

NormalCrafter: Learning Temporally Consistent Normals from Video Diffusion Priors

Supplementary Material

In this supplementary, we first elaborate implementation details in Sec. A. Then, we show more qualitative comparison in Sec. B and Sec. C.

A. More Details about Implementation

Data Resampling. Since our maximum number of frames is 14, we further divide long videos into shorter clips to ensure balanced sampling during training. For hypersim, after obtaining 613 videos, we segment them into 1,780 shot clips, each containing between 30 and 60 frames. In order to get a comparable number of samples with single-image datasets. We upsample 1,780 clips 40 times and get 71,200 clips. For MatrixCity, we divide the 2,316 videos to 7601 short clips and then upsample 10 times, yielding 76010 clips. For Replica, 3D Ken Burns and Objaverse, we do not resample.

Augmentation. During training, the input image goes through the following set of data augmentation (p : the probability of applying each augmentation).

- **Random Horizontal Flipping** ($p = 0.5$). Horizontally flip the clips.
- **Random Cropping** ($p = 0.3$). Crop the image with varying aspect ratios.
- **Color** ($p = 0.1$). Distort the color of input clips by changing brightness, contrast, saturate and hue.
- **Grayscale** ($p = 0.2$). Change the image into grayscale.

Note, **Random Cropping**, **Color** and **Grayscale** are mutually exclusive.

B. Effectiveness of the Two-stage Training Strategy

In “index.html”, we show more qualitative comparison of **Ours w/o Stage1** and **Ours**.

C. More Results

In “index.html”, we show more qualitative with **StableNormal** and **Marigold-E2E-FT** on the DAVIS dataset and Sora-generated videos.