Scaling CNNs for High Resolution Volumetric Reconstruction from a Single Image

Adrian Johnston Ravi Garg Gustavo Carneiro Ian Reid
Anton van den Hengel

Australian Centre for Visual Technologies
The University of Adelaide, Australia
{adrian.johnston, ravi.garg, gustavo.carneiro, ian.reid, anton.vandenhengel}@adelaide.edu.au

Abstract

One of the long-standing tasks in computer vision is to use a single 2-D view of an object in order to produce its 3-D shape. Recovering the lost dimension in this process has been the goal of classic shape-from-X methods, but often the assumptions made in those works are quite limiting to be useful for general 3-D objects. This problem has been recently addressed with deep learning methods containing a 2-D (convolution) encoder followed by a 3-D (deconvolution) decoder. These methods have been reasonably successful, but memory and run time constraints impose a strong limitation in terms of the resolution of the reconstructed 3-D shapes. In particular, state-of-the-art methods are able to reconstruct 3-D shapes represented by volumes of at most 32^3 voxels using state-of-the-art desktop computers. In this work, we present a scalable 2-D single view to 3-D volume reconstruction deep learning method, where the 3-D (deconvolution) decoder is replaced by a simple inverse discrete cosine transform (IDCT) decoder. Our simpler architecture has an order of magnitude faster inference when reconstructing 3-D volumes compared to the convolution-deconvolutional model, an exponentially smaller memory complexity while training and testing, and a sub-linear runtime training complexity with respect to the output volume size. We show on benchmark datasets that our method can produce high-resolution reconstructions with state of the art accuracy.

1. Introduction

Volumetric reconstruction of objects from images has been one of the most studied problems in computer vision [12]. Multi-view reconstruction approaches based on shape by space carving [17] and level-set reconstruction [37] have led to reasonable quality 3-D reconstructions.

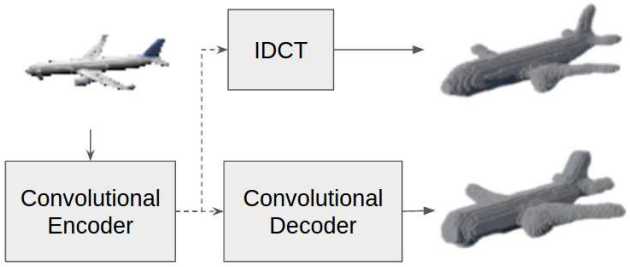

Figure 1. We propose a new convolution-deconvolution deep learning model, where the traditional 3-D deconvolutional decoder (bottom) is replaced by an efficient IDCT decoder (top) for high resolution volumetric reconstructions.

Systems like KinectFusion [25] and DynamicFusion [24] have opened the possibilities for various applications in the field of augmented and virtual reality by providing high quality reconstruction with the help of cheap sensors like Kinect. These multi-view and Kinect based systems work in constrained environments and disregard scene semantics. It has been long believed that a successful estimation of the semantic class, 3-D structure and pose of the objects in the scene can be immensely helpful for holistic visual understanding of images [22]. In fact, this estimation would allow intelligent systems to be more effective at interacting with the scene, but one important requirement, particularly regarding the 3-D structure of objects, is to obtain the highest possible 3-D representation resolution at the smallest computational cost – this is precisely the aim of this paper.

Recent success of convolutional neural networks (CNNs) [16, 19] has led to many approaches tackling the challenging problem of volumetric reconstruction from a single image to move towards full 3-D scene understanding [3, 31, 38, 39, 40, 42] However, most of these methods reconstructs object at very low resolution ranging from 20^3 to 32^3 voxels – thereby limiting the practical applica-

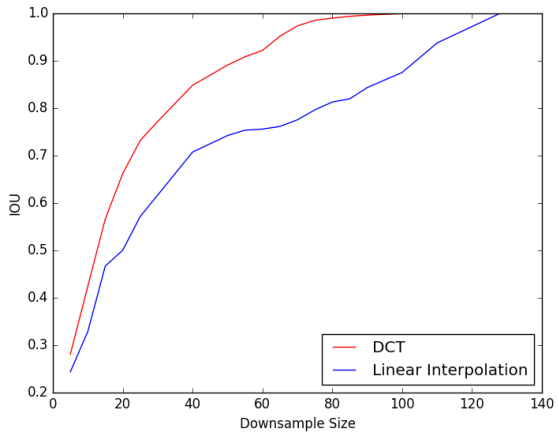

Figure 2. Comparison of low frequency 3-D DCT compression accuracy to simple interpolation at various compression rates on a subset of ShapeNet volumes [2]. 128^3 volumes are compressed using (i) nearest neighbour interpolation (blue curve) or (ii) by truncating the high frequency of DCT basis (red curve) and upscaled with respective inverse operations to compute mean IOU.

