

Representation Similarity Analysis for Efficient Task taxonomy & Transfer Learning

Kshitij Dwivedi

Singapore University of Technology and Design

kshitij_dwivedi@mymail.sutd.edu.sg, gemma.roig@sutd.edu.sg

Abstract

Transfer learning is widely used in deep neural network models when there are few labeled examples available. The common approach is to take a pre-trained network in a similar task and finetune the model parameters. This is usually done blindly without a pre-selection from a set of pre-trained models, or by finetuning a set of models trained on different tasks and selecting the best performing one by cross-validation. We address this problem by proposing an approach to assess the relationship between visual tasks and their task-specific models. Our method uses Representation Similarity Analysis (RSA), which is commonly used to find a correlation between neuronal responses from brain data and models. With RSA we obtain a similarity score among tasks by computing correlations between models trained on different tasks. Our method is efficient as it requires only pre-trained models, and a few images with no further training. We demonstrate the effectiveness and efficiency of our method for generating task taxonomy on Taskonomy dataset. We next evaluate the relationship of RSA with the transfer learning performance on Taskonomy tasks and a new task: Pascal VOC semantic segmentation. Our results reveal that models trained on tasks with higher similarity score show higher transfer learning performance. Surprisingly, the best transfer learning result for Pascal VOC semantic segmentation is not obtained from the pre-trained model on semantic segmentation, probably due to the domain differences, and our method successfully selects the high performing models.

1. Introduction

For an artificial agent to perform multiple tasks and learn in a life-long manner, it should be able to re-utilize information acquired in previously learned tasks and transfer it to learn new tasks from a few examples. A solution to the aforementioned setting is to use transfer learning. Transfer

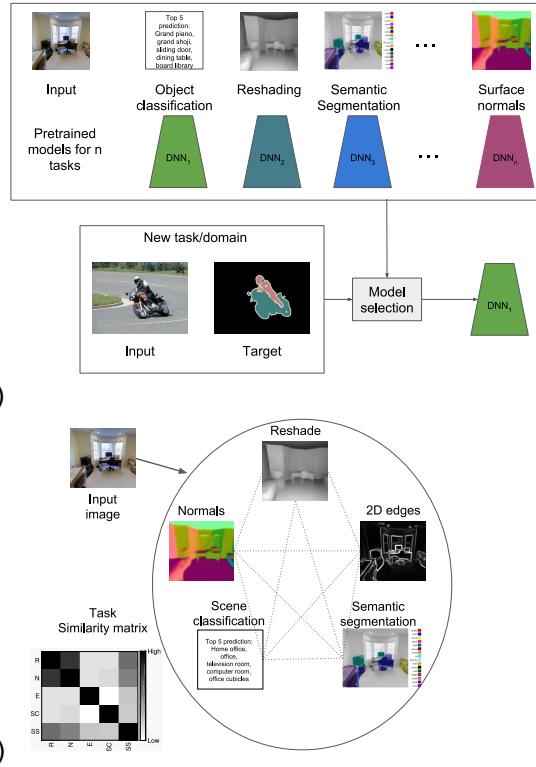


Figure 1. **Aims of this paper:** a) Deploy a strategy for model selection in transfer learning by b) Finding relationship between visual tasks.

learning allows to leverage representations learned from one task to facilitate learning of other tasks, even when labeled data is expensive or difficult to obtain. [30, 3, 23, 10].

With the recent success of deep neural networks (DNN), these have become the *ipso facto* models for almost all visual tasks [20, 32, 14, 35, 13, 34]. The deployment of DNN has become possible mostly due to a large amount of available labeled data, as well as advances in computing resources [20, 32, 14]. The need for data is a limita-

tion that researchers have overcome by introducing transfer learning techniques. Transfer learning in DNN commonly consists of taking a pre-trained model in a similar task or domain, and finetune the parameters to the new task. For instance, [30, 10] used a pre-trained model on ImageNet and finetuned it for object detection on Pascal VOC.

With a large number of pre-trained models (Figure 1a) available, trained on a variety of vision tasks, it is not trivial how to select a pre-trained representation suitable for transfer learning. To devise a model selection strategy, it is crucial to understand the underlying structure and relationship between tasks (Figure 1b). If the relationship between different tasks is known, the model selection can be performed by evaluating similarity rankings of different tasks with a new task, using available pre-trained models.

In a recent work, [34] modeled the relationship between tasks with a fully computational approach. They also introduce a dataset called Taskonomy, which contains labels of different visual tasks, ranging from object classification to edge occlusions detection. In this paper, we use the term Taskonomy for both the approach and the dataset from [34].

Taskonomy approach successfully computes the relationship between tasks. Yet, the relationship between a new task with an existing set of tasks is calculated with the transfer learning performance, which is tedious and computationally expensive. The performance on the new task is referred to transfer learning performance. To obtain the relationship of all previous tasks with the new task, Taskonomy approach also needs to compute the transfer learning performance on all the previous tasks using a model trained on the new task as a source. This defeats the purpose of not training a model from scratch for the new task, and all the procedure is computationally demanding as it is repeated for all the existing set of specific-task models. In this work, we address the above limitations by providing an alternative method to find the relationship between tasks.

We propose a novel approach to obtain task relationships using representation similarity analysis (RSA). In computational neuroscience, RSA is widely used as a tool to compare brain responses with computational and behavioral models. Motivated by the success of RSA in neuroscience [18, 4, 16, 1, 5, 25, 11], we investigate the application of RSA in obtaining task similarities (Figure 1b) and in transfer learning (Figure 1a). Our approach relies on the assumption that the representations of the models that perform a related task will be more similar as compared to tasks that are not related, which we validate in our analysis.

In our approach, we compute the similarity scores using pre-trained task-specific models and a few examples. Thus, our RSA method only requires the representations of a few randomly selected images for all the tasks to compute the similarity, and we do not need to obtain transfer learning performance by finetuning on previous tasks' models.

Further, we show in our results on Taskonomy dataset that task ranking similarity is independent of model size. Using small models trained with few samples for the existing tasks show similar results as the high performing models trained with all images. This allows to save computational time and memory, as well as it is more scalable to new tasks compared to Taskonomy approach.

