Definition of Tasks in Taskonomy Task Bank

Task definitions are presented in alphabetical order:

Autoencoding PCA is a widely used method of understanding data by finding a low-dimensional latent representation. Autoencoding [12] is a nonlinear generalization of PCA that was originally proposed with transfer learning in mind: better downstream performance through autoencoding pretraining.

Colorization [43] Colorization requires taking a grayscale image and predicting the original colors. It is an unsupervised task, but also one that is semantic-aware[5]. For example, predicting the color of a fruit is simple once the fruit is identified.

Context Encoding Context Encoding was first introduced by Pathak et al. [26] and is a version of autoencoding where a large portion of the input is masked from the model. In order to fill in the occluded area, the model must reason about the scene geometry and semantics. Similar to colorization, it is an unsupervised-yet-still-semantic task.

Content Prediction (Jigsaw)[25] A discriminative version of Context Encoding, Jigsaw [25] requires a network to unscramble a permuted tiling of the input image.

Curvature Estimation Curvature-based features are excellent for identification because they are invariant under rigid transformations. Curvature is known to be important in visual processing–so much so that the Macaque visual cortex has a dedicated curvature processing region [41].

Denoising It is desirable for similar inputs to have similar representations, but representations learned by autoencoding are excessively sensitive to perturbations in the input. Denoising [36] (autoencoding) encourages limited invariance by mapping slightly perturbed inputs to the unperturbed input.

Depth Estimation, Euclidean Depth estimation is an important task, useful for detecting proximity to obstacles and items of interest. It is also a useful intermediate step for agents to localize themselves in 3D space. Euclidean depth is measured as the distance from each pixel to the camera's optical center.

Depth Estimation, Z-Buffer As opposed to Euclidean depth estimation, researchers typically use z-buffer depth which is defined as the distance to the camera plane. This is not the way that humans typically perceive depth, but is included because this is the standard formulation and all of our depth-derived tasks are derived from *z-buffer*.

Edge Detection (2D) Edge detection is historically a fundamental task in computer vision. Edges are commonly used as an intermediate representation or as a feature in a larger processing pipeline. We include the output of a Canny edge detector without nonmax suppression (to make the task learnable by neural networks).

Edge Detection (3D) As opposed to 2D edges, we define 3D edges as "occlusion edges," or edges where an object in the foreground obscures something behind it. 2D edges respond to changes in texture, but 3D edges are features which depend only on the 3D geometry and are invariant to color and lighting.

Keypoint Detection (2D) Keypoint detection has a long history in computer vision, and is useful for many, many tasks. Keypoint algorithmns usually consist of two parts, both a keypoint detector and some local patch descriptor which is invariant across multiple images [20, 3, 28]. 2D keypoint detection encourages the network to identify locally important regions of an image, and *point matching* encourages the network to learn feature descriptors. Identifying keypoints is frequently still a first step in a larger visual pipeline. We use the output of SURF [3] (before nonmax suppression) as our ground-truth.

Keypoint Detection (3D) 3D keypoints are similar to 2D keypoints except that they are derived from 3D data and therefore account for scene geometry. They are often invariant to informative (but possibly distracting) features such as textures [44, 35, 21, 42, 14]. We use the output of the NARF [35] algorithm (before nonmax suppression) as our 3D keypoint ground-truth.

Point Matching Deep networks trained for point matching learn feature descriptors that prove useful for downstream tasks. Point matching has applications in fine-grained classification [39] and object recognition [19], multi-view reconstruction [32] and structure from motion [23], wide baseline matching [37], SLAM [31] and visual odometry[46].

Relative Camera Pose Estimation, Non-Fixated The famous "Kitten Carousel" experiment by Held and Hein [10] suggested that taking action is crucial for strong perception. Although more recent works call the original conclusion into question [27], the ability to localize oneself remains important for locomotion. For two different views with the same optical centers, we try to predict the 6-DOF

relative camera pose (yaw, pitch, roll, xyz translation) between them.

Relative Camera Pose Estimation, Fixated We also include a simpler variant of camera pose estimation for which the center pixel of the two inputs is always the same physical 3D point. This problem is simpler in the sense that there are only five degrees of freedom.

Relative Camera Pose Estimation, Triplets (Egomotion) Videos are a common object of study in computer vision (e.g. visual odometry [8, 24]) and they provide dense data with high redundancy. We therefore include camera pose matching for input triplets with a fixed center point. With three images, models have a greater ability to match points for accurate localization.

