# A. Appendix

## A.1. Dataset

In this section, we provide supplementary information and visualizations of the debris flow dataset. The representative shapes of the debris-flow surface are summarized in Fig. A1. The terrain of the monitoring site is shown in Fig. A2.

(a) (b) (c)

Figure A1. Three stages of debris flow: (a) pre-event, (b)arrival of boulder front and (c) fine-grained slurry fluid.

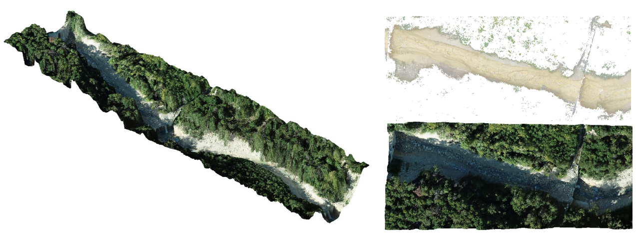


Figure A2. **3D reconstruction of the debris-flow monitoring site [11].** Overview of the scene on the left. Reconstruction of the channel before the event on the upper right, and after the event on the bottom right.

In the debris flow dataset, we release:

- 6000 high-resolution images (1920×1080)
- 6000 high-accuracy point clouds
- Camera calibration matrices
- Rotation matrices between camera and LiDAR
- Translation vectors between camera and LiDAR

## A.2. Evaluation Metrics

**SSIM**. Structural similarity index [50] is used in our optical flow loss to assess the similarity between the target image and warped image:

$$\operatorname{SSIM}(x, y)=\frac{\left(2 \mu_x \mu_y+c_1\right)\left(2 \sigma_{x y}+c_2\right)}{\left(\mu_x^2+\mu_y^2+c_1\right)\left(\sigma_x^2+\sigma_y^2+c_2\right)}, \quad (15)$$

where $\mu_x$ and $\mu_y$ represent the mean pixel values of images $x$ and $y$, and $\sigma$ denotes the corresponding standard deviation.

**Census Transform Loss**. We use ternary census transform loss [17,33,42] as the second metric to evaluate optical flow performance:

$$\mathbf{CT}\left(p, p^{\prime}\right)=\begin{cases}-1 & \text{if } p^{\prime}-p \geq \epsilon \\ +1 & \text{if } p-p^{\prime} \geq \epsilon \\ 0 & \text{if } \left|p-p^{\prime}\right|<\epsilon\end{cases} \quad (16)$$

Given two input images, we compute the corresponding census-transformed images and compute the average difference between them as the loss.

## A.3. LiDAR to Range Image

Since point cloud-based networks [37, 38, 54] are computationally demanding and complex to train, we convert the 3D scan points to sparse range maps with the help of the camera-LiDAR transformation by projecting 3D points onto the image plane with known camera intrinsics $\mathbf{K}$ and camera pose parameters $\mathbf{R}$ and $\mathbf{t}$:

$$\mathbf{K}[\mathbf{R} \mid \mathbf{t}]=\left[\begin{array}{ccc} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{array}\right]\left[\begin{array}{cccc} r_1 & r_2 & r_3 & t_1 \\ r_4 & r_5 & r_6 & t_2 \\ r_7 & r_8 & r_9 & t_3 \end{array}\right]. \quad (17)$$

The 2D projection $\mathbf{p}=(u, v, 1)^T$ of 3D point $\mathbf{P}=(x, y, z, 1)^T$ is computed as

$$\mathbf{p}=\mathbf{K}[\mathbf{R} \mid \mathbf{t}] \mathbf{P} \quad (18)$$

and rounded to the closest integer pixel coordinate.

