A. Baseline implementation details

In this section, we provide more implementation details for autoencoder, egomotion and shapecode baselines.

**Autoencoder** [2] is trained to reconstruct an input image $x$, with dimension $224 \times 224 \times 3$. We adopt the autoencoder architecture design of VGG16 from [3]. However, the architecture in [3] is similar to Unet [5], which has intermediate connections between different intermediate feature outputs. We remove those intermediate connections to resemble autoencoder, whose latent feature has 4096 dimension. L2 loss is used to measure the difference between ground truth $y = x$ and the reconstructed image $\hat{y}$, which can be formulated as.

$$
\mathcal{L} = ||\hat{y} - y||^2.
$$

**Egomotion** [1] predicts the camera motion between 2 images. Assuming only $V$ viewpoints exist in the dataset, the model will output $V - 1$ probabilities, corresponding to the $V - 1$ viewpoints differences. Given a pair of images $x_1$ and $x_2$, two 4096 dimension features $f_1$ and $f_2$ are extracted from the last layer of VGG16. These 2 feature are then concatenated into a 8192 dimension feature, which is then fed into a stacked fully connected layers ($8192 \rightarrow 4096 \rightarrow 1024 \rightarrow V - 1$) to predict the relative view point difference using softmax.

**ShapeCode** [4]. We again use architecture similar to autoencoder, but instead of outputting a reconstructed image $\hat{y}$, the network outputs $\{\hat{y}_j\}_{j=1}^V$, where $V$ is the total number of images associated to the same object. Specifically, given the input image $x$ with $224 \times 224 \times 3$ dimension, the output of the network has dimension $224 \times 224 \times 3 \times V$ and the loss function becomes

$$
\mathcal{L} = \sum_{j=1}^{V} ||\hat{y}_j - y_j||^2
$$

Note that these $V$ images are organized sequentially and support missing views as suggested in [4].
References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by mov-
ing. In 2015 IEEE International Conference on Computer Vi-


[3] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image seg-

supervised learning through one-shot image-based shape re-