

CoCoNets: Continuous Contrastive 3D Scene Representations

Shamit Lal*, Mihir Prabhudesai*, Ishita Mediratta, Adam W. Harley, Katerina Fragkiadaki
 Carnegie Mellon University
 {shamitl, mprabhud, imedirat, aharley, katef} @cs.cmu.edu

Abstract

This paper explores self-supervised learning of amodal 3D feature representations from RGB and RGB-D posed images and videos, agnostic to object and scene semantic content, and evaluates the resulting scene representations in the downstream tasks of visual correspondence, object tracking, and object detection. The model infers a latent 3D representation of the scene in the form of 3D feature points, where each continuous world 3D point is mapped to its corresponding feature vector. The model is trained for contrastive view prediction by rendering 3D feature clouds in queried viewpoints and matching against the 3D feature point cloud predicted from the query view. Notably, the representation can be queried for any 3D location, even if it is not visible from the input view. Our model brings together three powerful ideas of recent exciting research work: 3D feature grids as a neural bottleneck for view prediction, implicit functions for handling resolution limitations of 3D grids, and contrastive learning for unsupervised training of feature representations. We show the resulting 3D visual feature representations effectively scale across objects and scenes, imagine information occluded or missing from the input viewpoints, track objects over time, align semantically related objects in 3D, and improve 3D object detection. We outperform many existing state-of-the-art methods for 3D feature learning and view prediction, which are either limited by 3D grid spatial resolution, do not attempt to build amodal 3D representations, or do not handle combinatorial scene variability due to their non-convolutional bottlenecks.

1. Introduction

Understanding the three-dimensional structure of objects and scenes may be a key for success of machine perception and control in object detection, tracking, manipulation and navigation. Exciting recent works have explored learn-

ing representations of objects and scenes from multiview imagery and capture the three-dimensional scene structure implicitly or explicitly with 3D binary or feature grids [49, 48, 43], 3D point feature clouds [53], implicit functions that map continuous world coordinates to 3D point occupancy [5, 11, 44, 28, 34, 4], as well as 1D or 2D feature maps [7]. These methods typically evaluate the accuracy of the inferred 3D scene occupancy [5, 33, 49, 28, 34, 4] and the fidelity of image views rendered from the 3D representation [25, 29, 7, 53, 44], as opposed to the suitability of representations for downstream semantic tasks. Methods that indeed focus on rendering photo-realistic images often give up on cross-scene generalization [29, 39], or focus on single-object scenes [44]. Methods that instead focus on learning semantically relevant scene representations are expected to generalize across scenes, and handle multi-object scenes. In the 2D image space, contrastive predictive coding has shown to generate state-of-the-art visual features for correspondence and recognition [52, 14], but does not encode 3D scene structure. In 3D voxel feature learning methods [16, 15], convolutional latent 3D feature grids encode the 3D structure and a view contrastive objective learns semantically useful 3D representations, but the grid resolution limits the discriminability of the features learnt. Recent exciting works combine 3D voxel grids and implicit functions and learn to predict 3D scene and object 3D occupancy from a single view with unlimited spatial resolution [35, 36]. The model proposed in this work brings together these two powerful ideas: 3D feature grids as a 3D-informed neural bottleneck for contrastive view prediction [16], and implicit functions for handling the resolution limitations of 3D grids [35].

We propose Continuous Contrastive 3D Networks (CoCoNets), a model that learns to map RGB-D images to infinite-resolution 3D scene feature representations by contrastively predicting views, in an object and scene agnostic way. Our model is trained to predict views of static scenes given 2.5D (color and depth; RGB-D) video streams as input, and is evaluated on its ability to detect and recognize objects in 3D. CoCoNets map the 2.5D input streams into 3D feature grids of the depicted scene. Given a target view

*Equal contribution

Project page: https://mihirp1998.github.io/project_pages/coconets/

and its viewpoint, the model first warps its inferred 3D feature map from the input view to a target view, then queries point features using their continuous coordinates, and pulls these features closer to the point features extracted from the target view at the same 3D locations (Figure 1). We use a contrastive loss to measure the matching error, and back-propagate gradients end-to-end to our differentiable modular architecture. At test time, our model forms plausible 3D completions of the scene given a *single RGB-D image* as input: it learns to fill in information behind occlusions, and infer the 3D extents of objects.

We demonstrate the advantages of combining 3D neural bottleneck, implicit functions and contrastive learning for 3D representation learning by comparing our model against state-of-the-art self-supervised models, such as i) contrastive learning for pointclouds [54], which shares a similar loss but not the amodal predictive ability of our model, ii) contrastive neural mapping [15], which can amodally inpaint a 3D discrete feature grid but suffers from limited spatial grid resolution, and iii) dense ObjectNets [8], which self-learns 2D (instead of 3D) feature representations with a triangulation-driven supervision similar to (i). Our experimental results can be summarized as follows: **(1)** 3D object tracking and re-identification (Figure 3): We show that scene representations learnt by CoCoNets can detect objects in 3D across large frame gaps better than the baselines [54, 15, 8]. **(2)** Supervised 3D object detection: Using the learnt 3D point features as initialization boosts the performance of the state-of-the-art Deep Hough Voting detector of [37]. **(3)** 3D cross-view and cross-scene object 3D alignment: We show that the learnt 3D feature representations can infer 6DoF alignment between the same object in different viewpoints, and across different objects of the same category, better than [54, 15, 8]. We further show that our model can predict image views (with or without depth as input) and 3D occupancies that outperform or are on par with the state-of-the-art view and occupancy prediction models [36, 50, 7].

In summary, the main contribution of this paper is a model that learns infinite-resolution 3D scene representations from RGB-D posed images, useful for tracking and corresponding objects in 3D, pre-training 3D object detectors, and predicting views and 3D occupancies. We set a new state-of-the-art in self-supervision of 3D feature representations.

2. Related work

Learning to 3D reconstruct objects and scenes Learning to infer 3D reconstructions of objects and scenes from single images or videos has been the goal of recent deep geometrical methods, that have explored a variety of explicit 3D representations, such as 3D point clouds [32, 26], 3D binary voxel occupancies [49, 55], or 3D meshes [22, 12].

Since detailed 3D supervision is only possible in large scale in simulation, many approaches attempt supervision from RGB or depth map prediction through differentiable rendering of the inferred 3D reconstruction [22, 12]. To handle limitations of spatial resolution of 3D voxel grids, recent approaches represent 3D object occupancy with implicit functions parameterized by deep neural networks trained to map continuous world 3D coordinates to the corresponding point 3D occupancy values [5, 28, 34, 4]. While these methods train one 3D shape implicit function per object, by assigning a 1D latent embedding to each object, some works [36, 35] train 3D grids of functions where each function takes care of estimating the occupancy of points in the vicinity of the corresponding 3D voxel centroid. These latter methods enjoy the generalization of 3D convolutions and can scale to multi-object scenes, while implicit functions parametrized by non-convolutional, fully connected networks are mostly limited to single object scenes [5]. Our approach also employs 3D grids of functions which predict feature embeddings for the corresponding continuous 3D world coordinates as opposed to merely occupancy.

Neural image synthesis with 3D inductive biases Deep image generative networks have shown compelling results in generating photorealistic images that match the image statistics of the unlabelled image collections they are trained on [13]. They are based on variational autoencoders [24], generative adversarial networks [2], generative flows [23], or autoregressive image pixel generators [51]. However, these 2D generative models learn to parameterize the manifold of 2D natural images, and struggle to generate images that are multi-view consistent, since the underlying 3D scene structure cannot be exploited. Generative models trained from multi-view data can render arbitrary views of an input scene [46, 21, 43, 45, 40]. Such multi-view generative models often restrict themselves to single-object scenes [44, 31, 42] or to a single complex scene without aiming at cross-scene generalization [29, 39], with the goal of generating high fidelity photorealistic images, replacing hand-engineered graphics engines. Their architectures incorporate many inductive biases of graphics engines, such as 3D-to-2D rendering modules [46] and explicit feature transformations to handle viewpoint changes [50]. Their lack of cross-scene generalization or their limitation to single object scenes makes it hard to adopt their inferred feature representations for visual recognition.