We first validate the transfer learning applicability of our method on Taskonomy dataset. We find that for 16 out 17 Taskonomy tasks, the best model selected using RSA is in top-5 according to transfer learning performance. We also report results on Pascal VOC semantic segmentation task by analyzing the relationship of RSA similarity scores and the transfer learning performance. Our results show a strong relationship between RSA similarity score and transfer learning performance. We note that semantic segmentation model from Taskonomy dataset showed a lower similarity score than most of the 3D and semantic tasks, and a similar trend was observed in transfer learning performance. Our results suggest that in domain-shift, a model trained on the same task may not be the best option for transfer learning, and using our similarity score one can find a better model to achieve better performance. Using our RSA similarity scores method, we can select models with better transfer learning performance.

2. Related Works

Here, we discuss the works that are most closely related to the aim of this paper, namely transfer learning in DNNs and Taskonomy. Then, we briefly introduce the computational neuroscience literature that motivated our work.

2.1. Transfer Learning

The usual transfer learning approach in deep neural networks (DNNs) is to take a model pre-trained on a large dataset with annotations as an initialization of a part of the model. Then, some or all of the parameters are finetuned with backpropagation for a new task. The finetuning is performed because for most of the tasks there are insufficient annotations to train a DNN from scratch, which would lead to overfitting. Most of the works in the literature generally initialize the model parameters from a model pre-trained on Imagenet [6] dataset for image classification [20, 32, 14, 31, 22]. For example, [30] use Imagenet initialized models for object detection on Pascal VOC, [23] use Imagenet initialized models for semantic segmentation.

It has been noted in multiple works [24, 33, 28], that the initialization plays a significant role in performance in transfer learning. Hence, a strategy is required to select models for initialization. Our proposed similarity-based ranking approach offers a solution to this problem, and as we discuss in the rest of the paper, tackles the limitations

from Taskonomy [34], which is one of the first attempts to tackle the model selection for transfer learning in DNN.

2.2. Taskonomy

Our work is most closely related to Taskonomy [34], where the aim is to find the underlying task structure by computing the transfer performance among tasks. To achieve this goal, they create a dataset of indoor scene images with annotations available for 26 vision tasks. The task set, which they refer as task dictionary, covers common 2D, 3D, and semantics computer vision tasks. Then, task-specific independent models are trained in a fully supervised manner for each task in the task dictionary. They obtain a task similarity score by comparing the transfer learning performance from each of the task-specific models and computing an affinity matrix using a function of transfer learning performance. In this paper, instead of transfer learning performance, we rely on the similarity of the feature maps of the pre-trained models. Thus, we avoid additional training on pre-trained models to obtain transfer learning performance, saving computational time and memory, and still obtaining a meaningful relation with transfer learning performance as we will see in the results section.

2.3. Similarity of computational models and brain responses

In computational neuroscience, representation similarity analysis (RSA) is widely used to compare a computational or behavioral model with the brain responses. In [18], RSA is used to compute similarities between brain responses in different regions of visual cortex with categorical models and computational vision models. In [16], the authors use several unsupervised and supervised vision models to show that supervised models explain IT cortical area better than unsupervised models, and [25] uses RSA to correlate the dynamics of the visual system with deep neural networks. We note that as the approach can be used to assess the similarity between a computational model and brain data, the approach can also be utilized to assess similarities between two computational models. RSA has been rarely used in the pure computational domain. Only in [26] the RSA was introduced as a loss function for knowledge distillation [15], and in [27], the consistency of RSA correlations with different random initialization seeds within the same model trained on CIFAR-10 [19] dataset is explored. However, RSA is still unexplored in comparing DNNs for assessing similarity among them. Our work introduces, for the first time, the use of RSA as a similarity measure to find the relationship between tasks, and we believe it opens a new research line for the deep learning and computer vision.

We use RSA similarity measure for two applications namely task taxonomy and transfer learning. Our approach is not limited to only these two applications and can be

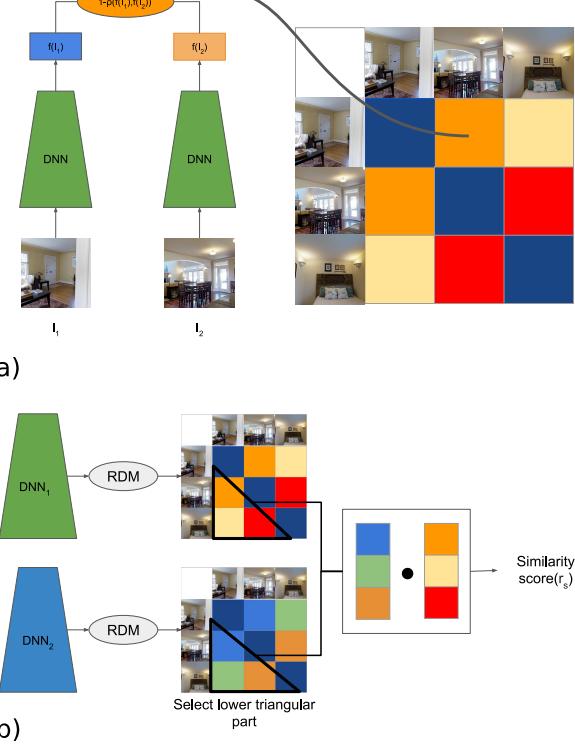


Figure 2. Representation Similarity Analysis (RSA): a) Representation dissimilarity matrices (RDMs) are generated by computing the pairwise dissimilarity ($1 - \text{Pearson's correlation}$) of each image pair in a subset of selected images. b) Similarity score: Spearman's correlation (r_s) (denoted with ●) of the low triangular RDMs of the two models is used as the similarity score. Here DNN₁ and DNN₂ refer to the models trained on task 1 and 2 respectively.

further applied in other computer vision problems. For instance, in multi-task learning [17, 13, 7, 21, 8] RSA could be used for deciding different branching out locations for different tasks, depending on their similarity with the representations at different depth of the shared root.

3. Representation Similarity Analysis (RSA)

Representation Similarity Analysis (RSA) [18], illustrated in Figure 2, is a widely used data-analytical framework in the field of computational neuroscience to quantitatively relate the brain activity measurement with computational and behavioral models. In RSA, a computational model and brain activity measurements are related by comparing representation-activity dissimilarity matrices. The dissimilarity matrices are obtained by comparing the pairwise dissimilarity of activity/representation associated with each pair of conditions.

In this work, we introduce RSA as a tool to quantify the relationship between DNNs and its application in transfer

learning for model selection. We explain the steps to obtain the dissimilarity matrix for a computational model such as DNN in the following paragraph.

Representation Dissimilarity Matrix (RDM) We first select a subset of images as conditions for dissimilarity computation. For a given DNN, we then obtain the representation of each image by performing a forward pass through the model. For each pair of conditions (images), we compute a dissimilarity score $1 - \rho$, where ρ is the Pearson's correlation coefficient. The RDM for this subset of conditions is then populated by the dissimilarity scores for each pair of conditions, see Figure 2a.

