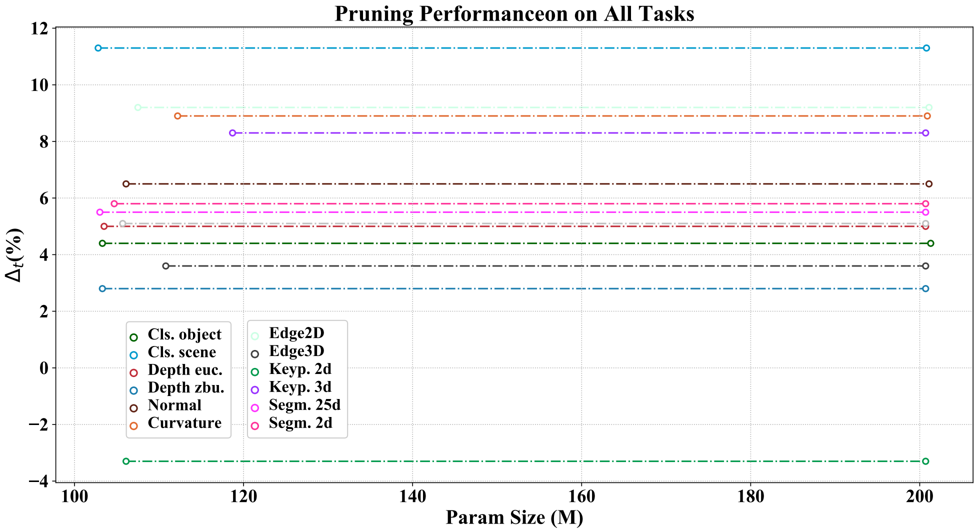
Supplementary Material for
Mod-Squad: Designing Mixtures of Experts As Modular Multi-Task Learners

Table A1. Metric for more tasks on the taskonomy dataset. The experiment section in the paper demonstrates what metric we use for each task.

Model	Depth Zbu.						Edge2D	Keyp.3D	Segm. 2.5D
	$RMSE\downarrow$	Error $\downarrow$		δ, within $\uparrow$			L1 dis. $\downarrow$	L1 dis. $\downarrow$	L1 dis. $\downarrow$
		Abs.	Rel.	1.25	1.25^2	1.25^3			
STL	6.90	0.086	1.71	92.6	97.0	98.7	0.00500	0.072	0.155
MTL	6.75	0.083	1.30	93.1	96.8	98.9	0.00497	0.071	0.152
M^3ViT	6.73	0.081	1.30	**93.4**	97.3	98.9	0.00490	0.070	0.150
Mod-Squad	**6.70**	**0.080**	**1.28**	93.3	**97.5**	**99.1**	**0.00482**	**0.068**	**0.147**

Figure A1. **Visualization of the pruning results on each task.** We set $\theta = 0.1\%$ and do the pruning. Every tasks keep the same performance while reducing the parameters size from over 200M to lower than 120M.



A1. Metric for more task on the Taskonomy.

We provide more results with the task metric in Table. A1. Other tasks are shown in the paper already. Mod-Squad consistently have the best performance on these metrics.

A2. Pruning on different tasks.

We show the pruning results for all tasks on the Taskonomy in Figure. A2. We set $\theta = 0.1\%$ and removing all experts have an activation frequency lower than θ. Every tasks keep the same performance while reducing the parameters size from over 200M to lower than 120M. This demonstrate the effectiveness of Mod-Squad on all tasks.