Learning visual feature representations by self-supervised view prediction Recent methods learn neural scene representations by predicting views of a scene under known egomotion [7, 44, 16, 43]. View prediction, as opposed to the related and very effective objective of feature learning via triangulation [54], results in explicitly

or implicitly *amodal* representations, i.e., representations that can predict information missing from the input observations [30], as opposed to simply featurizing the visible image pixels [8]. Different view prediction methods for learning representations vary with respect to the amount of their reasoning regarding geometry and the underlying 3D structure of the scene [47]. The generative query network (GQN) of Eslami *et al.* [7] showed that it can predict alternative views of toy simulated scenes without explicit 3D structure inference, and demonstrated the usefulness of the inferred representations as pre-training for reinforcement behaviour learning tasks. Geometry-aware recurrent networks of Tung *et al.* [50] use a latent 3D feature map in their bottleneck for view prediction and demonstrate superior generalization, granted from the 3D convolutional inductive bias. Harley *et al.* [16] uses a similar 3D latent feature grid but optimizes for contrastive prediction as opposed to RGB regression, and demonstrate its usefulness for 3D object tracking and 3D moving object segmentation. Our work optimizes a contrastive view prediction objective similar to [16] but uses a 3D grid of implicit functions as its latent bottleneck. We empirically show that the emergent 3D feature representations are more accurate in 3D object tracking and visual correspondence than the features obtained from existing state-of-the-art 3D feature learning methods.

3. Continuous Contrastive 3D Networks (CoCoNets) for Learning Amodal Visual Representations

We consider a mobile agent that can move about the scene and has access to its egomotion. The agent has a color camera with known intrinsics, and a depth sensor registered to the camera’s coordinate frame. We use groundtruth depth provided by the simulation environment, and we will show in Sec. 4 that the learned models generalize to the real world, where (sparser) depth is provided by a LiDAR unit.

CoCoNets learn 3D visual feature representations by collecting posed images in static scenes and doing contrastive view prediction. We describe the architecture in Sec. 3.1. We then evaluate the correspondability of the resulting 3D feature representations in 3D object re-identification and tracking in dynamic scenes (Section 4.1), as pre-training for 3D object detection (Section 4.2), and cross-object semantic visual correspondence (Section 4.3).

3.1. Continuous Contrastive 3D Networks (CoCoNets)

Our model’s architecture is illustrated in Figure 1. It is a neural network with a three-dimensional neural bottleneck $\mathbf{M} \in \mathbb{R}^{w \times h \times d \times c}$, which has three spatial dimensions (width w , height h , and depth d) and a feature dimension (c

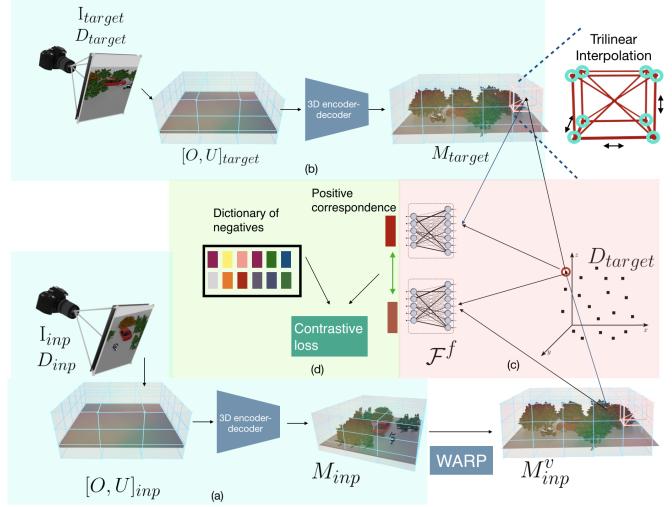


Figure 1: Continuous Convolutional Contrastive 3D Networks (CoCoNets) are trained to lift 2.5D images to 3D feature function grids of the scene by optimizing for view-contrastive prediction. (a) In the top-down path, the model encodes RGB-D images into a 3D feature map $\mathbf{M} \in \mathbb{R}^{w \times h \times d \times c}$, and uses explicit 3D feature transformations (translation and 3D rotation) to account for changes of viewpoint between the input and target views. (b) In the bottom-up path, we encode the RGB-D of the target viewpoint into a 3D feature cloud. (c) Given continuous 3D world coordinates (X, Y, Z) and its embedded code $v_{(X,Y,Z)}$ inferred via trilinear interpolation, a fully connected network maps the coordinates and the embedded code, to the feature vector of the 3D point at location (X, Y, Z) . (d) Metric learning losses in 3D tie the two point cloud representations together.

channels per grid location).

The latent state aims to capture an informative and geometrically-consistent 3D deep feature map of the world space. Therefore, the spatial extent corresponds to a large cuboid of world space, defined with respect to the camera’s position at the first timestep. Each voxel in the 3D feature map \mathbf{M} corresponds to a cuboid in the 3D scene depicted in the RGB-D image.

To be able to generate features at infinite spatial resolution, i.e., to featurize continuous 3D physical points within a voxel, we use implicit function parametrization. Our model’s architecture of interpolating within a voxel grid resembles that of Peng *et al.* [35]. Let (X, Y, Z) denote the continuous world coordinate of a 3D point whose feature we wish to infer. First, we trilinearly interpolate the feature grid to obtain a c -dimensional feature vector at point (X, Y, Z) . Denoting this trilinearly interpolated feature vector as p for input point (X, Y, Z) , we further obtain a

3d location conditioned feature vector using $\phi(p, x)$, where ϕ is a small fully-connected network. Finally we obtain our 32-dimensional embedding vector using $f_\theta(p, \phi(p, x))$, where f_θ represents a multi-block fully connected ResNet. Further details about the network architecture can be found in the supplementary.

A similar fully-connected ResNet parametrization provides a point’s binary occupancy $o_{(X, Y, Z)}$, and its RGB color value $c_{(X, Y, Z)}$. The three ResNets that predict features, occupancy and RGB values of 3D points do not share weights. During training, we use the point cloud of the target viewpoint to query our model for features, occupancies and colors in the corresponding visible 3D point locations and propagate gradients in an end-to-end manner to the same feature voxel 3D map M . We denote the operation of obtaining point features, occupancies and colors by querying the feature map at continuous locations by $\mathcal{F}^f(M, (X, Y, Z))$, $\mathcal{F}^o(M, (X, Y, Z))$, and $\mathcal{F}^c(M, (X, Y, Z))$. At test time, we query the model in both visible and non visible 3D point locations, to obtain an amodal completed 3D point feature cloud.

CoCoNets is made up of differentiable modules that go back and forth between 3D feature space and 2D image space. It can take as input a variable number of RGB-D images both at training time and test time. For simplicity, in our experiments we encode only a single view input at both train and test time. More details on each of the following modules are included in the supplementary file.

2D-to-3D unprojection (Figure 1 (a, b)) This module converts the input RGB image $I \in \mathbb{R}^{w \times h \times 3}$ and depth map $D \in \mathbb{R}^{n \times 3}$ where, n is the size of the pointcloud) into 3D tensors using available camera intrinsics. The RGB is “un-projected” into a 3D tensor $U \in \mathbb{R}^{w \times h \times d \times 3}$ by filling each 3D grid location with the RGB value of its corresponding subpixel. The pointcloud is converted to a 3D occupancy grid $O \in \mathbb{R}^{w \times h \times d \times 1}$, by assigning each voxel a value of 1 or 0, depending on whether or not a point lands in the voxel. We then convert the concatenation of these tensors into a 3D feature tensor $M \in \mathbb{R}^{w \times h \times d \times c}$, via a 3D convolutional encoder-decoder network with skip connections. We L_2 -normalize the feature in each grid cell.

3D-to-2D projection This module first warps the 3D feature map M to align it to the target viewpoint, yielding M^V . It then generates a 3D feature point cloud by querying the model at the 3D point locations visible from the target viewpoint, provided by the target depth map D_{target} . It also generates a 2D feature map by computing for each visible 3D point the corresponding 2D pixel location using the camera intrinsics, $(x, y) = f_Z^X, f_Z^Y$, and copying the 3D point feature into that location in the 2D feature map. Note that each 3D feature point is mapped independently to 2D, in contrast to neural renderers that convolutionally map feature maps to image views [43, 7]. This independent

point-by-point rendering is closer to the spirit of graphics operations [44].

