# RayTran: 3D pose estimation and shape reconstruction of multiple objects from videos with ray-traced transformers

## *Supplemental Material*

M. J. Tyszkiewicz[2⋆], K.-K. Maninis[1], S. Popov[1], and V. Ferrari[1]

[1]Google Research [2]EPFL

## A Preliminary results on novel view synthesis

In the main paper, we showed that the RayTran backbone is an effective approach for 3D pose estimation and shape reconstruction from videos, and leads to state-of-the-art results. On top of the main task, the architecture enables two additional auxiliary tasks: 3D occupancy prediction, and 2D foreground-background segmentation.

Similarly, the features created by RayTran and their dual $3D \Leftrightarrow 2D$ interpretation enable other different tasks that use posed images as input and require geometric reasoning. In this section we re-purpose the backbone for the task of novel view synthesis. Despite being a fundamentally different task, we are able to approach novel view synthesis with minimal modifications to the same end-to-end trainable architecture.

We enable novel view synthesis by projecting the 3D voxel grid $V$ to a $30 \times 40$ 2D grid $P_i$ aligned with the novel (query) view, by using a new camera pose. To this end, we use an additional $3D \rightarrow 2D$ transformer block. We then use a simple decoder comprising of transpose convolutions with non-linearities and normalization layers to recover images at the original input resolution ($480 \times 640$), and we supervise with the MSE loss. We ask our model to predict new query views, unseen during training. The encoder-decoder setup is such that the encoder does not have access to the query frame parameters, forcing it to infuse the 3D representation with information relevant to all possible query viewpoints.

We show preliminary qualitative results in Fig. 1 and Fig. 2. Our method predicts the overall structure of the images rather well, which suggests that it builds features for sufficient geometric reasoning. It is less accurate in terms of high frequency details, likely because the resolution of the 2D feature grid is too low for the simple network that we use for upscaling. In contrast to NeRF-like architectures [1], however, it does not require any training at test time.

⋆ Work done while at Google Research.

## B Additional Qualitative Results on 3D pose and shape.

Fig. 3 and Fig. 4 illustrate additional results of RayTran in Scan2CAD for the task of 3D pose estimation and shape reconstruction, including successful reconstructions, as well as typical failure cases.

