Supplementary Material
Duality Diagram Similarity:
a generic framework for initialization selection in task transfer learning

Kshitij Dwivedi[1,3][0000−0001−6442−7140], Jiahui Huang[2][0000−0002−0389−1721], Radoslaw Martin Cichy[3][0000−0003−4190−6071], and Gemma Roig[1][0000−0002−6439−8076]

[1] Department of Computer Science, Goethe University Frankfurt, Germany
kshitijdwivedi93@gmail.com, roig@cs.uni-frankfurt.de
[2] ISTD, Singapore University of Technology and Design, Singapore
jiahui_huang@sutd.edu.sg
[3] Department of Education and Psychology, Free University Berlin, Germany
rmcichy@zedat.fu-berlin.de

We provide the following items in the supplementary material, which complement the results reported in the main paper:

S1 RSA and CKA as a special case of duality diagram similarity (DDS).

S2 Different normalizations in DDS Framework.

S3 Results on DDS's dependence on number of images.

S4 Results on model selection using coarse task representations.

S5 Quantitative and qualitative results of layer selection using a ImageNet/Places365 pre-trained encoder.

S6 Effect of unbiased centering.

S7 Results with Spearman's correlation as g.

S8 DDS Results on Taskonomy and Pascal VOC for all distance/kernels as f.

S9 DDS Results for 17 Taskonomy tasks.

S10 Precision and Recall curves for DDS.

S11 DDS dependences on image dataset choice.

S1 RSA and CKA as special cases of duality diagram similarity (DDS)

The duality diagram of a matrix $\mathbf{X} \in \mathbb{R}^{n \times d_1}$ can be calculated by the product of $\mathbf{Q_X}$, $\mathbf{X}$ and $\mathbf{D}$, where $\mathbf{Q} \in \mathbb{R}^{d_1 \times d_1}$ is a matrix that quantifies dependencies between the individual feature dimensions, and $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a matrix that assigns weights on the observations.

Let $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ be the duality diagrams obtained from two different models (layers), the duality diagram similarity (DDS) between those two can be calculated by first computing pairwise distance matrices, $\mathbf{M_X}$, $\mathbf{M_Y}$, using a distance

function, f, then use another function, g, to compare $\mathbf{M_X}$ and $\mathbf{M_Y}$ to obtain the final similarity score, $\mathbf{S}$. The formulation of DDS can be written as:

$$\mathbf{S} = g\left(f\left(\mathbf{DXQ_X}\right), f\left(\mathbf{DYQ_Y}\right)\right) \tag{1}$$

RSA as DDS. To compute RSA, one needs to obtain for each model (layer) the Representation Dissimilarity Matrices (RDMs), which is populated by computing a dissimilarity score $1 - \rho$, where ρ is the Pearson's correlation coefficient between each pair of images (observations). Once the RDMs for each model (layer) is computed, then Spearman's correlation of the upper triangular part of the 2 RDMs is used to compute the final similarity score between the two RDMs. Here, one can observe the connection between RSA and DDS. In Equation 1, RDMs are the above-mentioned pairwise distance matrices, $\mathbf{M_X}$ and $\mathbf{M_Y}$, the distance function f used in RSA is the dissimilarity score $1 - \rho$. If no normalization is used, matrix $\mathbf{D}$ and matrix $\mathbf{Q}$ are both identity matrices, $\mathbf{I}$ (ones in the diagonal and the rest of the elements in the matrix are zeros). In [1], they use a centering matrix $\mathbf{C}$ ($\mathbf{C} = \mathbf{I_n} - \frac{1}{n}\mathbf{1_n}$, where $\mathbf{1}$ is the $n \times n$ matrix of all ones) as $\mathbf{D}$ in the formulation of the duality diagram, and $\mathbf{Q}$ as the identity matrix. The final similarity score, g, used in RSA is the Spearman's correlation between lower/upper triangular part of the two RDMs. Finally, RSA as a special case of DDS can be written as:

$$\mathbf{S} = r_s^t\left(1 - \rho\left(\mathbf{CXI}\right), 1 - \rho\left(\mathbf{CYI}\right)\right) \tag{2}$$

where r_s^t denotes the Spearman's correlation of the upper traingular part of the two input matrices, the dissimilarity score $1 - \rho$ is computed with the Pearson's correlation, $\mathbf{C}$ is the centering matrix for $\mathbf{X}$ and $\mathbf{Y}$, and $\mathbf{I}$ is the identity matrix.