3D contrastive learning of point features Given a pair of input RGB images (I_{inp}, I_{target}), depth maps (D_{inp}, D_{target}), and the camera pose change between input and target views V , we consider two types of representations for the target view:

- a top-down one, M_{inp}^V (Figure 1 (a)) by encoding the input RGB-D image (I_{inp}, D_{inp}) and orienting the feature map to the target viewpoint, and predicting the features for the target 3D points in D_{target} by querying the functions in M_{inp}^V , obtaining the feature cloud $\{(X, Y, Z, \mathcal{F}(M_{inp}^V, (X, Y, Z)))\}$ for $(X, Y, Z) \in D_{target}$ (Figure 1 (c)).
- a bottom-up one, M_{target} (Figure 1 (b)) by simply encoding the target RGB-D image I_{target}, D_{target} and predicting the features for the target 3D points in D_{target} , obtaining the feature cloud $\{(X, Y, Z, \mathcal{F}(M_{target}, (X, Y, Z)))\}$ for $(X, Y, Z) \in D_{target}$ (Figure 1 (c)).

We use a contrastive InfoNCE loss [52] (Figure 1 (d)) to pull corresponding top-down and bottom-up point features close together in embedding space and push non-corresponding ones farther away:

$$\mathcal{L} = -\log \frac{\exp(\frac{q \cdot k_+}{\tau})}{\sum_{i=0}^K \exp \frac{q \cdot k_i}{\tau}}, \quad (1)$$

where τ is a temperature hyper-parameter and the sum in the denominator is over one positive and K negative samples. In the numerator, q represents $\mathcal{F}^f(M_{inp}, (i, j, k))$, and k_+ is its corresponding positive sample $\mathcal{F}^f(M_{target}, (m, n, o))$. The dot product computes a similarity. In practice, we randomly sample corresponding points from the top-down or bottom-up feature clouds at each training iteration. We also maintain a large pool of negative pairs in a dictionary, using the approach proposed by He *et al.* [18].

Our metric learning loss, if applied only on **3D points visible from both input and target views**, coincides with the point contrastive metric learning of Xie *et al.* [54], a state-of-the-art 3D point feature learning method. CoCoNets can handle input and target views that actually have few or no points in common. We compare against Xie *et al.*’s state-of-the-art 3D feature learning model in our experiments, and demonstrate the importance of amodal completion for feature learning.

Occupancy prediction and RGB view regression We train a separate CoCoNet model to predict occupancy and RGB in novel viewpoints, which we denote CoCoNets-*OccRGB*. We use a standard binary cross entropy loss

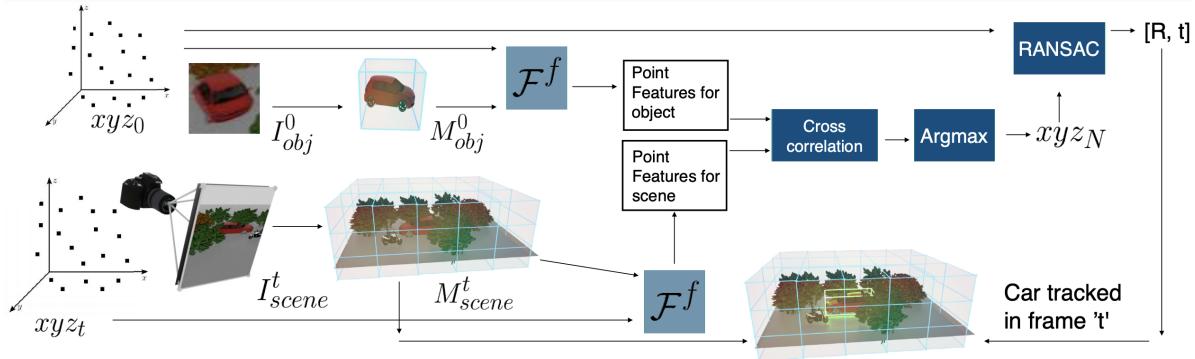


Figure 2: 3D object tracking using CoCoNets. Given the cropped RGB-D image I_{obj}^0 of the object to track at $t = 0$, our model infers the 3D object feature map M_{obj}^0 , and queries it using xyz_0 , the point cloud of the object, to obtain object point features. Similarly, it obtains the point features of the entire scene at timestep t . Finally, it does cross correlation between these features to get xyz_N , where each i^{th} point in xyz_N is the point from the scene whose feature matched best with the feature for i^{th} point in xyz_0 . We then apply RANSAC on xyz_0 and xyz_N to obtain the location of the car at timestep t .

for the occupancy, and a regression loss for RGB. The model predicts point clouds in target viewpoints and infers their occupancy and color from the implicit function grids $\mathcal{F}^o, \mathcal{F}^c$. Specifically, to predict a target RGB view and its occupancy given a source RGB-D image, after encoding and orienting the 3D feature map M to the target viewpoint, we predict the point occupancies of the target view, and predict the point RGB colors for the occupied points. These are then projected to the image space in the target view.

Jointly training RGB and occupancy prediction with contrastive view prediction did not improve discriminability of our feature representations. The novelty of our paper is in using voxel-implicit architectures for amodal contrastive 3D feature learning as opposed to 3D occupancy prediction [36, 35], or in addition to it.

RGB view regression without depth input Additionally, we also make use of NeRF’s volumetric renderer [29], with features extracted from M_{input} as priors, to render novel views using just a single RGB image (no depth information) as input. We call this CoCoNets^{NoDepthRGB} and explain the setup in the supplementary file. Qualitative results from this model can be seen in Figure 5.

4. Experiments

Useful visual feature representations are expected to correspond visual entities across variations in pose and appearance, as well as across intra-category variability. We evaluate the correspondability of our 3D scene representations in 3D object tracking and object re-identification, 3D pose estimation and cross-object correspondence. We further evaluate our features as pre-training for 3D object detector for the state-of-the-art method of [37]. Our experiments aim to answer the following questions:

1. Is amodal completion important for learning scene representations useful for visual correspondence, across time and across scenes? To this end, we compare 3D feature representations learned by CoCoNets against contrastive point clouds [54] that do not consider amodal completion but share a similar contrastive objective for 3D points visible in both scenes.
2. Is our 3D function parameterization important for learning discriminative visual representations? To this end, we compare with 3D contrastive neural mapping of Harley *et al.* [16] that does not consider continuous functions, but rather uses discrete 3D feature maps as the neural bottleneck.
3. Does the performance of 3D object detectors in point-clouds improve when using CoCoNets as initialization of the point features, and by how much? To this end, we compare with VoteNet[37], a model that uses PointNet++ [38] as their feature extractor and deep Hough voting for 3D object detection in point clouds.
4. Is the continuous 3D convolutional bottleneck of CoCoNets useful for view prediction? To this end, we compare image views rendered from our model in novel scenes with images from GQN [7], which does not consider 3D neural bottleneck, and GRNN [50], which has a discrete 3D convolutional bottleneck.

We train and test CoCoNets in the simulated datasets of CARLA [6], ShapeNet [3] and test them further—without training—in the real-world KITTI dataset [9]. CARLA is an open-source photorealistic simulator of urban driving scenes, which permits moving the camera to any desired viewpoint in the scene. ShapeNet is a large scale 3D object repository which again permits camera placement at will.

KITTI is an urban driving scene dataset collected from a camera mounted on a car. More information on datasets is available in the supplementary file.