bility. Almost all these deep networks, designed for single view volumetric reconstructions, rely on a convolution-deconvolution architecture, as shown in Fig. 1. In this setup, a traditional 2-D convolution network (often used in classifiers) encodes a large patch of the image into an abstract feature (i.e., an embedded low-dimensional representation), which is then converted into a volume by successive deconvolution operations. These convolution-deconvolution architectures are based on the success of deconvolution networks for semantic segmentation [21, 26] that shows that the loss of resolution due to strided convolutions/pooling operations can be recovered by learning deconvolution filters. These convolution - deconvolution architectures give reasonably accurate reconstructions at low resolution (typically 32^3 voxels or less) from a single image, but do not scale well to high resolution volumetric reconstructions. The main reason behind this issue lies in the successive deconvolution to upscale a coarse reconstruction, which requires intermediate volumetric representations to be learned in succession in a coarse to fine manner, where each deconvolution layer upscales the predictions by a factor of two. Although deconvolution layers have very few parameters, the memory and the time required to process volumes (both for training and inference) in this coarse-to-fine fashion via deconvolution grows rapidly and is intractable. Table 1 (see baseline-32 and baseline-128 results) reports how the 3-D resolution affects traditional convolution-deconvolution architectures in terms of memory required for training as well as training and inference running time.

In this work, we explore a simple option in the design of a novel deep learning model that can reconstruct high-resolution 3-D volumes from a single 2-D single view. In particular, our main goal is to have a model that scales well with an increase in resolution of the 3-D volume reconstruction with respect to memory, training time and inference time. One straightforward approach is to learn a linear model (e.g., principal component analysis [13]) or a non-linear model (e.g., Gaussian Process latent variable model [18]) to represent the shapes of the objects and use it in place of the deconvolution network. However, this will make (i) the reconstruction methods sensitive to the 3-D volumetric data used for training, which is not available in abundance and (ii) would not be easily adaptable to semi-supervised methods [27], which does not require 2-D image-volumetric model pairs for training. An alternative solution is the use of the low frequency coefficients computed from the discrete cosine transform (DCT) or Fourier basis, which are in general good linear bases to represent smooth signals. In fact, the DCT basis has already been shown to be a robust volume representation [28], as evidenced in Fig. 2, which shows that for a representative set of volumetric object shapes taken from ShapeNet [2], the low-frequency DCT basis is much more information preserving then that of the commonly used local interpolation methods in CNNs for up-sampling low resolution predictions. It is important to note that while being generic, the DCT basis is almost as information preserving as a linear PCA basis when the variability in the dataset increases.

Therefore, we propose a model that extends the convolution-deconvolution network by replacing the computationally expensive deconvolution network by a simple inverse DCT (IDCT) linear transform, as shown in Fig. 1, where this IDCT transform reconstructs the low-frequency signal at the desired resolution. Our proposed extension has profound impact in terms of the computational cost involved in training and inference. In particular, we show through extensive experiments on benchmark datasets that our proposed framework:

- presents an inference time that is one order of magnitude faster than equivalent convolution-deconvolution networks,
- shows a slightly more accurate 3-D object shape prediction than equivalent convolution-deconvolution networks;
- scales gracefully with increase in resolution of the output 3-D volume in terms of training memory requirements, training time, and inference time,
- allows a 3-D volume recovery at a much larger resolution compared to previously proposed approaches in the field.