CKA as DDS. The formulation of CKA [3] can be written as:

$$\text{CKA}\left(\mathbf{X}, \mathbf{Y}\right) = \text{tr}\left(\mathbf{KHLH}\right) / \sqrt{\text{tr}\left(\mathbf{KHKH}\right)\text{tr}\left(\mathbf{LHLH}\right)}, \tag{3}$$

in which $\mathbf{K}$ and $\mathbf{L}$ are the output matrices after applying either the RBF or linear kernel (kernel function k) on data matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively. Mathematically, $\mathbf{K}$ and $\mathbf{L}$ can be expressed as , $\mathbf{K_{ij}} = k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma_2||\mathbf{x}_i - \mathbf{x}_j||^2)$, $\mathbf{L_{ij}} = k(\mathbf{y}_i, \mathbf{y}_j) = \exp(-\gamma_2||\mathbf{y}_i - \mathbf{y}_j||^2)$ for RBF, and, $\mathbf{K_{ij}} = k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$, $\mathbf{L_{ij}} = k(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{y}_i^T\mathbf{y}_j$ for the linear kernel. Here $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the ith column of $\mathbf{X}$ and $\mathbf{Y}$ respectively, $\mathbf{H}$ is the centering matrix($\mathbf{H} = \mathbf{I_n} - \frac{1}{n}\mathbf{1_n}$, where $\mathbf{1}$ is the $n \times n$ matrix of all ones) , and T denotes the transpose. To obtain the equivalent formulation in DDS, in the following equations, we substitute $\mathbf{KH}$ and $\mathbf{LH}$ with $\hat{\mathbf{K}}$ and $\hat{\mathbf{L}}$ for simplification. Since $\hat{\mathbf{K}}_{\mathbf{ij}} = \hat{\mathbf{K}}_{\mathbf{ji}}$, we can get $\mathbf{K^T} = \mathbf{K}$, similarly, $\mathbf{L^T} = \mathbf{L}$, thus the above equation can be written as:

$$\text{CKA}\left(\mathbf{X}, \mathbf{Y}\right) = \sum_{ij} \hat{\mathbf{K}}_{\mathbf{ij}}\hat{\mathbf{L}}_{\mathbf{ij}} / \sqrt{\sum_{\mathbf{ij}} \hat{\mathbf{K}}_{\mathbf{ij}}^2 \sum_{\mathbf{ij}} \hat{\mathbf{L}}_{\mathbf{ij}}^2}, \tag{4}$$

which corresponds to the cosine similarity between $\hat{\mathbf{K}}$ and $\hat{\mathbf{L}}$. Since $\mathbf{K_{ij}} = k(x_i, x_j)$ and $\mathbf{L_{ij}} = k(y_i, y_j)$, we can treat them as pairwise distance matrix $\mathbf{M_X}$, $\mathbf{M_Y}$, respectively, in the formulation of DDS. In both cases, $\mathbf{Q}$ and $\mathbf{D}$ are both identity matrix here, and the distance function f is the linear or the RBF kernel. The final similarity function g used here is the cosine distance combined with the multiplication of the input matrices with the centering matrix $\mathbf{H}$. From above, we can derive CKA as a special case of DDS, and it can be written as:

$$\mathbf{S} = cos\Big(k\big(\mathbf{IXI}\big)\mathbf{H}, k\big(\mathbf{IYI}\big)\mathbf{H}\Big) \tag{5}$$

where cos is the cosine distance, k is either the linear or the RBF kernel, and $\mathbf{I}$ is the identity matrix.

S2   Different normalizations in DDS framework

In Table S1, we show how different normalizations used in deep learning and z-scoring can be reformulated in the DDS framework. Let $\mathbf{X} \in \mathbb{R}^{n \times c \times h \times w}$ be the output feature map of a convolutional layer with number of channels c, height h, width w for n input images. By swapping axes and reshaping the feature map $\mathbf{X}$, all the normalizations investigated in this work can be described in Duality Diagram setup. It is crucial to note that after reshaping the feature map, $\mathbf{D}$ and $\mathbf{Q}$ no longer represent weighing image and feature dimensions.