4.1. Self-supervised 3D object tracking and re-identification

Setup We evaluate the 3D feature representations of CoCoNets in their ability to track a 3D object over time, as well as across large frame gaps, a task known as object re-identification. Given the 3D bounding box of an object in the first frame, we are interested in tracking the object in subsequent frames. We first infer 3D features for object points visible in the first frame. For subsequent timesteps, we predict the amodal *complete* 3D feature cloud of the scene from a single view by querying our model in a set of visible and invisible 3D point locations. We found this uniform random sampling of query point set strategy to suffice for our tasks, though more elaborate strategies that focus on querying points close to object surfaces can be used, such as Marching Cubes [27], which has been used by previous 3D implicit shape function models for learning object occupancy to obtain a watertight mesh [11, 35, 36]. Given the object feature point cloud and the amodal scene feature cloud, we obtain a rigid body transformation (3D rotation and translation) by solving the orthogonal Procrustes problem [41], and use RANSAC to find the transformation that satisfies most inliers as shown in Figure 2, where inlier point matches are obtained by thresholding inner products at 0 between the object 3D point features and 3D point features of the target scene. We find the object 3D box in the target scene by warping the box b_0 using the obtained transformation. Note that we do not restrict our model to search locally only for a matching object, but search over the whole scene which makes the task harder. Such object matching across frame gaps is important for re-detecting an object after occlusions. It is also a good test for the discriminability of the features, since they are not assisted by temporal continuity or local search regions.

Evaluation and comparison with baselines We compare CoCoNets against the following baselines (we use names indicative to the differences and characteristics): i) 3DVoxContrast of Harley *et al.* [15], ii) 3DPointContrast of Xie *et al.* [54] iii) 2.5DDenseObjectNets of Florence *et al.* [8], which uses triangulation supervision to train 2D deep feature maps, then lifts them to 3D point feature clouds using the available depth map.

We follow the experimental setup of Harley *et al.* [15], which uses video sequences of 10 timesteps. For 3DVoxContrast, we first use multiview RGB-D data with known egomotion to train a model using 3D contrastive losses. Upon training, 3DVoxContrast maps a single RGB-D image to a discrete 3D feature map. Then, 3DVoxContrast

uses this model for tracking by solving orthogonal procrustes problem with RANSAC between voxel features as opposed to point features. For 3DPointContrast, upon training the model, we compute RANSAC on the cross-correlation scores between point features from the visible 3D points on the object at frame 0 and the *visible* point cloud at each subsequent frame. Our method also considers amodal completed point cloud at each frame. For 2.5DDenseObjectNets, upon training the model, we apply RANSAC much the same way as in the 3DPointContrast model.

We show quantitative tracking results for our model and baselines on the CARLA and KITTI datasets in Table 1. Figure 3 visualizes 3D object trajectories over time from an overhead view. Note that we do not use any locality search, i.e., each model searches across the whole target frame for the configuration of the object. Our model achieves superior performance on 3D object tracking than all the baselines. It dramatically outperforms 3DVoxContrast, which highlights the importance of high spatial resolution for 3D feature learning and correspondence. It also outperforms 3DPointContrast, which highlights the importance of amodal completion supported by our model. Lastly, all 3D methods outperform 2.5DDenseObjectNets which highlights the effectiveness of learning features in a metric 3D feature space.

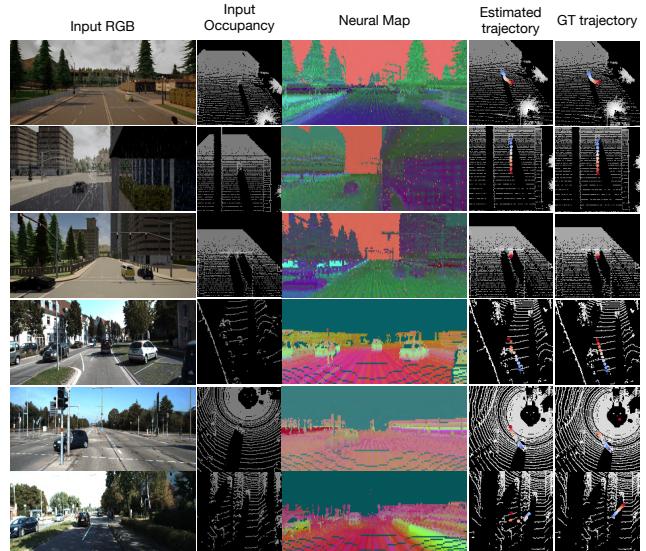


Figure 3: Self-supervised 3D object tracking. In the 1st and 2nd column we visualize the RGB and depth from the first frame, which is given as input to our model, along with a 3D box specifying the object to track. In the 3rd column we visualize our inferred point features by projecting them to the same RGB image and then doing PCA compression. In the last 2 columns we show the estimated and ground truth trajectories. The top three rows show our results on CARLA; the bottom three rows show our KITTI results.

Method	CARLA	KITTI
CoCoNets (ours)	0.61	0.54
3DPointContrast [54]	0.55	0.48
3DVoxContrast [15]	0.37	0.23
2.5DDenseObjectNets [8]	0.24	0.19

Table 1: 3D object tracking mean IOU across timesteps.

Method	0.25	0.30	0.40	0.50
VoteNet [37]	0.32	0.26	0.24	0.20
CoCoNets-VoteNet (Ours)	0.51	0.47	0.41	0.32

Table 2: Comparing VoteNet with CoCoNets-VoteNet on mAP for different IOUs. Pre-training the 3D point cloud features with CoCoNets significantly boosts performance.

4.2. Supervised 3D object detection in point clouds

Setup We test the 3D feature representations obtained by CoCoNets in their ability to boost the performance of a supervised 3D object detector when used as initialization. We use VoteNet [37], a state-of-the-art 3D object detector that combines the classic ideas of Hough voting in a deep neural network architecture that computes point features, uses them to vote for locations and sizes of the 3D object boxes, and aggregates the votes to propose 3D object bounding boxes with associated confidence, in an end-to-end differentiable framework. We modify VoteNet [37] by replacing their point cloud feature extractor backbone with a trained CoCoNet; we will refer to this as CoCoNets-VoteNet. The original backbone in VoteNet uses PointNet++ [38] to compute point features. Both the VoteNet backbone and CoCoNets-VoteNet backbone sample 1024 points from the dense point cloud using farthest point sampling. The subsequent voting and proposal modules of VoteNet then operate on these sampled point features to make the final predictions. In both cases, upon initialization, the 3D point features are trained end-to-end supervised for the 3D object detection task. We show below that our pre-training gives a significant boost to 3D object detection performance.

Evaluation and comparison with baselines We evaluate both models on the CARLA dataset in the task of predicting 3D axis aligned car bounding boxes. We show the mean average precision (mAP) scores for both the models for different IoU thresholds in Table 2. CoCoNets-VoteNet outperforms VoteNet across all IoU thresholds.

4.3. Cross-view and cross-object 3D alignment

Setup We evaluate the correspondability of our 3D feature representations by testing how well they can align dif-

ferent views of the same instance, and different instances that belong to the same object category. Given the amodal 3D feature point clouds extracted for two objects in random viewpoints, we use orthogonal Procrustes problem with RANSAC to estimate the 3D rigid body transformation that aligns them, similar to tracking a car in Section 4.1. Estimating fine grained correspondences between objects, once their rigid alignment has been estimated, is obtained using iterative closest point method [1], and we show such results in the supplementary file.

Evaluation and comparison with baselines We investigate two different setups: i) Given two viewpoints of the same object, our goal is to estimate the relative 3D rotation between them. ii) Given two objects of the same category in random viewpoints, our goal is to estimate a rigid 3d alignment between them. We compare our model against baselines 3DVoxContrast and 3DPointContrast, for which we use RANSAC on voxel and point features, respectively, to estimate a 3D rigid transformation for the two objects.

Method	cross-object	cross-view
3DVoxContrast [15]	0.18	0.21
3DPointContrast [54]	0.09	0.14
CoCoNets (Ours)	0.18	0.58

Table 3: Cross-object and cross-view 3D alignment accuracies in Shapenet dataset (mean over 4 classes: Aeroplane, Mug, Car, Chair).

The inferred 3D alignment is considered correct if each of the yaw, roll, and pitch angles are within 10° of their respective ground truth values. In Table 3 we show the cross-view (same-object) and cross-object 3D alignment accuracy for our model and the baselines. Our model performs the same as 3DVoxContrast on cross-object alignment, and outperforms both models on cross-view alignment. Note that cross-object alignment is much harder than cross-view for all three models.