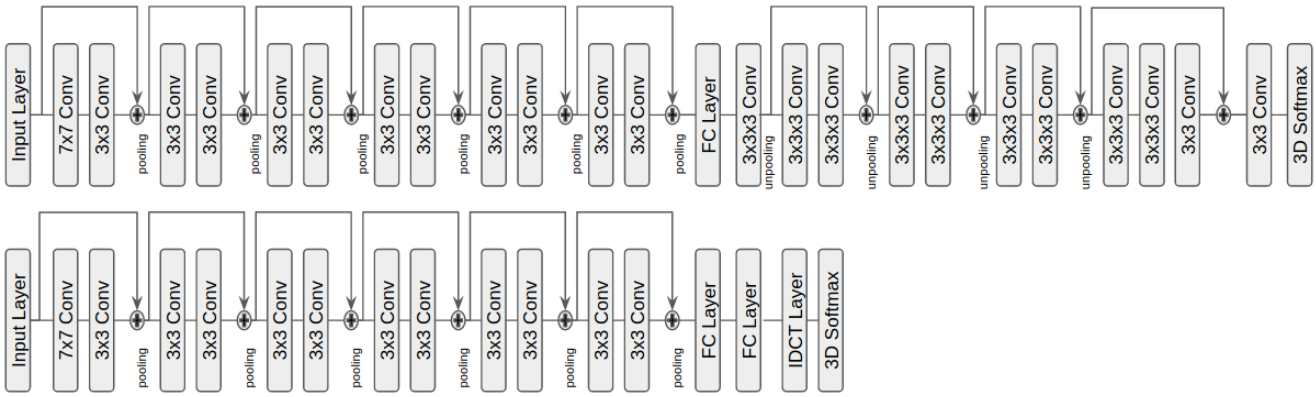

Figure 3. Network Architectures: Top: Baseline Network mimicking 3-D R2N2[3] without RNN/3-D GRU. Bottom: Our Network utilizing the IDCT Layer.

2. Related Work

The problem of reconstructing the 3-D shape of an object from a single image has recently received renewed attention from the field with the use of traditional computer vision methods [35] (e.g., structure-from-motion, optimisation of the visual hull representation, etc.). However, with the advent of deep learning techniques [16] and new datasets containing 3-D model annotations of images containing particular visual objects, the field has moved towards the application of these deep learning models to the task of 3-D reconstruction from images [2, 40]. In particular, the seminal paper by Wu *et al*. [40] is the first to propose a deep learning methodology that reconstructs 3-D volumes from depth maps, which has led to several extensions [3, 23].

The more recently proposed methods replaced depth maps by the RGB image, with the same goal of recovering the 3-D shape of the object from a single or multiple views of it. For instance, Girdhar *et al*. [8] used a 3-stage training process to perform 3-D reconstruction from single images: 1) train a 2-D classifier with mixed synthetic and real images; 2) train a 3-D auto-encoder for learning a representation of their 3-D volumes; and 3) merge the two by minimizing the Euclidean distance between the 2-D and 3-D codes. In parallel, Choy *et al*. [3] developed a recurrent neural network model which aims to use multiple views of a single object to perform 2-D to 3-D reconstruction (the reasoning behind the use of multiple views was to enable the encoding of more information about the object). The use of a projective transformer network that can align the visual object and its projected image allows the unsupervised modelling of 3-D shape reconstruction approaches from single images, as shown by Yan *et al*. [42]. Adversarial training methods for deep learning models [9] have also influenced the development of 3-D shape reconstruction approaches from single images. Wu *et al*. [39] applied a variational encoder and an adversarial decoder for the task of 3-D shape reconstruction from single images. Rezende *et al*. [31] introduced an unsupervised learning framework for recovering 3-D shapes from 2-D projections, with results on the the recovery of only simple 3-D primitives using reinforcement learning. These methods above are based on a relatively similar underlying convolution-deconvolution network, so they have the same limitations discussed in Sec. 1.

State-of-the-art deep learning semantic segmentation models are also based on a similar convolution-deconvolution architectures [7, 21, 26], so it is useful to understand the functionality of such approaches and assess their applicability for the problem of recovering the 3-D shape of the object from a single view. In particular, these approaches show that fully trainable convolution-deconvolution architectures [26], the exploration of a Laplacian reconstruction pyramid to merge predictions from multiple scales [7], and the use of skip connections [21] can produce state-of-the-art semantic segmentation results. However, it is unclear how to extend these ideas in a computationally efficient manner for the the case of volumetric predictions from images, given the explosion of the number of parameters required to generate volumes at high resolutions.

