Taskonomy: Disentangling Task Transfer Learning

——By Wei Gao

Questions

1 too many data with fewer annotations

2 tasks have latent relation

3 designing specific model to solve problem
brings redundant computation cost

Task Relation

1 Surface Normal can derive from Depth

2 Semantic Segmentation is related with Occlusion Edge Det

3 ImageNet trained model as a pretrained model for many tasks

Related work

Self-supervised learning methods leverage the inherent relationships between tasks to learn a desired expensive one (e.g. object detection) via a cheap surrogate (e.g. colorization) [68, 72, 17, 103, 100, 69]. Specifically, they use a manually-entered local part of the structure in the task space (as the surrogate task is manually defined). In contrast, our approach models this large space of tasks in a computational manner and can discover obscure relationships.

Unsupervised learning is concerned with the redundancies in the input domain and leveraging them for forming compact representations, which are usually agnostic to the downstream task [8, 49, 20, 9, 32, 77]. Our approach is not unsupervised by definition as it is not agnostic to the tasks. Instead, it models the space tasks belong to and in a way utilizes the *functional* redundancies among tasks.

Meta-learning generally seeks performing the learning at a level higher than where conventional learning occurs, e.g. as employed in reinforcement learning [21, 31, 28], optimization [2, 82, 48], or certain architectural mechanisms [27, 30, 87, 65]. The motivation behind meta learning has similarities to ours and our outcome can be seen as a computational meta-structure of the space of tasks.

Multi-task learning targets developing systems that can provide multiple outputs for an input in one run [50, 18]. Multi-task learning has experienced recent progress and the reported advantages are another support for existence of a useful structure among tasks [93, 100, 50, 76, 73, 50, 18, 97, 61, 11, 66]. Unlike multi-task learning, we explicitly model the relations among tasks and extract a meta-structure. The large number of tasks we consider also makes developing one multi-task network for all infeasible.

Related work

Domain adaption seeks to render a function that is developed on a certain domain applicable to another [44, 99, 5, 80, 52, 26, 36]. It often addresses a shift in the *input* domain, e.g. webcam images to D-SLR [47], while the task is kept the same. In contrast, our framework is concerned with *output* (task) space, hence can be viewed as *task/output adaptation*. We also perform the adaptation in a larger space among many elements, rather than two or a few.

Task Definition and Purpose

Definition:

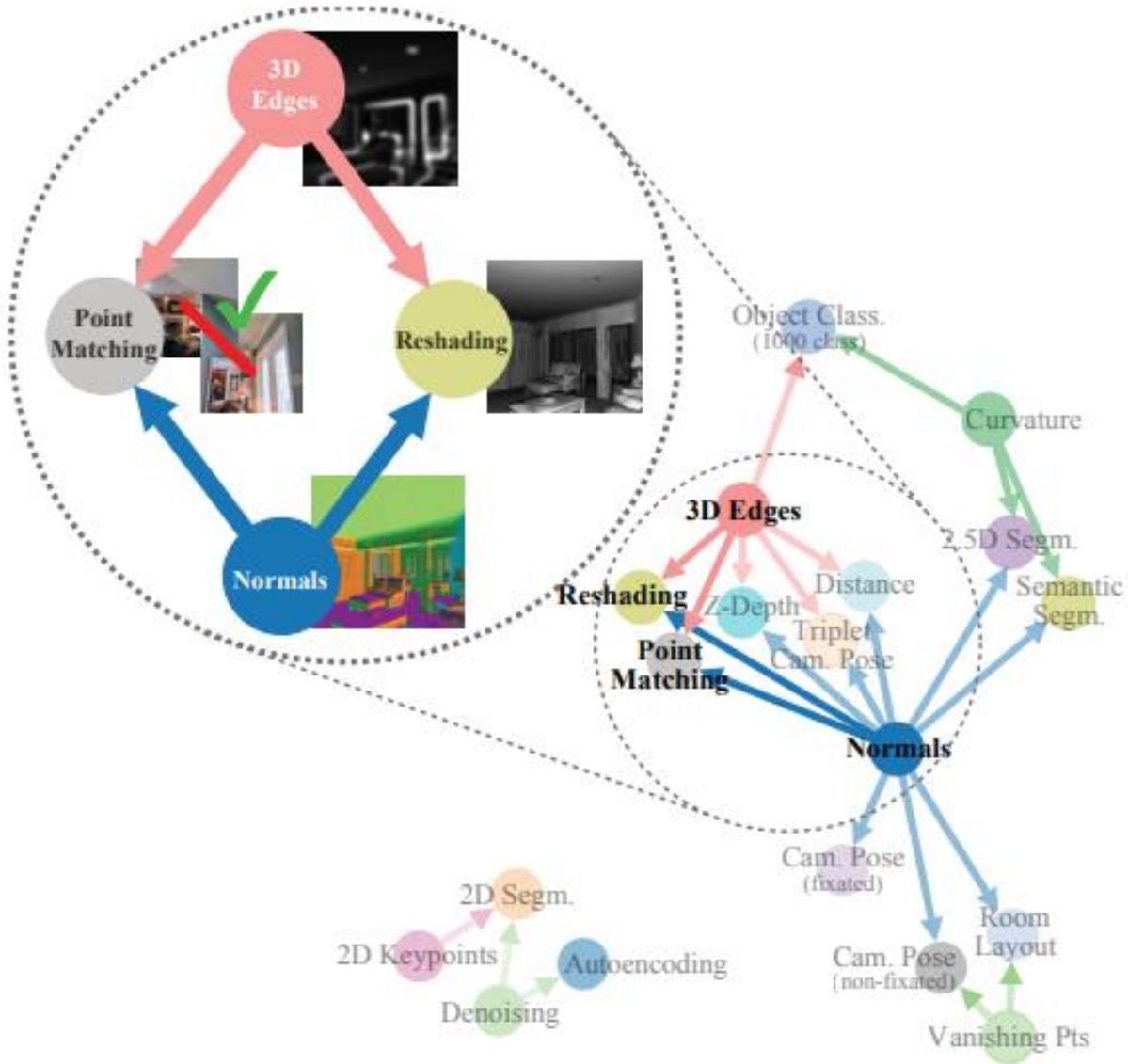
(1) maximize the overall performance of all tasks by using the least annotated data.