Norm	D	X	Q
Z-score	$\mathbf{I}_{n \times n} - \mathbf{1}_{n \times n}/n$	$\mathbf{X}_{n \times chw}$	$\mathbf{S}_{chw \times chw}$
Batch Norm	$\mathbf{I}_{nhw \times nhw} - \mathbf{1}_{nhw \times nhw}/nhw$	$\mathbf{X}_{nhw \times c}$	$\mathbf{S}_{c \times c}$
Group Norm	$\mathbf{I}_{\frac{c}{g}hw \times \frac{c}{g}hw} - \mathbf{1}_{\frac{c}{g}hw \times \frac{c}{g}hw}/\frac{c}{g}hw$	$\mathbf{X}_{\frac{c}{g}hw \times ng}$	$\mathbf{S}_{ng \times ng}$

Table S1. *Different normalizations in DDS framework.* Here $\mathbf{I}$ denotes an identity matrix, $\mathbf{1}$ denotes matrix filled with all 1's, $\mathbf{X}$ is the output feature map of a convolutional layer with number of channels c, height h, width w for n input images, g is group size for group normalization, and $\mathbf{S}$ is a diagonal matrix with diagonal values equal to standard deviation of $\mathbf{X}$ calculated across its rows. For each normalization, $\mathbf{X}$ is reshaped as indicated in the table. Layer and Instance normalization can be described by setting the group size g in Group norm to 1 and c respectively.

S3   DDS's dependence on number of images

To calculate similarity between 2 Deep Neural Networks (DNNs) using DDS, we need to perform feedforward pass through both DNNs on a selected set of images. Here we analyse the impact of the number of images selected to compute

the similarity measure. We varied the number of images from 10 to 500, in increments of 10, in a randomly selected set of Taskonomy images for Taskonomy benchmark and Pascal VOC images for Pascal VOC transfer learning benchmark, to calculate DDS. We plotted the correlation with transfer learning on Taskonomy tasks and on Pascal VOC semantic segmentation task in Figure S1a and Figure S1b respectively. We show the results for DDS with different f, namely Laplacian, RBF, linear, cosine, Euclidean and Pearson's correlation. We observe from the plots that correlation value with transfer learning saturates at around 200 images for all functions. For this reason, in all the experiments reported in the main paper we use 200 images in the selected sets.

S4   Model selection using a coarse task representation

Using task affinities as a method for source model selection, which is common also in all related works [1,5,7], requires a pre-trained model on the new task itself to measure affinities. In [1], it was proposed to train a small model on the new task, instead of a full large model, because it can be trained faster. The small models learn a coarse representation of the new task, and the task affinities to the source models can be compared faster. We use the small model from [1], and compare the correlation with transfer learning performance for $DDS(f = Laplacian)$ using small models and big models. We show the comparison in Figure S2, and we observe that correlation with transfer learning performance using small model is very close to the correlation using fully trained Taskonomy type model. Further, we observe that using $DDS(f = Laplacian)$ even with small model we outperform baseline RSA [1] that uses fully trained Taskonomy type models. Overall, the above results suggest that even with a coarse representation obtained by training a small model on new task can assist in model selection using the similarity measures proposed in this work.

S5   Results of Layer selection(ImageNet/Places pre-trained encoder)

In addition to the experiments conducted with ImageNet pre-trained encoder, reported in the main paper, here we also provide results for an encoder pretrained on Places365 [8]. The representation type of different blocks of Places pre-trained model, as shown in Figure S3, is similar to what we observed in Imagenet pretrained model, reported in the main paper. From Table S2 and Table S3, we observe that our similarity measure successfully predicted best branching location for 5 out of 6 cases. Only exception is NYUv2 depth estimation task, where the transfer learning performance of block 3, selected by the method, is slightly lower than the best branching location (block 4). Overall from the above results combined with ImageNet results from the main text, we find that the proposed method reliably selects high performing branching locations to transfer to new tasks.

Block \ Task	Edge (MAE)	Normals (mDEG_DIFF)	Semantic (mIOU)
1	**0.680**	17.89	0.244
2	0.777	15.62	0.368
3	1.012	**14.35**	0.532
4	1.002	14.73	**0.616**

Table S2. *Transfer learning performance of branching Places pre-trained encoder on 3 tasks on Pascal VOC dataset.* The results indicate that branching out from block 1, 3, 4 of the encoder have better performances on edge, normals and semantic tasks, respectively.