Table A2. **Ablation study of Top-K on MoE attention layer and MoE MLP layer.**

	FLOPs(G)	Params(M)	Hidden Dim	Δ_t
K=2	5.2	75	192	-0.4
K=4	5.2	75	96	1.3
K=6	5.2	75	64	**3.5**
K=9	5.2	75	42	3.1
K=12	5.2	75	32	2.2

(a) **Ablation on Top-K in the MoE attention layer.**

	FLOPs(G)	Params(M)	Hidden Dim	Δ_t
K=1	5.2	75	1536	3.3
K=2	5.2	75	768	3.3
K=4	5.2	75	384	**3.5**
K=6	5.2	75	256	3.2
K=8	5.2	75	192	3.1

(b) **Ablation on Top-K in the MoE MLP layer.**

Table A3. **Ablation study of expert number E on MoE attention layer and MoE MLP layer.**

	Params(M)	Δ_t
E=6	66	2.0
E=9	69	2.7
E=12	72	3.2
E=15	75	**3.5**
E=18	78	**3.5**

(a) **Ablation on E in the MoE attention layer.**

	Params(M)	Δ_t
E=4	33	2.2
E=8	47	2.7
E=12	61	3.3
E=16	75	**3.5**
E=20	90	**3.5**

(b) **Ablation on E in the MoE MLP layer.**

A3. Ablation on Top-K.

As showin in Table. A2, we explore the effect on Top-K for both MoE attention layer and MoE MLP layer. The experiment setting is the same as in § A4. To control the FLOPs to be the same for different Top-K, the hidden dimension of attention experts and the mlp experts is divided by K. All experiments have the same parameter size and the same FLOPs.

A4. Ablation on number of experts.

As showin in Table. A3, we explore the effect on number of experts E for both MoE attention layer and MoE MLP layer. For quick experiments, we use ViT-small as the backbone network. In default, we add MoE attention layer with 15 experts and Top-K as 6 as well as MoE MLP layer with 16 experts and toop-k as 4. The MoE modules are added at every layer. We use Δ_t to evaluate the multi-task performance and report the parameters size. The Δ_t compare various version of models to the vanilla single task learning baseline with ViT-small. The experiments show that increasing experts number bring extra performance but the gain gradually diminish as the E goes up. This conclusion hold true for both MoE attention layer and MoE MLP layer.

A5. More visualization on Mod-Squad.

We visualize Mod-Squad based on ViT-small as shown in Figure. A2 and Figure. A3. To better understand the relation between experts and task in all layers, we insert MoE attention layer and MoE MLP layer on every transformer block. In Figure. A2, the activation frequencies of MoE attention modules are shown in all transformer blocks with 15 experts and Top-K as 6. In Figure. A3, the activation frequencies of MoE MLP modules are shown in all transformer blocks with 16 experts and Top-K as 4. Both visualization demonstrates the sparsity of Mod-Squad in all layers for all tasks.

A6. Task relation from different layers of Mod-Squad.

In the paper, we define the similarity between tasks as the mean of the percentage of experts that they are sharing given the same input. We put all experts into the calculation of task similarity. However, we can also calculate the task similarity for each layer, by only put the experts from that specific layer into the calculation. As shown in Figure. A4, we visualize the task similarity for the first two layers, the middle two layers, and the last two layers of Mod-Squad. Mod-Squad is trained on the Taskonomy with ViT-base as the backbone and follow the same setting in our Taskonomy experiments in the paper. We notice that (1) all 3d tasks are very similar in the first layer; (2) all tasks are more similar to classification tasks in the middle two layers; (3) there is less similarity between tasks in the last two layers.

Figure A2. **Visualization of the frequency that experts being selected for each task in the MoE attention layer.** We visualize Mod-Squad based on ViT-small. The activation frequencies of MoE attention modules are shown in all transformer blocks with 15 experts and Top-K as 6. The y-axis represents the tasks and the x-axis represents the experts. It demonstrates the sparsity of Mod-Squad in all layers for all tasks.