The high memory, training and inference complexities in processing volumes by an encoder (i.e., the convolutional part of the architecture) has also been addressed in the field [20, 32]. Li *et al*. [20] proposed to replace convolutional layers by field probing layers, which is a type of filter that can efficiently extract features from 3-D volumes. However, this method is focused on discriminative features and is not invertible, so it would not be suitable for 3-D reconstruction. Similarly, a memory and run-time efficient processing of 3-D input data has been proposed by Riegler *et al*. [32] with a method focused on the classification and segmentation of volumes and point clouds. That work relies on the use of specialized convolution, pooling and unpooling layers based on the Octree data structure, and shows

excellent results on scaling up 3-D classification and point cloud segmentation. Nevertheless, in order to be applicable for the problem of 3-D reconstruction from 2-D views, this approach would need to be extended to be able to receive 2-D data as input (instead of 3-D) and output a 3-D representation.

There have been many examples of methods that explore 3-D shape representations, consisting of a relatively small set of principal component analysis (PCA) [1] or DCT [4] components that can be further reduced with Gaussian Process Latent Variable Models (GPLVM) [18]. These methods are successful at several tasks, ranging from object shape reconstruction [1, 4], image segmentation and tracking [29], etc. Finally, Zheng *et al*. [43] show that the use of such low-dimensional pre-learned representations are useful for the task of object detection from a single depth image.

3. Network Architecture

Our main contribution is in exchanging the decoder with a simple IDCT layer which is compatible with any 2-D encoder architecture. To show the impact of the proposed frequency based representation, we extensively analyze the performance of our IDCT decoder against a deconvolution baseline. We adapt the state-of-the-art convolutional - deconvolution network for volumetric reconstruction called 3-D-R2N2, proposed by Choy *et al*. [3]. The 3-D-R2N2 model [3] iteratively refines reconstructed volumes by using a recurrent module to fuse the 2-D information coming from multiple views, which is then passed to the deconvolution decoder to generate volumetric reconstructions. To restrict the experiments for single-view training and testing, we remove the recurrent module from 3-D-R2N2 and replace it with a single fully connected layer. The result is a simpler convolutional-deconvolutional baseline network, shown in Figure 3, as a direct replacement of 3-D-R2N2, for single view reconstruction. In the encoder, we use standard max pooling layers for down sampling, while leaky rectified units are used for the activations with residual connections [11].

Our proposed IDCT decoder uses the same baseline encoder defined above to predict the low frequency DCT coefficients, which our decoder converts to solid volumes. The DCT/IDCT function can be efficiently implemented by utilizing the symmetry and separability properties of the nD-DCT function[30]. That is to say that we can pre-compute the 1D-DCT matrix and apply it independently across each axis of the volume. The Discrete Cosine Function has several variants (e.g DCT-I through DCT-VIII)[30]. In this work we will refer to DCT-II as the DCT function and DCT-III as the IDCT function. The DCT-III function is the inverse of the DCT-II function, furthermore, when the DCT matrix is orthogonal the DCT-III/IDCT is the transpose of the DCT-II matrix[30]. The orthogonal 1D DCT-II is given by:

$$X_k = \left(\frac{2}{N}\right)^{\frac{1}{2}} \sum_{i=0}^{N-1} \Lambda(i) cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)\right] x_i \quad (1)$$

where x_i is the input signal at a given index i, X_k is the output coefficient at index k and Λ is the scaling constant applied to x_0 used to make the transform orthogonal, as defined by

$$\Lambda(i) \begin{cases} \frac{1}{\sqrt{2}} & \text{if } i = 0 \\ 1 & \text{otherwise} \end{cases}. \quad (2)$$

In this work, we use the transpose of the DCT-II matrix as our IDCT matrix, however it could also be implemented directly using the DCT-III equation [30].

As our baseline is modeled after 3-D-R2N2, we keep the same loss function defined by the sum of voxel Cross-Entropy[3]:

$$L = \sum_{i,j,k} \{y_{(i,j,k)} \log(p_{(i,j,k)}) + (1 - y_{(i,j,k)} \log(1 - p_{(i,j,k)})\} \quad (3)$$

where $p_{(i,j,k)}$ represents the predicted object occupancy probabilities, $y_{(i,j,k)} \in \{0, 1\}$ denotes the given label for voxel (i, j, k)

We use the voxel intersection over union metric [3] to evaluate the quality of our 3-D reconstructions, defined by:

$$\text{IoU} = \frac{\sum_{i,j,k}[I(p_{(i,j,k)} > t) I(y_{(i,j,k)})]}{\sum_{i,j,k}[I(p_{(i,j,k)} > t) + I(y_{(i,j,k)})]}, \quad (4)$$

where t is the voxelization threshold and $I(.)$ is the indicator function.