4.4. Generalization in view prediction

We evaluate CoCoNets-OccRGB in its ability to predict plausible images in scenes with novel number of objects, novel object appearance, and novel arrangements. We compare against state-of-the-art view prediction models. We further evaluate its ability for 3D occupancy prediction and completion. Extensive results of occupancy and RGB predictions on the ShapeNet dataset for our model can be found in Sectionn 8.

We compare CoCoNets in RGB prediction against generative query network (GQN) [7] and geometry-aware recurrent networks (GRNN) [50]. We adapted GQN code to

use posed RGB-D images as input. GQN uses a 2D latent feature space to encode the RGB-D input image, and GRNN uses a discrete latent 3D feature map of the scene. Our model uses a grid of functions and does not suffer from resolution limitations of the 3D feature map.

CoCoNets predicts RGB images for target viewpoints as described in Sec. 3.1. Figure 4 shows the view generation results for CoCoNets, GQN, and GRNN. For qualitative results on RGB prediction and 3D occupancy prediction, details on the datasets used, architectural details and video results of our self-supervised tracking, please see our supplementary file.

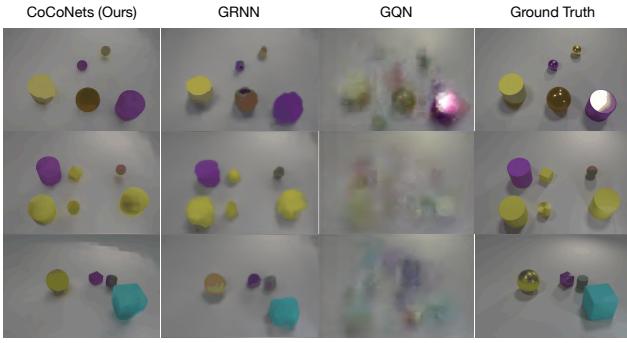


Figure 4: **Novel view prediction.** We compare CoCoNets, GQN [7], and GRNN [50] on the CLEVR [20] dataset.



Figure 5: **RGB view regression** on the ShapeNet dataset using CoCoNets^{NoDepthRGB}. See the supplementary file for details on the experimental setup.

Limitations/Extensions CoCoNets can be extended to operate without the availability of depth maps at either train or test time using a differential rendering module [44], as shown in Figure 5. Our supplementary material further contains results from variations of our model that do not assume depth available at test time, yet use it at training time.

5. Conclusion

We present a method for learning 3D visual feature representations by self-supervised view and depth prediction from posed RGB and RGB-D images. Our networks lift input 2.5D images of objects and scenes into latent three-dimensional function grids, which can be decoded to infinite-resolution 3D occupancy and 3D feature predictions of the object or scene. Our networks are trained by predicting views using a contrastive mutual information maximization objective. We evaluate the emergent 3D visual feature representations in 3D object tracking in dynamic scenes and in cross-view and cross-object alignment estimation. We empirically demonstrate that the features are semantically meaningful and outperform popular point-based supervision [54] which does not consider 3D completion, and discrete voxel 3D latent feature maps of previous works [16, 15] which are limited by the spatial resolution of the 3D feature grid. Moreover, our models can better generalize to novel scenes with unseen number and appearance of objects than networks that do not encode 3D structure [7] or do not incorporate 3D convolutional modules [44]. They make a step towards self-supervised learning of amodal 3D feature representations, which we show are useful for 3D object tracking and correspondence. Avenues for future work include learning such representations directly from dynamic videos, and relaxing the egomotion and depth supervision.

6. Acknowledgement

This work has been funded by Sony AI, DARPA Machine Common Sense, a NSF CAREER award, the Air Force Office of Scientific Research under award number FA9550-20-1-0423. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

References

- [1] Leopoldo Armesto, Javier Minguez, and Luis Montesano. A generalization of the metric-based iterative closest point technique for 3d scan matching. In *2010 IEEE International Conference on Robotics and Automation*, pages 1367–1372. IEEE, 2010. 7
- [2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. 2
- [3] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. *arXiv:1512.03012*, 2015. 5, 11
- [4] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 1, 2
- [5] Julian Chibane, Thiendo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion, 2020. 1, 2
- [6] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CORL*, pages 1–16, 2017. 5, 11
- [7] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 1, 2, 3, 4, 5, 7, 8
- [8] Peter R. Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 373–385. PMLR, 29–31 Oct 2018. 2, 3, 6, 7
- [9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 5
- [10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012. 11
- [11] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape, 2020. 1, 6
- [12] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. *CoRR*, abs/1906.02739, 2019. 2
- [13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *NIPS*, pages 2672–2680, 2014. 2
- [14] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding, 2019. 1
- [15] Adam W Harley, Shrinidhi K Lakshminarayanan, Paul Schmid, and Katerina Fragkiadaki. Tracking emerges by looking around static scenes, with neural 3d mapping. *arXiv preprint arXiv:2008.01295*, 2020. 1, 2, 6, 7, 8
- [16] Adam W Harley, Fangyu Li, Shrinidhi K Lakshminarayanan, Xian Zhou, Hsiao-Yu Fish Tung, and Katerina Fragkiadaki. Learning from unlabelled videos using contrastive predictive neural 3d mapping. In *ICLR*, 2020. 1, 2, 3, 5, 8
- [17] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 11
- [18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 4
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 12
- [20] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 8, 11
- [21] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *NIPS*, 2017. 2
- [22] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. *CoRR*, abs/1711.07566, 2017. 2
- [23] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018. 2
- [24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013. 2
- [25] Hoang-An Le, Thomas Mensink, Partha Das, and Theo Gevers. Novel view synthesis from single images via point cloud transformation, 2020. 1
- [26] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction, 2017. 2
- [27] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, Aug. 1987. 6
- [28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 1, 2, 13
- [29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 1, 2, 5, 12
- [30] Bence Nanay. The importance of amodal completion in everyday perception. *i-Perception*, 9(4):2041669518788887, 2018. PMID: 30109014. 3
- [31] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised

- learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2
- [32] David Novotný, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. *CoRR*, abs/1705.03951, 2017. 2
- [33] David Novotny, Roman Shapovalov, and Andrea Vedaldi. Canonical 3d deformers maps: Unifying parametric and non-parametric methods for dense weakly-supervised category reconstruction, 2020. 1
- [34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 1, 2
- [35] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks, 2020. 1, 2, 3, 5, 6, 12
- [36] Stefan Popov, Pablo Bauszat, and Vittorio Ferrari. Corenet: Coherent 3d scene reconstruction from a single rgb image, 2020. 1, 2, 5, 6
- [37] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 5, 7
- [38] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 5, 7
- [39] Gernot Riegler and Vladlen Koltun. Free view synthesis, 2020. 1, 2
- [40] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigi, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv:1905.05172*, 2019. 2
- [41] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 6
- [42] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *arXiv preprint arXiv:2007.02442*, 2020. 2
- [43] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning persistent 3D feature embeddings. In *CVPR*, 2019. 1, 2, 4
- [44] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations, 2020. 1, 2, 4, 8
- [45] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Single-view to multi-view: Reconstructing unseen views with a convolutional network. In *ECCV*, 2016. 2
- [46] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, Rohit Pandey, Sean Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B Goldman, and Michael Zollhöfer. State of the art on neural rendering, 2020. 2
- [47] Josh Tobin, OpenAI Robotics, and Pieter Abbeel. Geometry-aware neural rendering, 2019. 3
- [48] Shubham Tulsiani, Saurabh Gupta, David F. Fouhey, Alexei A. Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. *CoRR*, abs/1712.01812, 2017. 1
- [49] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. *CoRR*, abs/1704.06254, 2017. 1, 2
- [50] Hsiao-Yu Fish Tung, Ricson Cheng, and Katerina Fragkiadaki. Learning spatial common sense with geometry-aware recurrent networks. In *CVPR*, 2019. 2, 3, 5, 7, 8
- [51] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NIPS*, pages 4790–4798, 2016. 2
- [52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 1, 4
- [53] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1
- [54] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding, 2020. 2, 4, 5, 6, 7, 8
- [55] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision, 2017. 2
- [56] Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. *arXiv preprint arXiv:2002.12687*, 2020. 13
- [57] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2020. 12

Supplementary File

The structure of this supplementary file is as follows: Section 7 covers the details on dataset preparation. Section 8 elucidates the implementation details, including inputs to our model, architecture details, and hyperparameter values used. Finally, Section 9 provides a more detailed analysis of the results.

