Supplementary materials:
Representation Similarity Analysis for Efficient Task Taxonomy & Transfer Learning

Kshitij Dwivedi Gemma Roig
Singapore University of Technology and Design
kshitij_dwivedi@mymail.sutd.edu.sg, gemma_roig@sutd.edu.sg

Here we report the additional details and results which we left in the main text to the supplementary material. In the first section, we provide details about the small models used and report the results and comparison with the Taskonomy pretrained models. In the second section, we compare the task similarity matrix and clustering using our RSA approach with that of Taskonomy[1] approach. In the third section, we report the consistency of RSA based similarity ranking and transfer learning performance for all the tasks.

S1. Small models for task taxonomy

We select the tasks (a total of 14 tasks) which can be optimized using only L1/L2/triple-metric loss and the output of the task is spatial such that all the tasks can have the same decoder except the final layer. The architecture of the small model is reported in Table S1.

Layer	Kernel size	# Channels	Stride
Encoder			
Conv1	3×3	16	2
Conv2	3×3	32	2
Conv3	3×3	64	2
Conv4	3×3	64	2
Conv5	3×3	8	1
Decoder			
Conv6	3×3	32	1
$Upscale \times 2$			
Conv7	3×3	16	1
$Upscale \times 2$			
Conv8	3×3	4	1
$Upscale \times 2$			
Conv9	3×3	4	1
$Upscale \times 2$			
Conv10	3×3	n	1

Table S1. Small model architecture.The number of channel in Conv10 n was task-specific

We show the task similarity comparison results (Figure S1) of all the selected tasks. We note that for most of the 2D tasks the correlation (Pearson's ρ) of similarity rankings between small vs. Taskonomy models is very high (>0.97 except segment2d) and visually look similar. Although the correlation for all the 3D tasks is still high (>0.77), correlation values are relatively lower than 2D tasks.

We also evaluated the predicted output of 3D tasks and 2D tasks visually. We observed that for the tasks where the predicted output looks more similar to the target, the correlation is higher (Figure S2). The difference in correlation could also be attributed to different training setting of Taskonomy and small models as it was not possible to exactly replicate the Taskonomy training with small models because the training code is not publicly available, and the small models are trained using only a subset of the whole dataset. We computed the task similarity matrix for the selected tasks using both small models and Taskonomy models. Although the similarity ranking using small models on 3D task did not show as high correlation with the Taskonomy models, we found that the Pearson's correlation between them is high (0.8510). On visual inspection of both similarity matrices (Figure S3), 2D tasks of small models show similar scores as with Taskonomy models. The 3D tasks although show higher similarity with corresponding 3D tasks rather than 2D tasks but similarity scores within 3D tasks are lower and therefore matrix looks lighter as compared to the similarity matrix with Taskonomy models.

S2. Taskonomy[1] vs RSA(Our approach)

We show the clustering obtained using Taskonomy approach and compare it our approach in Figure S4. From the figure, we observe that almost all of the 20 single image task we select for our paper (except room layout and denoise) belong in the same cluster as using Taskonomy approach. It is also possible that the difference in clustering arises due to different clustering method, which was not specified, used in [1].

