Voxel-informed Language Grounding

Rodolfo Corona   **Shizhan Zhu**   **Dan Klein**   **Trevor Darrell**

Computer Science Division, University of California, Berkeley

{rcorona, shizhan_zhu, klein, trevordarrell}@berkeley.edu

Abstract

Natural language applied to natural 2D images describes a fundamentally 3D world. We present the Voxel-informed Language Grounder (VLG), a language grounding model that leverages *3D geometric information* in the form of voxel maps derived from the visual input using a volumetric reconstruction model. We show that VLG significantly improves grounding accuracy on SNARE (Thomason et al., 2021), an object reference game task. At the time of writing, VLG holds the top place on the SNARE leaderboard,[1] achieving SOTA results with a 2.0% absolute improvement.

1 Introduction

Embodied robotic agents hold great potential for providing assistive technologies in home environments (Pineau et al., 2003), and natural language provides an intuitive interface for users to interact with such systems (Andreas et al., 2020). For these systems to be effective, they must be able to reliably ground language in perception (Bisk et al., 2020; Bender and Koller, 2020).

Despite typically being paired with 2D images, natural language that is grounded in vision describes a fundamentally 3D world. For example, consider the grounding task in Figure 1, where the agent must select a target chair against a distractor given the description "the swivel chair with 6 wheels." Although the agent is provided with multiple images revealing all of the wheels on each chair, it must be able to properly aggregate information across images to successfully differentiate them, something that requires reasoning about their *3D geometry* at some level.

In this work, we show how language grounding performance may be improved by leveraging 3D prior knowledge. Our model, Voxel-informed Language Grounder (VLG), extracts 3D voxel maps using a pre-trained *volumetric reconstruction model*,
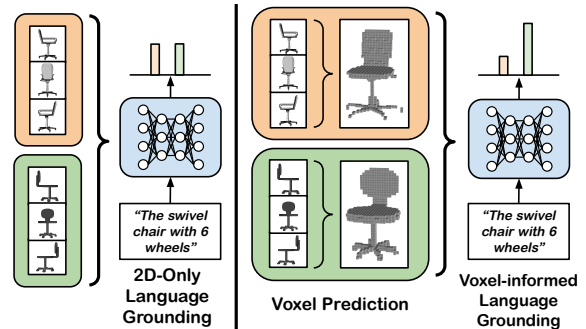

Figure 1: **Voxel-informed Language Grounder.** Our VLG model leverages explicit 3D information by inferring volumetric voxel maps from input images, allowing the agent to reason jointly over the geometric and visual properties of objects when grounding.

which it fuses with multimodal features from a large-scale vision and language model in order to reason jointly over the visual and 3D geometric properties of objects.

We focus our investigation within the context of SNARE (Thomason et al., 2021), an object reference game where an agent must ground natural language describing common household objects by their geometric and visual properties, showing that grounding accuracy significantly improves by incorporating information from predicted 3D volumes of objects. At the time of writing, VLG achieves SOTA performance on SNARE, attaining an absolute improvement of 2.0% over the next closest baseline. Code to replicate our results is publicly available.[2]

2 Related Work

Prior work has studied deriving structured representations from images to scaffold language grounding. However, a majority of systems use representations such as 2D regions of interest (Anderson et al., 2018; Wang et al., 2020) or symbolic graph-

---

[1]https://github.com/snaredataset/snareleaderboard

[2]https://github.com/rcorona/voxel_informed_language_grounding

based representations (Hudson and Manning, 2019; Kulkarni et al., 2013), which do not encode 3D properties of objects.

Most prior work tying language to 3D representations has largely focused on generating 3D structures conditioned on language, rather than using them as intermediate representations for language grounding as we do here. Specifically, prior work has performed language conditioned generation at the scene (Chang et al., 2014, 2015a), pose (Ahuja and Morency, 2019; Lin et al., 2018), or object (Chen et al., 2018) level. More recently, a line of work has explored referring expression grounding in 3D by mapping referring expressions of objects to 3D bounding boxes localizing them in point clouds of indoor scenes (Achlioptas et al., 2020; Chen et al., 2020; Zhao et al., 2021; Roh et al., 2022). Standard approaches follow a two-tiered process where an object proposal system will first provide bounding boxes for candidate objects, and a scoring module will then compute a compatibility score between each box and the referring expression in order to ground it. At a more granular level, Koo et al. (2021) learn alignments from language to object parts by training agents on a reference game over point cloud representations of objects.