Block \ Task	Edge (MAE)	Depth (log RMSE)	Semantic (mIOU)
1	**1.027**	0.320	0.125
2	1.188	0.286	0.167
3	1.183	0.223	0.216
4	1.120	**0.201**	**0.291**

Table S3. *Transfer learning performance of branching Places pre-trained encoder on 3 tasks on NYUv2 dataset.* The results are mostly consistent with branching location prediction based on DDS.

We show qualitative results on Pascal VOC[2] and NYUv2[4] datasets in Figure S4 and Figure S5. Here we illustrate branching results of 3 tasks: Edge Detection, Surface Normal (Depth) Prediction and Semantic Segmentation. For each task, ImageNet pre-trained encoder results are shown on the upper row, and Places 365 pre-trained encoder results are shown on the lower row. We observed some visual quality degradation in the results of non-optimal branching locations predicted by our similarity measures: Edge contours become blurry as the branching location goes deeper; semantic segmentation maps become closer to the ground truth at deeper layers.

S6   Effect of unbiased centering

We report in Table S4 the effect of applying unbiased centering (eq. 3.1 in [6]) to M_X and M_Y on the correlation with transferability. We observe that for all cases unbiased centering improves the correlation with transfer learning, and hence, in all the reported results in the main paper we used unbiased centering.

S7   Results with Spearman's correlation as g

In the main text, we reported the results with g as Pearson's correlation between unbiased centered (dis)similarity matrices M_X and M_Y. Here, in Table S5, we

Centering $(\mathbf{M_X}, \mathbf{M_Y})$ f	kernels			distances		
	linear	Laplacian	RBF	Pearson	euclidean	cosine
No centering	0.818	0.691	0.690	0.776	0.613	0.792
Unbiased centering	0.842	0.860	0.841	0.856	0.850	0.864

Table S4. *Effect of unbiased centering.* We report the results of comparison with transferability on Taskonomy transfer learning for with and without unbiased centering on pairwise (dis)similarity matrices $\mathbf{M_X}$ and $\mathbf{M_Y}$

$\mathbf{Q}$ f	kernels			distances		
	linear	Laplacian	RBF	Pearson	euclidean	cosine
Identity	0.778	0.828	0.803	0.816	0.798	0.803
Z-score	0.858	0.864	0.846	0.844	0.862	0.860

Table S5. *Spearman's as g.* We report the results of comparison with transferability on Taskonomy transfer learning benchmark for with and without z-scoring when using Spearman's as g.

report results when g is the Spearman's correlation between upper/lower triangular part of unbiased centered (dis)similarity matrices $\mathbf{M_X}$ and $\mathbf{M_Y}$, as in [1], on Taskonomy transfer learning benchmark. We observe that the results show similar trend with Spearman's correlation (improvement on applying z-scoring on $\mathbf{X}$,$\mathbf{Y}$) as using Pearson's correlation as g, shown in main text Table 2 .

S8   DDS results for all distance/kernels as f

In Table 3 and Table 4 of main paper we reported the best f selected using the results in Table 2. Here we report the complete results for Table 3 and Table 4 with all investigated functions as f. Due to efficiency of our method it was possible to perform multiple bootstrap to calculate standard deviation in correlation with transfer learning. In the tables below (Table S6 and Table S7), we report bootstrap mean and standard deviation of correlation with transfer learning for Taskonomy tasks and Pascal VOC semantic segmentation task. We observe that $\mathrm{DDS}(f = Laplacian)$ is the most robust (in Top 1,2) measure across both Taskonomy benchmark and Pascal VOC semantic segmentation tranfer learning.

S9   DDS similarity measure comparison for 17 Taskonomy tasks

In the main paper, we reported the mean correlation of similarity measures with transfer learning across 17 Taskonomy tasks. In Figure S6, we provide the detailed results on all tasks. We find that almost on all the tasks our proposed similarity measures outperform the state-of-the-art method [5,1].

