Cerberus Transformer: Joint Semantic, Affordance and Attribute Parsing

Xiaoxue Chen[1], Tianyu Liu[2], Hao Zhao[3,4], Guyue Zhou[1], Ya-Qin Zhang[1]
[1]AIR, Tsinghua University [2]HKUST [3]Peking University [4]Intel Labs
{chenxiaoxue, zhouguyue, zhangyaqin}@air.tsinghua.edu.cn
tianyu.liu@connect.ust.hk, zhao-hao@pku.edu.cn, hao.zhao@intel.com

Abstract

Multi-task indoor scene understanding is widely considered as an intriguing formulation, as the affinity of different tasks may lead to improved performance. In this paper, we tackle the new problem of joint semantic, affordance and attribute parsing. However, successfully resolving it requires a model to capture long-range dependency, learn from weakly aligned data and properly balance sub-tasks during training. To this end, we propose an attention-based architecture named Cerberus and a tailored training framework. Our method effectively addresses aforementioned challenges and achieves state-of-the-art performance on all three tasks. Moreover, an in-depth analysis shows concept affinity consistent with human cognition, which inspires us to explore the possibility of weakly supervised learning. Surprisingly, Cerberus achieves strong results using only $0.1\% - 1\%$ *annotation. Visualizations further confirm that this success is credited to common attention maps across tasks. Code and models can be accessed at* https://github.com/OPEN-AIR-SUN/Cerberus.

1. Introduction

Understanding indoor scenes is a fundamental computer vision topic, with many applications in intelligent robots and metaverse. To achieve a holistic understanding, many sub-tasks need to be addressed and it is widely believed and evidenced that jointly addressing them lead to more accurate results [46] [9] [35] [40] [10]. Different from former arts, we study a new and challenging formulation: joint semantic, affordance, and attribute parsing from a single image. As shown in Fig. 1, these three tasks cover a wide spectrum of human recognition and cognition abilities. The attribute of an object (like *wood* or *Glossy*) is a low-level physical property. The semantic category of a region (like *floor* or *sofa*) is a recognition-level concept. Affordance prediction (like *movable* or *walkable*) is a cognition-level problem. These three tasks are closely associated, since objects with specific semantics tend to have specific attribute or affordance. Parsing them jointly is a natural yet unexplored formulation.

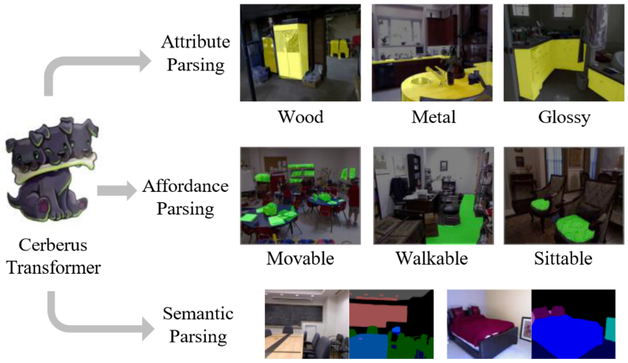


Figure 1. **Cerberus Transformer**. Given a single image, Cerberus parses attribute, affordance and semantics simultaneously. The cartoon is credited to https://www.redbubble.com/i/sticker/Baby-Cerberus-by-ArtOfBianca/48150266.EJUG5.

This new formulation brings both challenges and opportunities. In order to resolve three tasks with a single model, we need to learn shared representations that effectively serve all of them. Meanwhile, the representations are expected to model long range dependency in inputs in a principled manner. In order to simultaneously meet these two requirements, we resort to the transformer architecture [36], which has a global receptive field at each layer. The proposed architecture is named as Cerberus.

Our formulation is challenged by another uncommon issue: weakly aligned data. During the historical development of scene understanding techniques, attribute [47] and affordance [27] annotations are gradually added to the original NYUd2 semantic parsing dataset [34]. Unfortunately, their image-annotation pairs are only weakly aligned in the spatial domain. This is in contrast to former multi-task scene understanding methods which exploit aligned *one-input-multi-output* datasets. To this end, we develop a tailored training framework that treats three datasets as separate sources and leverages a gradient projection technique on pre-computed task-wise gradient tensors. It unleashes the power of multi-task learning and boosts the quantitative results of all three tasks to a state-of-the-art level.

As mentioned before, opportunities come along with challenges. We first conduct in-depth analyses to investigate concept affinity in our three tasks. Interestingly, we observe concept affinity matrices well aligned with human cognitive commonsense. For example, if a pixel is predicted as *floor*, naturally it should be labelled as *walkable*. This finding inspires us to leverage task affinity for weakly supervised learning. During the training of Cerberus, we reduce the annotation amount of a specific sub-task to only $0.1\% - 1\%$ and rely upon representations learnt by other sub-tasks. It is shown that Ceberus consistently out-performs baselines by significant margins in these settings. What's more, we visualize attention maps and validate that the capability of weakly-supervised learning is indeed enabled by shared attention weights. We argue this is a human-like learning feature: If one (e.g., an infant) knows what is floor, then she can learns where is walkable using very few examples.

We have following contributions: (1) We propose a novel multi-task dense prediction transformer named Cerberus, for joint semantic, affordance and attribute parsing in indoor scenes; (2) Cerberus achieves state-of-the-art results for all three tasks while requiring a single forward pass, with the help of a task weight balancing framework that learns from weakly-aligned data. (3) Via extensive analyses, we show that Cerberus learns task affinity consistent with human cognition and achieves strong weakly-supervised learning performance using only 0.1% annotation.

