The human visual system allows us to effortlessly construct a picture of the world from very little input, often in the form of inferring higher dimensions from lower dimensional representations. We understand 2D surfaces as lines, 3D structure as surfaces, and motion and spatial relations through 3D structures. The motivation for my research is to automatically infer the spatial context outside the image domain by bridging the gap between such representational dualities in our visual world — in particular, learning to generate 3D shapes through projective supervision. We consider the problem of recovering the 3D shape of an object from a single depth image and silhouette under the assumption that the underlying volume is not available at training time. Reconstructing the complete 3D shape of an object from one or a few images (e.g. predicting the four legs of a table when only the top portion is visible) is considered ill-posed in general without strong contextual cues due to self-occlusions and ambiguities in the object's pose and scale. A common formulation is to predict binary labels in a volumetric grid using synthetic datasets. In contrast, humans rely on view-based observations to derive 3D shapes without ever "seeing through" the underlying volume. We are interested in evaluating the two aforementioned representations on volumetric shape transfer across unseen object views, instances, and categories. Studies on the human visual system suggest that humans process 3D information through viewer-centered representations by remembering many different viewpoints of an object and representing each view as a 2.5D sketch. To apply this idea to 3D shape generation, we train a neural net to output 3D shapes represented as multiple depth images and silhouettes. In addition, we observe the general problem that 3D pose alignment is not always uniquely defined across categories.

As addressed in the work of Tulsiani et al. [1], cross-category pose induction is a difficult task on its own, so we choose a viewer-centered frame of reference for the 3D representation (rather than an object-centered or shared coordinate system). This is also related to biological findings that the human visual reference frame is tied to viewer-centered coordinates [2]. As a result, our network learns a view-specific 3D representation which can potentially generalize better to novel object categories because new shapes can be interpolated from alignment-independent representations. In summary, our formulation is viewer-centered in that: (1). we learn to generate 3D shapes represented as 2.5D projections (2). in a reference frame relative to the viewer.

REFERENCES

[1] S. Tulsiani, J. Carreira, and J. Malik, “Pose induction for novel object categories,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 64–72, 2015.

[2] M. J. Tarr and S. Pinker, “When does human object recognition use a viewer-centered reference frame?,” Psychological Science, vol. 1, no. 4, pp. 253–256, 1990.