Method	Affinity	Winrate
DDS($f = pearson$)	0.853 ± 0.090	0.851 ± 0.090
DDS($f = euclidean$)	0.852 ± 0.076	0.855 ± 0.079
DDS ($f = cosine$)	0.862 ± 0.076	0.863 ± 0.078
DDS($f = linear$)	0.837 ± 0.084	0.841 ± 0.088
DDS ($f = Laplacian$)	0.862 ± 0.072	0.861 ± 0.072
DDS($f = rbf$)	0.854 ± 0.086	0.854 ± 0.088

Table S6. *Correlation (Bootstrap mean ±standarddev) of DDS based affinity matrices with Taskonomy affinity and winrate matrix, averaged for 17 Taskonomy tasks.* Top 2 scores are shown in green, and blue respectively. For this experiment, **Q** is set to z-scoring and **D** to the identity matrix, in all DDS tested frameworks.

Method	Taskonomy	Pascal VOC	NYUv2
DDS ($f = Pearson$)	0.534 ± 0.063	0.726 ± 0.049	0.505 ± 0.033
DDS ($f = euclidean$)	0.534 ± 0.055	0.746 ± 0.051	0.518 ± 0.030
DDS ($f = cosine$)	0.525 ± 0.057	0.722 ± 0.049	0.518 ± 0.034
DDS ($f = linear$)	0.496 ± 0.063	0.718 ± 0.062	0.515 ± 0.033
DDS ($f = Laplacian$)	0.577 ± 0.050	0.765 ± 0.038	0.521 ± 0.029
DDS ($f = RBF$)	0.591 ± 0.053	0.753 ± 0.051	0.534 ± 0.030

Table S7. *DDS correlation with transfer learning for Pascal VOC Semantic Segmentation.* Here each row represents DDS with a particular distance/kernel function as f, and each column represents the dataset from which the images were selected to get similarity scores. The values in the table are bootstrap mean correlation and standard deviation of a particular similarity measure computed using the image from a particular dataset. Top score is shown in green.

S10 Precision and Recall curve for DDS

In the main text, we used correlation of similarity measure based source model rankings with transfer learning performance based rankings as our evaluation criteria. Song *et al.* [5] used precision and recall of selecting top-5 source tasks as the evaluation criteria. We use the evaluation code provided by [5], and we plot precision and recall curve for one of our most robust proposed method, DDS($f = Laplacian$), against state-of-the-art methods [5,1], in Figure S7. In Figure S7, we plot results using 200 Taskonomy images for all the similarity measures that we compared. We further add the results of the methods from Song *et al.* [5] using 1000 images from indoor dataset used in Song *et al.* [5] that showed best performance in their paper. We observe from Precision and Recall plots in Figure S7 that DDS($f = Laplacian$) outperforms the state-of-the-art methods.

S11 DDS dependences on image dataset choice

In this section, we investigate the effect of image dataset used to calculate Duality Diagrams. We report the results of DDS correlation with Taskonomy winrate in Table Table S8when images from Taskonomy, Pascal VOC, and NYUv2 were used to calculate Duality Diagrams. We observe a slight drop in DDS's correla-

Method	Taskonomy	Pascal VOC	NYUv2
DDS ($f = Pearson$)	0.856	0.818	0.842
DDS ($f = euclidean$)	0.850	0.815	0.805
DDS ($f = cosine$)	0.864	0.818	0.822
DDS ($f = linear$)	0.842	0.811	0.801
DDS ($f = Laplacian$)	0.860	0.811	0.818
DDS ($f = RBF$)	0.841	0.807	0.792

Table S8. *DDS correlation with transfer learning on Taskonomy Tasks.* Here each row represents DDS with a particular distance/kernel function as f, and each column represents the dataset from which the images were selected to obtains similarity scores.

tion with Taskonomy winrate matrix when using images from Pascal VOC and NYUv2 dataset.

These results are consistent with [5] where they show that their method is robust to choice of images used to compute similarity between neural networks. In the aforementioned results, both source and target tasks were trained using the same training dataset, i.e. Taskonomy, and we believe that is the reason we, as well as [5], do not observe much difference.

However, when we compare transferability on Pascal VOC, source models are trained on Taskonomy dataset and target task is on Pascal VOC, which has significantly different statistics than Taskonomy. In this more challenging setting, we observe the impact of using images from different datasets, as reported in Section 6.3 in the main text.