(2) quantify the relation of different tasks.

2 Purpose

liberation of alchemist and development of computer vision community

Expected Result



Pipeline

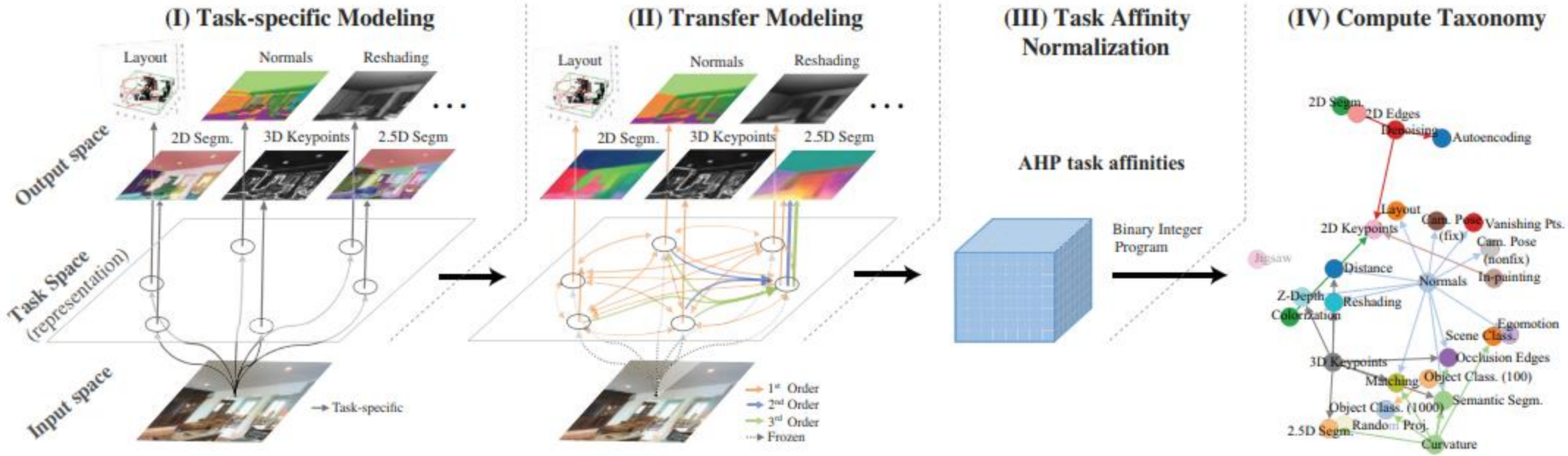


Figure 2: **Computational modeling of task relations and creating the taxonomy.** From left to right: I. Train task-specific networks. II. Train (first order and higher) transfer functions among tasks in a latent space. III. Get normalized transfer affinities using AHP (Analytic Hierarchy Process). IV. Find global transfer taxonomy using BIP (Binary Integer Program).