2. Related Works

Transformer [36] has transformed natural language processing since its advent. Due to its strong power to model long-range dependency and capture contextual information, transformer has been proven effective for both 2D [18] [26] and 3D [19] [4] scene understanding problems. Apart from this established advantage, we think transformer is well suited for another potential scenario: multi-task dense prediction. The intuition is that related tasks naturally share attention weights, e.g., *floor* and *walkable*. Interestingly, we validate this point using both strong weakly-supervised learning results and intuitive visualizations.

Scene understanding has long been addresed in a multi-task setting, even before the advent of deep learning. A joint probabilistic formulation can incorporate priors and allow physically more plausible understanding [11] [30] [6]. Incorporating deep representations leads to compelling holistic understanding capabilities including layout, object and human [5] [21] [41]. Semantic scene completion naturally entangles reconstruction and semantic labelling [35] [42] [43] [2]. [44] exploits semantics-layout concept affinity for effective representation learning from unbalanced data. [39] [20] demonstrate improved robustness of deep models via exploiting multi-task consistency. While semantics, affordance, and attribute serve as three fundamental tasks in scene understanding, previous works [22] [15] [24] [8] [29] [33] [16] [14] [28] address them separately. To our knowledge, Cerberus for the first time addresses joint semantic, affordance, and attribute parsing in this large literature. New challenges addressed and new opportunities captured, as mentioned above, distinguish this study from former ones.

3. Method

In this paper, we aim at parsing semantics, affordance and attribute jointly. Semantics (e.g., *sofa* or *cabinet*) describes object/stuff categories in an indoor scene. Affordance means an object's capability to support a certain human action, for instance *walkable* or *sittable*. And attribute refers to object material like *metal* or surface properties like *shiny*. Via predicting these labels, an agent understands an indoor scene in a comprehensive manner. We define $\mathcal{O} = \{o_1, o_2, ..., o_x\}$ as the semantic label set, $\mathcal{F} = \{f_1, f_2, ..., f_y\}$ as the affordance label set, and $\mathcal{T} = \{t_1, t_2, ..., t_z\}$ as the attribute label set. Given an image I, for each pixel $\mathrm{I_i}$, the task is formally stated as a mapping

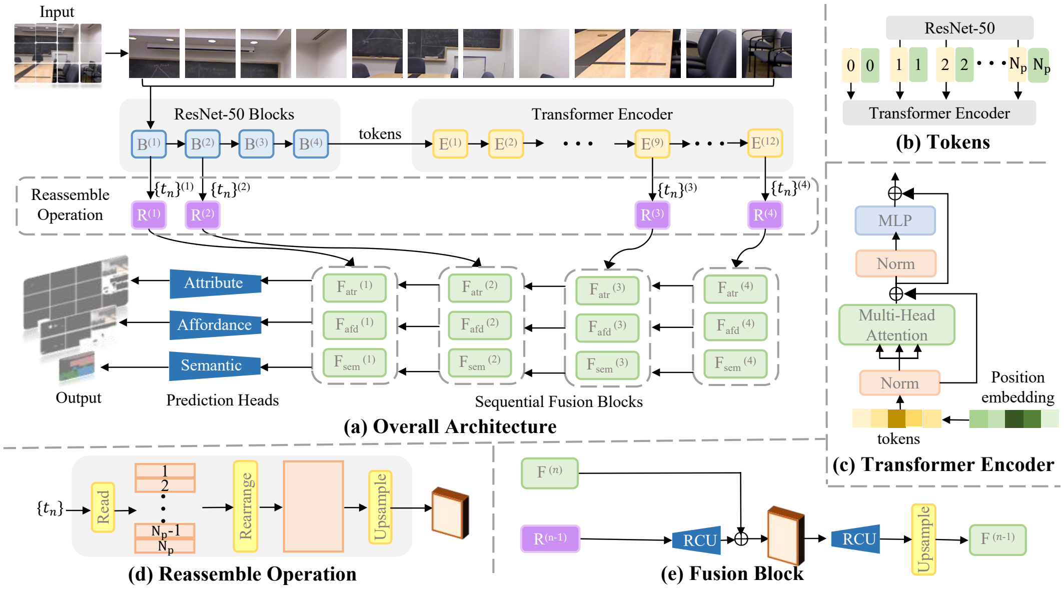
$$\mathrm{I_i} \rightarrow \mathcal{O} \times \mathcal{P}(\mathcal{F}) \times \mathcal{P}(\mathcal{T}) \tag{1}$$

where $\mathcal{P}$ is the power-set operator, and $\times$ is the Cartesian product operator. This means each pixel corresponds to one semantic label, j affordance labels and k attribute labels, where $0 \leq j \leq y, 0 \leq k \leq z$.

3.1. Network Architecture of Cerberus

Intuitively, these three tasks are not independent, e.g., pillows are intrinsically movable. We believe parsing them with a single network can improve performance by exploiting inductive biases between different tasks. However, what is the best architecture for multi-task dense prediction remains an open problem. Generic principles do exist: such an architecture should capture long-range dependency within visual inputs and learn shared representations that effectively serve several tasks. Our observation is that transformer well meets these two requirements: the attention operator has a global receptive field and learning attention focused on a region naturally facilitates representation sharing if different sub-task labels coexist in this region. Hence, we propose the first multi-task dense prediction transformer for joint semantic, affodance and attribute parsing, which is named Cerberus and depicted in Fig.2.