References

1. Dwivedi, K., Roig, G.: Representation similarity analysis for efficient task taxonomy & transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12387–12396 (2019)
2. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (Jun 2010). https://doi.org/10.1007/s11263-009-0275-4, `http://dx.doi.org/10.1007/s11263-009-0275-4`
3. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2661–2671 (2019)
4. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
5. Song, J., Chen, Y., Wang, X., Shen, C., Song, M.: Deep model transferability from attribution maps. In: Advances in Neural Information Processing Systems (2019)
6. Székely, G.J., Rizzo, M.L., et al.: Partial distance correlation with methods for dissimilarities. The Annals of Statistics **42**(6), 2382–2412 (2014)
7. Zamir, A.R., Sax, A., Shen, W.: Taskonomy: Disentangling task transfer learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
8. Zhou, B., Khosla, A., Lapedriza, À., Torralba, A., Oliva, A.: Places: An image database for deep scene understanding. CoRR **abs/1610.02055** (2017)
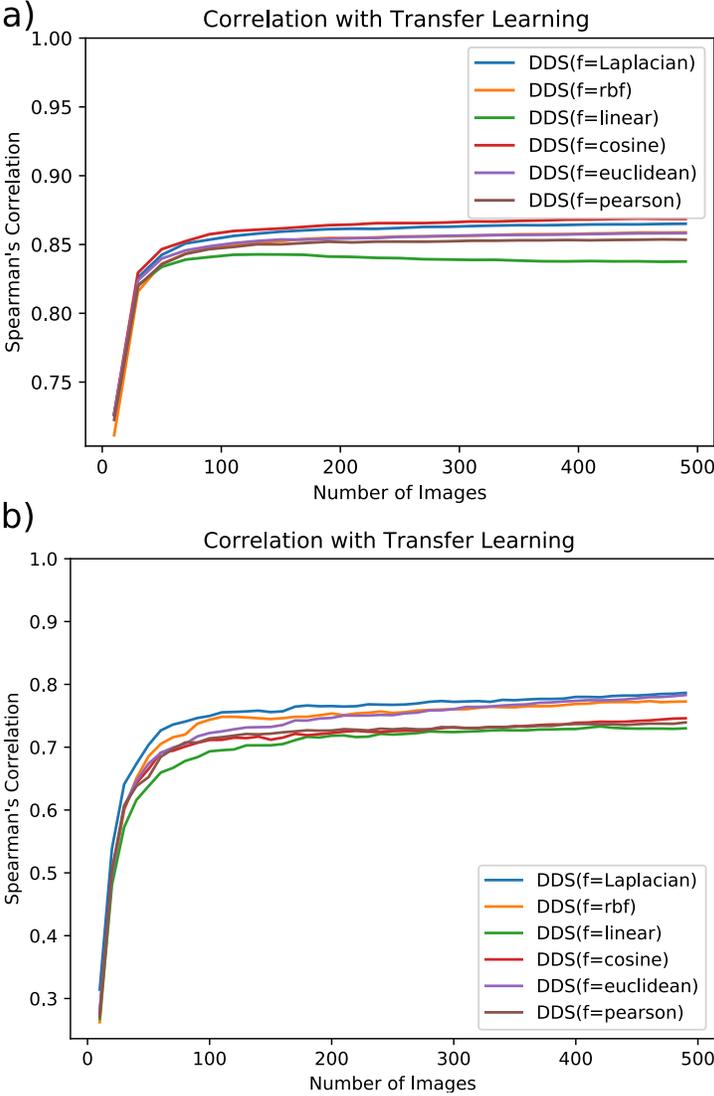
a)

Correlation with Transfer Learning

b)