Problem Description

1 Symbol:

S:source task,

T: target task

2 Task Dictionary:

sample set, not all existing task

3 Dataset:

four million images of indoor scenes from about 600 buildings every image has an annotation for every task. Missing labels are annotated by state-of-art models.

Task-Specific Modeling

train similar network architecture for every task S

Encoder-Decoder have enough express power for task

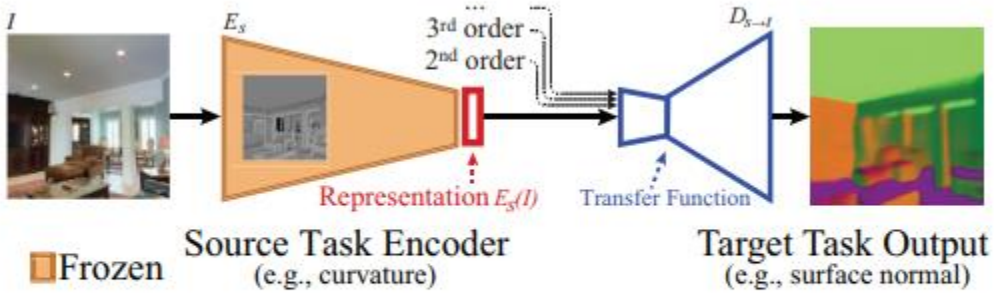


Figure 4: **Transfer Function**. A small readout function is trained to map representations of source task's frozen encoder to target task's labels. If order> 1, transfer function receives representations from multiple sources.

Transfer Modelling

1 Learn readout function

adopt low capacity nn to measure the transferability.

$$D_{s\to t} := \arg\min_{\theta} \mathbb{E}_{I\in\mathcal{D}}\Big[L_t\Big(D_\theta\big(E_s(I)\big), f_t(I)\Big)\Big],$$

2 High Order Transfer

Select top K (K<=5), Beam search (K > 5)

3 Transitive Transfer

Doesn’t work

Ordinal Normalization using AHP

1 different tasks have different loss with vastly different scales

2 simple linear normalize [0, 1] fails because some tasks output quality have relations with loss.

3 AHP

We clip this intermediate pairwise matrix W_t to be in $[0.001, 0.999]$ as a form of Laplace smoothing. Then we divide $W'_t = W_t / W_t^T$ so that the matrix shows how many times better s_i is compared to s_j. The final tournament ratio matrix is positive reciprocal with each element $w'_{i,j}$ of W'_t:

$$w'_{i,j} = \frac{\mathbb{E}_{I \in \mathcal{D}_{test}}[D_{s_i \to t}(I) > D_{s_j \to t}(I)]}{\mathbb{E}_{I \in \mathcal{D}_{test}}[D_{s_i \to t}(I) < D_{s_j \to t}(I)]}. \quad (2)$$

Computing the Global Taxonomy

Problem: Given the normalized task affinity matrix, we need to devise a global transfer policy which maximizes collective performance across all tasks, while minimizing the used supervision.

Our transfers (edges), E, are indexed by i with the form $(\{s_1^i, \ldots, s_{m_i}^i\}, t^i)$ where $\{s_1^i, \ldots, s_{m_i}^i\} \subset \mathcal{S}$ and $t^i \in \mathcal{T}$. We define operators returning target and sources of an edge:

$$\left(\{s_1^i, \ldots, s_{m_i}^i\}, t^i\right) \xmapsto{sources} \{s_1^i, \ldots, s_{m_i}^i\}$$

$$\left(\{s_1^i, \ldots, s_{m_i}^i\}, t^i\right) \xmapsto{target} t^i.$$

Solving a task t by fully supervising it is denoted as $\left(\{t\}, t\right)$. We also index the targets $\mathcal{T}$ with j so that in this section, i is an edge and j is a target.