4. Experiments

To clearly demonstrate the usefulness of our IDCT decoder based volumetric reconstruction method, in this section we first compare the runtime and memory requirement of both deconvolutional and IDCT architectures at two different resolutions of 32^3 and 128^3. To estimate 128^3 volumetric reconstructions with deconvolutional network we simply add two extra deconvolution blocks to the deconvolution baseline of Fig. 3. An appropriate IDCT basis function is replaced to generate 128^3 volumes from 20^3 coefficients for the proposed method. Table 1 shows the training time[1], inference time and the peak GPU memory required to train the baseline and the proposed IDCT based network to reconstruct volumes at both resolutions from 127×127 images[2].

[1]Both training and test times are estimated after the data is loaded to the GPUs

[2]Nvidia Titan X (Maxwell), with Intel i7 4970k was used for these experiments.

Method	Resolution	Batch Size	Forward Time (Hz)	Train time (Hz)	Memory (GB)
DCT-32 - 20^3 coeff	32^3	24	294(4x)	80.75(6.3x)	1.7
Baseline-32	32^3	24	66.83(1x)	12.63(1x)	4.5
DCT-128 - 20^3 coeff	128^3	24	30.48(0.45x)	22.99 (1.8x)	2.2
Baseline-128	128^3	2	2.82 (0.04x)	0.19 (0.015x)	10.4

Table 1. Performance indicators using deconvolution and IDCT networks at different resolutions.

Due to the large reduction in the depth of the our IDCT decoder, our proposed network is approximately four times faster for inference and over six times faster during training, when compared with our baseline model at a smaller resolution of 32^3 with batch size of 24. Furthermore the memory requirements during training are drastically reduced as the intermediate coarser volumes are not predicted by our decoder. When the resolution is increased by a factor of four (in each of the three dimensions), to be 128^3, it becomes evident that the traditional 3-D deconvolution networks become intractable. Already approximately seven times slower and three times more memory hungry deconvolution networks now can only be trained with a batch size of 2 on a 12 GB GPU card. Per-image training goes up by a factor of over 50 compared to 32^3 resolution deconvolution baseline and the test time performance degrades equally drastically making this baseline unusable. Conversely, a single layer IDCT decoder is only three times slower to train when the resolution is increased by a factor of four (in each of the three dimensions) – however it still remains faster to train when compared to the deconvolutional network reconstructing volumes at 32^3 resolution. The memory required for training this IDCT decoder only grows by the size needed to store the high resolution predictions. Training the network for high resolution volumes becomes feasible with a much higher batch size while the number of parameters required remains constant.

To validate the 3-D reconstruction accuracy with the proposed IDCT decoder, we compare the single view reconstruction accuracies on both synthetic (ShapeNet[2]) and real (PascalVOC 3-D+[41]) datasets. We show that using our single IDCT layer as decoder does not degrade the quality of low-resolution predictions but enables substantially faster training and gives better high resolution reconstructions.

4.1. Experiments on Synthetic Dataset

Following Choy *et. al.*[3], we use synthetically rendered images of resolution 127×127 provided by the authors containing a 13 class subset of the original ShapeNet [2]. This subset (ShapeNet13) consists of approximately 50,000 2-D-3-D pairs, with a split of 4/5 for training and 1/5 for testing, exactly as defined in [3]. For all experiments on ShapeNet dataset, we use Theano [33] and Lasagne [6] libraries for our implementations. In addition, the training procedure uses mini-batches of size 24 and learning rate of 10^{-5} with Adam [14] optimizer.

We compare the mean IoU error (Table 2) of the baseline deconvolution architecture against the proposed IDCT decoder architecture in Table 2. As our baseline can be seen as a simpler version of [3] with one view training, for completeness, we report results for the entire test-set for our baseline deconvolutional network alongside that of [3]. As expected, our baseline using only single-view to predict volumes against five views used in [3] gives marginally lower reconstruction accuracies than that of [3]. However it is important to note that our IDCT decoder could also be integrated with the RNN as proposed in [3]. For simplicity, we limit our experiments to the one-view training and testing paradigm.