7. Dataset Preparation

CARLA dataset. We use the CARLA simulator [6] to generate multi-view data for training, and tracking data for testing. CARLA is an open-source simulator for urban driving scenes, which allows multiple cameras to be placed at arbitrary positions. We generate episodes 300 frames long, with a framerate of 10 FPS. Each episode is captured from 6 random camera locations, sampled from a set of 18 cameras that lie on a hemisphere. We treat Towns 0-3 as the training set, and Towns 4-5 as the test set. In total, we use 388 scenes for self-supervised multi-view training and 188 scenes for tracking evaluation. The resolution of our RGB and depth images is 128x384.

KITTI dataset. We use the KITTI benchmark [10] to evaluate our model on the tracking task. We use the “tracking” subset of KITTI to test our model. This subset has 8008 frames across twenty labelled sequences. We create 8-frame sub-sequences of this data. We make sure that all the eight frames have a valid label for the target object. The egomotion information in the “tracking” data is only approximate.

Shapenet dataset. We use the meshes from ShapeNet-Core [3]. We first normalize the ShapeNet meshes. We then render RGB-D data for them using Blender, where we place 24 cameras around the object. The cameras are placed at a distance of 2m, with azimuth ranging from 0° - 360° at intervals of 45° , and elevation ranging from 0° - 80° at intervals of 20° .

CLEVR dataset. We build upon the CLEVR Blender simulator [20]. Our scenes consist of cubes, spheres, and cylinders. We create scenes as follows: each object model is randomly rotated (0° to 360° along vertical axis), translated (randomly within a sphere of radius 10.5 units) and scaled (0.25 to 1.25 times the actual size). Each scene contains 2 – 10 objects. We randomly vary the lighting of each scene. We render each scene by placing 28 RGB-D cameras at elevations ranging from 26° to 80° with 13° increments and azimuths ranging from 0° to 360° with 45° increments. Each camera is placed within a sphere of radius 1.5 metres from the center of the scene.

8. Implementation details

Code, training details and computation complexity. Our model is implemented in Python/Pytorch. We use a batch size of 2 and our learning rate is set at 10^{-4} . We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. Our model takes 24 hours (approximately 100k iterations) of training for convergence on a single V100 GPU. As far as the running time of the proposed method for tasks in the experiments is concerned, extraction of point features from an RGB-D image takes 0.3 seconds for our model on a V100 GPU. A single iteration of Object tracking (Section 4.1) takes 5 seconds, Object Detection (Section 4.2) takes 0.4 seconds, Object Alignment (Section 4.3) takes 6 seconds and View Prediction using CoCoNets-*OccRGB* (Section 4.4) takes 0.6 seconds.

Inputs. We randomly select images from 2 views which includes the RGB-D and relative ground-truth egomotion (query view and target view) as input for training our model, while we use a single view for testing.

2.5D-to-3D lifting. Our 2.5D-to-3D unprojection module takes as input RGB-D images and converts it into a 4D tensor $\mathbf{U} \in \mathbb{R}^{w \times h \times d \times 4}$, where w, h, d is 64, 64, 64. We use perspective (un)projection to fill the 3D grid with samples from 2D image. Specifically, using pinhole camera model [17], we find the floating-point 2D pixel location that every cell in the 3D grid, indexed by the coordinate (i, j, k) , projects onto from the current camera viewpoint. This is given by $[u, v]^T = \mathbf{KS}[i, j, k]^T$, where \mathbf{S} , the similarity transform, converts memory coordinates to camera coordinates and \mathbf{K} , the camera intrinsics, convert camera coordinates to pixel coordinates. Bilinear interpolation is applied on pixel values to fill the grid cells. We obtain a binary occupancy grid $\mathbf{O} \in \mathbb{R}^{w \times h \times d \times 1}$ from the depth image \mathbf{D} in a similar way. This occupancy is then concatenated with the unprojected RGB to get a tensor $[\mathbf{U}, \mathbf{O}] \in \mathbb{R}^{w \times h \times d \times 4}$. This tensor is then passed through a 3D encoder-decoder network, the architecture of which is as follows: 4-2-64, 4-2-128, 4-2-256, 4-0.5-128, 4-0.5-64, 1-1- F . Here, we use the notation $k\text{-}s\text{-}c$ for kernel-stride-channels, and F is the feature dimension, which we set to $F = 32$. We concatenate the output of transposed convolutions in decoder with same resolution feature map output from the encoder. The concatenated tensor is then passed to the next layer in the decoder. We use leaky ReLU activation and batch normalization after every convolution layer, except for the last one in each network. We obtain our 3D feature map \mathbf{M} as the output of this process. Even though we do not use multiple views as input in the paper, our system can handle that easily by orienting the feature map \mathbf{M}_v obtained from each view v to a reference view to cancel the egomotion, and then

fusing them to obtain the final feature map \mathbf{M} .

Implicit function parameterization. To predict the features, RGB, and occupancy values for different 3D physical points within a voxel grid, we use implicit functions parameterized by neural networks. As mentioned in the main paper, we denote the operation of obtaining point features, occupancies and colors by querying the feature map \mathbf{M} at continuous locations (X, Y, Z) by $\mathcal{F}^f(\mathbf{M}, (X, Y, Z))$, $\mathcal{F}^o(\mathbf{M}, (X, Y, Z))$, $\mathcal{F}^c(\mathbf{M}, (X, Y, Z))$ respectively. We parameterize these three functions using neural networks with identical architectures, with the exception that the output of \mathcal{F}^f , \mathcal{F}^o , \mathcal{F}^c is 32-D, 1-D and 3D, respectively. There is no weight sharing between these three neural networks. For brevity, we will use \mathcal{F} when explaining properties common to all the three networks/functions.

The architecture of \mathcal{F} is similar to that of [35]. Given the point (X, Y, Z) that we want to featurize, we first infer the trilinearly-interpolated feature vector at point (X, Y, Z) from the 8 neighbouring voxel features, which we denote as c . We encode the coordinate (X, Y, Z) into a 32-D feature vector using a linear layer. We denote this vector as p . The inputs c and p are then processed as follows:

$$\begin{aligned} \text{out} = & RN_i(RN_{i-1}(\cdots RN_1(p + FC_1(c)) \\ & \cdots) + FC_{i-1}(c)) + FC_i(c)). \end{aligned} \quad (2)$$

We set $i = 5$ for our usecase. FC_i is a linear layer which gives a 32-dimensional output. RN_i is a 2 layer ResNet block [19]. The architecture of ResNet block is as follows: ReLU, 32-32, ReLU, 32-32. Here, we use the notation $i_u - o_u$ to represent a linear layer, where i_u is the input dimension, and o_u is the output dimension of that layer. Each ResNet block finally generates the output by adding the input to the output of the above layers. out is then passed through a ReLU activation function followed by a linear layer, the output dimension of which depends on the property of the point we are trying to predict.

CoCoNets on RGB images during inference for 3D alignment. Our model can operate without depth maps at inference time. We achieve this by replacing our bottom-up 3D convolutional encoder-decoder architecture with a ResNet-18 [19] which directly operates on RGB image from the target view I_{target} . Our ResNet-18 encodes I_{target} to a 1D feature vector \mathbf{x}_{target} . During training, our top-down mapping of (I_{inp}, D_{inp}) to \mathbf{M}_{inp}^V and then to feature cloud $\{(X, Y, Z, \mathcal{F}(\mathbf{M}_{inp}^V, (X, Y, Z)))\}$ remains the same, we instead replace the bottom-up feature point clouds obtained from \mathbf{M}_{target} with feature point cloud

$\{(X, Y, Z, \mathcal{F}(\mathbf{x}_{target}, (X, Y, Z)))\}$ obtained by interpolating the 1D feature vector \mathbf{x}_{target} . We then train these feature clouds using the same contrastive loss discussed in Section 3. Our bottom-up network gives us an option to operate without depth images during inference time. In Section 9.5, we show the comparison on 3D alignment of our model operating on RGB versus RGB-D images.