Transformer encoder. Given an image of $H \times W$ pixels, we divide it into $N_p = \frac{HW}{p^2}$ non-overlapping square patches of size p^2. As illustrated in Fig.2 (b), the set of



(a) Overall Architecture

(b) Tokens

(c) Transformer Encoder

(d) Reassemble Operation

(e) Fusion Block

Figure 2. **Overall network architecture of Cerberus.** Given an image, ResNet-50 extract features from the input image to form a set of tokens. The tokens are processed by a transformer encoder and decoded by reassemble operations and fusion blocks. Through three prediction heads, the feature maps are turned into final attribute, affordance and semantic parsing results.

patches is flattened into a vector of length N_p then passed through a ResNet-50 backbone to form N_p embeddings. The embeddings are denoted as a set of tokens: $\{t_n\}, n = 1, ..., N_p$. Learnable position embeddings are concatenated with the tokens to retain positional information. Following [26], an extra learnable token t_0 is added to the sequence, which serves for attention visualization in Fig. 7. It aggregates information from the entire sequence and is named as a readout token. All the N_p+1 tokens are then fed into sequential blocks of multi-head self-attention, which learn shared representations for different tasks.

Reassemble operation. After processing a set of tokens $\{t_n\}, n = 0, ..., N_p$ with transformer encoder, we then assemble them into image-like feature representations at various resolutions which is illustrated in Fig.2 (d).

First, We get N_p embeddings by concatenating t_0 to all other tokens and project the embeddings to D-dimensional features using a fully connected layer. Then, we rearrange the new N_p features by placing them according to the position of the initial patch and get a feature map $\mathrm{F_{rearrange}} \in \mathbb{R}^{\frac{\mathrm{H}}{\mathrm{p}} \times \frac{\mathrm{W}}{\mathrm{p}} \times \mathrm{D}}$. Next, we use a spatial unsampling layer to resize $\mathrm{F_{rearrange}}$ to $\mathrm{F_{upsample}} \in \mathbb{R}^{\frac{\mathrm{H}}{\mathrm{s}} \times \frac{\mathrm{W}}{\mathrm{s}} \times \hat{\mathrm{D}}}$. We reassemble tokens from the outputs of four different stages (the first and second ResNet-50 blocks, layer 9 and layer 12 of transformer encoder) into four image-like representations with different resolutions.

Fusion block. After generating four feature maps from aforementioned stages, Cerberus uses RefineNet-style [17] feature fusion blocks to progressively upsample them. The fusion block is depicted in Fig.2 (e). In the n^{th} fusion stage, We use a residual convolutional unit (RCU) to process the reassembled feature $\mathrm{R_{n-1}}$ first, then fuse it with the previous feature $\mathrm{F_n}$ via another RCU after element-wise summation. Then we upsample the result by a factor of two and get the new fused feature map $\mathrm{F_{n-1}}$. We use the final fused feature map to generate task-specific predictions.

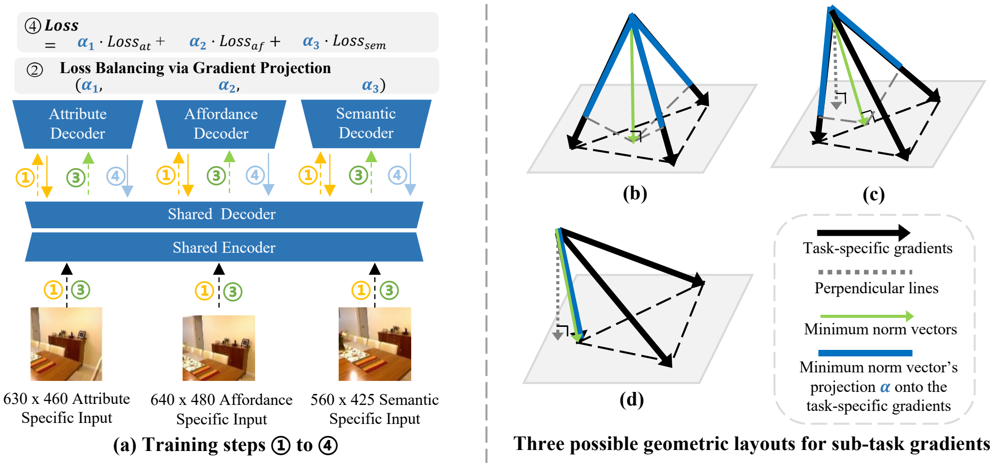
Prediction head. We use three separate prediction heads to produce the final dense prediction results. Each head is composed of two parts: (1) a fully connected layer to generate the semantic, affordance or attribute map, (2) an interpolation function to upsample the predicted map to the original image resolution. For affordance and attribute, we get maps of size $\mathrm{y \times H \times W}$ and $\mathrm{z \times H \times W}$ respectively, where y and z are the number of label classes. And for semantics, the size of predicted map is $\mathrm{H \times W}$, where each pixel corresponds to a semantic class. We use y binary cross entropy losses to supervise affordance, z binary cross entropy losses for attribute and a x-way cross entropy loss for semantics.