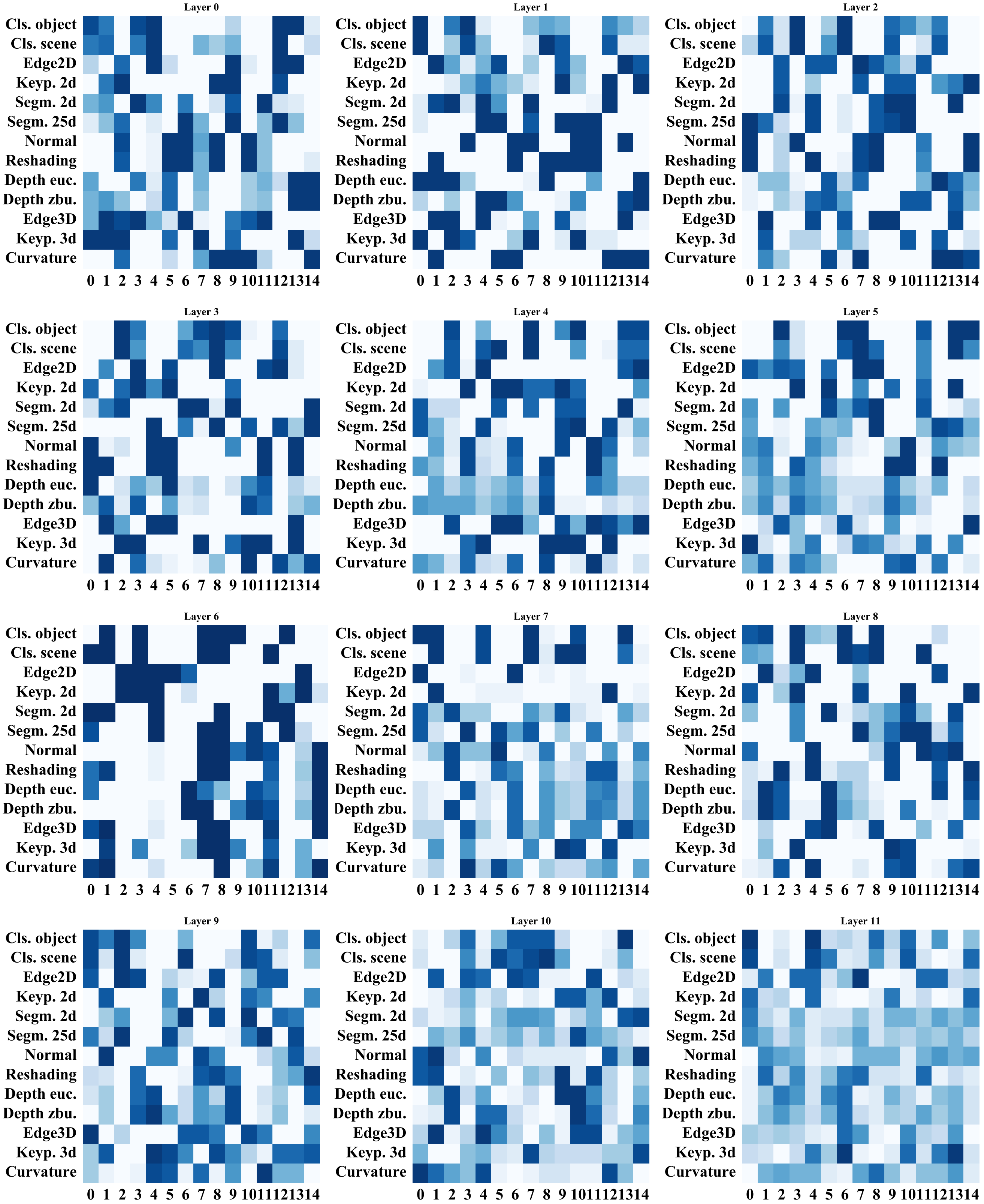


Figure A3. **Visualization of the frequency that experts being selected for each task in the MoE MLP layer.** We visualize Mod-Squad based on ViT-small. The activation frequencies of MoE MLP modules are shown in all transformer blocks with 16 experts and Top-K as 4. The y-axis represents the tasks and the x-axis represents the experts. It demonstrates the sparsity of Mod-Squad in all layers for all tasks.