Reshading One way to infer scene geometry is "shape from shading" [2] using the intrinsic image decomposition $I = A \cdot S$, where S is a shading function parameterized by lighting and depth. This decomposition is thought to be useful in human visual perception [1]. We define reshading as follows: Given an RGB image, the label is the shading function S that results from having a single point light at the camera origin, and S is multiplied by a constant fixed albedo.

Room Layout Estimation Estimating and aligning a 3D bounding box is a mid-level task that includes vanishing point estimation as a sub-problem, and has applications for robotic navigation [34], scene reconstruction [13], and augmented reality [7]. A variant of room layout estimation was used in the LSUN room layout challenge [40], but that formulation is ill-posed when there is camera roll or when no room corners are in view. Instead, we offer a formulation that remains well-defined regardless of the camera pose and field of view. This task includes some semantic information such as 'what constitutes a room' while simultaneously also including scene geometry.

Segmentation, Unsupervised (2D)) Gestalt psychologists proposed principles of grouping as a mechanism through which humans learn to perceive the world as a set of coherent objects [38]. Normalized cuts [33] are one method for segmenting images into perceptually similar groups, and we include this Gestalt task in our dictionary.

Segmentation, Unsupervised (2.5D)) Segmentation 2.5D uses the same algorithm as 2D, but the labels are computed jointly from the RGB image, the aligned depth image, and the aligned surface normals image. Therefore the 2.5D segmentation applies the principles not just to the world as it seems (in the RGB image), but also to the world as it is (ground-truth 3D). 2.5D segmentation incorporates information about the scene geometry that is not directly present in the RGB image but that is readily inferred by humans.

Surface Normal Estimation Surface normal estimation is thought to be crucial for spatial cognition. For example, objects can only be placed on surfaces with upwards-facing normals. Even for locomotion, a point with horizontal-facing normals indicates that it cannot be easily traversed. Surface normals are computed directly from the 3D mesh.

Vanishing Point Estimation An consequence of perspective, vanishing points offer useful information about the scene geometry [22, 15] and are well studied. Vanishing points prove particularly useful in a Manhattan world [6, 43, 4] where there are three dominant vanishing points corresponding to an X, Y, and Z axis. This assumption is usually met in urban environments. For each model we analytically find these three vanishing points and include them as labels.

Semantic Learning through Knowledge distillation While our dataset does not include semantic annotations, semantic understanding comprises a large and important component of modern computer vision. Therefore, we add pseudo-semantic annotations through knowledge distillation [11]. We distill knowledge from state-of-the-art models [9, 16] trained on ImageNet [29] and MS-COCO [17] by using them to annotate our dataset and then supervising models with those annotations. See section 1 for the details of knowledge distillation process.

1. **Classification, Semantic (1000-classes)** Semantic object recognition is a fundamental component of visual perception. Children learn at an early age to classify objects after seeing just a handful of examples [**?**]. For semantic classification, we distill knowledge form a pretrained ResNet-152 [9] (trained on ImageNet) by supervising our model with the ResNet activations.
2. **Classification Semantic (100-classes)** Many of the classes in ImageNet never appear in our dataset (e.g. animals, sports). We therefore manually select classes which appear in our dataset – and this happened to be 100 classes. For 100-way classification, we distill knowledge from these 100 activations only.
3. **Segmentation, Semantic** Models and agents need to know more than just what they are looking at – they also need to be able to locate the object. Therefore, we include knowledge distillation from the semantic segmentation model in [16], trained on MS COCO. [30]

1. Pseudo-semantics Annotations

As we do not have semantic annotations on our dataset, we gathered pseudo-semantic annotations by using knowledge distillation [11] approach. That is, we labeled our dataset using the output of state-of-the-art large fully-supervised network (ResNet-151) for semantic objects (using ImageNet, 100 applicable classes), scene categories (using MIT Places [45], 63 indoor workplace and home classes in MIT Places scene hierarchy), and semantic object segmentation (using COCO [18], 17

applicable classes). In a user study, human annotators indicated that only 6.3% of images have no plausible label in their top-5 classes in Scene categories, showing that such labels are reliable enough and effective for being a gateway to modeling semantics in task space. The quality of the semantic labels can be seen in the provided sample of qualitative results or in the video. The list of selected classes for COCO and ImageNet are available in https://github.com/StanfordVL/taskonomy/tree/master/taskbank/assets/web_assets/pseudosemantics.