Volumetric rendering using CoCoNets on RGB images. We leverage our model’s continuous representations on the task of rendering novel views on the ShapeNet dataset without depth as input and call it *CoCoNets^{NoDepthRGB}*. In this experiment, since we do not use depth information, we only pass the RGB image from the input view, I_{inp} , as an unprojected 3D tensor $U \in R^{w \times h \times d \times 3}$, to the top-down 3D-CNN encoder-decoder (Figure 1(a)) to get M_{inp} . Then we take the camera pose change between the target (the view for which we want to render the image) and the input view, defined as V , and sample $w \times h$ number of rays, sampling 32 points on each ray, as per the method given in NeRF [29]. We use these sampled points, defined as D_{target}^V , to query M_{inp} using trilinear interpolation. Thereafter, we use the MLP architecture given in [29] and pass as input the positional encoding of the query point, its viewing direction and the corresponding interpolated feature vector, getting (r, g, b, σ) as the output for each query point. Extracted (r, g, b, σ) of all query points on each ray are then passed into the volume rendering module of [29] to get the final RGB image, \hat{I}_{target} . We use MSE as the rendering loss between the ground truth RGB, I_{target} and the RGB image rendered by NeRF, \hat{I}_{target} . Figure 5 shows the qualitative results for this experiment. A concurrent work to this experiment is pixelNeRF [57], with the difference being that we operate on a 3D feature tensor, whereas [57] operates on a 2D feature tensor.

9. Results

In this section, we show more results on occupancy prediction, RGB view prediction, dense correspondence, vehicle tracking and pose alignment when operating on RGB input.

9.1. Occupancy prediction

Figure 6 shows the occupancies predicted by our model on ShapeNet objects. The first three columns show the renderings of the mesh created from predicted occupancies from three views. The last three columns show the corresponding ground truth. We also show multiview renderings of the predicted occupancies and the ground truth meshes as GIFs in the supplementary video on our project page and urge the readers to refer that. At inference time, we extract meshes by applying Multiresolution IsoSurface Extraction

(MISE) [28]. Table 4 shows the IOU for occupancy prediction for our model on CARLA and ShapeNet datasets. For ShapeNet, we report the mean IOU over all the classes.

Dataset	IOU
CARLA	0.79
ShapeNet	0.56

Table 4: **Occupancy prediction evaluation.** Metric used is IOU.

9.2. RGB view prediction

Figure 7 shows qualitative results on the RGB novel view prediction task by our model on the ShapeNet dataset using depth as input. On the other hand, Figure 5 shows the views rendered without using depth via the NeRF volumetric renderer. Given an input view (first column), our model can render the scene from an arbitrary viewpoint (third column). We compare this rendering with the ground truth RGB for that view (second column).

9.3. Self-supervised 3D object tracking

Figure 8 shows more qualitative tracking results on CARLA and KITTI dataset. We have further attached videos of estimated and ground truth trajectories on our project page, and we urge the readers to have a look at those videos for a better understanding of the dataset complexity and our model’s capabilities.

9.4. Dense correspondence

In this section, we evaluate our features on the task of establishing correspondence between two instances, I_1 (the Source entity) and I_2 (the Target entity), belonging to the same category. We use the keypoints provided by You *et al.* [56] on the ShapeNet dataset for qualitative evaluation. We assume that the input objects are already aligned. Given the dense pointcloud for both the objects, we featurize the points using the approach proposed in Section 3. Then, for a point (x, y, z) with feature $f_{x,y,z}$ in I_1 , we find the corresponding point (x', y', z') in I_2 whose feature $f'_{x',y',z'}$ has the highest cosine similarity with $f_{x,y,z}$. Since the instances are already aligned, we limit this search to a local neighborhood of radius 10cm. Figure 9 shows the cross-object correspondence results achieved using the features learned by our model.

9.5. Using only RGB for pose estimation

In this section, we show the comparison for cross-scene and cross-object 3D alignment, while operating our model on RGB versus RGB-D input. For this we follow the same experimental setup as of Section 4.3. We refer our model

operating on RGB as CoCoNets-RGB. We show the comparison with CoCoNets in Table 5.

Method	cross-object	cross-view
CoCoNets	0.18	0.58
CoCoNets-RGB	0.16	0.39

Table 5: Cross-object and cross-view 3D alignment accuracies in ShapeNet dataset (mean over 4 classes: Aeroplane, Mug, Car, Chair).

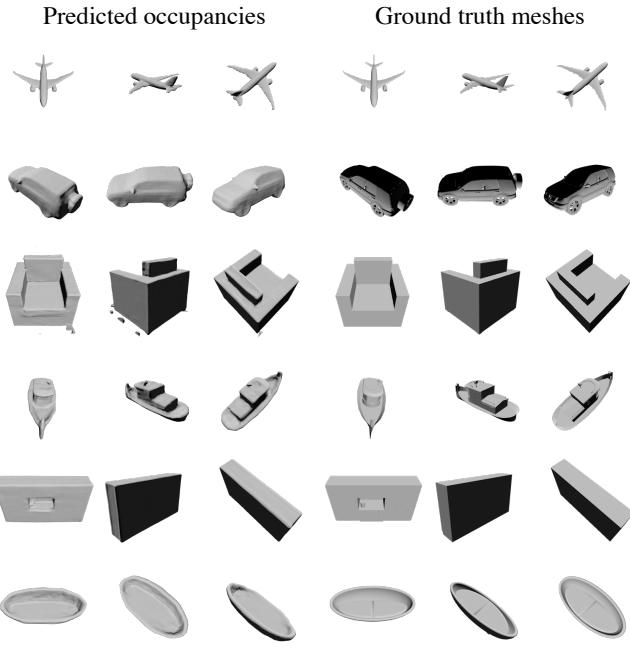


Figure 6: **Occupancies predicted by our model for ShapeNet objects.** First three columns show the renderings of the mesh created from predicted occupancies from three views. Last three columns show the same for ground truth mesh.

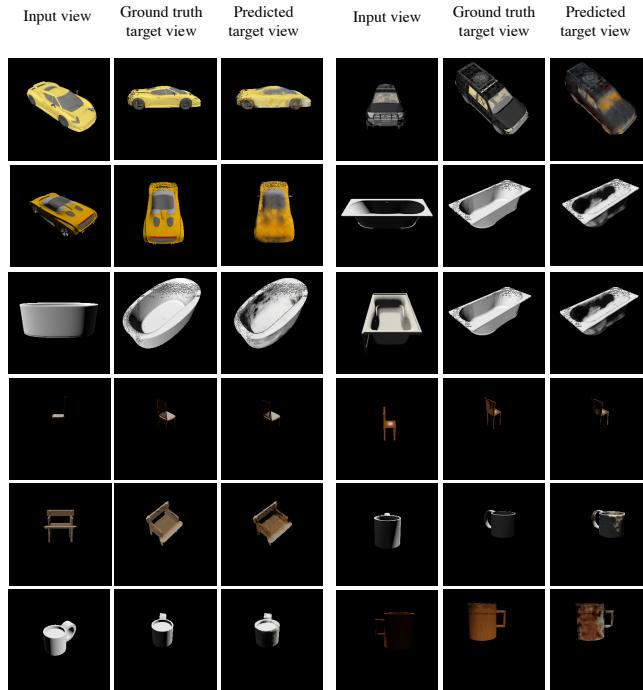


Figure 7: **Neural renders on ShapeNet dataset from CoCoNets-*OccRGB*.** The first column shows the view given as input to our model. The second column shows the ground truth target view. The third column shows the predicted target view. Columns 4-6 follow the same convention.

