COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Richard Higgins

Learning Dense Visual Features for the Sun and Natural Scenes

Thursday, January 16, 2025
12:00pm – 2:00pm
3725 Beyster
Hybrid – Zoom Passcode: 966387

ABSTRACT: We demonstrate simple methods for visual representation learning in two data-rich settings: the sun and natural scenes. We produce dense representations that are applicable to the following tasks: (1) improved solar magnetic field estimation; (2) improved hand and object segmentation for video understanding; (3) 3D-aware image/scene editing for compositional generation.

First, we train UNets to fast and accurately estimate the per-pixel magnetic field of the sun by inverting polarized light measurements from the Solar Dynamics Observatory/Helioseismic and Magnetic Imager (SDO/HMI). We then create SynthIA, a synthetic instrument trained on paired data from two satellites: SDO and Hinode. SDO/HMI images the entire solar disk, while the Hinode/SOT spectropolarimeter captures only small regions at a higher spectral and spatial resolution. We pair the co-observed input data from SDO/HMI with MERLIN inversions from the Hinode/SOT spectropolarimeter to create a cross-satellite dataset. After training, SynthIA is able to generate Hinode MERLIN-like inversions from only SDO/HMI input data, synthetically expanding Hinode's spatially-limited but high-quality inversion results to the full disk.

For natural scenes, we introduce responsibility, a means of ascribing scene motion (estimated via optical flow) to hands. We pair responsibility with an off-the-shelf person segmentation system and use the resulting pseudolabels with a three-way contrastive loss to train a UNet that segments people, held-objects, and background pixels. We next extend this gestaltist idea of motion and shared fate to learn a self-supervised image segmentation system. We first estimate a fundamental matrix between two ego-centric video frames. We then create pseudolabels by decomposing pixels that disagree with this camera motion model into hands and held-objects. We then use these pseudolabels to train an HRNet that class-agnostically segments scenes.

Finally, we perform 3D-aware image/scene edits by conditioning latent diffusion using a sequence of neural nouns and verbs as a visual prompt. We edit scenes by applying verbs in an object-centric manner and then recompose the scene with a background. This factorization affords test-time compositionality, allowing us to compose edited objects from multiple datasets in the same scene while preserving camera control.

CHAIR: Prof. David Fouhey