Computing the Global Taxonomy

Formulation

The BIP is parameterized by a vector x where each transfer and each task is represented by a binary variable; x indicates which nodes are picked to be source and which transfers are selected. The canonical form for a BIP is:

$$\text{maximize } c^T x \,,$$

$$\text{subject to } Ax \preceq b$$

$$\text{and } x \in \{0,1\}^{|E|+|V|} \,.$$

Each element c_i for a transfer is the product of the importance of its target task and its transfer performance:

$$c_i := r_{target(i)} \cdot p_i \,. \tag{3}$$

Computing the Global Taxonomy

Constrain 1: if a transfer is included in the subgraph, all of its source nodes/tasks must be included too

Constraint I: For each row a_i in A we require $a_i \cdot x \leq b_i$, where

$$a_{i,k} = \begin{cases} |sources(i)| & \text{if } k = i \\ -1 & \text{if } (k - |E|) \in sources(i) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$b_i = 0. \quad (5)$$

Computing the Global Taxonomy

Constrain 2: each target task has exactly one transfer in

Constraint II: Via the row $a_{|E|+j}$, we enforce that each target has exactly one transfer:

$$a_{|E|+j,i} := 2 \cdot \mathbb{1}_{\{target(i)=j\}}, \qquad b_{|E|+j} := -1. \tag{6}$$

Constraint III: the solution is enforced to not exceed the

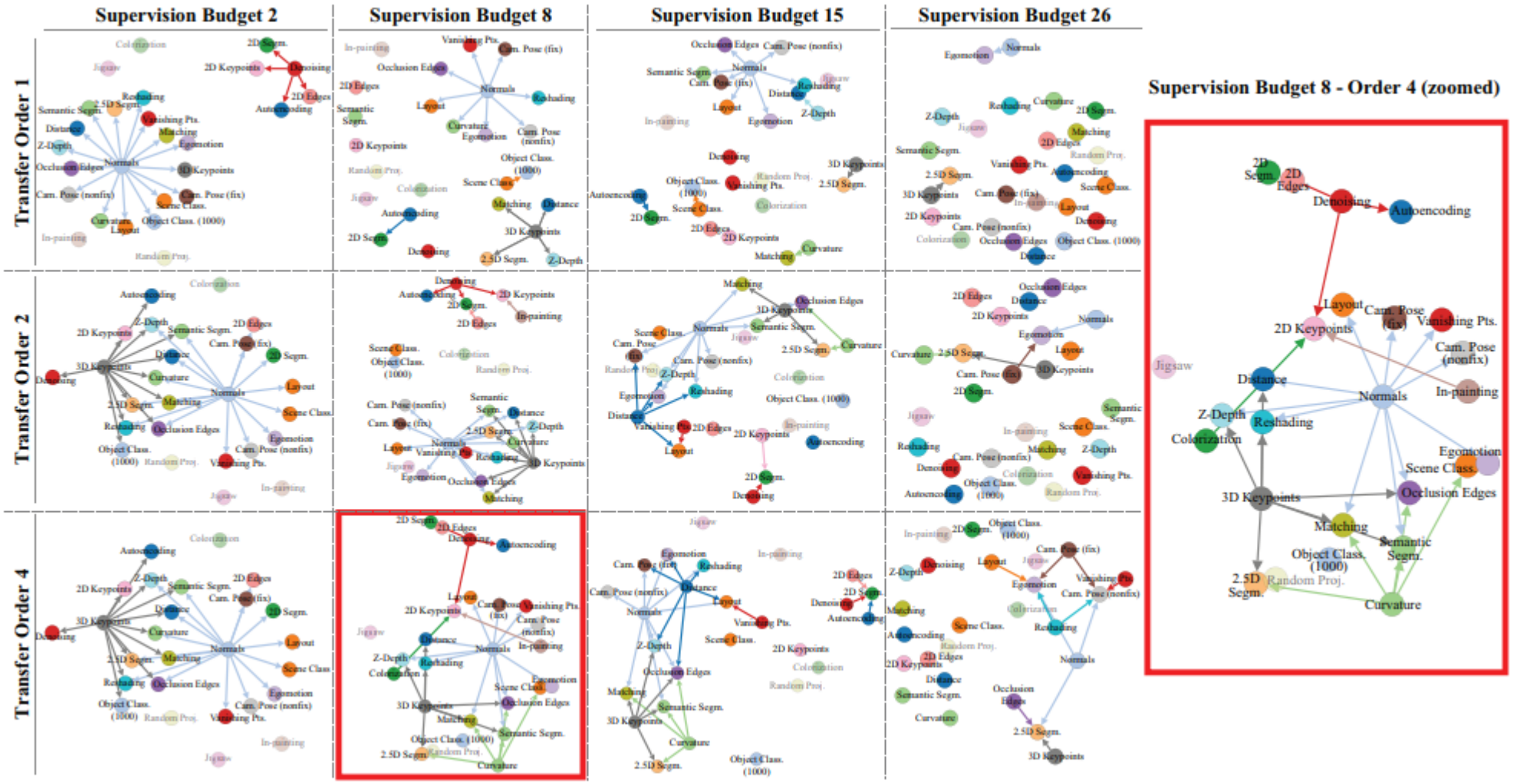
Computing the Global Taxonomy

Constrain 3: supervision budget is not exceeded

Constraint III: the solution is enforced to not exceed the budget. Each transfer i is assigned a label cost ℓ_i, so

$$a_{|E|+|\mathcal{V}|+1,i} := \ell_i, \qquad b_{|E|+|\mathcal{V}|+1} := \gamma. \tag{7}$$