In contrast, in this work we focus on augmenting language grounding over 2D RGB images using structured 3D representations derived from them. For the task of visual language navigation, prior work has shown how a persistent 3D semantic map may be used as an intermediate representation to aid in selecting navigational waypoints (Chaplot et al., 2020; Blukis et al., 2021). The semantic maps, however, represent entire scenes with individual voxels representing object categories, rather than their geometry. In this work, we show how a more granular occupancy map representing objects' geometry can improve language grounding performance.

Closest to our work is that of Prabhudesai et al. (2020), which presents a method for mapping language to 3D features within scenes from the CLEVR (Johnson et al., 2017) dataset. Their system generates 3D feature maps inferred from images and then grounds language directly to 3D bounding boxes or coordinates. Their method assumes, however, that dependency parse trees are provided for the natural language inputs, and it is trained with supervised alignments between noun phrases and the 3D representations, which VLG does not require.

3 Voxel-informed Language Grounder

We consider a task where an agent must correctly predict a target object v^t against a distractor v^c given a natural language description $w^t = \{w_1, ..., w_m\}$ of the target. For each object, the agent is provided with n 2D views $v = \{x_1, ..., x_n\}, x_i \in \mathbb{R}^{3 \times W \times H}$.

An agent for this task is represented by a scoring function $s(v, w) \in [0, 1]$, computing the compatibility between the target description and the 2D views of each object, and is used to select the maximally scoring candidate. We first use unimodal encoders to encode the language description into $e_w = h(w)$ and the object view images into a single aggregate visual embedding $e_v = g(v)$ before fusing them with a visiolinguistic module $e_{vw} = f_{vw}([e_v; e_w])$. Prior approaches to this problem (Thomason et al., 2021) directly input this fused representation to a scoring module to produce a score $s(e_{vw})$. They do not explicitly reason about the 3D properties of the observed objects, requiring the models to learn them implicitly.

In contrast, our Voxel-informed Language Grounder augments the scoring function s with explicit 3D volumetric information $e_o = o(v)$ extracted from a pre-trained multiview reconstruction model. The volumetric information (in the form of a factorization of a voxel occupancy map in $\mathbb{R}^{W \times H \times D}$) is first fused into a joint representation with the language using a multimodal voxel-language module $e_{ow} = f_{ow}([e_o; e_w])$. The scoring function then produces a score based on all three modalities $s([e_{vw}; e_{ow}])$.

3.1 Model Architecture

VLG (Figure 2) consists of two branches: a visiolinguistic module for fusing language and 2D RGB features, and a voxel-language module for fusing language with 3D volumetric features. A scoring function is then used to reason jointly over the output of the two branches, producing a compatibility score.

Visiolinguistic Module. The architecture of our visiolinguistic module f_{vw} (left panel, Figure 2) largely mirrors the architecture of MATCH from Thomason et al. (2021). A pre-trained CLIP-ViT (Radford et al., 2021) model is used to
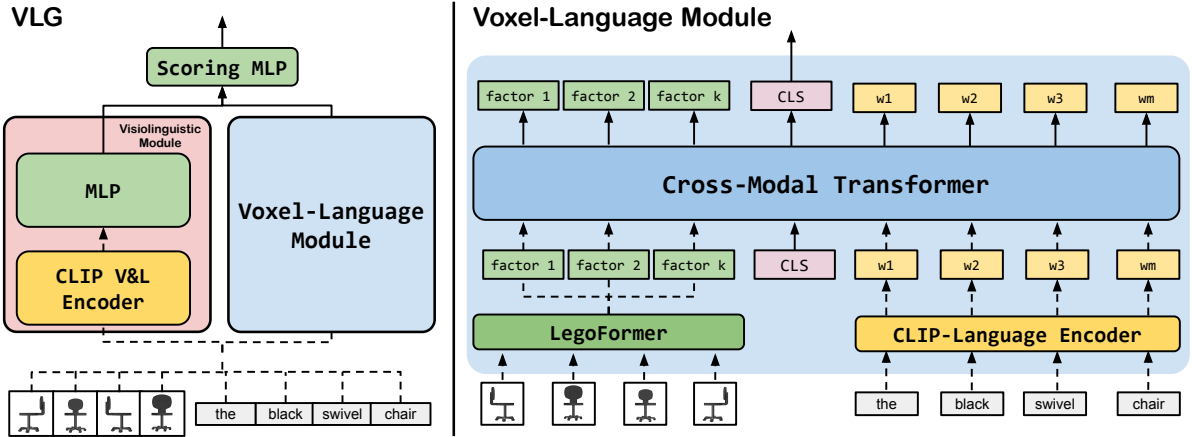
Figure 2: **VLG Architecture.** (Left) Our VLG model consists of a visiolinguistic module which produces a joint embedding for text and images using CLIP (Radford et al., 2021) and a voxel-language module for jointly embedding language and volumetric maps. (Right) The voxel-language module uses a cross modal transformer to fuse word embeddings from CLIP with voxel map factors extracted from LegoFormer (Yagubbayli et al., 2021). During training, gradients only flow through solid lines.