In our method, the RDMs computed for DNNs are used for obtaining the similarity between two computer vision tasks. Note that by using RDMs, the representation for different tasks can be of different length. The similarity is computed with the Spearman's correlation (r_s) between the upper or lower triangular part of the RDMs of the two DNNs. This is: $r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$, where d_i is the difference between the ranks of i_{th} elements of the lower triangular part of the two RDMs in Figure 2b, and n are the number of elements in the lower triangular part of the RDM.

The Spearman's correlation provides a quantitative measure of similarity between the task the DNNs were optimized for (Figure 2b). We explore the application of this similarity score in obtaining the relationship between computer vision tasks [34], and in transfer learning.

4. RSA for Task Taxonomy and Transfer Learning

In this section, we introduce our RSA approach for getting a task taxonomy of computer vision tasks, as well as its application in transfer learning. We show the effectiveness of RSA for obtaining task similarity by answering three questions: 1) we investigate if we can group tasks into meaningful clusters based on task type using RSA on pre-trained task-specific models; 2) we analyze if the performance is important for computing task similarity or we can use a smaller subset of data with smaller suboptimal models; and 3) we investigate if the similarity we obtain using RSA is related to transfer learning.

4.1. Is task similarity related to task type?

We validate our hypothesis that tasks similar according to RSA are grouped into clusters according to task type, for instance, 2D, 3D, semantic. To do so, we randomly select 500 images from the Taskonomy dataset, and select 20¹ tasks from the task dictionary. Then, we compute the RDMs of the pre-trained models for each of the 20 tasks using the

¹we exclude Jigsaw task as it is unrelated to all other tasks

task-specific representations of the 500 sampled images, as described in section 3. The task-specific representations are obtained by doing a forward pass on the pre-trained task-specific DNN models. With the resulting RDMs per task, we compute a pairwise correlation of RDMs of each task with the 19 other tasks to get a 20×20 task similarity matrix (Figure 3a). We perform a hierarchical clustering from the similarity matrix, to visualize if the clustering groups the tasks according to the task type or some other criteria. We report the results in the experiments section and compare it with the clustering obtained with the Taskonomy approach.

We note that RSA is symmetric, as compared to the transfer performance based metric in Taskonomy [34]. Yet, symmetry does not affect task similarity rankings, as the positions of the tasks in the rankings are computed by relative comparison, and therefore, independent of symmetry.

4.2. Does ranking using RSA depends on dataset and model size?

We analyze whether RSA based task similarity depends on the model size and amount of training data. Intuitively, it should be independent of model and dataset size, because our method is based on relative similarities. To investigate this, we select a subset of Taskonomy tasks (details in supp. material section S1) and trained smaller models, one per task, with fewer parameters than the models provided by Taskonomy, and on a small subset of Taskonomy data. First, we evaluate if we obtain a similar task clustering using the small models on the selected tasks. Then, for each small model, we compute the similarity score with the pre-trained Taskonomy models on all 20 tasks. The same analysis is repeated with pre-trained Taskonomy model trained on the same task, and we compare the relative similarity based rankings of the small and Taskonomy high-performing models. If the relative rankings of both small and Taskonomy model are similar, then the result suggests that for a completely new task one can train a small model and compute similarity scores to rank them.

4.3. Is RSA related to transfer performance?

We investigate if RSA based task similarity can be applied to transfer learning problem. We first compute the correlation between each column of Taskonomy affinity matrix with RSA matrix after removing the diagonal. As the Taskonomy affinity matrix is populated by raw losses/evaluations, it is indicative of transfer learning performance [34]. We next select a task and dataset different from Taskonomy and obtained the similarity scores of a model trained on the new task with Taskonomy pre-trained models. The pre-trained models were ranked according to the similarity score. We then use the pre-trained models for initializing the model and add the last task dependent layers on top of the initialized model to train on the new task. The

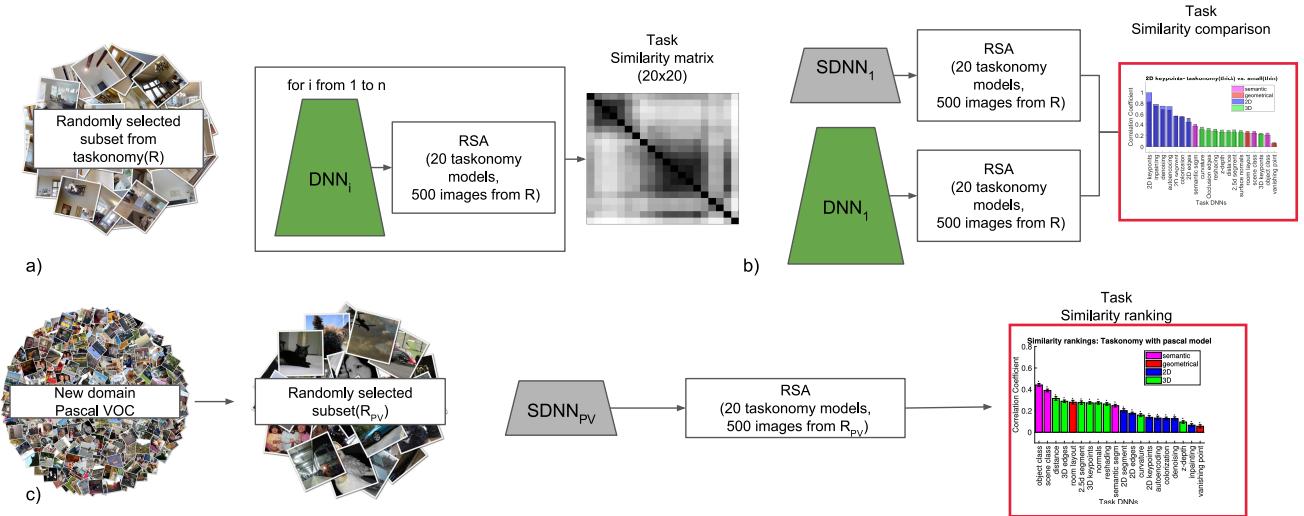


Figure 3. **Our approach:** a) RSA of task-specific pre-trained DNN models (from Taskonomy) to compute a task similarity matrix, b) RSA of small model (SDNN) trained on small datasets and comparison with Taskonomy pre-trained models. c) RSA of small model (SDNN_{PV}) trained on new task (Pascal VOC semantic segmentation) with Taskonomy pretrained models.

ranking based on the transfer performance is compared with the ranking based on RSA to evaluate the relation between transfer performance and RSA. As we will see in the results, RSA can be used to select the high performing models for transfer learning.