One other advantage of our approach over Taskonomy is

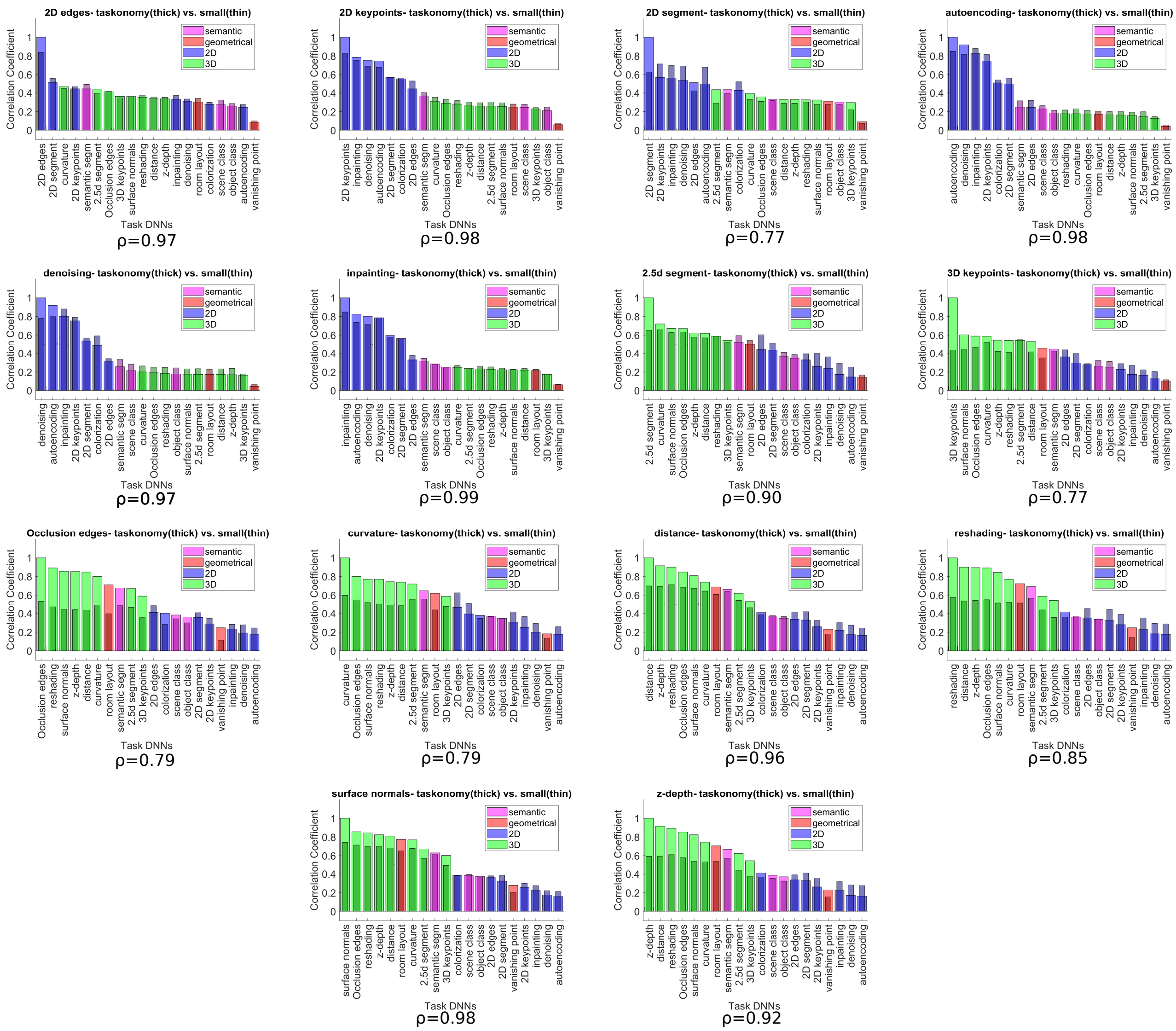

Figure S1. **Similarity ranking with taskonomy model vs small models for 14 tasks.** The ρ value below each plot specifies the Pearson's correlation coefficient between the two similarity rankings.

that our similarity scores lie between -1 and 1 and thus similarity matrix is easy to visualize and evaluate. In Taskonomy approach, an exponential scaling of the similarity score has to be performed to bring them in a good range for visualization. Figure S5 shows both the similarity matrix without any scaling.

S3. Transfer learning in Pascal VOC

In the first three subsections below, we show the consistency of RSA with varying number of iterations, the model size, and the number of images selected for RDM computation. In the last subsection, we report the transfer learning performance of all the task DNNs used for initialization.

S3.1. Consistency with training stage

We show in Figure S6 that even at $1/10$ of the final training stage the Pearson's correlation with the final stage is 0.88 and after $1/2$ of the training the correlation with the final stage stays above 0.99. This shows that one can also use models from an early stage of training for task similarity using RSA.

S3.2. Consistency with model size

We show in Figure S7 the comparison of task similarity obtained using a small encoder (thin bars) vs. task similarity obtained using taskonomy encoder architecture (thick bars). A high correlation (ρ = 0.95, r_s = 0.96) suggests that we