When compared at 32^3 resolution, our approach with IDCT decoder gives marginally better volumetric reconstructions (with 20^3 DCT coefficients) compared to the baseline. However, it is trained in a day and half whereas the baseline takes more than a week to train. A significant boost in accuracy can be seen at 128^3 reconstructions when we fine-tune our network with high resolution ground truth. As shown in Figure 4, the reconstructions produced by the baseline approach after upscaling with linear interpolation overestimates the foreground objects, leading to less accurate and blocky reconstructions. On the other hand, our proposed method is able to preserve a significant amount of shape details.

Method	Resolution	Mean IoU
R2N2 (5V train, 5V test) [3]	32^3	0.634
R2N2 (5V train, 1V test) [3]	32^3	0.6096
Baseline (1V train, 1V test)	32^3	0.5701
DCT - 20^3 coeff	32^3	0.5791
Baseline Upscaled	128^3	0.3988
DCT - 20^3 coeff	128^3	0.4174

Table 2. Volumetric shape prediction IoU errors on ShapeNet 3-D.

4.2. Experiment with Real Images

Most of the CNNs based volumetric reconstruction approach [3, 8, 39] use an intermediate step of training the network with a semi-synthetic dataset by augmenting the syn-

Figure 4. Examples of 3-D reconstructions from single view images using the Synthetic ShapeNet13 dataset [3, 2]. First Row: Input Image, Second Row: Ground truth shape, Third Row:32^3 Volumetric prediction using deconvolutional decoder upscaled to 128^3, Bottom Row: Volumetric predictions at 128^3 using the proposed IDCT decoder.

	Resolution	**aero**	**bike**	**boat**	**bus**	**car**	**chair**	**mbike**	**sofa**	**train**	**tv**	**mean**
DCT - 20^3 Coeff	32^3	**0.5552**	**0.4893**	**0.5231**	**0.7756**	0.6221	**0.2497**	**0.6561**	0.4624	**0.5739**	**0.5492**	**0.5474**
Deconvolution Baseline	32^3	0.5492	0.4516	0.5011	0.7593	**0.6345**	0.244	0.6437	**0.546**	0.5675	0.5161	0.5419
DCT - 20^3 Coeff	128^3	**0.4502**	**0.2606**	**0.4067**	**0.6942**	**0.561**	**0.1836**	**0.5509**	0.4311	**0.4273**	**0.5105**	**0.4496**
Baseline upscaled	128^3	0.2824	0.1263	0.336	0.6167	0.5126	0.181	0.4377	**0.4654**	0.3287	0.4095	0.3671

Table 3. Per category and mean volumetric shape prediction IoU errors on PASCAL VOC 3-D+ at 32^3 and 128^3 resolutions.

thetically rendered object instances with real backgrounds. We choose to directly fine-tune both the deconvolutional and IDCT decoder based networks on real images from PASCAL VOC 3-D+ dataset (specifically we use v1.1 with ImageNet[5] augmentation) [41]. We prune the object instances that are classified as either difficult or truncated, leaving approximately 11400 image instances, which we will use as our training samples. The same pruning strategy is applied to the testing set. Object instances were cropped from the real images to the regions corresponding to 20% dilated bounding boxes for training. Padding with white background was used along the shortest image axis to maintain the aspect ratio when resizing the cropped objects to the input resolution for our network (127x127). Only horizontal flips of images were used for data augmentation while fine tuning.

Our setup of directly fine-tuning the synthetic shapenet model onto PASCAL VOC 3-D+ can be considered to be more challenging compared to other methods due to lack of training data and amount of background clutter and occlusion. These issues make the training more difficult. Following [21], the pre-trained models evaluated in Section 4.1 were fine-tuned with a batch size of 1, using stochastic gradient descent (SGD) with higher Nesterov momentum of 0.99 and learning rate of 10^{-5}. Furthermore, in order to reduce over-fitting, we also added dropout to all models as well as weight decay of 10^{-4}.

The IoU errors are compared in Table 3 at both 32^3 and 128^3 resolutions. As observed in the synthetic dataset, results for 32^3 resolution with both deconvolution and IDCT decoder methods are similar. Despite the truncation of predictions to 20^3 coefficients, we observe that with the exception of car and sofa, IDCT decoder based reconstruction outperforms the deconvolutional network by narrow margin. More drastic performance gains are observed when high resolution volumes are used for training our IDCT decoder with mean IoU increasing by $\sim 22\%$.