Correlation with Transfer Learning

Fig. S1. *Spearman's correlation of DDS and transfer learning performance* on a) Taskonomy tasks, and b) Pascal VOC semantic segmentation task. The above plots shows how Spearman's correlation of DDS with transfer learning varies with the number of images used to compute similarity using DDS with different distance/kernel functions as f. The images are randomly sampled from the Pascal VOC dataset.
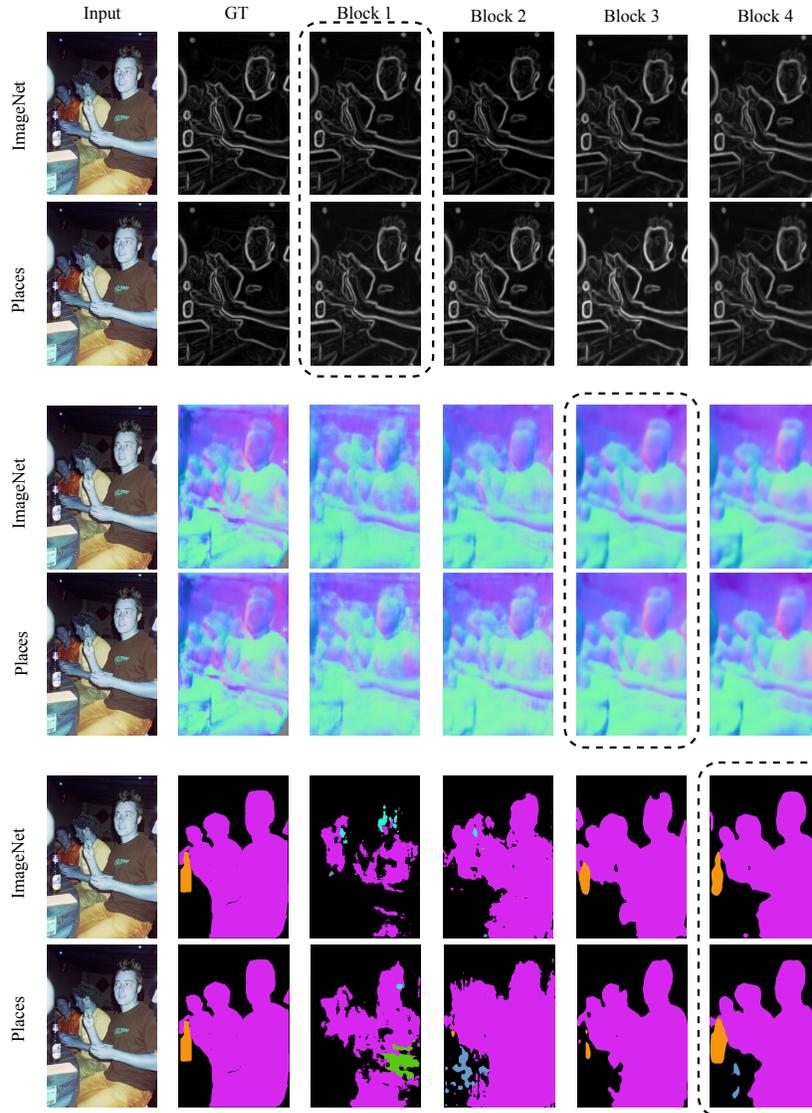
Fig. S2. *Correlation of DDS(f = Laplacian) based rankings obtained using a small model trained on Pascal VOC semantic segmentation task with the transfer learning performance.* We show how correlation varies with different stages of training. We further compare the results with the upper bound obtained by using the large trained Taskonomy type model on Pascal VOC semantic segmentation as the task representation and also a baseline RSA using the large model.

Fig. S3. *Block selection using DDS* on pre-trained encoder on Places, and with DNNs trained on Taskonomy dataset on different tasks.