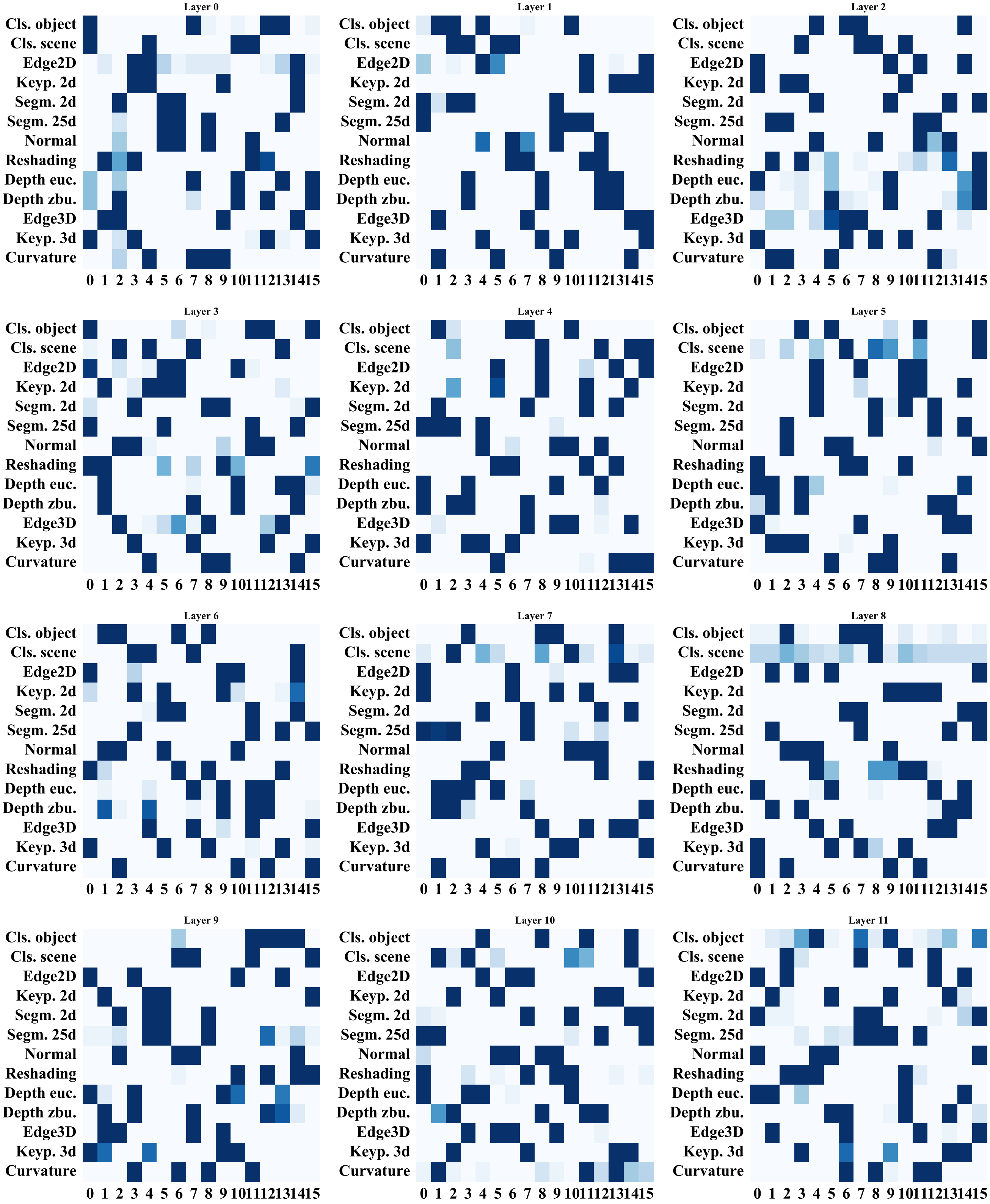