encode the language description and view images into vectors in $\mathbb{R}^{512}$. The image embeddings are max-pooled and concatenated to the description embedding before being passed into an MLP which generates a fused representation.

Voxel-Language Module. We use representations extracted from a ShapeNet (Chang et al., 2015b; Wu et al., 2015) pre-trained Lego-FormerM (Yagubbayli et al., 2021), a multi-view 3D volumetric reconstruction model, as input to our voxel-language module f_{ow}. LegoFormer is a transformer (Vaswani et al., 2017) based model whose decoder generates volumetric maps factorized into 12 parts. Each object factor is represented by a set of three vectors $x, y, z \in \mathbb{R}^{32}$, which we concatenate to use as input tokens for our voxel-language module. A triple cross-product over x, y, z may be used to recover a 3D volume $\mathcal{V} \in \mathbb{R}^{32 \times 32 \times 32}$ for each factor. The full volume for the object is generated by aggregating the factor volumes through a sum operation. For more details on LegoFormer, we refer the reader to Yagubbayli et al. (2021). We use a cross-modal transformer (Vaswani et al., 2017) encoder to fuse the language and object factors (Figure 2, right). The cross-modal transformer takes as input language tokens, in the form of CLIP word embeddings, and the 12 object factors output by the LegoFormer decoder, which contain the inferred geometric occupancy information of the object. We use a CLS token as an aggregate

representation of the language and object factors.

Scoring Function. The scoring function is represented by an MLP which takes as input the concatenation of the visiolinguistic module output and the cross-modal transformer's CLS token.

4 Language Grounding Evaluation

Evaluation. We test our method on the SNARE benchmark (Thomason et al., 2021). SNARE is a language grounding dataset which augments ACRONYM (Eppner et al., 2021), a grasping dataset built off of ShapeNetSem (Savva et al., 2015; Chang et al., 2015a), with natural language annotations of objects.

SNARE presents an object reference game where an agent must correctly guess a target object against a distractor. In each instance of the game, the agent is provided with a language description of the target as well as multiple 2D views of each object. SNARE differentiates between **visual** and **blindfolded** object descriptions. Visual descriptions primarily include attributes such as *name*, *shape*, and *color* (e.g. "classic armchair with white seat"). In contrast, blindfolded descriptions include attributes such as *shape* and *parts* (e.g. "oval back and vertical legs"). The train/validation/test sets were generated by splitting over (207 / 7 / 48) ShapeNetSem object categories, respectively containing (6,153 / 371 / 1,357) unique object instances and (39,104 / 2,304 / 8,751) object pairings with referring expressions. Renderings are provided for each object

Model	VALIDATION			TEST		
	Visual	Blind	All	Visual	Blind	All
ViLBERT	89.5	76.6	83.1	80.2	**73.0**	76.6
MATCH	89.2 (0.9)	75.2 (0.7)	82.2 (0.4)	83.9 (0.5)	68.7 (0.9)	76.5 (0.5)
MATCH*	90.6 (0.4)	75.7 (1.2)	83.2 (0.8)	-	-	-
LAGOR	89.8 (0.4)	75.3 (0.7)	82.6 (0.4)	84.3 (0.4)	69.4 (0.5)	77.0 (0.5)
LAGOR*	89.8 (0.5)	75.0 (0.4)	82.5 (0.1)	-	-	-
VLG (Ours)	**91.2** (0.4)	**78.4**†(0.7)	**84.9**†(0.3)	**86.0**	71.7	**79.0**

Table 1: **SNARE Benchmark Performance.** Object reference game accuracy on the SNARE task across validation and test sets. Performance on models with an asterisk are our replications of the baselines in Thomason et al. (2021). Standard deviations over 3 seeds are shown in parentheses. MATCH corresponds to the max-pool variant from Thomason et al. (2021) since no test set results are provided for the mean-pool variant. Our VLG model achieves the best overall performance. Due to leaderboard submission restrictions, we were not able to get test set results for the MATCH* and LAGOR* replications. † denotes statistical significance with $p < 0.1$.

instance over 8 canonical viewing angles.

