520 A.6 Probing experiment on normal estimation from albedo prediction

521 We conduct a probing experiment with a low-resolution (64 by 64) diffusion UNet that is trained
 522 to infer the object albedo given a conditional image. We are interested in whether the model learns
 523 useful representations for predicting object surface normals. To do this, inspired by prior work [17],
 524 we first pre-train the albedo prediction UNet and then insert probes similar to the DPT decoder [45]
 525 using multiscale convolutions at intermediate layers. We train the lightweight probes to estimate
 526 shape from latent features using a small dataset of 10K ground-truth pairs for five epochs. At test
 527 time, we directly read out from the inserted probes for surface normal estimates.

528 As shown in Figure 9, the model trained for albedo prediction can indeed produce plausible shape
 529 estimates through probing. As a result, we hypothesize that building a unified framework for both
 530 shape and material estimation can leverage a shared representation and enable a more computationally
 531 efficient architecture.

532 A.7 Network Architecture Details

533 We use the following hyperparameters for the U-ViT3D-Mixer model.

```
534 channels          = [96, 192, 384, 768],
535 block_dropout     = [0, 0, 0.1, 0.1],
536 block_type        = ['Local3D'(1), 'Local3D'(1), 'Transformer'(3), 'Transformer'(8)],
537 noise_embedding_channels = 768,
538 attention_num_heads = 6,
539 patch_size        = 2,
540 local_attention_window_size = 7,
541 channel_mixer_expansion_factor = 3,
542 loss_type         = v-prediction (MSE)
```

543 with the following training setup,

```
544 batch_size        = 64,
545 optimizer          = 'AdamW',
546 adam_betas         = (0.9, 0.99),
547 adam_weight_decay  = 0.01
548 learning_rate      = 1e-4,
549 mixed_precision    = 'bfloat16',
550 max_train_steps    = 400k
```

Appendix References

- [23] Xinran Han, Todd Zickler, and Ko Nishino. Multistable shape from shading emerges from patch diffusion. *Advances in Neural Information Processing Systems*, 37:34686–34711, 2024.
- [60] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. *arXiv preprint arXiv:2311.17977*, 2023.
- [61] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36, 2024.
- [62] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. *arXiv preprint arXiv:2311.16473*, 2023.
- [63] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024.
- [39] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024.
- [65] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [66] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [67] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025.
- [68] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.
- [69] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.
- [57] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024.