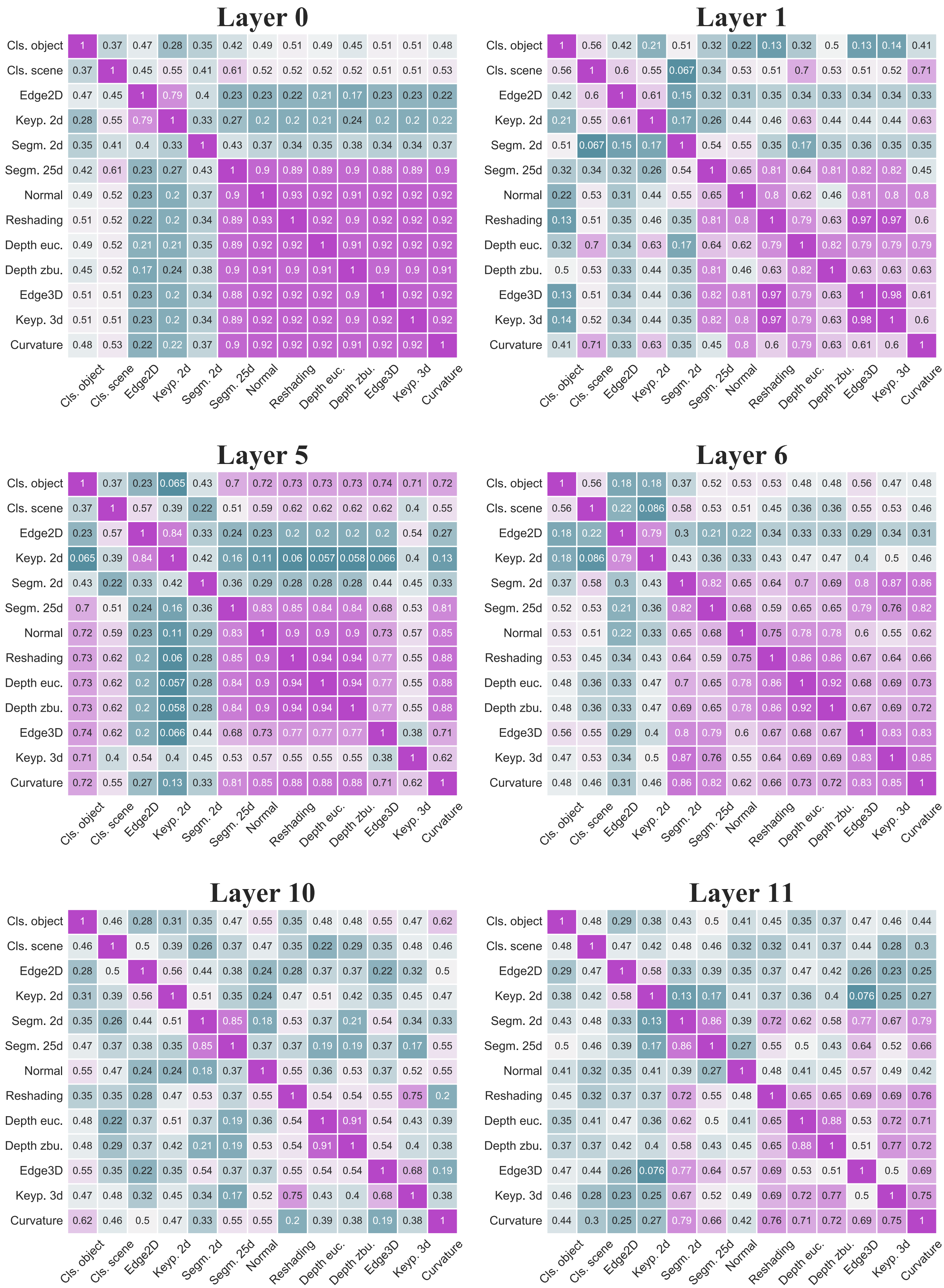