Because ShapeNet and ShapeNetSem represent different splits of the broader ShapeNet database, we pre-train the LegoFormerM model on a modified dataset to avoid dataset leakage. Specifically, any objects which appear in both datasets are re-assigned within the pre-training dataset used to train LegoFormerM to match its split assignment from SNARE.

ShapeNetSem images are resized to 224×224 when inputting them to LegoFormerM in order to match its ShapeNet pre-training conditions.

Baselines. We compare VLG against the set of models provided with SNARE.[3] All SNARE baselines except **ViLBERT** use a CLIP ViT-B/32 (Radford et al., 2021) backbone for encoding both images and language descriptions:

> **MATCH** first uses CLIP-ViT to embed the language description as well as each of the 8 view images. Next, the view embeddings are mean-pooled and concatenated to the description embedding. Finally, a learned MLP is used over the concatenated feature vector in order to produce a final compatibility score.

> **ViLBERT** fine-tunes a 12-in-1 (Lu et al., 2020) pre-trained ViLBERT(Lu et al., 2019) as the backbone for MATCH instead of using CLIP-ViT. Each object is presented to ViLBERT in the form of a single tiled image containing all 14 views from ShapeNetSem, instead of just the canonical 8 presented in the standard task. ViLBERT tokenizes images by

extracting features from image regions, with the ground truth bounding boxes for each region (i.e. view) being provided. Because this baseline is not open-source, we report the original numbers from Thomason et al. (2021).

LAGOR (**La**nguage **G**rounding through **O**bject **R**otation) fine-tunes a pre-trained MATCH module and is additionally regularized through the auxiliary task of predicting the canonical viewing angle of individual view images, which it predicts using an added output MLP head. Following Thomason et al. (2021), the LAGOR baseline is only provided with 2 random views of each object both during training and inference.

For more details on the baseline models, we refer the reader to Thomason et al. (2021).

Training Details. Apart from the dataset split re-assignments mentioned in Section 4, we use the code[4] and hyperparameters presented by Yagubbayli et al. (2021) to train LegoFormerM.

For training on SNARE, we follow Thomason et al. (2021) and train all models with a smoothed binary cross-entropy loss (Achlioptas et al., 2019).

We train each model for 75 epochs, reporting performance of the best performing checkpoint on the validation set. For our replication of the SNARE MATCH and LAGOR baselines, we use the code and hyperparameters provided by Thomason et al. (2021). For all variants of our VLG model we use the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 1e-3 and a linear learning rate warmup of 10K steps.

---

[3]https://github.com/snaredataset/snare

[4]https://github.com/faridyagubbayli/LegoFormer

Model	Visual	Blind	All
VGG16	**91.4** (0.5)	76.5 (0.9)	84.0 (0.2)
MLP	91.1 (0.8)	77.9 (0.9)	84.6 (0.1)
no-CLIP	71.0 (0.6)	65.8 (0.7)	68.4 (0.1)
VLG	91.2 (0.4)	**78.4** (0.7)	**84.9** (0.3)

Table 2: **Ablation Study.** SNARE reference game accuracy across ablations of our model on the validation set. We show performance when replacing LegoformerM object factors with **VGG16** features, replacing the cross-modal transformer with an **MLP**, and when foregoing the use of CLIP features (**no-CLIP**).

5 Results

We present test set performance for VLG and the SNARE baselines reported by Thomason et al. (2021). We also present average performance for trained models over 3 seeds with standard deviations on the validation set.

5.1 Comparison to SOTA

In Table 1 we can observe reference game performance for all models. VLG achieves SOTA performance with an absolute improvement on the test set of 2.0% over LAGOR, the next best leaderboard model. Although there is a general improvement of 1.7% in **visual** reference grounding, there is an improvement of 2.3% in **blindfolded** (denoted as **Blind** in tables to conserve space) reference grounding. This suggests that the injected 3D information provides a greater boost for disambiguating between examples referring to geometric properties of target objects. VLG generally improves over all baselines and conditions for blindfolded examples, with the exception of ViLBERT, which may be due to the additional information ViLBERT receives in the form of 14 viewing angles of each object instead of 8.

Improvements on the Blind and All conditions of the validation set are statistically significant with $p < 0.1$ under a Welch's two-tailed t-test.