Figure 5 shows the visual comparison of the results for our proposed IDCT decoder based network and the deconvolution baseline. We observe that due to the challenging background clutter, occlusion and significant truncation of the training and test instances, both the IDCT and deconvolutional decoder networks are thrown off (see Figure 6 for failures). However, for most of the successful reconstruction scenarios, the IDCT decoder based reconstruction were more accurate while preserving details in the object structures evident from images. For example, 3D deconvolutional reconstruction fails to pick up the back of the car and depth of the computer monitor evident in the image to reconstruct the pick-up car or flat-screen whereas proposed

Figure 5. Examples of volumetric reconstructions on instances of PASCAL VOC 3-D+ dataset. From left to right: Input image, ground truth volume at 32^3, ground truth volume at 128^3 resolutions, IDCT decoder based reconstruction at 32^3, IDCT decoder based reconstruction at 128^3 and the baseline 32^3 reconstruction with deconvolutional decoder upscaled to 128^3 respectively.

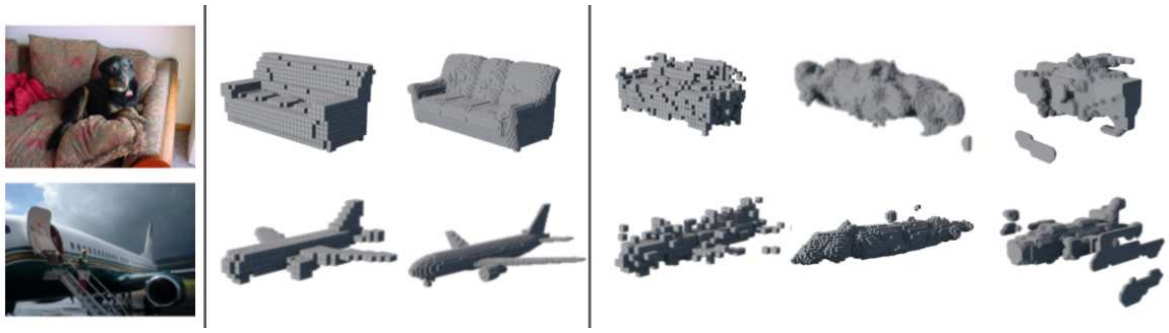

Figure 6. Failure Cases: Truncated and cluttered background throwing off the volumetric reconstructions. From left to right: Input image, ground truth volume at 32^3, ground truth volume at 128^3 resolutions, IDCT decoder based reconstruction at 32^3, IDCT decoder based reconstruction at 128^3 and the baseline 32^3 reconstruction with deconvolutional decoder upscaled to 128^3 respectively.

method correctly reconstruct the objects. Also note in Figure 5 that the 128^3-voxel reconstructions from real images with IDCT often contains much richer details, even though our network was still restricted to estimate 20^3 low frequency DCT coefficients like reconstruction of aeroplane, train, motorbike.

As discussed in Tulsiani *et al*. [34], it is important to note that the PASCAL VOC 3D+ dataset was not originally intended for the purposed of evaluating supervised volumetric reconstruction. The dataset contains a limited number of ground truth CAD models/volumes that are shared in both the training and the test sets. This means that instead of learning to interpolate in the manifold of possible 3D shapes from ShapeNet, neural network with reconstruction loss might over-fit to retrieve the nearest volumetric shape in the training set for every image. An evidence of this can be seen in 128^3 reconstruction of the chair in Figure 5 where the style of chair-back is hallucinated or in the reconstruction of sofa which is reconstructed to be a two-seater without evidence in the image. However, in the absence of a better alternative to test on real data and for fair comparison with existing volumetric reconstruction methods, we still use PASCAL VOC 3D+ dataset for evaluation. The aforementioned over-fitting problem can be avoided to some extent by fine tuning on real data in a weakly supervised manner instead of using direct volume supervision with limited CAD models. A perspective projection layer with segmentation loss of projected volumes is used for this purpose in [42, 10, 44, 34]. These weakly supervised modules can be easily deployed with our IDCT decoder to facilitate faster training for high resolution volumetric reconstructions.

Finally, thin structures like bike wheels, chair legs are found missing at times in our 128^3-voxel reconstructions, which potentially can be recovered using fully connected CRFs [15] or object connectivity priors [36].