3.2. Weakly-aligned Training with Optimal Weights

Motivation How to train our multi-task dense prediction transformer in an effective manner? A straightforward idea is to use a naïve combination of different task losses:

$$\mathcal{L}_{\text{multi\_task}} = \sum_{t=1}^{T} w_t \mathcal{L}_{\mathrm{t}}(\theta) \qquad (2)$$

T is the number of tasks, w_t is the loss weights of tasks t and θ is the parameters of the network. As shown in former


④ Loss = $\alpha_1 \cdot Loss_{at} + \alpha_2 \cdot Loss_{af} + \alpha_3 \cdot Loss_{sem}$
② Loss Balancing via Gradient Projection (α_1, α_2, α_3)
Attribute Decoder
Affordance Decoder
Semantic Decoder
①
③
④
Shared Decoder
Shared Encoder
① ③
630 x 460 Attribute Specific Input
640 x 480 Affordance Specific Input
560 x 425 Semantic Specific Input
(b)
(c)
(d)
Task-specific gradients
Perpendicular lines
Minimum norm vectors
Minimum norm vector's projection α onto the task-specific gradients

(a) Training steps ① to ④

Three possible geometric layouts for sub-task gradients

Figure 3. The illustration of **training framework** (left) and layouts of **gradient vectors** (right).

studies (e.g., Fig.2 in [13]), the performance of the model is sensitive to the selection of weights. Tuning these weights manually is difficult and expensive. Moreover, the optimal weights might change during the training process, which is verified in our experiment (Fig. 5).

We are faced with another challenge: weakly aligned data. Though we use the same dataset to train semantics, affordance and attribute, the annotation of the three tasks have a spatial shift problem. For example, in Fig.3 (a), the input images for three tasks are taken from the same scene, however they are not strictly aligned and even have different resolutions. This means we can't use the one-input-multi-output training paradigms like [31], but have to do forward propagation for each task. To resolve this issue while avoiding manual weight tuning, we resort to the original MGDA formulation [7] which is naturally compatible.

Preliminaries A solution θ_1 dominates another solution θ_2 if $\forall$t, $\mathcal{L}_{\rm t}\left(\theta_1\right) \leq \mathcal{L}_{\rm t}\left(\theta_2\right)$ and $\exists$t, $\mathcal{L}_{\rm t}\left(\theta_1\right) < \mathcal{L}_{\rm t}\left(\theta_2\right)$. We call a solution θ^* Pareto-optimal, if there is no other solution dominates θ^*. And a solution θ^* is said to be Pareto-stationary, if:

$$\sum_{t=1}^{T} \alpha_t \nabla \mathcal{L}_{\rm t}\left(\theta^*\right) = 0, st. \quad \sum_{\rm t=1}^{\rm T} \alpha_{\rm t} = 1, \alpha_{\rm t} \geq 0, \forall {\rm t}. \tag{3}$$

$\nabla \mathcal{L}_{\rm t}\left(\theta^*\right)$ is the gradient of $\mathcal{L}_{\rm t}\left(\theta^*\right)$. If solution θ^* is Pareto-optimal, it is Pareto-stationary, but the reverse isn't always true [7]. We consider the following optimization problem for Pareto-stationary:

$$min_{\alpha_1 \ldots \alpha_T} \| \sum_{t=1}^{T} \alpha_t \nabla \mathcal{L}_{\rm t}\left(\theta\right) \|_2 = 0, \tag{4}$$

$$st. \quad \sum_{\rm t=1}^{\rm T} \alpha_{\rm t} = 1, \alpha_{\rm t} \geq 0, \forall {\rm t}. \tag{5}$$

This is equivalent to finding the vector of minimum norm in the convex hull of the input gradient set.

As for our attention-based model, the effectiveness of this optimization scheme is not clear, since self-attention instead of network parameters play the major role in a transformer. We take the solution of Eq.4 as **optimal weights** to train our multi-task transformer, and empirical evidence shows that they effectively balance Cerberus (Tab. 4).

Formulation Considering the case of Cerberus, T is equal to 3. For notational simplicity, we denote $\nabla \mathcal{L}_1\left(\theta\right)$ as g_1, $\nabla \mathcal{L}_2\left(\theta\right)$ as g_2 and $\nabla \mathcal{L}_3\left(\theta\right)$ as g_3. The geometric illustration is shown in Fig. 3 (right). The minimum norm vector w is either perpendicular to the convex hull or in an boundary case. If the minimum norm is a perpendicular vector which is illustrated in Fig. 3 (b), we have:

$$w = \alpha_1 g_1 + \alpha_2 g_2 + \alpha_3 g_3, \tag{6}$$

$$w \perp \left(g_1 - g_2\right), w \perp \left(g_1 - g_3\right). \tag{7}$$

which is equal to solving three ternary linear equations with constraints that interested variables are greater than zero:

$$\begin{bmatrix} g_1^{\rm T} - g_2^{\rm T} \\ g_1^{\rm T} - g_3^{\rm T} \end{bmatrix} \begin{bmatrix} g_1 & g_2 & g_3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = 0, \tag{8}$$

$$st. \quad \sum_{\rm t=1}^{3} \alpha_{\rm t} = 1, \alpha_{\rm t} \geq 0, \forall {\rm t}. \tag{9}$$

If there is an analytical solution $\alpha_1^*, \alpha_2^*, \alpha_3^*$, the minimum norm vector is $w = \alpha_1^* g_1 + \alpha_2^* g_2 + \alpha_3^* g_3$. Otherwise, the minimum norm vector points to an edge, and it must be a convex combination of two gradient-vectors, which is illustrated in Fig.3 (c). Choose two gradient vectors with smaller norms, for instance g_1 and g_2, then there is:

$$w = \alpha_1 g_1 + \alpha_2 g_2, \quad w \perp \left(g_1 - g_2\right). \tag{10}$$

Method	**mIoU (%)**
J-CRF [47]	14.4
JH-CRF [47]	15.1
PSPNet [45]	36.7
DeepLab V3 [1]	38.1
Ours (single)	44.2
Cerberus	**45.3**

Table 1. Attribute quantitative results on NYUd2.