Figure A4. **Visualization of task similarity from the first, the middle, and the last layers.** For each layer L_i, we evaluate the similarity between a task pair as the mean of the percentage of experts in L_i that task pairs are sharing with the same input.



Layer 0

	Cls. object	Cls. scene	Edge2D	Keyp. 2d	Segm. 2d	Segm. 25d	Normal	Reshading	Depth euc.	Depth zbu.	Edge3D	Keyp. 3d	Curvature
Cls. object	1	0.37	0.47	0.28	0.35	0.42	0.49	0.51	0.49	0.45	0.51	0.51	0.48
Cls. scene	0.37	1	0.45	0.55	0.41	0.61	0.52	0.52	0.52	0.52	0.51	0.51	0.53
Edge2D	0.47	0.45	1	0.79	0.4	0.23	0.23	0.22	0.21	0.17	0.23	0.23	0.22
Keyp. 2d	0.28	0.55	0.79	1	0.33	0.27	0.2	0.2	0.21	0.24	0.2	0.2	0.22
Segm. 2d	0.35	0.41	0.4	0.33	1	0.43	0.37	0.34	0.35	0.38	0.34	0.34	0.37
Segm. 25d	0.42	0.61	0.23	0.27	0.43	1	0.9	0.89	0.89	0.9	0.88	0.89	0.9
Normal	0.49	0.52	0.23	0.2	0.37	0.9	1	0.93	0.92	0.91	0.92	0.92	0.92
Reshading	0.51	0.52	0.22	0.2	0.34	0.89	0.93	1	0.92	0.9	0.92	0.92	0.92
Depth euc.	0.49	0.52	0.21	0.21	0.35	0.89	0.92	0.92	1	0.91	0.92	0.92	0.92
Depth zbu.	0.45	0.52	0.17	0.24	0.38	0.9	0.91	0.9	0.91	1	0.9	0.9	0.91
Edge3D	0.51	0.51	0.23	0.2	0.34	0.88	0.92	0.92	0.92	0.9	1	0.92	0.92
Keyp. 3d	0.51	0.51	0.23	0.2	0.34	0.89	0.92	0.92	0.92	0.9	0.92	1	0.92
Curvature	0.48	0.53	0.22	0.22	0.37	0.9	0.92	0.92	0.92	0.91	0.92	0.92	1

Layer 1

	Cls. object	Cls. scene	Edge2D	Keyp. 2d	Segm. 2d	Segm. 25d	Normal	Reshading	Depth euc.	Depth zbu.	Edge3D	Keyp. 3d	Curvature
Cls. object	1	0.56	0.42	0.21	0.51	0.32	0.22	0.13	0.32	0.5	0.13	0.14	0.41
Cls. scene	0.56	1	0.6	0.55	0.067	0.34	0.53	0.51	0.7	0.53	0.51	0.52	0.71
Edge2D	0.42	0.6	1	0.61	0.15	0.32	0.31	0.35	0.34	0.33	0.34	0.34	0.33
Keyp. 2d	0.21	0.55	0.61	1	0.17	0.26	0.44	0.46	0.63	0.44	0.44	0.44	0.63
Segm. 2d	0.51	0.067	0.15	0.17	1	0.54	0.55	0.35	0.17	0.35	0.36	0.35	0.35
Segm. 25d	0.32	0.34	0.32	0.26	0.54	1	0.65	0.81	0.64	0.81	0.82	0.82	0.45
Normal	0.22	0.53	0.31	0.44	0.55	0.65	1	0.8	0.62	0.46	0.81	0.8	0.8
Reshading	0.13	0.51	0.35	0.46	0.35	0.81	0.8	1	0.79	0.63	0.97	0.97	0.6
Depth euc.	0.32	0.7	0.34	0.63	0.17	0.64	0.62	0.79	1	0.82	0.79	0.79	0.79
Depth zbu.	0.5	0.53	0.33	0.44	0.35	0.81	0.46	0.63	0.82	1	0.63	0.63	0.63
Edge3D	0.13	0.51	0.34	0.44	0.36	0.82	0.81	0.97	0.79	0.63	1	0.98	0.61
Keyp. 3d	0.14	0.52	0.34	0.44	0.35	0.82	0.8	0.97	0.79	0.63	0.98	1	0.6
Curvature	0.41	0.71	0.33	0.63	0.35	0.45	0.8	0.6	0.79	0.63	0.61	0.6	1