Evaluation of Computed Taxonomies



Evaluation of Computed Taxonomies

How good are the trained task-specific networks? *Win rate (%)* is the proportion of test set images for which a baseline is beaten. Table 1 provides win rates of the task-

Gain: win rate (%) against a network trained from scratch using the same training data as transfer networks'. That is, the best that could be done if transfer learning was not utilized. This quantifies the *gained* value by transferring.
Quality: win rate (%) against a fully supervised network trained with 120k images (gold standard).

Evaluation of Computed Taxonomies

How good are the trained task-specific networks? *Win rate (%)* is the proportion of test set images for which a baseline is beaten. Table 1 provides win rates of the task-

Gain: win rate (%) against a network trained from scratch using the same training data as transfer networks'. That is, the best that could be done if transfer learning was not utilized. This quantifies the *gained* value by transferring.
Quality: win rate (%) against a fully supervised network trained with 120k images (gold standard).

Evaluation of Computed Taxonomies

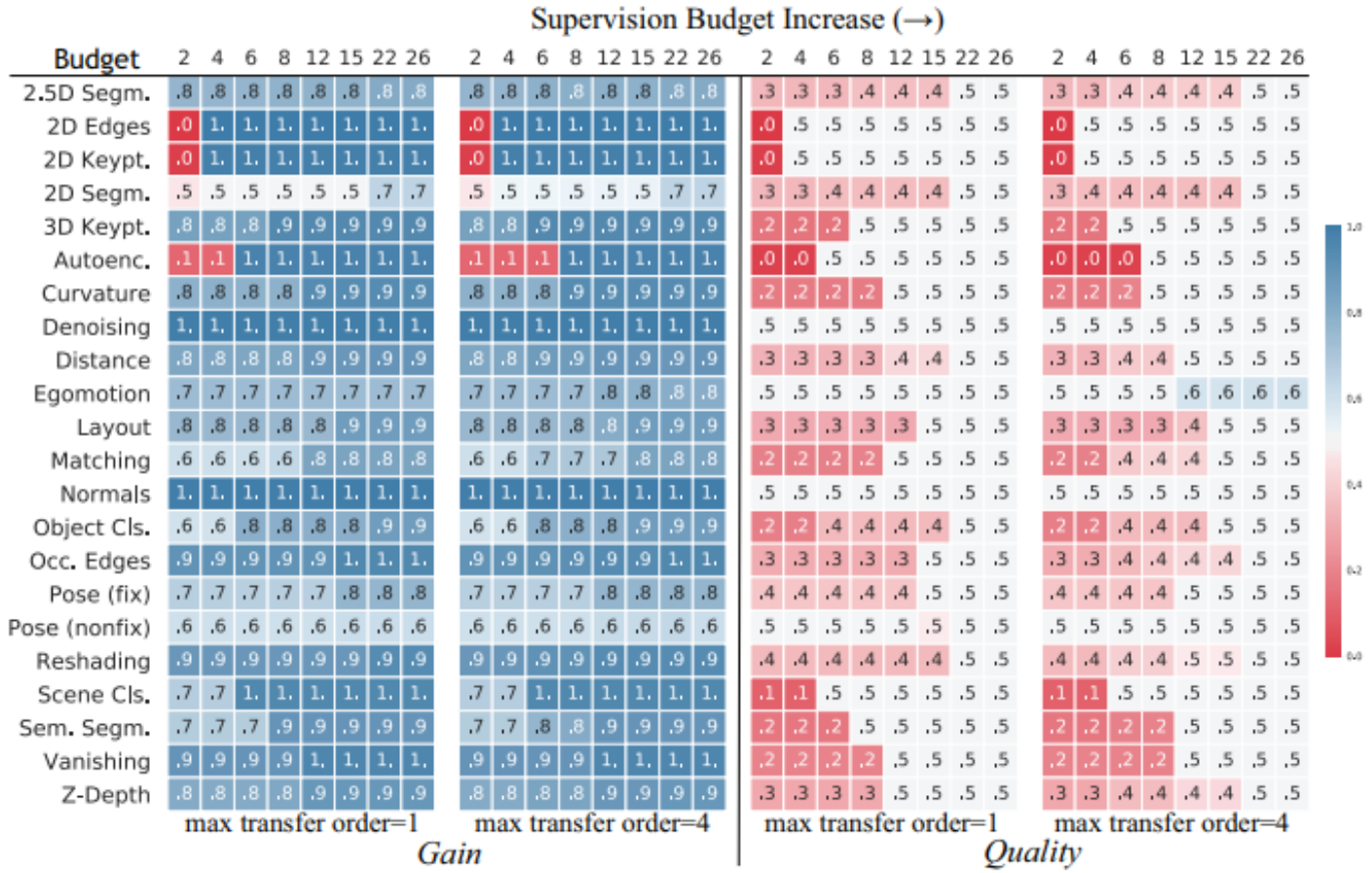
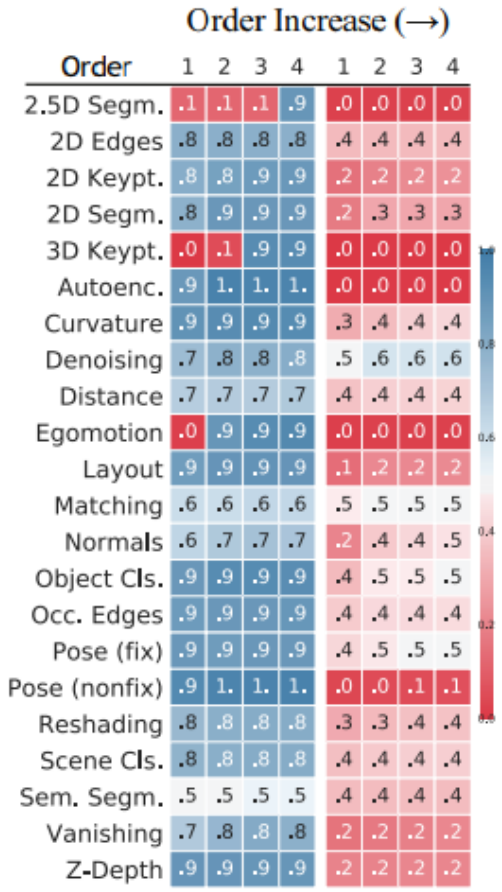