5. Experimental set-up

We first provide the details of datasets used for the experiments, followed by the details of the models’ architecture.

5.1. Datasets

Taskonomy dataset It includes over 4 million indoor images from 500 buildings with annotations available for 26 image tasks. 21 of these tasks are single image tasks, and 5 tasks are multi-image tasks. For this work, we select 20 single image task for obtaining task similarities¹.

We randomly selected 500 images from the Taskonomy training dataset as 500 different conditions to perform RSA. These images are used as input to generate representations of different task-specific models to compute the RDMs.

To analyze the dependency of RSA on dataset and model size used for training, we select one building (Hanson) from Taskonomy dataset, which contains 12138 images. We divide them into 10048 training and 2090 validation images.

Pascal VOC semantic segmentation To evaluate the application of RSA in transfer learning, we select the Pascal VOC [9, 12] dataset for semantic segmentation task. It has pixelwise annotations for 10, 582 training images, 1, 449 validation and 1, 456 test images. We argue that this task is different from the Taskonomy semantic segmentation as the images are from a different domain.

5.2. Models

Below, we provide details of the network architectures of pre-trained Taskonomy models, small models trained for Taskonomy tasks, and models used for Pascal VOC.

Taskonomy models The Taskonomy models² consist of an encoder and decoder. The encoder for all the tasks is a Resnet-50 [14] model followed by convolution layer that compresses the channel dimension of the encoder output from 2048 to 8. The decoder is task-specific and varies according to the task. For classification tasks and tasks where the output is low dimensional the decoder consists of 2-3 fully connected (FC) layers. For all the other tasks, the decoder consists of 15 layers (except colorization with 12 layers) consisting of convolution and deconvolution layers.

We select the final compressed output of the encoder as the representation for RSA as in [34]. In Taskonomy approach, the compressed output of the encoder was used as an input to transfer function to evaluate the transfer learning performance. Selecting the compressed output of the encoder ensures that the architecture for all the task is the same, and the differences in representation can only arise due to the task that the model was optimized for, as images are also the same for all tasks.

We also explore the representation of earlier layers of the encoder and the task labels as the representation for computing RSA based similarity score. We perform this analysis to investigate how task specificity varies across the depth in the network and if the task’s labels are enough to understand the relationship between tasks.

²publicly available at <https://github.com/StanfordVL/taskonomy>

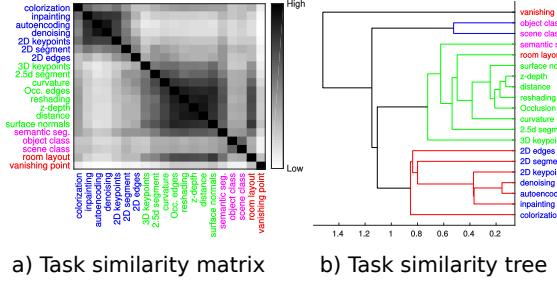


Figure 4. **Task similarity using RSA:** a) Similarity matrix of the 20 Taskonomy tasks, b) Agglomerative clustering using RDM.

Small models The smaller version of the models follows a similar style to Taskonomy and consists of an encoder and decoder. The encoder consists of 4 convolution layer each with a stride of 2 to generate a final feature map with the dimensions same as that of Taskonomy encoder. For this experiment, we select the tasks which require a fully-convolution decoder structure and use 4 convolution layers each followed by an upsampling layer. The models are trained on Hanson subset of Taskonomy dataset.

Pascal VOC Models We use two types of models for Pascal VOC semantic segmentation task: 1) a small model to compute similarity score with pre-trained Taskonomy models; 2) models initialized with pre-trained Taskonomy encoders to evaluate transfer learning performance. The small model consists of an encoder and a decoder. The encoder consists of 4 convolution layer each with a stride of 2 to generate a final feature map with the dimensions same as that of Taskonomy encoder. The decoder is an Atrous Spatial Pyramid Pooling (ASPP) [2], which contains convolution layers that operate in parallel with different dilations. The model is trained on Pascal VOC training set with learning rate 10^{-4} for 200,000 iterations. The encoder representation of the small model trained on Pascal VOC is then used to compute similarity with Taskonomy pre-trained models. The models for evaluating transfer learning performance consists of an encoder with similar architecture as Taskonomy models and an ASPP decoder. The encoder part is initialized by the pre-trained Taskonomy models of the corresponding task.

Implementation and evaluation details We use the publicly available tensorflow implementation³ of deeplabv3 [3] and modify the code for transfer learning experiments. We use RSA Matlab toolbox [29] for RSA related analysis⁴. We refer to the supplementary material for further details.

6. Results

Here, we present the results of RSA for computing task similarity and its relation to transfer learning performance.

³https://github.com/sthalles/deeplab_v3

⁴Code available at <https://github.com/kshitijd20/RSA-CVPR19-release>

We follow the same nomenclature of task type as in [34], and color code 2D, 3D, semantic and geometric tasks.

6.1. Task similarity using RSA

Figure 4a shows the similarity matrix of the tasks computed using RSA with the compressed encoder output as the task representation. Recall that we compute the 20×20 similarity matrix using RSA with given task-specific representations for all the randomly selected 500 images. To visualize the relationship between tasks, we applied agglomerative hierarchical clustering to the similarity matrix. The resulting dendrogram from this clustering is shown in Figure 4b. We can see that the tasks are clustered following visual criteria of 2D, 3D, and semantic tasks.

We further investigate the task similarity using RSA at different depths in the encoder architecture and task labels. Figure 5 shows the task similarity matrix for different depths of the Resnet-50 encoder, namely blocks 1, 2, 3 and 4. We also compare the similarity matrix computed using the tasks’ labels. We observe, in Figure 5, that at block 1 all the similarity values are very high implying that at initial layers representations of most of the tasks are similar irrespective of the task type. As we go deeper, the similarity score between tasks starts decreasing, and in compressed encoder output, we can see three dark blocks corresponding to 2D, 3D, and semantic tasks. The above results further validate our choice of using compressed encoder output as the task-specific representation for assessing the similarity between tasks. Interestingly, the clustering using task labels does not group into tasks of the same type, and most of the similarity scores are low. Instead, the labels clustering follows the output structure of the labels, independently of the task type. This is because the labels contain only limited information about the task, and it depends on the annotator criteria on how to represent the output.