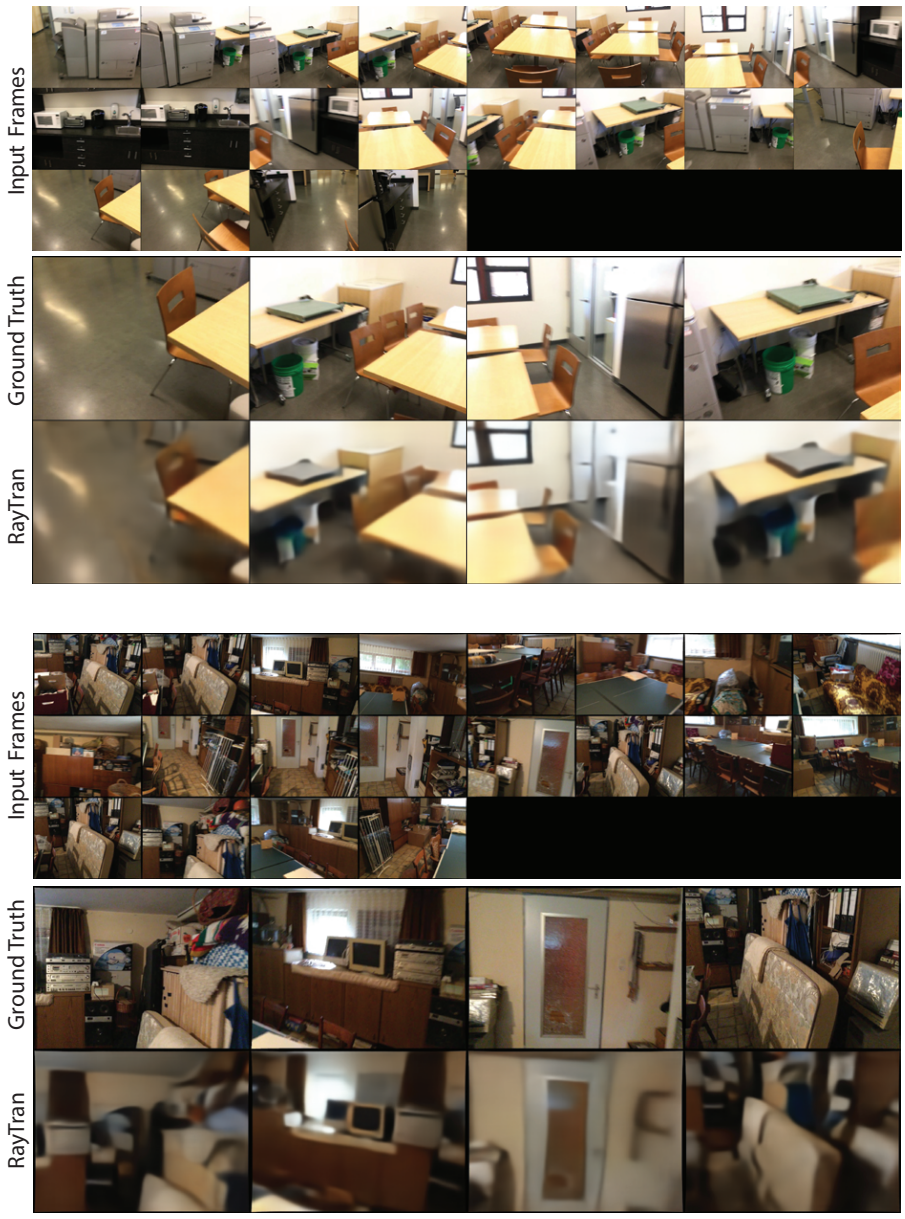


Fig. 1: **Novel view synthesis (NVS).** Qualitative results for 2 different scenes from the test set. For each scene, we show the input frames to the network (top), the ground-truth views from new query camera viewpoints (middle), and the corresponding views predicted by the novel view synthesis head on top of the RayTran backbone (bottom). The $2D \Rightarrow 3D$ attention in our model recovers the view's features at a resolution of $30 \times 40$, which is then upsampled to $480 \times 640$.