Method	**mIoU (%)**
Roy et al. [27]	49.6
Roy et al. (w/ GT) [27]	53.2
PSPNet [45]	60.4
DeepLab V3 [1]	61.4
Ours (single)	65.2
Cerberus	**66.3**

Table 2. Affordance quantitative results on NYUd2.

Method	**Input**	**mIoU (%)**
3DGNN [25]	RGB-D	43.1
RDF-101 [23]	RGB-D	49.1
ACNet [12]	RGB-D	48.3
PSPNet [45]	RGB	43.1
DeepLab V3 [1]	RGB	44.7
OCNet [38]	RGB	44.5
FastFCN [37]	RGB	45.4
VarReg [32]	RGB	**50.7**
Ours (single)	RGB	48.8
Cerberus	RGB	50.4

Table 3. Semantic quantitative results on NYUd2.

and the analytical solution is:

$$\alpha_1^* = \frac{(g_2 - g_1)^{\mathrm{T}} g_2}{\|g_2 - g_1\|^2} \tag{11}$$

if $0<\alpha_1^*<1$, then $w = \alpha_1^* g_1 + (1 - \alpha_1^*) g_2$, otherwise the minimum norm vector points to a vertex which is depicted in Fig.3 (d). If $\alpha_1^* \geq 1$, $w = g_1$, else if $\alpha_1^* \leq 0$, $w = g_2$.

After solving the gradient-based problem Eq.(4), we get the optimal weights $\alpha_1^*, \alpha_2^*, \alpha_3^*$ for Cerberus:

$$\mathcal{L}_{\text{Cerberus}} = \alpha_1^* \mathcal{L}_{\text{at}} + \alpha_2^* \mathcal{L}_{\text{af}} + \alpha_3^* \mathcal{L}_{\text{sem}} \tag{12}$$

$\mathcal{L}_{\text{at}}$ is the loss for attribute, $\mathcal{L}_{\text{af}}$ is the loss for affordance and $\mathcal{L}_{\text{sem}}$ is the loss for semantics. With this joint loss, our model gradually converges to a Pareto-stationary solution.

Implementation. To resolve weakly-aligned data and collect gradients for Eq.4, we propose a tailored training framework which is demonstrated in Fig.3 (a). There is a forward propagation for each task, which is followed by a backward pass to calculate the task-specific gradient (step 1). Then we solve the gradient-based problem and update the optimal weights (step 2). Afterwards, we do forward passes again (step 3) to get prediction results. Finally, we calculate the joint loss with optimal weights and update the network with backward propagation (step 4).

4. Experiment

4.1. Comparisons with State-of-the-art Methods

Evaluation details. We benchmark our multi-task dense prediction model on NYUd2 [34] dataset. NYUd2 contains 1449 RGB-D images of indoor scenes with 40 object categories. [27] augments NYUd2 with additional five affordance maps: *sittable*, *walkable*, *lyable*, *reachable* and *movable*. Furthermore, [47] annotates the dataset with 11 additional attribute labels. We train and evaluate Cerberus only with RGB input, using 795 images for training and 654 images for testing. For comparison, we additionally train two widely-used CNN-based dense prediction network PSPNet [45] and DeepLab V3 [1]. Following previous works, we choose the mean intersection over union (mIoU) score as the evaluation metric for both one-label semantic parsing and multi-label affordance/attribute parsing.

Attribute. We show our attribute prediction results in Tab.1. We compare Cerberus against best published results. In contrast to DeepLab v3, our attribute parsing mIoU is significantly promoted from 38.1% to 45.3%. Besides, we train a single-task attribute parsing transformer denoted as Ours (single), to investigate the effect of joint learning. As shown in Tab.1, the mIoU of Cerberus outperforms Ours (single) by 1.1%. This shows that cues from the other two tasks help regularize the shared attention and improve the performance of attribute parsing.

Affordance. We provide the performance of affordance prediction in Tab.2. These results suggest a large improvement over the previous state-of-the-art. We achieve a 13.1% boost over Roy et al. (w/ GT) [27], which uses ground truth cues. Similar to attribute, Cerberus also obtains superior performance over separately trained model, verifying the effectiveness of multi-task learning.

Semantics. Tab.3 provides the performance of Cerberus on the NYUd2 semantic parsing task. Our model outperforms most of the previous state-of-the-arts and achieves comparable performance with SceneParsing [32]. The baseline results reported here are mainly evaluated by [3]. Like the other two tasks, Cerberus outperforms the separately trained model by 1.6%, indicating that multi-task training is of crucial importance.