We next compare our approach with Taskonomy approach⁵. We use hierarchical clustering to visually compare the dendograms obtained using both the methods in Figure 6. For quantifying the similarity, we compute the correlation of Taskonomy similarity matrix with RSA similarity matrix ($\rho = 0.62$, $r_s = 0.65$). The results show that both approaches group the tasks into similar clusters with few exceptions. Room layout is grouped with the vanishing point in Taskonomy approach and in 3D tasks with our approach. Denoising is clustered with inpainting and autoencoding using our approach, which are related tasks. We argue that our results are plausible.

⁵We show 17 tasks as we had access to only affinity values of these tasks. For comparison with figure 13 in [34], please refer to section S2 of supplementary material

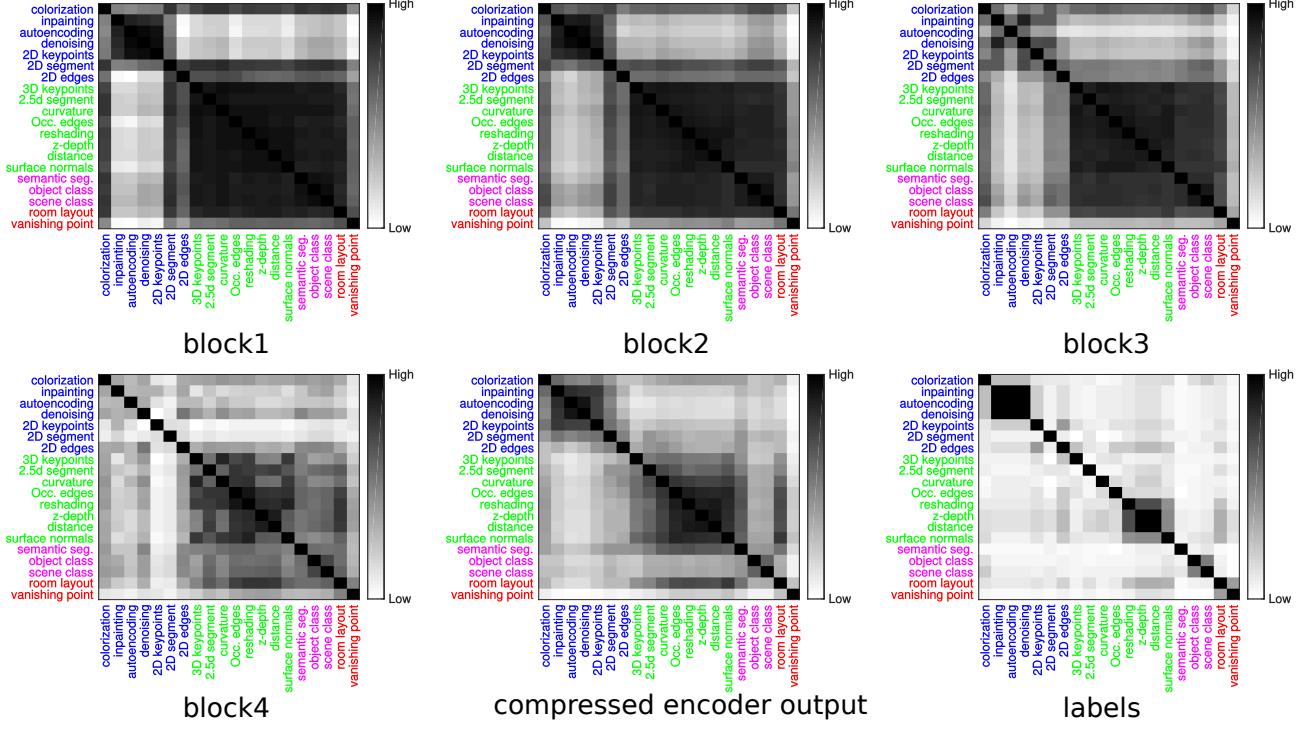


Figure 5. **Task taxonomy using RSA:** 1 – 5) Similarity matrix of 20 Taskonomy tasks using features at different depth in the model as task-specific representations 6) Similarity matrix of 20 Taskonomy tasks using labels as task-specific representations.

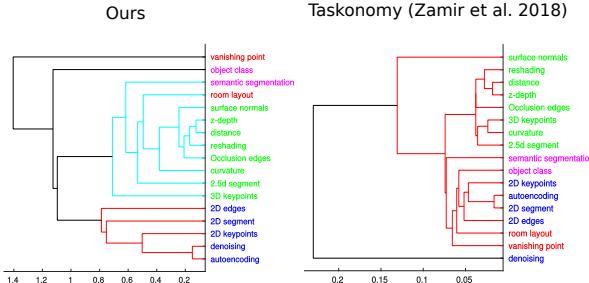


Figure 6. **RSA vs Taskonomy:** Clustering comparison.

6.2. Does model size impact similarity score?

In this experiment, we investigate how the model and dataset size affect task similarity. We show the results of similarity rankings for 2 tasks: 2D keypoints and surface normals (for other tasks, please see section S1 in supplementary material). We compare the similarity rankings obtained using the small model trained on Hanson subset of Taskonomy data with the Taskonomy model trained on the same task. As we visually observe from the comparison (Figure 7) in both the tasks the ranking look similar. For all the tasks considered in the above comparison the mean correlation is high ($\rho = 0.84$, $r_s = 0.85$).

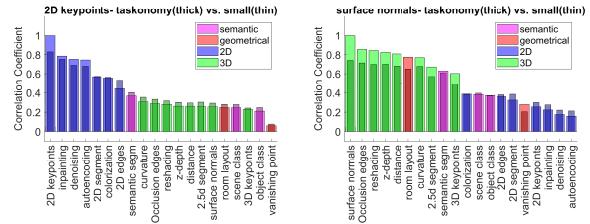


Figure 7. **Task taxonomy using small models:** Similarity ranking of (a) keypoint2d Taskonomy model vs small model. (b) surface normals Taskonomy model vs small model.

Next, we also computed task similarity matrices by comparing a small model with small models trained on other tasks. We find that the correlation ($\rho = 0.85$, $r_s = 0.88$) between task similarity matrices (Figure S3) using Taskonomy model and small model is comparable to previous correlation results. The above results together provide strong evidence that the model and dataset size do not have much effect on the similarity score.

6.3. Model selection for transfer learning

We first report the model selection using RSA for Taskonomy tasks and then on Pascal VOC semantic seg-

Top-1	Top-3	Top-5
7/17	14/17	16/17

Table 1. Number of tasks for which best model selected for transfer learning using RSA is in top-n models according to transfer performance for 17 tasks

mentation task.