Layer 5

	Cls. object	Cls. scene	Edge2D	Keyp. 2d	Segm. 2d	Segm. 25d	Normal	Reshading	Depth euc.	Depth zbu.	Edge3D	Keyp. 3d	Curvature
Cls. object	1	0.37	0.23	0.065	0.43	0.7	0.72	0.73	0.73	0.73	0.74	0.71	0.72
Cls. scene	0.37	1	0.57	0.39	0.22	0.51	0.59	0.62	0.62	0.62	0.62	0.4	0.55
Edge2D	0.23	0.57	1	0.84	0.33	0.24	0.23	0.2	0.2	0.2	0.2	0.54	0.27
Keyp. 2d	0.065	0.39	0.84	1	0.42	0.16	0.11	0.06	0.057	0.058	0.066	0.4	0.13
Segm. 2d	0.43	0.22	0.33	0.42	1	0.36	0.29	0.28	0.28	0.28	0.44	0.45	0.33
Segm. 25d	0.7	0.51	0.24	0.16	0.36	1	0.83	0.85	0.84	0.84	0.68	0.53	0.81
Normal	0.72	0.59	0.23	0.11	0.29	0.83	1	0.9	0.9	0.9	0.73	0.57	0.85
Reshading	0.73	0.62	0.2	0.06	0.28	0.85	0.9	1	0.94	0.94	0.77	0.55	0.88
Depth euc.	0.73	0.62	0.2	0.057	0.28	0.84	0.9	0.94	1	0.94	0.77	0.55	0.88
Depth zbu.	0.73	0.62	0.2	0.058	0.28	0.84	0.9	0.94	0.94	1	0.77	0.55	0.88
Edge3D	0.74	0.62	0.2	0.066	0.44	0.68	0.73	0.77	0.77	0.77	1	0.38	0.71
Keyp. 3d	0.71	0.4	0.54	0.4	0.45	0.53	0.57	0.55	0.55	0.55	0.38	1	0.62
Curvature	0.72	0.55	0.27	0.13	0.33	0.81	0.85	0.88	0.88	0.88	0.71	0.62	1

Layer 6

	Cls. object	Cls. scene	Edge2D	Keyp. 2d	Segm. 2d	Segm. 25d	Normal	Reshading	Depth euc.	Depth zbu.	Edge3D	Keyp. 3d	Curvature
Cls. object	1	0.56	0.18	0.18	0.37	0.52	0.53	0.53	0.48	0.48	0.56	0.47	0.48
Cls. scene	0.56	1	0.22	0.086	0.58	0.53	0.51	0.45	0.36	0.36	0.55	0.53	0.46
Edge2D	0.18	0.22	1	0.79	0.3	0.21	0.22	0.34	0.33	0.33	0.29	0.34	0.31
Keyp. 2d	0.18	0.086	0.79	1	0.43	0.36	0.33	0.43	0.47	0.47	0.4	0.5	0.46
Segm. 2d	0.37	0.58	0.3	0.43	1	0.82	0.65	0.64	0.7	0.69	0.8	0.87	0.86
Segm. 25d	0.52	0.53	0.21	0.36	0.82	1	0.68	0.59	0.65	0.65	0.79	0.76	0.82
Normal	0.53	0.51	0.22	0.33	0.65	0.68	1	0.75	0.78	0.78	0.6	0.55	0.62
Reshading	0.53	0.45	0.34	0.43	0.64	0.59	0.75	1	0.86	0.86	0.67	0.64	0.66
Depth euc.	0.48	0.36	0.33	0.47	0.7	0.65	0.78	0.86	1	0.92	0.68	0.69	0.73
Depth zbu.	0.48	0.36	0.33	0.47	0.69	0.65	0.78	0.86	0.92	1	0.67	0.69	0.72
Edge3D	0.56	0.55	0.29	0.4	0.8	0.79	0.6	0.67	0.68	0.67	1	0.83	0.83
Keyp. 3d	0.47	0.53	0.34	0.5	0.87	0.76	0.55	0.64	0.69	0.69	0.83	1	0.85
Curvature	0.48	0.46	0.31	0.46	0.86	0.82	0.62	0.66	0.73	0.72	0.83	0.85	1