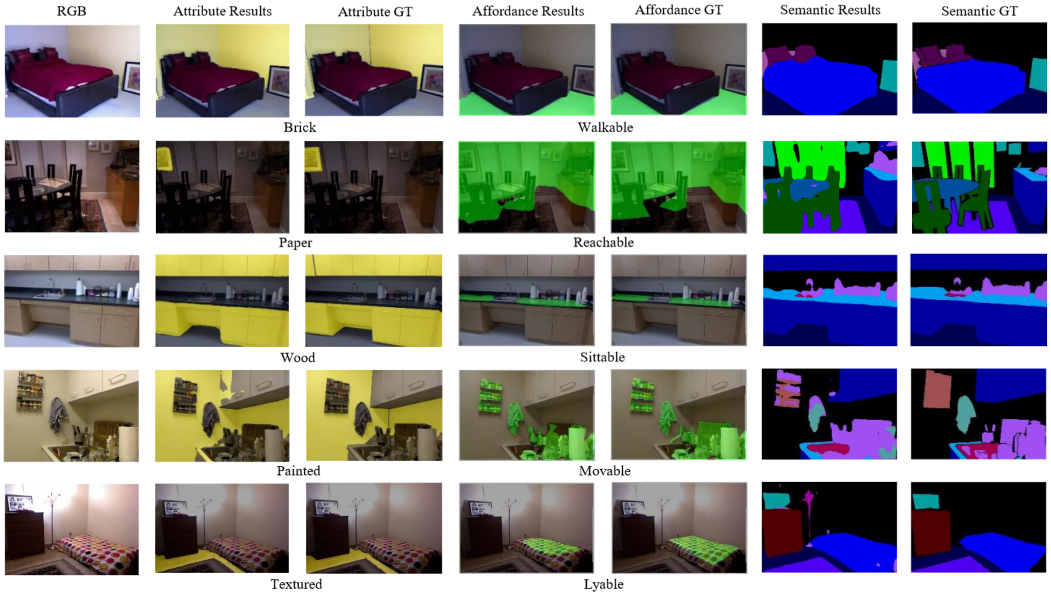
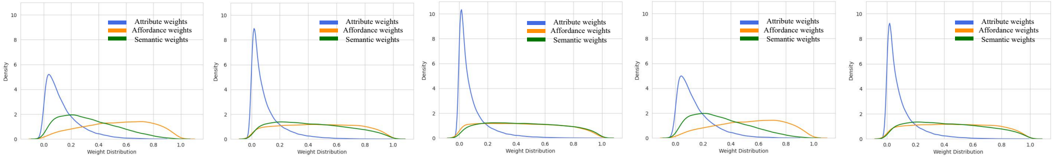


Figure 4. Qualitative prediction results on NYUd2 for three tasks addressed.



(a) Lyable IoU: 53.14 (b) Movable IoU: 58.75 (c) Reachable IoU: 88.85 (d) Sittable IoU: 44.22 (e) Walkable IoU: 86.39

Figure 5. **Weight distribution** (kernel density estimation) for different affordance classes during the training of Cerberus.

Architecture	Weights	Attr.	Aff.	Sem.
PSPNet	single	36.7	60.4	43.1
PSPNet	uniform	38.3	60.3	42.4
PSPNet	optimal	**38.6**	**61.3**	**43.2**
DeepLab V3	single	38.1	61.4	44.7
DeepLab V3	uniform	41.1	62.5	43.6
DeepLab V3	optimal	**42.2**	**63.2**	**44.8**
Cerberus	single	44.2	65.2	48.8
Cerberus	uniform	44.5	63.9	48.3
Cerberus	optimal	**45.3**	**66.3**	**50.4**

Table 4. Quantitative results on NYUd2 with or without using the proposed optimal weights balancing scheme.

Qualitative results. Fig.4 shows our joint parsing results on NYUd2. Our model can predict attribute, affordance and semantic maps precisely in diverse indoor scenes. With all the predicted semantics and properties, one may have an internal image of the scene even without seeing the original RGB image. From the prediction results in the third row of Fig.4, we know there is a cabinet made of wood, and its surface is *sittable*. And in the fourth row, we could see there are many *movable* objects hanging on the painted wall. These results are beneficial to a range of applications, like intelligent service robots and augmented reality.

4.2. Experiments on Optimal Weights

Optimal weights effectively balance Cerberus. We conduct experiments to understand the effect of using optimal weights. For comparison, we train a Cerberus using a uniformly weighted multi-task loss, and the results are shown in Tab.4. We also conduct experiments with two CNN-based models under the same settings to further show the role of optimal weights. It is clear that training with optimal weights gets superior performance. Specifically, for Cerberus, the mIoU of attribute is increased by 0.8%, affordance performance increases 2.3% and mIoU of semantic is also raised by 2.1%. Notably, using a uniformly weighted loss even results in lower performance on certain sub-tasks, when compared with separately trained models. This indicates that there are inconsistent gradient directions between tasks during training and using uniform weights cannot successfully leverage task affinity to resolve them, while optimal weights can better unleash the power of multi-task training to achieve that.

Model	Attribute	Affordance	Semantic
single (at)	38.3	n/a	n/a
Uniform (at)	43.2	65.2	46.8
Cerberus (at)	**44.1**	65.2	47.1
single (af)	n/a	60.9	n/a
Uniform (af)	44.7	**64.2**	46.9
Cerberus (af)	44.6	64.1	47.2
single (sem)	n/a	n/a	23.9
Uniform (sem)	44.3	63.1	37.6
Cerberus (sem)	44.7	65.9	**42.7**

Table 5. 1% weakly-supervised experiments on NYUd2.

Method	Attribute	Affordance	Semantic
single (at)	36.4	n/a	n/a
Uniform (at)	41.8	64.3	47.8
Cerberus (at)	**43.5**	65.3	49.3
single (af)	n/a	57.5	n/a
Uniform (af)	45.0	62.4	46.2
Cerberus (af)	43.9	**64.1**	48.7
single (sem)	n/a	n/a	18.6
Uniform (sem)	43.3	64.4	32.3
Cerberus (sem)	44.3	65.4	**39.1**

Table 7. 0.1% weakly-supervised experiments on NYUd2.