Taskonomy We obtain high mean correlation ($\rho = 0.70$, $r_s = 0.76$) between RSA and transfer learning for 17 tasks from the Taskonomy dataset. We also report in Table 1 that for 16 out of 17 tasks, the best model selected by RSA for transfer learning is in top-5 models selected using Taskonomy approach (transfer learning performance).

Pascal VOC We show the relation of similarity score using RSA with transfer learning by selecting a new task (semantic segmentation in Pascal VOC). We compare the transfer learning performance of models initialized by different task-specific pre-trained models from Taskonomy dataset. Then we compare the transfer learning performance based ranking with similarity score ranking. Here we select the small Pascal model to compute the similarity with the Taskonomy models. We report the robustness of similarity ranking using RSA with respect to model size, number of images used for RSA analysis, and different training stages in supplementary section S3.

We show the similarity score based ranking in Figure 8. Surprisingly, semantic segmentation model from Taskonomy shows a lower similarity score as compared to other models trained on semantic (scene class, object class) and 3D tasks (occlusion edges, surface normals). Most of the 2D tasks show low similarity scores.

To investigate if similarity scores are related to transfer learning performance we evaluated the models initialized with task-specific Taskonomy models, finetuned with Pascal VOC training set, and compared the performance on Pascal VOC test set. Table 1 shows the comparison of transfer learning performance for models with initialization from a set of selected tasks (For a complete comparison refer to section S3 in the supplementary material). The tasks are listed in the order of their similarity scores. We note from the table that the tasks on the top (object class, scene class, occlusion edges, and semantic segmentation) shows higher performance while autoencoder and vanishing point performance is even less than model trained from scratch (random in Table 2). We note that our results are comparable to the results (64.81%) reported in [3], when they use Resnet-50 trained on Imagenet for initialization. The results provide evidence that the similarity score obtained using RSA provide an estimate of the expected transfer performance.

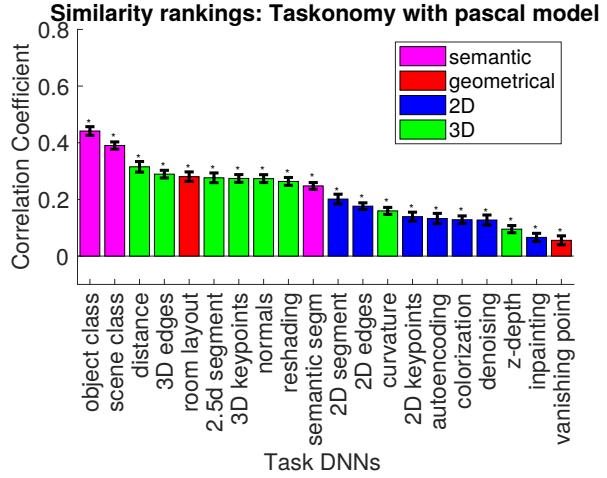


Figure 8. RSA based similarity of scores of pre-trained Taskonomy models with the small model trained on Pascal VOC.

Initialization(Task)	mIoU
Object class	0.6492
Scene class	0.6529
Occlusion edges	0.6496
Semantic segmentation	0.6487
Autoencoder	0.5901
Vanishing point	0.5891
Random(Taskonomy encoder)	0.6083
Random(Small encoder)	0.4072

Table 2. Transfer learning performance on Pascal VOC test set.

7. Conclusion

We presented an efficient alternative approach to obtain the similarity between computer vision models trained on different tasks using their learned representations. Our approach uses RSA, and it is suitable for obtaining task similarity by just using the pre-trained models without any further training, as opposed to the earlier state of the art method Taskonomy for this problem.

We provided strong evidence that for obtaining the similarity, the model and training dataset size does not play a significant role and we can obtain a task similarity relative ranking using small models as well as state of the art models with few data samples. This comes with computational and memory savings.

We also showed the relationship of the task similarity using RSA with the transfer learning performance and its applicability. We demonstrated on both, Taskonomy and Pascal VOC semantic segmentation, that the transfer learning performance is closely related to the similarity obtained with RSA. The above results showed that for domain shift the model trained on the same task might not be the best fit for transfer learning and our proposed approach can help in

model selection for transfer learning. Our method is applicable to a wide range of potential problems, such as multi-task models, architecture selection.

Acknowledgements This work was funded by the SUTD-MIT IDC grant (IDG31800103). K.D. was also funded by SUTD Presidents Graduate Fellowship. We thank Taskonomy authors for the support and the code.

References

- [1] Michael F Bonner and Russell A Epstein. Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology*.
- [2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.
- [3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [4] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(June):1–13, 2016.
- [5] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455, 2014.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [7] Thanuja Dharmasiri, Andrew Spek, and Tom Drummond. Joint prediction of depths, normals and surface curvature from rgb images using cnns. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 1505–1512. IEEE, 2017.
- [8] Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. Blitznet: A real-time deep network for scene understanding.
- [9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.
- [10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.
- [11] Iris IA Groen, Michelle R Greene, Christopher Baldassano, Li Fei-Fei, Diane M Beck, and Chris I Baker. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, 7:e32962, 2018.
- [12] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 991–998. IEEE, 2011.
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.
- [14] 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.
- [15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [16] Seyed Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11), 2014.
- [17] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory.
- [18] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.
- [19] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [21] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.
- [22] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.
- [23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [24] Arun Mallya and Svetlana Lazebnik. Piggyback: Adding multiple tasks to a single, fixed network by learning to mask. *arXiv preprint arXiv:1801.06519*, 2018.
- [25] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Oliva. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346–358, 2017.
- [26] Patrick McClure and Nikolaus Kriegeskorte. Representational distance learning for deep neural networks. *Frontiers in computational neuroscience*, 10:131, 2016.
- [27] Johannes Mehrer, Nikolaus Kriegeskorte, and Tim Kietzmann. Beware of the beginnings: intermediate and higher-

- level representations in deep neural networks are strongly affected by weight initialization. In *Conference on Cognitive Computational Neuroscience*, 2018.
- [28] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Tom Yan, Alex Andonian, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding.
 - [29] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.
 - [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
 - [31] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
 - [32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
 - [33] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
 - [34] Amir R Zamir, Alexander Sax, and William Shen. Taskonomy: Disentangling task transfer learning.
 - [35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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[34] 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.