Fig. S4. *Qualitative Results on Pascal VOC. Branching results of all locations on three tasks are shown: Edge Detection, Surface Normal Prediction and Semantic Segmentation. For each task, ImageNet pre-trained encoder are shown on the upper row, while Places 365 pre-trained encoder are shown on the lower row. Best results are circled with dotted lines.*
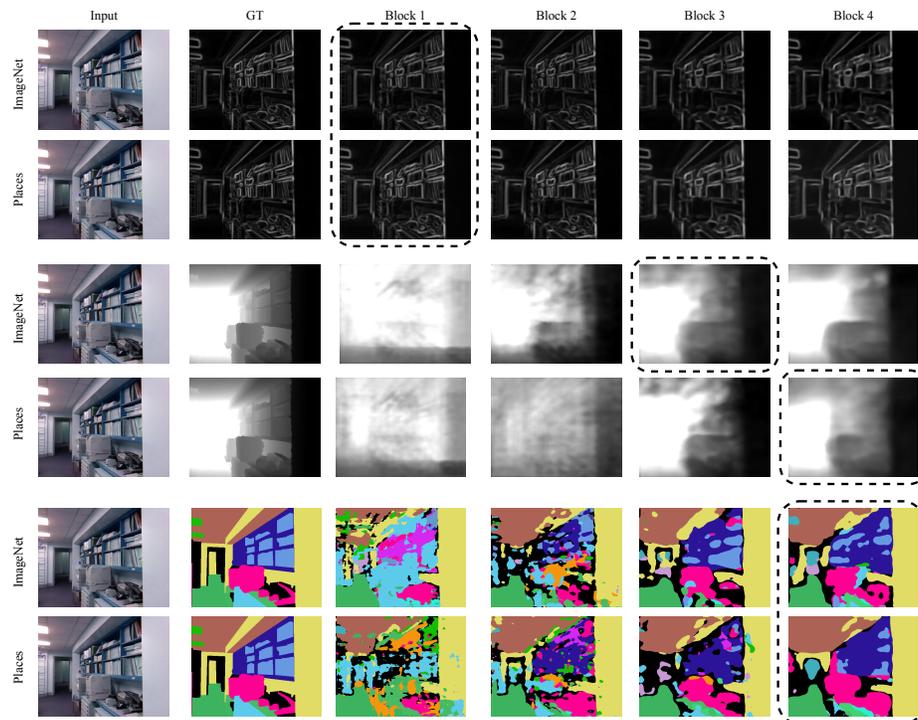
Fig. S5. Qualitative Results on NYUv2. *Branching results of all locations on three tasks are shown: Edge Detection, Depth Prediction and Semantic Segmentation. For each task, ImageNet pre-trained encoder are shown on the upper row, while Places 365 pre-trained encoder are shown on the lower row. Best results are circled with dotted lines.*
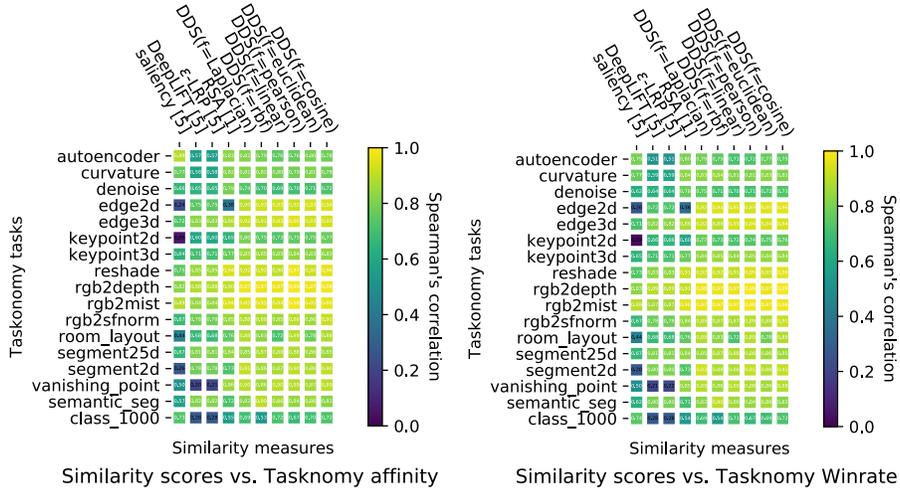
Fig. S6. *Similarity measures' comparison on Taskonomy Tasks.* Spearman's correlation of different similarity measure based rankings with transfer learning performance based rankings from Taskonomy affinity matrix (left), and Taskonomy winrate matrix (right) for 17 Taskonomy tasks as target. We show the results for 17 Taskonomy tasks (rows) for different similarity measures (columns). More yellow indicates higher the correlation, hence, is better.
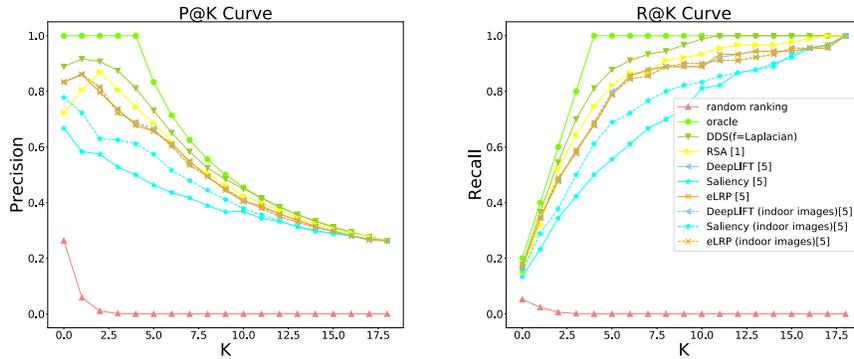

Fig. S7. *Precision and Recall Curve for comparing similarity measures.* The x-axis in all the plots above refers to the number of source tasks used for calculating precision and recall value.