Current non-rigid structure from motion (NRSfM) algorithms are limited with respect to: (i) the number of images, and (ii) the type of shape variability they can handle. This has hampered the practical utility of NRSfM for many applications within vision. Deep Neural Networks (DNNs) are an obvious candidate to help with such issue. However, their use has not been explored in recovering poses and 3D shapes from an ensemble of vector-based 2D landmarks. In this proposal, we present a novel deep neural network to recover camera poses and 3D points solely from an ensemble of 2D image coordinates. The proposed neural network is built upon our prior work on compressible structure from motion -- extending the original single-layer sparsity constraint to a multi-layer one. The network architecture is mathematically interpretable as a multi-layer block sparse dictionary learning problem.

Our network is capable of handling problems of unprecedented scale in terms of samples and parameterization -- allowing us to effectively recover 3D shapes deemed too complex by previous state-of-the-art. We further propose a generalization measure (based on the network weights) for guiding training to efficiently avoid over-fitting, circumventing the need for 3D ground-truth. Once the network’s weights are estimated (for a non-rigid object) we show how our approach can be used to recover 3D shape from a single image without 3D supervision. We shall propose how to extend our current framework to handle missing data, and identify new applications -- such as Structure from Category (SfC) -- where our approach can have substantial impact within computer vision.