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 Taskon-

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
<i>Upscale</i> $\times 2$			
Conv7	3×3	16	1
<i>Upscale</i> $\times 2$			
Conv8	3×3	4	1
<i>Upscale</i> $\times 2$			
Conv9	3×3	4	1
<i>Upscale</i> $\times 2$			
Conv10	3×3	n	1

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

omy 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[34] 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

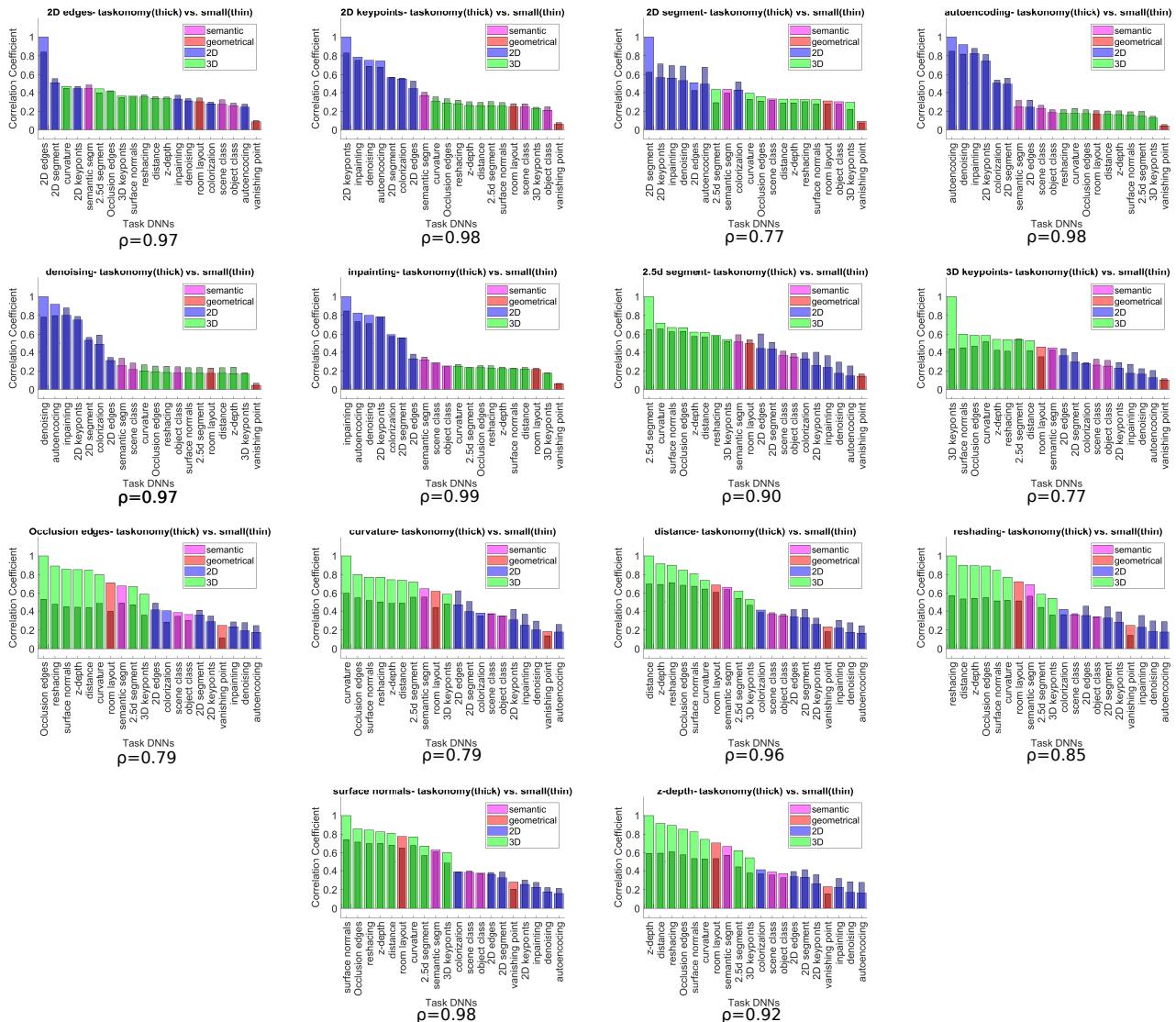


Figure S1. Similarity ranking with taskonomy model vs small models for 14 tasks. The ρ value below each plot specifies the Pearson's correlation coefficient between the two similarity rankings.

different clustering method, which was not specified, used in [34].

One other advantage of our approach over Taskonomy is 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.

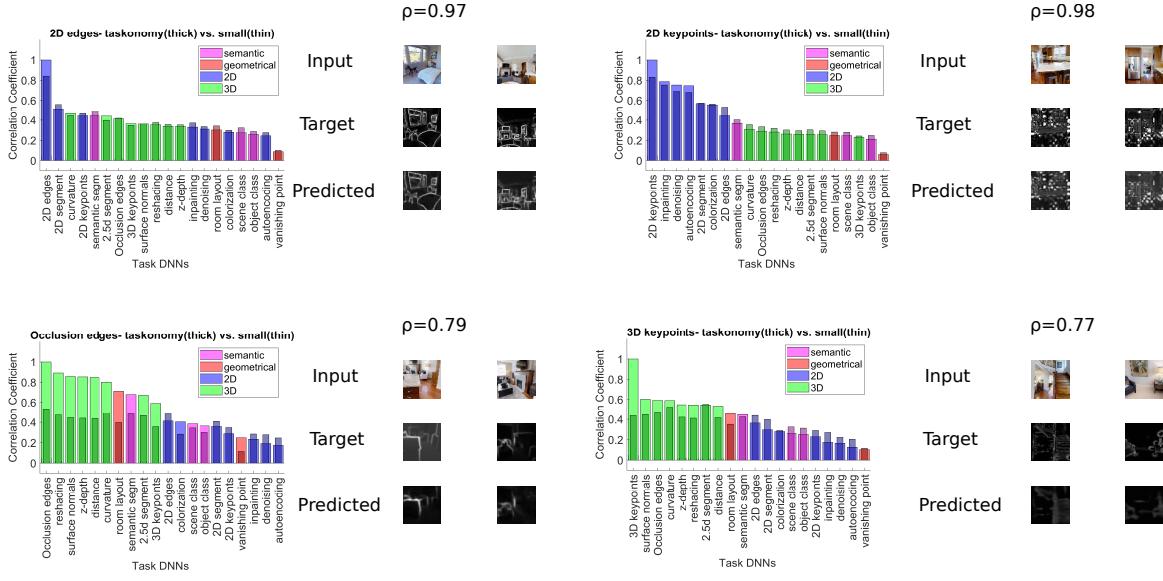


Figure S2. Is correlation related to visual similarity of the predicted output with the target?