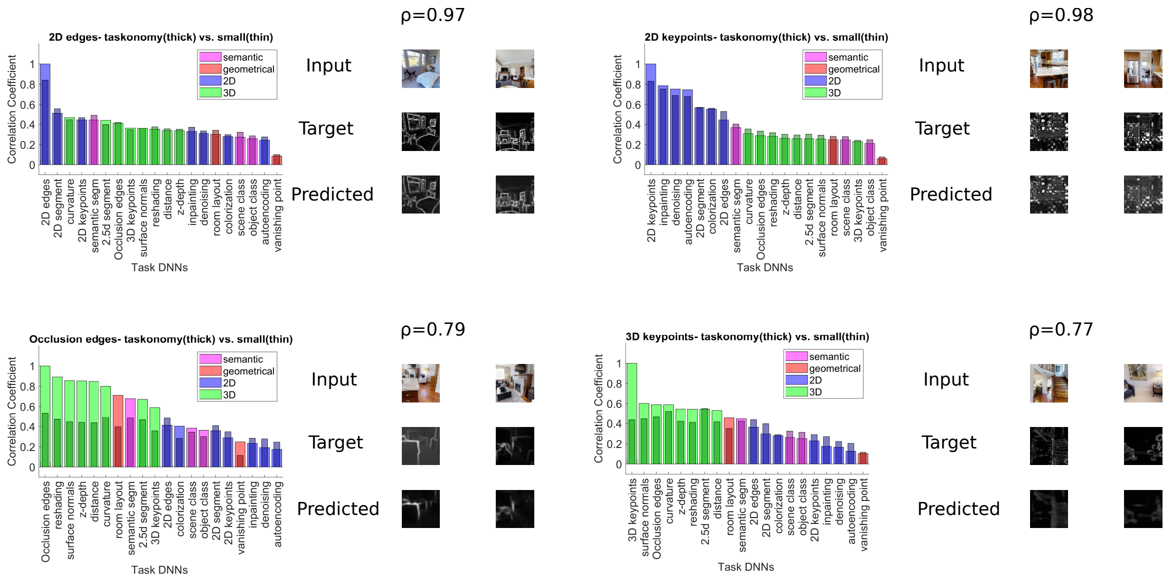

Figure S2. **Is correlation related to visual similarity of the predicted output with the target?**

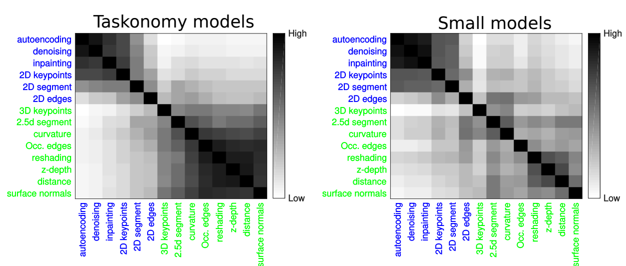

Figure S3. **Task similarity matrix using Taskonomy models vs small models.**

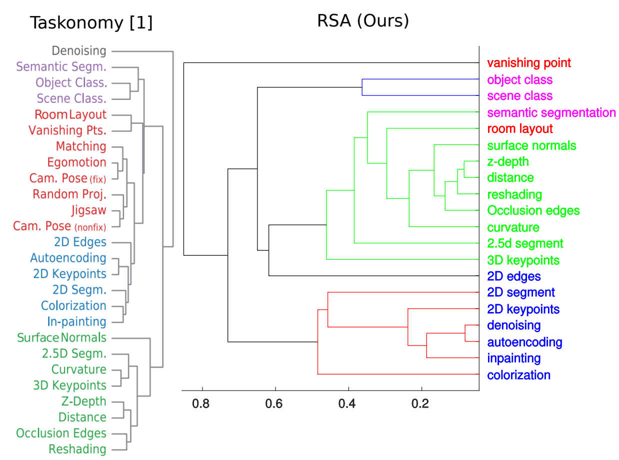

Figure S4. **Clustering: Taskonomy vs RSA (Ours)** Image source: Figure 13 from [1]

can use small models to train on a new task and use RSA to select a good model for initialization.

S3.3. Consistency with the number of images

We varied the number of images from 100 to 2000 and plot the Pearson's correlation of task similarity ranking obtained using n images with the task similarity ranking obtained using 2000 images (Figure S8). After 400 images the Pearson's correlation with the task similarity ranking is always above 0.99, thus suggesting that around 500 images are sufficient for RDM computation.

S3.4. Transfer learning performance for all the tasks

Figure S9 shows the transfer learning performance (mIoU) for 17 single image tasks [1] in the descending order of similarity rankings. The curve shows that the performance in most of the tasks seems to decrease as the similarity score decreases (although it is not a perfect monotonically decreasing curve).Also, generally the tasks with

[1] We ignore denoise, autoencoding, and colorization as these tasks require modified input