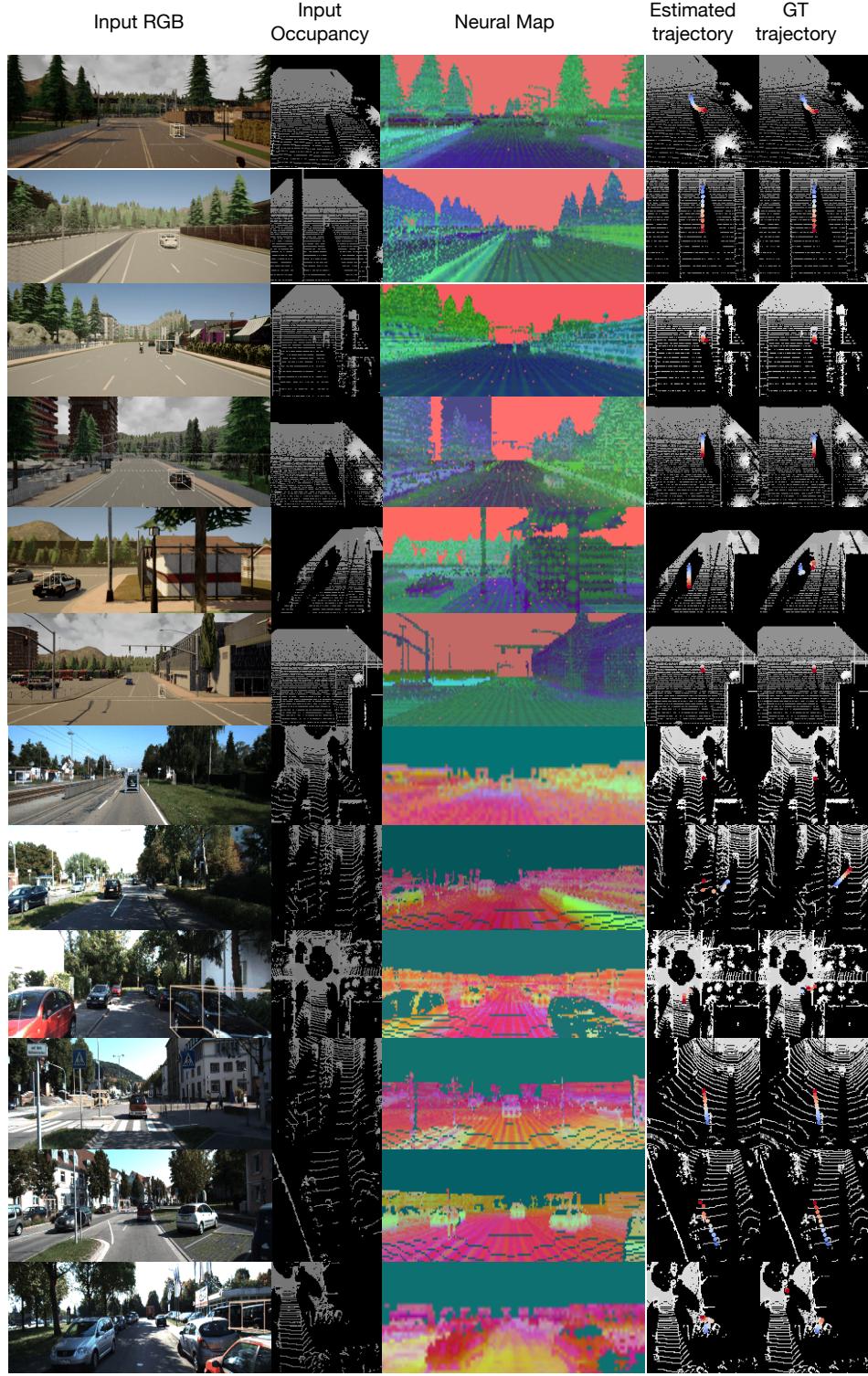


Figure 8: Self-supervised 3D object tracking. In the 1st and 2nd column, we visualize the RGB, the object to track and depth from the first time frame, which is given as input to our model. In the 3rd column, we visualize our inferred point features by projecting them to the same RGB image and then doing PCA compression. In the last two columns, we show the estimated and ground truth trajectories. The top six rows show our results on CARLA; the bottom six rows show our KITTI results.

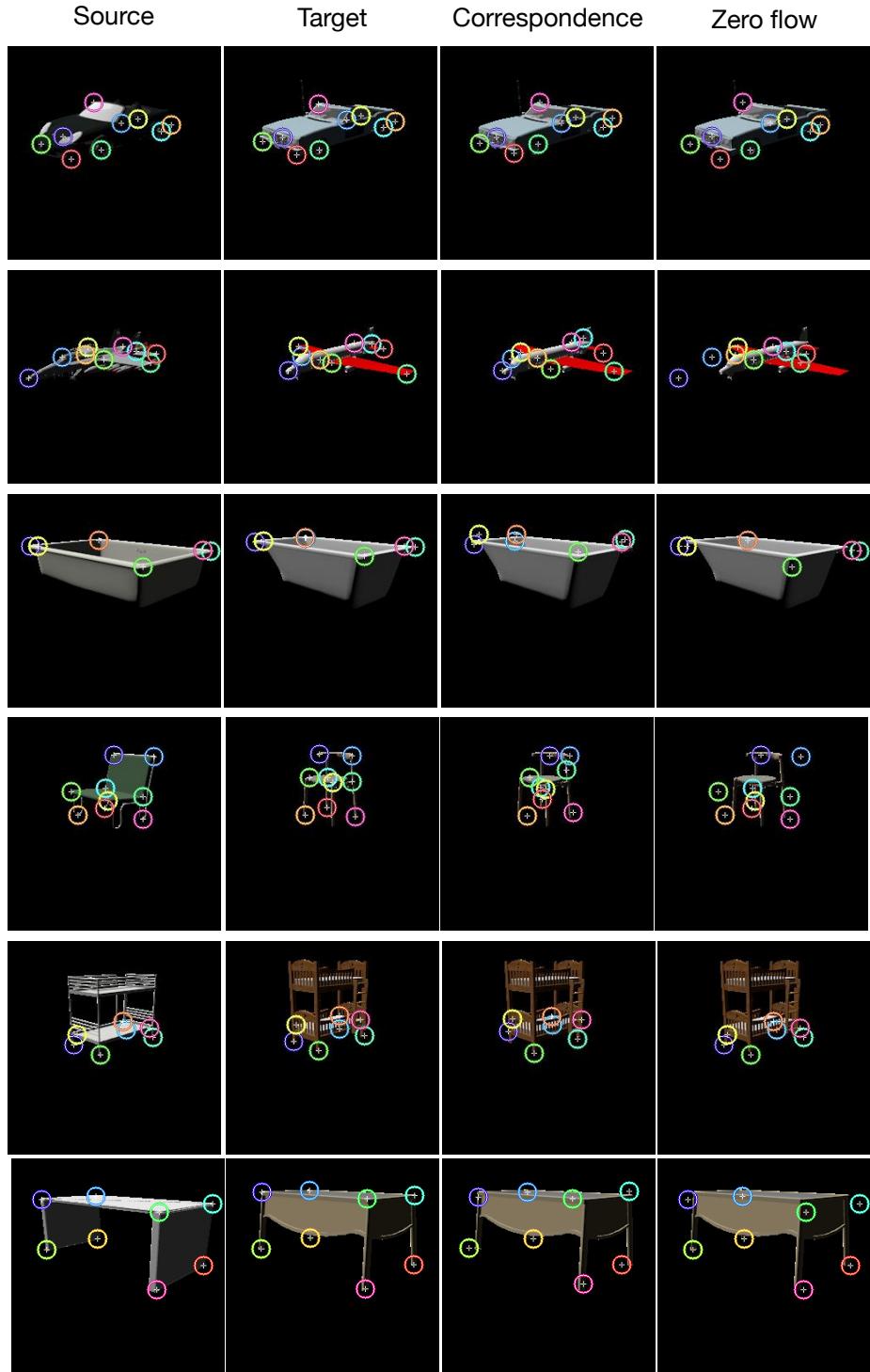


Figure 9: **Cross-object correspondence.** First column shows the Source entity. Second column shows the Target entity. The keypoints shown for these two columns are the ground truth keypoints. Third column shows the keypoints in Target entity inferred using correspondence from Source to Target. Last column shows the keypoint locations obtained assuming zero-flow correspondence from Source to Target, i.e. a point at location (x, y, z) in Source gets mapped to the same location in Target.