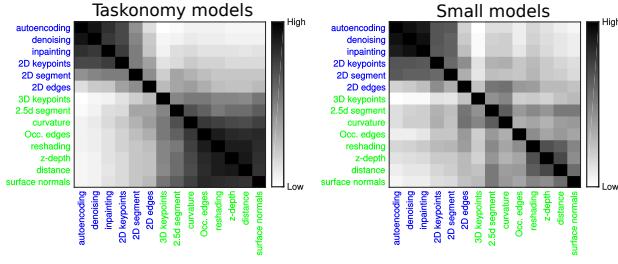


Figure S3. Task similarity matrix using Taskonomy models vs small models.

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 ($\rho = 0.95$, $r_s = 0.96$) suggests that we 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.

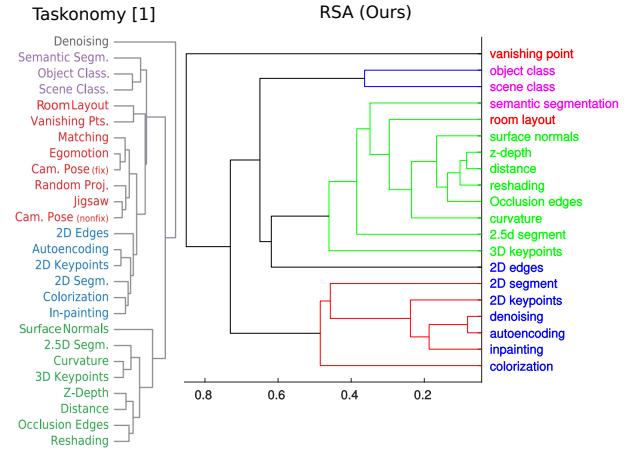


Figure S4. Clustering: Taskonomy vs RSA (Ours) Image source: Figure 13 from [34]

S3.4. Transfer learning performance for all the tasks

Figure S9 shows the transfer learning performance (mIoU) for 17 single image tasks¹ 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

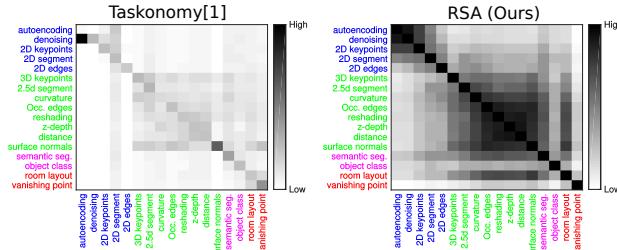


Figure S5. Similarity matrix: Taskonomy vs RSA(Ours)

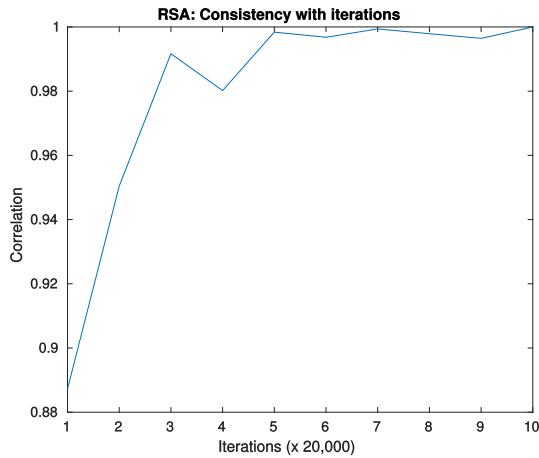


Figure S6. Consistency with training iterations

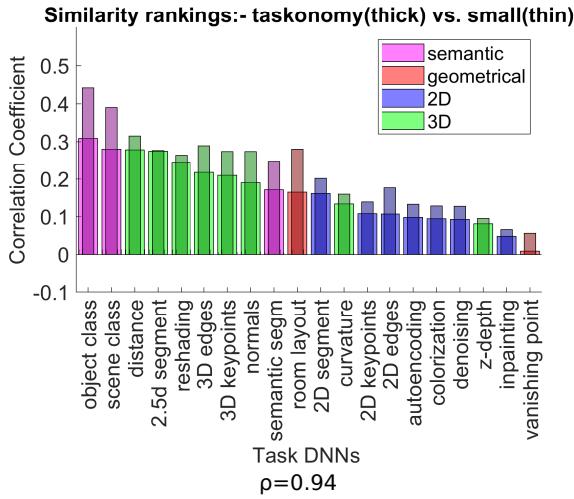


Figure S7. Consistency with model size

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.

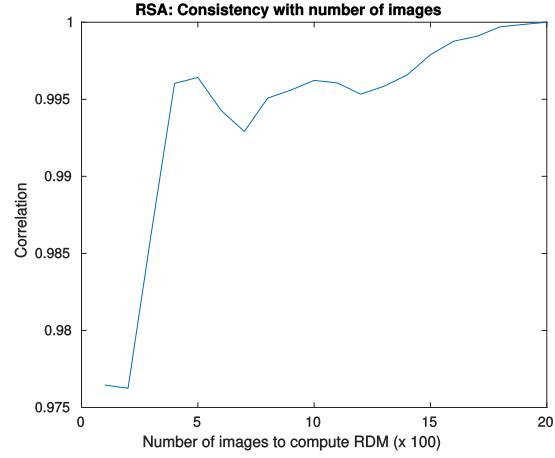


Figure S8. Consistency with number of images

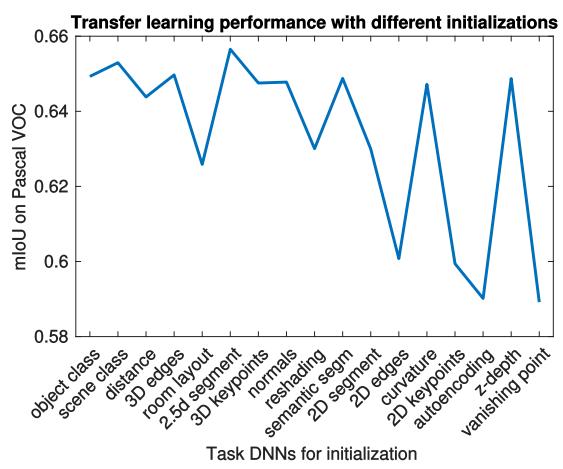


Figure S9. Transfer learning performance in descending order of similarity scores with task DNNs on the x-axis as initialization