References

[1] E. H. Adelson and A. P. Pentland. The perception of shading and reflectance. *Perception as Baysian Inference*, pages pp. 409–423, 1996. 2

[2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, Aug 2015. 2

[3] H. Bay, T. Tuytelaars, and L. Van Gool. *SURF: Speeded Up Robust Features*, pages 404–417. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. 1

[4] J. C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, and M. Pollefeys. Globally optimal line clustering and vanishing point estimation in manhattan world. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 638–645, June 2012. 2

[5] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin. Semantic colorization with internet images. *ACM Trans. Graph.*, 30(6):156:1–156:8, Dec. 2011. 1

[6] J. M. Coughlan and A. L. Yuille. The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 845–851. MIT Press, 2001. 2

[7] D. DeTone, T. Malisiewicz, and A. Rabinovich. Deep image homography estimation. *CoRR*, abs/1606.03798, 2016. 2

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2

[10] R. Held and A. Hein. Movement-produced stimulation in the development of visually guided behavior. *J Comp Physiol Psychol*, pages 872–876. 1

[11] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[12] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 1

[13] H. Izadinia, Q. Shan, and S. M. Seitz. IM2CAD. *CoRR*, abs/1608.05137, 2016. 2

[14] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. Van Gool. *Hough Transform and 3DSURF for Robust ThreeDimensional Classification*, pages 589–602. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 1

[15] H. Kong, J. Y. Audibert, and J. Ponce. Vanishing point detection for road detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 96–103, June 2009. 2

[16] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. *arXiv preprint arXiv:1611.07709*, 2016. 2

[17] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2

[18] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2

[19] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999. 1

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 1

[21] A. Mian, M. Bennamoun, and R. Owens. On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision*, 89(2):348–361, Sep 2010. 1

[22] O. Miksik. Rapid vanishing point estimation for general road detection. In *2012 IEEE International Conference on Robotics and Automation*, pages 4844–4849, May 2012. 2

[23] N. D. Molton, A. J. Davison, and I. D. Reid. Locally planar patch features for real-time structure from motion. In *Proc. British Machine Vision Conference*. BMVC, Sept. 2004. (To appear). 1

[24] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–652–I–659 Vol.1, June 2004. 2

[25] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 1

[26] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. 2016. 1

[27] N. Rader, M. Bausano, and J. E. Richards. On the nature of the visual-cliff-avoidance response in human infants. *Child Development*, 51(1):61–68, 1980. 1

[28] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society. 1

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[30] A. Sax*, W. B. Shen*, A. R. Zamir*, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. *CVPR*, 2017. submitted. 2

[31] S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks, 2002. 1

[32] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528, June 2006. 1

[33] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, Aug. 2000. 2

[34] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. *Indoor Segmentation and Support Inference from RGBD Images*, pages 746–760. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 2

[35] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard. Narf: 3d range image features for object recognition. 1

[36] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM. 1

[37] L. Wang, U. Neumann, and S. You. Wide-baseline image matching using line signatures. In *ICCV*, pages 1311–1318. IEEE Computer Society, 2009. 1

[38] M. Wertheimer. Laws of organization in perceptual forms. *Psycologische Forschung*, 4:301–350, 1923. 2

[39] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3466–3473, June 2012. 1

[40] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2

[41] X. Yue, I. S. Pourladian, R. B. H. Tootell, and L. G. Ungerleider. Curvature-processing network in macaque visual cortex. *Proceedings of the National Academy of Sciences*, 111(33):E3467–E3475, 2014. 1

[42] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud. Surface feature detection and description with applications to mesh matching. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 373–380, June 2009. 1

[43] L. Zhang, H. Lu, X. Hu, and R. Koch. Vanishing point estimation and line classification in a manhattan world with a unifying camera model. *International Journal of Computer Vision*, 117(2):111–130, Apr 2016. 1, 2

[44] Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 689–696, Sept 2009. 1

[45] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc., 2014. 2

[46] Z. Zhu, T. Oskiper, S. Samarasekera, R. Kumar, and H. S. Sawhney. Ten-fold improvement in visual odometry using landmark matching. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007. 1