Method	Attribute	Affordance	Semantic
single (at)	37.1	n/a	n/a
Uniform (at)	44.1	63.3	47.7
Cerberus (at)	**44.2**	65.4	46.3
single (af)	n/a	58.8	n/a
Uniform (af)	43.2	62.5	48.9
Cerberus (af)	44.5	**63.5**	48.4
single (sem)	n/a	n/a	20.2
Uniform (sem)	44.5	65.3	36.5
Cerberus (sem)	43.8	65.2	**39.9**

Table 6. 0.5% weakly-supervised experiments on NYUd2.

Optimal weights bias towards tougher tasks. To investigate how optimal weights benefit Cerberus training, we visualize the weight distributions during training in Fig.5, taking affordance as an example. For every affordance label, we collect a subset of the train set which contains all the samples in which the label exists. We save the corresponding optimal weights when encountering these samples. As shown in Fig.5, different labels correspond to different distributions. For *Lyable* and *Sittable*, the affordance weights are higher than in other affordance distributions. And these two classes have lower mIoU results. This reveals that when training tougher sub-tasks, optimal weights tend to be higher. And through dynamic balancing between different sub-tasks, Cerberus achieves superior performance.

4.3. Towards a Deeper Understanding of Cerberus

Task Affinity. Inspired by the semantic transfer technique in [44], we conduct experiments to explore the subtask relationships. Due to the fact of weakly aligned data, we can't use the annotation masks directly. Instead we use aligned predictions generated from a single image to explore task affinity learnt by Cerberus. And we quantify the affinity between label k_1 and k_2 by calculating the intersection over union score (IoU):

$$\mathrm{IoU} = \frac{\mathrm{M}_{\mathrm{k}_1} \cap \mathrm{M}_{\mathrm{k}_2}}{\mathrm{M}_{\mathrm{k}_1} \cup \mathrm{M}_{\mathrm{k}_2}} \tag{13}$$

$\mathrm{M}_{\mathrm{k}_\mathrm{i}}$ is the predicted mask for label k_i. We calculate the mIoU of all label pairs between different tasks on NYUd2 test set, and the task affinities are visualized in Fig.6.

As shown in the figure, the results of different pairs vary a lot but are still reasonable. For example, *Textured*, *Floor* and *Walkable* have high mIoU with each other, since floors are usually textured and walkable. And walls are often painted, leading to a high correlation between them. The high mIoU pairs in semantic-attribute affinity are also in line with common sense: Windows and pictures are usually made of glass, while cabinets and chairs are typically made of wood. The task affinities reveal that the three tasks are not independent of each other, and learning one of them may help the other two. That's the potential reason why our Cerberus out-performs separately trained models.

Cerberus performs well under weak supervision. Inspired by task affinity, we conduct a set of experiments to further unleash the potential of Cerberus. We train a set of Cerberus with one task supervised by $0.1\% - 1\%$ annotation while the other two by full supervision. For comparison, we also train single-task models with $0.1\% - 1\%$ annotation. We use a random mask to select the annotation. The results are shown in Tab. 5, 6 and 7. Our weakly-supervised Cerberus models outperform single-task models, and multi-task models trained with uniformly-weighted losses. For attribute and affordance, the weakly-supervised Cerberus are only slightly worse than the fully supervised Cerberus. Specially, on 0.1% weakly-supervised semantic task, Cerberus outperforms separately trained model by 20.5%.

Shared attention facilitates weakly-supervised learning. In order to explore how Cerberus actually helps with multi-task learning and weakly-supervised learning, we visualize the attention maps of readout token in Fig.7. Though the readout token is not grounded in the input image, it

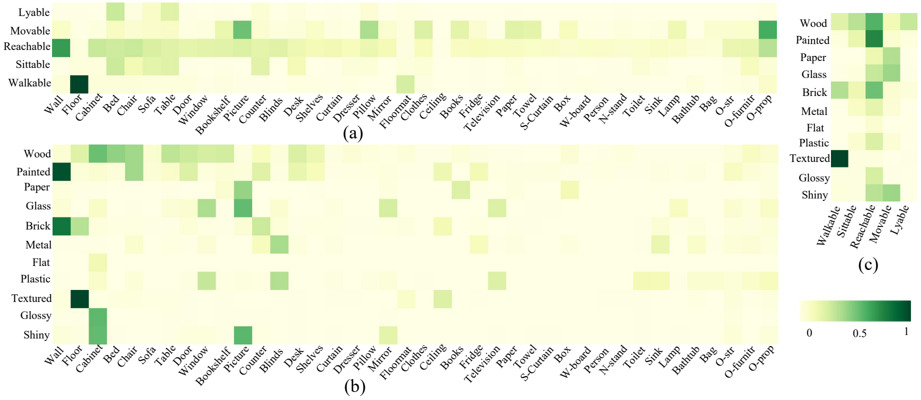


Figure 6. **Visualization of task affinity**. We calculate mIoU between different sub-task concepts, and visualize them with color maps.