5. Conclusion

In this paper we have presented a method for reconstructing high resolution 3-D volumes from single view 2-D images, using a decoder based on the inverse Discrete Cosine Transform. Our proposed method is shown to be an order of magnitude faster and require less memory than standard deconvolutional decoders and to be scalable in terms of memory and runtime complexities as a function of the output volume resolution. We also show that it is possible to compress the dimensionality of the prediction with generic DCT basis without losing important details. We observe that a simple dimensionality reduction with a generic basis not only allows for faster inference, but it makes training more stable. For future work, we will study the feasibility of processing both the input images and output volumes in the frequency domain. As most of the training and inference times as well as the memory required for high resolution reconstruction contributes to our loss layer, it will be fruitful to explore robust reconstruction loss in the frequency domain for further speedup.

6. Acknowledgements

This research was in part supported by the Data to Decisions Cooperative Research Centre (A.J). We are also grateful to the Australian Research Council which has supported this research through the Australian Research Council Centre of Excellence for Robotic Vision CE140100016 (G.C, I.R) and through a Laureate Fellowship FL130100102 (R.G, I.R). We gratefully thank the NVIDIA corporation for the donation of the Titan X GPU used in this research.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 408–416. ACM, 2005.

[2] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[3] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[4] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. D. Reid. Dense reconstruction using 3d object shape priors. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 1288–1295, 2013.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[6] S. Dieleman, J. Schlter, C. Raffel, E. Olson, S. K. Snderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, J. D. Fauw, M. Heilman, D. M. de Almeida, B. McFee, H. Weideman, G. Takcs, P. de Rivaz, J. Crall, G. Sanders, K. Rasul, C. Liu, G. French, and J. Degrave. Lasagne: First release., Aug. 2015.

[7] G. Ghiasi and C. C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*, pages 519–534. Springer, 2016.

[8] R. Girdhar, D. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] J. Gwak, C. B. Choy, A. Garg, M. Chandraker, and S. Savarese. Weakly supervised generative adversarial networks for 3d reconstruction. *arXiv preprint arXiv:1705.10904*, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[12] B. K. Horn and M. J. Brooks. *Shape from shading*. MIT press, 1989.

[13] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*, 2(3):4, 2011.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.

[18] N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Nips*, volume 2, page 5, 2003.

[19] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[20] Y. Li, S. Pirk, H. Su, C. R. Qi, and L. J. Guibas. Fpnn: Field probing neural networks for 3d data. In *Advances in Neural Information Processing Systems*, pages 307–315, 2016.

[21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[22] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences*, 200(1140):269–294, 1978.

[23] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 922–928. IEEE, 2015.

[24] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015.

[25] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011.

[26] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[27] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1742–1750, 2015.

[28] V. A. Prisacariu, A. V. Segal, and I. Reid. Simultaneous monocular 2d segmentation, 3d pose recovery and 3d reconstruction. In *Asian Conference on Computer Vision*, pages 593–606. Springer, 2012.

[29] V. A. Prisacariu, A. V. Segal, and I. Reid. Simultaneous monocular 2d segmentation, 3d pose recovery and 3d reconstruction. In *Computer Vision ACCV 2012*, volume 7724 of *Lecture Notes in Computer Science*, pages 593–606. 2013.

[30] K. R. Rao and P. Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[31] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In *Advances In Neural Information Processing Systems*, pages 4997–5005, 2016.

[32] G. Riegler, A. O. Ulusoy, and A. Geiger. Octnet: Learning deep 3d representations at high resolutions. *CoRR*, abs/1611.05009, 2016.

[33] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[34] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Computer Vision and Pattern Regognition (CVPR)*, 2017.

[35] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48, 2014.

[36] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE conference on*, pages 1–8. IEEE, 2008.

[37] R. T. Whitaker. A level-set approach to 3d reconstruction from range data. *International journal of computer vision*, 29(3):203–231, 1998.

[38] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision*, pages 365–382. Springer, 2016.

[39] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling.

[40] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015.

[41] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[42] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1696–1704. Curran Associates, Inc., 2016.

[43] S. Zheng, V. A. Prisacariu, M. Averkiou, M.-M. Cheng, N. J. Mitra, J. Shotton, P. H. Torr, and C. Rother. Object proposals estimation in depth image using compact 3d shape manifolds. In *German Conference on Pattern Recognition*, pages 196–208. Springer, 2015.

[44] R. Zhu, H. Kiani, C. Wang, and S. Lucey. Rethinking reprojection: Closing the loop for pose-aware shapereconstruction from a single image. *arXiv preprint arXiv:1707.04682*, 2017.