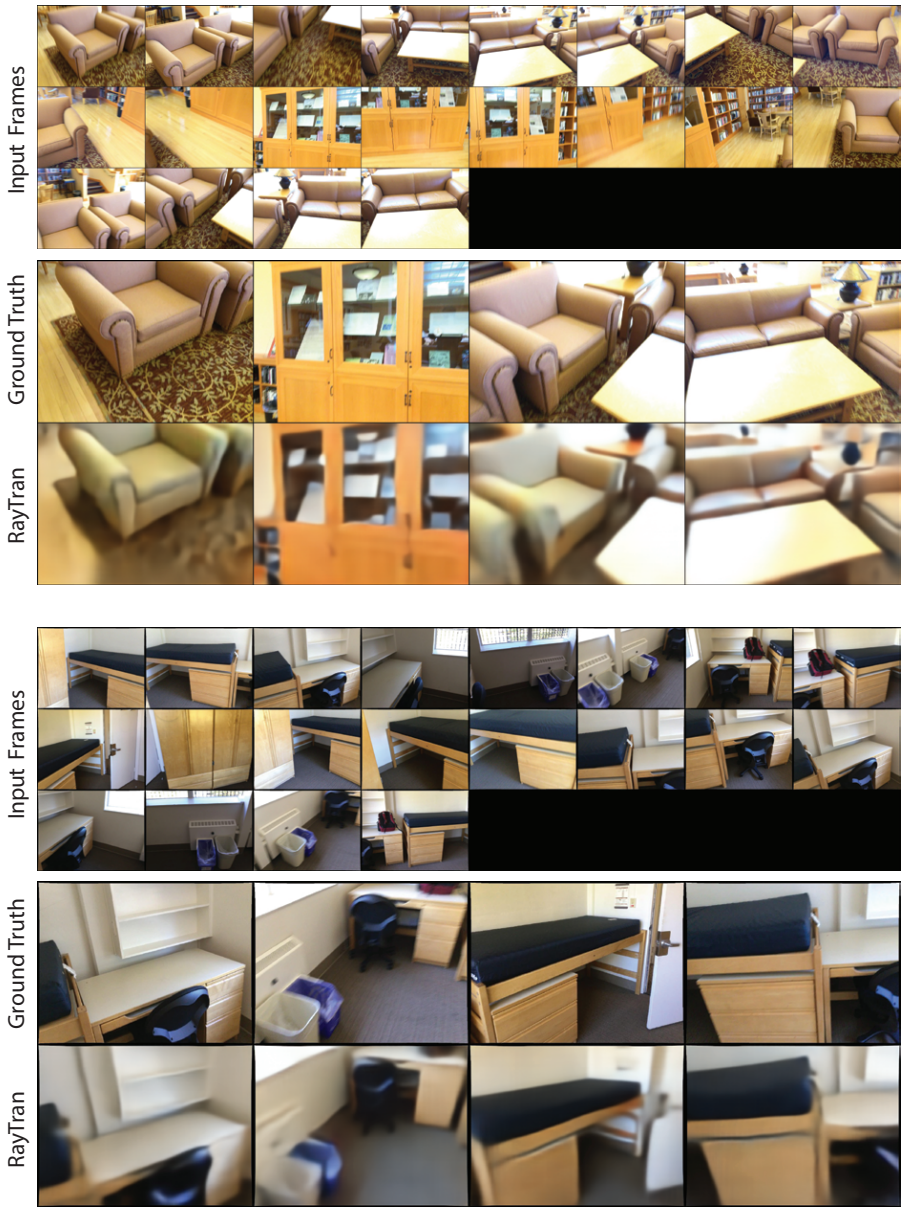


Fig. 2: **More results on novel view synthesis (NVS).** Qualitative results for 2 different scenes from the test set. For each scene, we show the input frames to the network (top), the ground-truth views from new query camera viewpoints (middle), and the corresponding views predicted by the novel view synthesis head on top of the RayTran backbone (bottom). The $2D \Rightarrow 3D$ attention in our model recovers the view's features at a resolution of $30 \times 40$, which is then upsampled to $480 \times 640$.

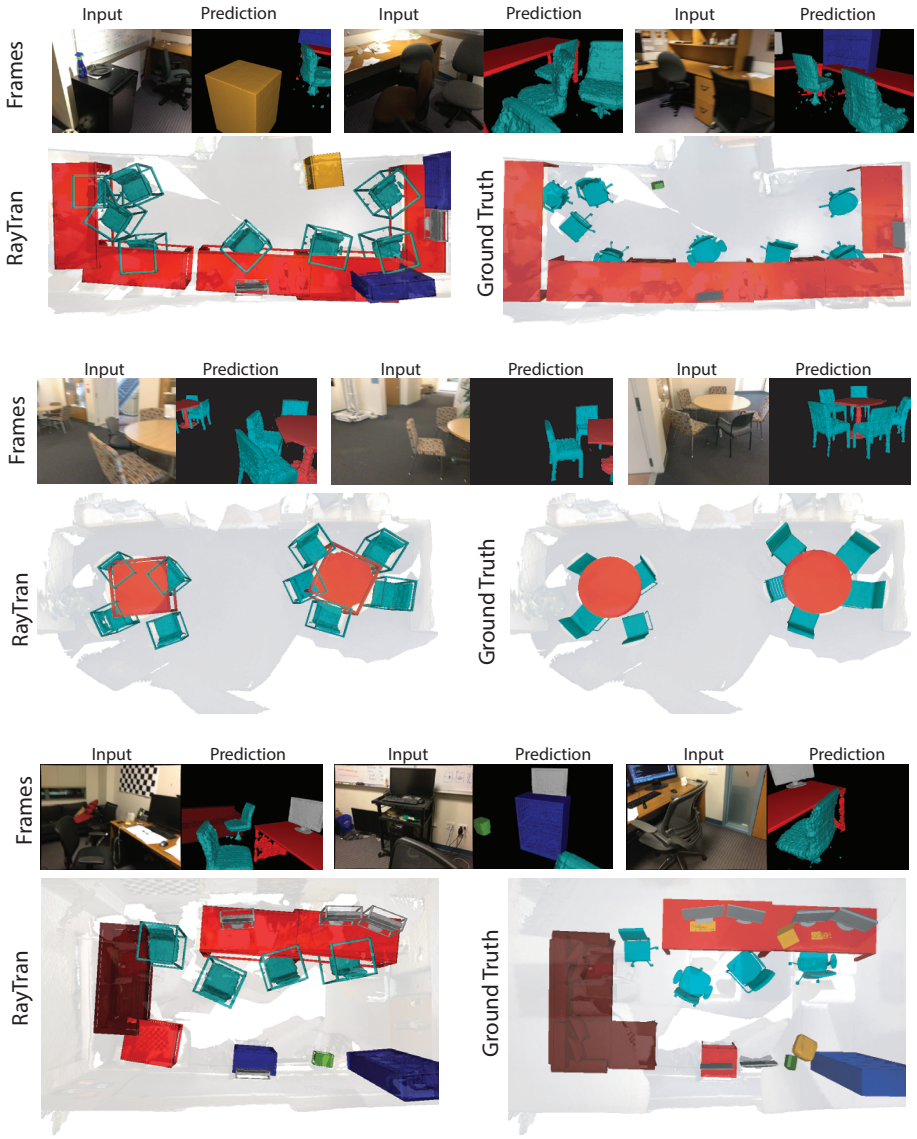


Fig. 3: **Qualitative Results for 3D pose and shape (with frame overlays)**: We show 3D pose estimation and shape reconstruction outputs of RayTran against the ground truth. We also show the results from the viewpoint of the images. The objects are colored by class. Notice that RayTran predicts objects that are omitted from the ground truth but appear in the video (eg. cabinet and the bookshelf of top row).