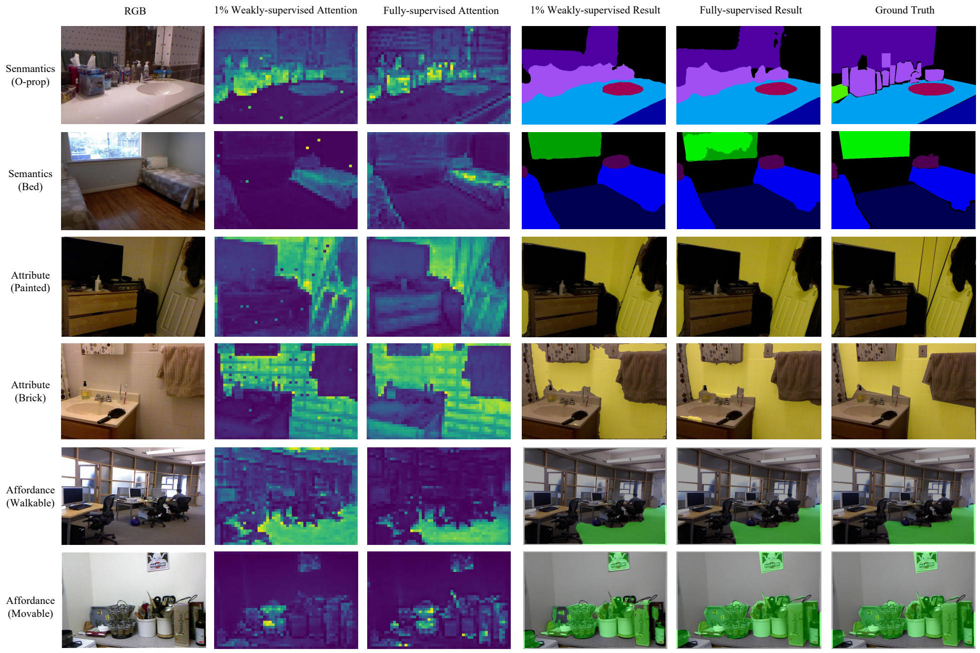


Figure 7. **Visualization of attention weights**. We analyze the self-attention weights of the readout token on different task heads. Interestingly, in the weakly-supervised setting, learned attention weights still well align with ground truth regions.

can aggregates information from other tokens. As can be seen, different heads of the last attention layer aggregate different task-specific features. And we observe that, even for weakly-supervised settings, Cerberus still successfully generates attention maps well aligned with ground truth regions, which appears very similar to the fully supervised attention maps. We believe that this is because that related sub-tasks help attention learning with little annotation. We consider this as a human-like learning capability: one can achieve weakly-supervised learning with the help of other related tasks. For example, if we know what is window, we can easily learn how glass looks like. That's why our attention-based model achieves strong performance under weak supervision.

5. Conclusion

In this paper, we propose a novel multi-task dense prediction transformer named Cerberus to parse semantics, affordance and attribute jointly. We successfully train our model from weakly-aligned data using a weight balancing framework to unleash the power of multi-task learning. Cerberus achieves state-of-the-art performance on all tasks and strong results under weak supervision (using as low as 0.1% annotation). We observe task affinity consistent with common sense and further demonstrate that shared attention between tasks facilitates weakly-supervised learning.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5

[2] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4193–4202, 2020. 2

[3] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 561–577. Springer, 2020. 5

[4] Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Pq-transformer: Jointly parsing 3d objects and layouts from point clouds. *IEEE Robotics and Automation Letters*, 2022. 2

[5] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8648–8657, 2019. 2

[6] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Indoor scene understanding with geometric and semantic contexts. *International Journal of Computer Vision*, 112(2):204–220, 2015. 2

[7] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012. 4

[8] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018. 2

[9] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 1

[10] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 1

[11] Varsha Hedau, Derek Hoiem, and David Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *European Conference on Computer Vision*, pages 224–237. Springer, 2010. 2

[12] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019. 5

[13] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 4

[14] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24:109–117, 2011. 2

[15] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pages 365–372. IEEE, 2009. 2

[16] L'ubor Ladickỳ, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Associative hierarchical crfs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 739–746. IEEE, 2009. 2

[17] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 3

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2

[19] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *arXiv preprint arXiv:2104.00678*, 2021. 2

[20] Chengzhi Mao, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. Multitask learning strengthens adversarial robustness. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 158–174. Springer, 2020. 2

[21] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 2

[22] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011. 2

[23] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 4980–4989, 2017. 5

[24] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012. 2

[25] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5199–5208, 2017. 5

[26] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision, pages 12179–12188, 2021. 2, 3

[27] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *European conference on computer vision*, pages 186–201. Springer, 2016. 1, 5

[28] Johann Sawatzky, Yaser Souri, Christian Grund, and Jurgen Gall. What object should i use?-task driven object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7605–7614, 2019. 2

[29] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2795–2804, 2017. 2

[30] Alexander G Schwing, Sanja Fidler, Marc Pollefeys, and Raquel Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 353–360, 2013. 2

[31] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *arXiv preprint arXiv:1810.04650*, 2018. 4

[32] Hengcan Shi, Hongliang Li, Qingbo Wu, and Zichen Song. Scene parsing via integrated classification model and variance-based regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5307–5316, 2019. 5

[33] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International journal of computer vision*, 81(1):2–23, 2009. 2

[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 1, 5

[35] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017. 1, 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2

[37] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816*, 2019. 5

[38] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 5

[39] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020. 2

[40] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 1

[41] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 2

[42] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 733–749, 2018. 2

[43] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7801–7810, 2019. 2

[44] Hao Zhao, Ming Lu, Anbang Yao, Yiwen Guo, Yurong Chen, and Li Zhang. Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10–18, 2017. 2, 7

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 5

[46] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3119–3126, 2013. 1

[47] Shuai Zheng, Ming-Ming Cheng, Jonathan Warrell, Paul Sturgess, Vibhav Vineet, Carsten Rother, and Philip HS Torr. Dense semantic image segmentation with objects and attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3214–3221, 2014. 1, 5