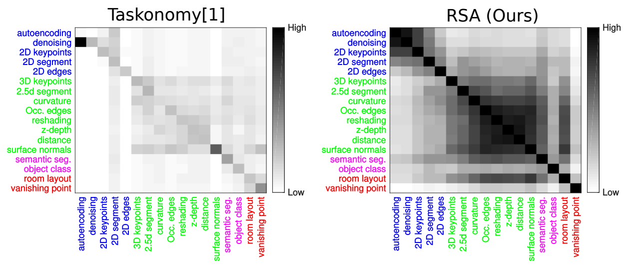

Figure S5. **Similarity matrix: Taskonomy vs RSA(Ours)**

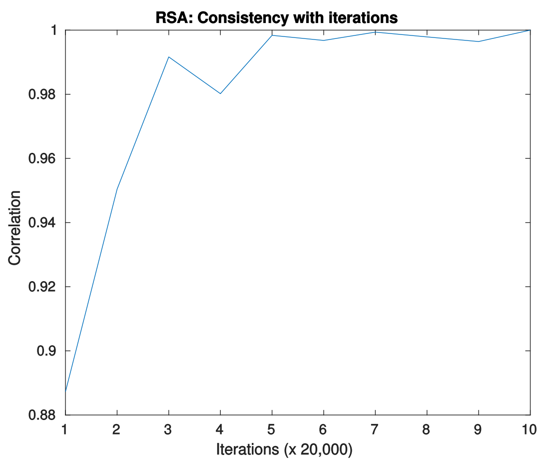

Figure S6. **Consistency with training iterations**

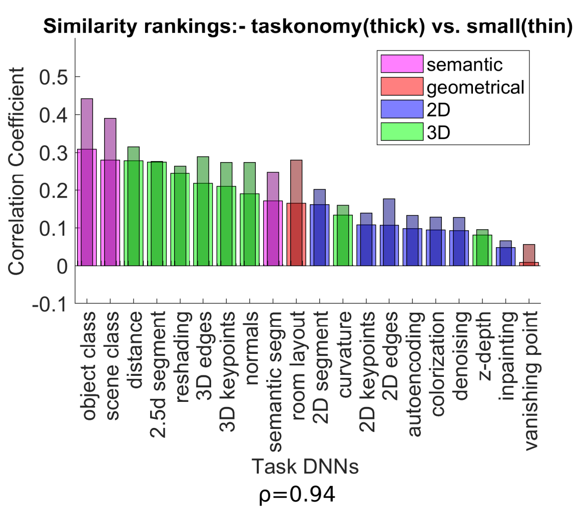

Figure S7. **Consistency with model size**

higher similarity ranking (object class, surface normals, segment25d) showed high transfer learning performance, and tasks with lower similarity score (autoencoding, vanishing point) showed lower performance.

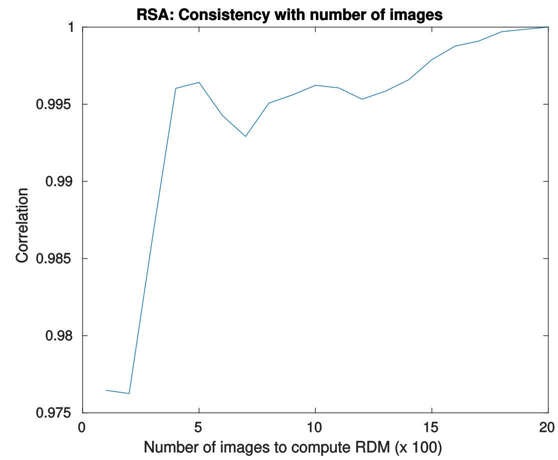

Figure S8. **Consistency with number of images**

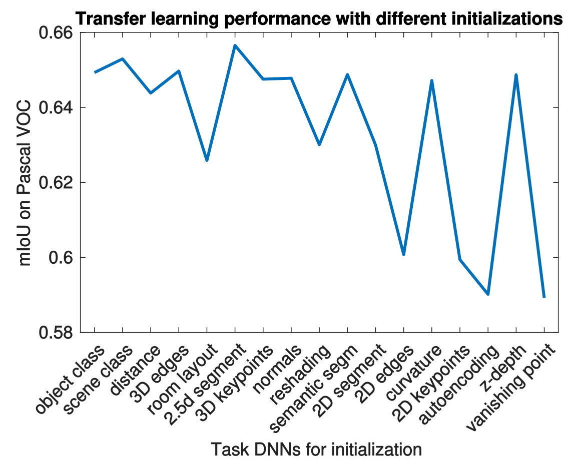

Figure S9. **Transfer learning performance in descending order of similarity scores with task DNNs on the x-axis as initialization**

References

[1] A. R. Zamir, A. Sax, and W. Shen. Taskonomy: Disentangling task transfer learning.