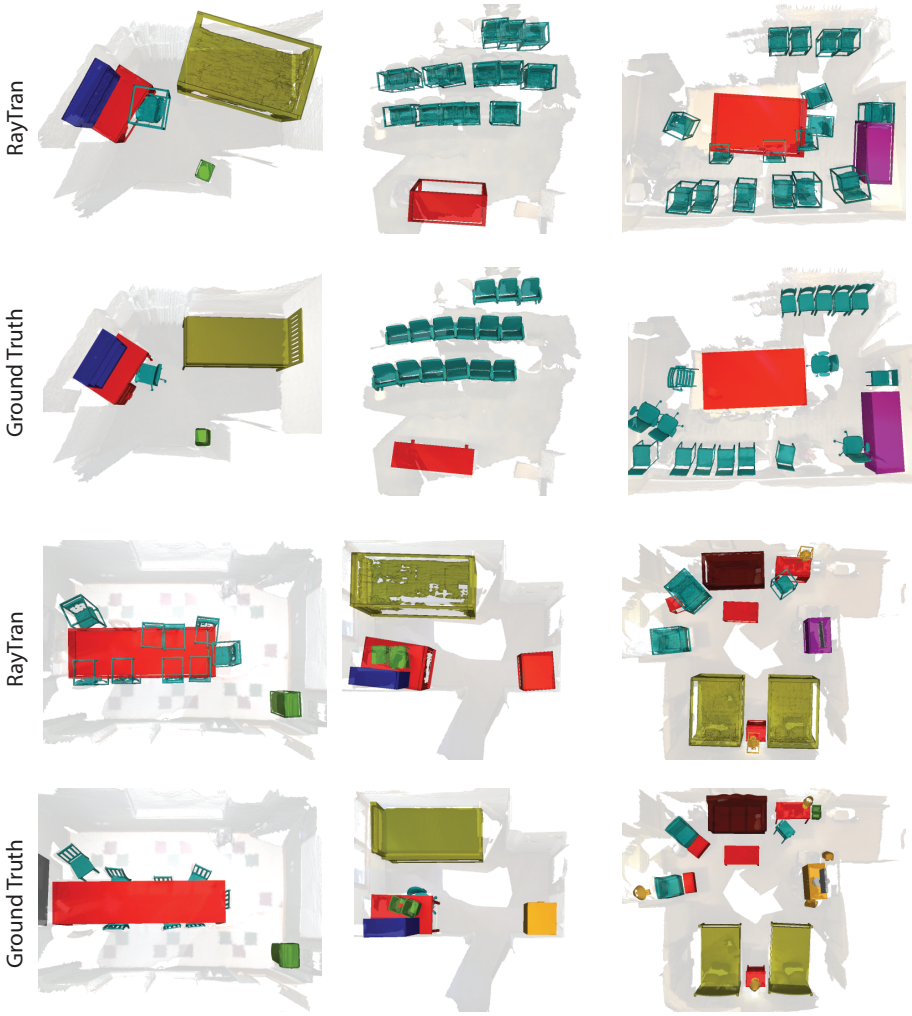


Fig. 4: **More qualitative Results for 3D pose and shape (top-view)**: We show the 3D pose estimation and shape reconstruction outputs for 6 example scenes. For each detected object we visualize its 3D oriented bounding box, as well as its reconstructed mesh. For reference, we also show the ground-truth alignments, as well as the RGB-D scan of the scene in the background (which we do not use). RayTran is able to reconstruct objects of very complicated scenes with densely placed chairs (top row, columns 2 and 3), and scenes with many different classes. It also reconstructs large objects that are often truncated in every individual frame of the video due to their size, and thus need multiple frames to be reconstructed (tables of top row column 3, bottom row column 1). Failure cases include predicting an object from a different class (eg. table instead of a similar-looking cabinet of bottom row, column 2), or missing objects entirely (eg. the left lamp of bottom row, column 3).

## References

1. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 1