Layer 10

	Cls. object	Cls. scene	Edge2D	Keyp. 2d	Segm. 2d	Segm. 25d	Normal	Reshading	Depth euc.	Depth zbu.	Edge3D	Keyp. 3d	Curvature
Cls. object	1	0.46	0.28	0.31	0.35	0.47	0.55	0.35	0.48	0.48	0.55	0.47	0.62
Cls. scene	0.46	1	0.5	0.39	0.26	0.37	0.47	0.35	0.22	0.29	0.35	0.48	0.46
Edge2D	0.28	0.5	1	0.56	0.44	0.38	0.24	0.28	0.37	0.37	0.22	0.32	0.5
Keyp. 2d	0.31	0.39	0.56	1	0.51	0.35	0.24	0.47	0.51	0.42	0.35	0.45	0.47
Segm. 2d	0.35	0.26	0.44	0.51	1	0.85	0.18	0.53	0.37	0.21	0.54	0.34	0.33
Segm. 25d	0.47	0.37	0.38	0.35	0.85	1	0.37	0.37	0.19	0.19	0.37	0.17	0.55
Normal	0.55	0.47	0.24	0.24	0.18	0.37	1	0.55	0.36	0.53	0.37	0.52	0.55
Reshading	0.35	0.35	0.28	0.47	0.53	0.37	0.55	1	0.54	0.54	0.55	0.75	0.2
Depth euc.	0.48	0.22	0.37	0.51	0.37	0.19	0.36	0.54	1	0.91	0.54	0.43	0.39
Depth zbu.	0.48	0.29	0.37	0.42	0.21	0.19	0.53	0.54	0.91	1	0.54	0.4	0.38
Edge3D	0.55	0.35	0.22	0.35	0.54	0.37	0.37	0.55	0.54	0.54	1	0.68	0.19
Keyp. 3d	0.47	0.48	0.32	0.45	0.34	0.17	0.52	0.75	0.43	0.4	0.68	1	0.38
Curvature	0.62	0.46	0.5	0.47	0.33	0.55	0.55	0.2	0.39	0.38	0.19	0.38	1

Layer 11

	Cls. object	Cls. scene	Edge2D	Keyp. 2d	Segm. 2d	Segm. 25d	Normal	Reshading	Depth euc.	Depth zbu.	Edge3D	Keyp. 3d	Curvature
Cls. object	1	0.48	0.29	0.38	0.43	0.5	0.41	0.45	0.35	0.37	0.47	0.46	0.44
Cls. scene	0.48	1	0.47	0.42	0.48	0.46	0.32	0.32	0.41	0.37	0.44	0.28	0.3
Edge2D	0.29	0.47	1	0.58	0.33	0.39	0.35	0.37	0.47	0.42	0.26	0.23	0.25
Keyp. 2d	0.38	0.42	0.58	1	0.13	0.17	0.41	0.37	0.36	0.4	0.076	0.25	0.27
Segm. 2d	0.43	0.48	0.33	0.13	1	0.86	0.39	0.72	0.62	0.58	0.77	0.67	0.79
Segm. 25d	0.5	0.46	0.39	0.17	0.86	1	0.27	0.55	0.5	0.43	0.64	0.52	0.66
Normal	0.41	0.32	0.35	0.41	0.39	0.27	1	0.48	0.41	0.45	0.57	0.49	0.42
Reshading	0.45	0.32	0.37	0.37	0.72	0.55	0.48	1	0.65	0.65	0.69	0.69	0.76
Depth euc.	0.35	0.41	0.47	0.36	0.62	0.5	0.41	0.65	1	0.88	0.53	0.72	0.71
Depth zbu.	0.37	0.37	0.42	0.4	0.58	0.43	0.45	0.65	0.88	1	0.51	0.77	0.72
Edge3D	0.47	0.44	0.26	0.076	0.77	0.64	0.57	0.69	0.53	0.51	1	0.5	0.69
Keyp. 3d	0.46	0.28	0.23	0.25	0.67	0.52	0.49	0.69	0.72	0.77	0.5	1	0.75
Curvature	0.44	0.3	0.25	0.27	0.79	0.66	0.42	0.76	0.71	0.72	0.69	0.75	1