5.2 Ablation Study

We present a variety of ablations on the validation set to investigate the contributions of each piece of our model. All results can be observed in Table 2.

VGG16 Embeddings. LegoFormer uses an ImageNet (Deng et al., 2009) pre-trained VGG16 (Simonyan and Zisserman, 2014) as a backbone for extracting visual representations, which is a different dataset and pre-training task

than what the CLIP-ViT image encoder is trained on. This presents a confounding factor which we ablate by performing an experiment feeding our model's scoring function VGG16 features directly instead of LegoFormer object factors (VGG16 in Table 2). Despite getting comparable results to VGG16 on visual reference grounding, VLG provides a clear improvement in blindfolded (and therefore overall) reference performance, suggesting that the extracted 3D information is useful for grounding more geometrically based language descriptions, with the VGG16 features being largely redundant in terms of visual signal.

Architecture. We ablate the contribution of our cross-modal transformer branch by comparing it against an MLP mirroring the structure of the SNARE MATCH baseline. This model (MLP in Table 2) max-pools the LegoFormer object factors and concatenates the result to the CLIP visual and language features before passing them to an MLP scoring function. The MLP model overall outperforms the SNARE baselines from Table 1, highlighting the usefulness of the 3D information for grounding, but does not result in as large an improvement as the cross-modal transformer. This suggests that the transformer is better able at integrating information from the multi-view input.

CLIP Visual Embeddings. Finally, we evaluate the contribution of the visiolinguistic branch of the model by removing it and only using the cross-modal transformer over language and object factors. As may be observed, there is a large drop in performance (16.5% overall), particularly for visual references (20.2%). These results suggest that maintaining visual information such as color and texture is critical for performing well on this task, since the LegoFormer outputs contain only volumetric occupancy information.

6 Discussion

We have presented the Voxel-informed Language Grounder (VLG), a model which leverages explicit 3D information from predicted volumetric voxel maps to improve language grounding performance. VLG achieves SOTA results on SNARE, and ablations demonstrate the effectiveness of using this 3D information for grounding. We hope this paper may inspire future work on integrating structured 3D representations into language grounding tasks.

Acknowledgements

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. 2020. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*.

Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. 2019. Shapeglot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8938–8947.

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, et al. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. 2021. A persistent spatial semantic representation for high-level natural language instruction execution. *arXiv preprint arXiv:2107.05612*.

Angel Chang, Manolis Savva, and Christopher D Manning. 2014. Learning spatial knowledge for text to 3d scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2028–2038.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015a. Shapenet: An information-rich 3d model repository. Cite arxiv:1512.03012.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015b. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 202–221. Springer.

Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2018. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian Conference on Computer Vision*, pages 100–116. Springer.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Clemens Eppner, Arsalan Mousavian, and Dieter Fox. 2021. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE.

Drew A Hudson and Christopher D Manning. 2019. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 2901–2910.

Juil Koo, Ian Huang, Panos Achlioptas, Leonidas Guibas, and Minhyuk Sung. 2021. Partglot: Learning shape part segmentation from language reference games. *arXiv preprint arXiv:2112.06390*.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903.

Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. 2018. Generating animated videos of human activities from natural language descriptions. *Learning*, 2018:1.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Joelle Pineau, Michael Montemerlo, Martha Pollack, Nicholas Roy, and Sebastian Thrun. 2003. Towards robotic assistants in nursing homes: Challenges and results. *Robotics and autonomous systems*, 42(3-4):271–281.

Mihir Prabhudesai, Hsiao-Yu Fish Tung, Syed Ashar Javed, Maximilian Sieb, Adam W Harley, and Katerina Fragkiadaki. 2020. Embodied language grounding with 3d visual feature representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2220–2229.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. 2022. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR.

Manolis Savva, Angel X. Chang, and Pat Hanrahan. 2015. Semantically-Enriched 3D Models for Common-sense Knowledge. *CVPR 2015 Workshop on Functionality, Physics, Intentionality and Causality*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. 2021. Language grounding with 3d objects. In *5th Annual Conference on Robot Learning*.

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

Ruocheng Wang, Jiayuan Mao, Samuel J Gershman, and Jiajun Wu. 2020. Language-mediated, object-centric representation learning. *arXiv preprint arXiv:2012.15814*.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920.

Farid Yagubbayli, Alessio Tonioni, and Federico Tombari. 2021. Legoformer: Transformers for block-by-block multi-view 3d reconstruction. *arXiv preprint arXiv:2106.12102*.

Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 2021. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2928–2937.