## A.4. Runtime and Model Size

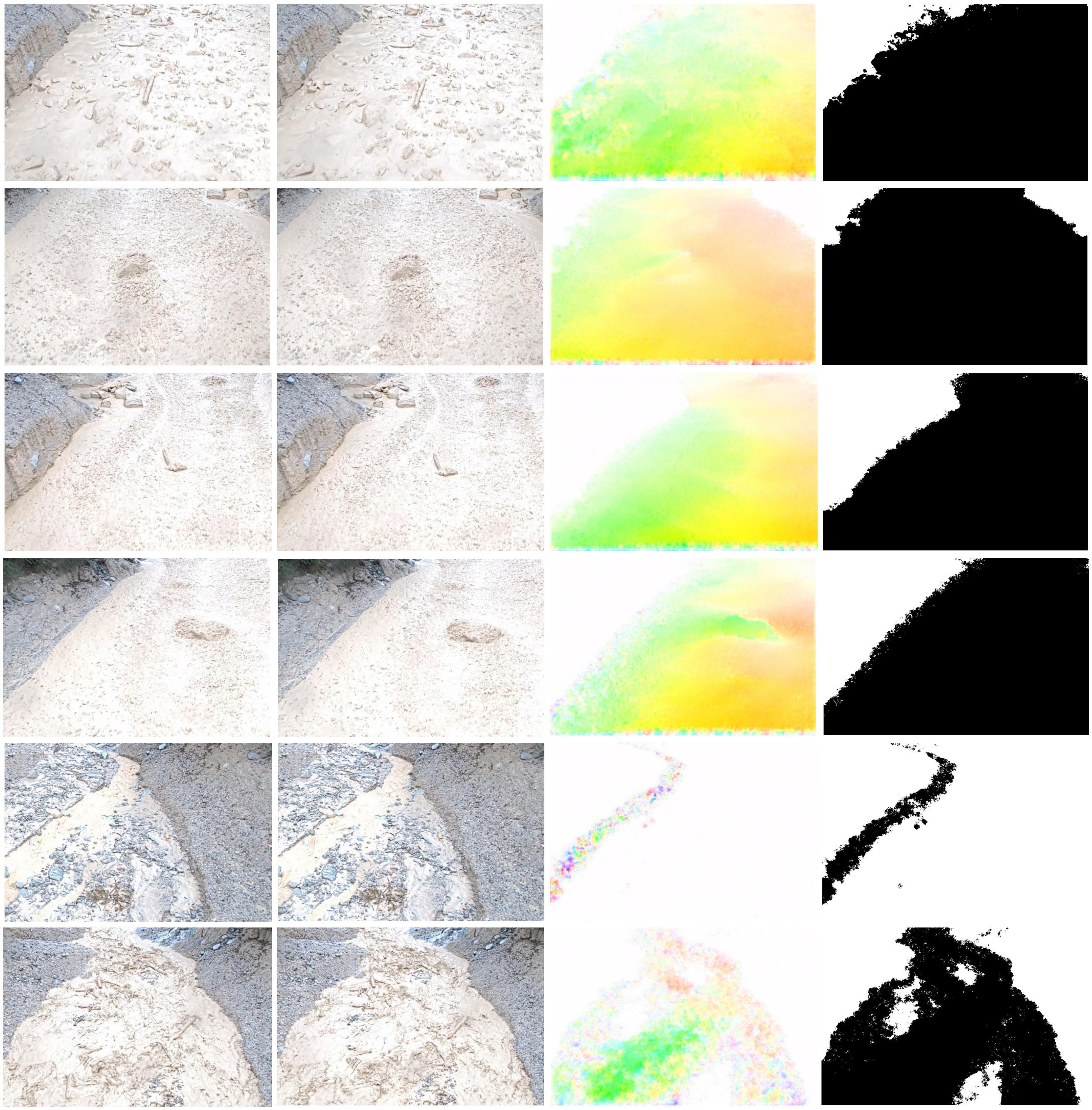
We report the runtimes and model sizes of RAFT, DeFlow-Cam, and DeFlow-Fusion in Tab. A1. Our camera-only baseline is significantly smaller and faster,since we follow the lightweight design of PWC-Net [43] and Mono-SF [21]. Our fusion model has similar model size and runtime as RAFT. The increase in runtime and model size compared to the camera-only baseline is caused by the additional depth encoder and the multi-level feature fusion, which in return offer marked performance gains.

| Method | Runtime [ms] | # params |
|---|---|---|
| RAFT [45] | 171.3 | 5.26 M |
| DeFlow-Cam (*Ours*) | 54.1 | 4.16 M |
| DeFlow-Fusion (*Ours*) | 191.9 | 4.99 M |

Table A1. Runtime and number of parameters for different models.

## A.5. Additional Qualitative Results

In Fig. A3, we present additional visualizations of optical flow results, and of the static pixel masks used to enforce a static sensor pose. To see the qualitative behaviour of our temporal smoothing module, *readers are encouraged to watch the video in the supplementary material.*



(a) (b) (c) (d)

Figure A3. Qualitative results of optical flow (c) and binary (static/dynamic) segmentation (d).