Figure 9: **Evaluation of taxonomy computed for solving the full task dictionary**. Gain (left) and Quality (right) values for each task using the

Evaluation of Computed Taxonomies



Task	scratch	ImageNet[51]	Wang.[96]	Agrawal.[1]	Zamir.[100]	Zhang.[103]	Noroozi.[68]	full sup.	Taxonomy
Depth	88	88	93	89	88	84	86	43	-
	.03	.04	.04	.03	.04	.03	.03	**.02**	.02
Scene Cls.	80	52	83	74	74	71	75	15	-
	3.30	2.76	3.56	3.15	3.17	3.09	3.19	**2.23**	2.63
Sem. Segm.	78	79	82	85	76	78	84	21	-
	1.74	1.88	1.92	1.80	1.85	1.74	1.71	**1.42**	1.53
Object Cls.	79	54	82	76	75	76	76	34	-
	4.08	3.57	4.27	3.99	3.98	4.00	3.97	**3.26**	3.46
Normals	97	98	98	98	98	97	97	6	-
	.22	.30	.34	.28	.28	.23	.24	**.12**	.15
2.5D Segm.	80	93	92	89	90	84	87	40	-
	.21	.34	.34	.26	.29	.22	.24	**.16**	.17
Occ. Edges	93	96	95	93	94	93	94	42	-
	.16	.19	.18	.17	.18	.16	.17	**.12**	.13
Curvature	88	94	89	85	88	92	88	29	-
	.25	.28	.26	.25	.26	.26	.25	**.21**	.22
Egomotion	79	78	83	77	76	74	71	59	-
	8.60	8.58	9.26	8.41	8.34	8.15	7.94	7.32	**6.85**
Layout	80	76	85	79	77	78	70	36	-
	.66	.66	.85	.65	.65	.62	.54	**.37**	.41

Figure 10: **Generalization to Novel Tasks**. Each row shows a novel test task. Left: Gain and Quality values using the devised "all-for-one" transfer policies for novel tasks for orders 1-4. Right: Win rates (%) of the