Inpaint2Learn: A Self-Supervised Framework for Affordance Learning

Lingzhi Zhang* Weiyu Du*† Shenghao Zhou Jiancong Wang Jianbo Shi
University of Pennsylvania

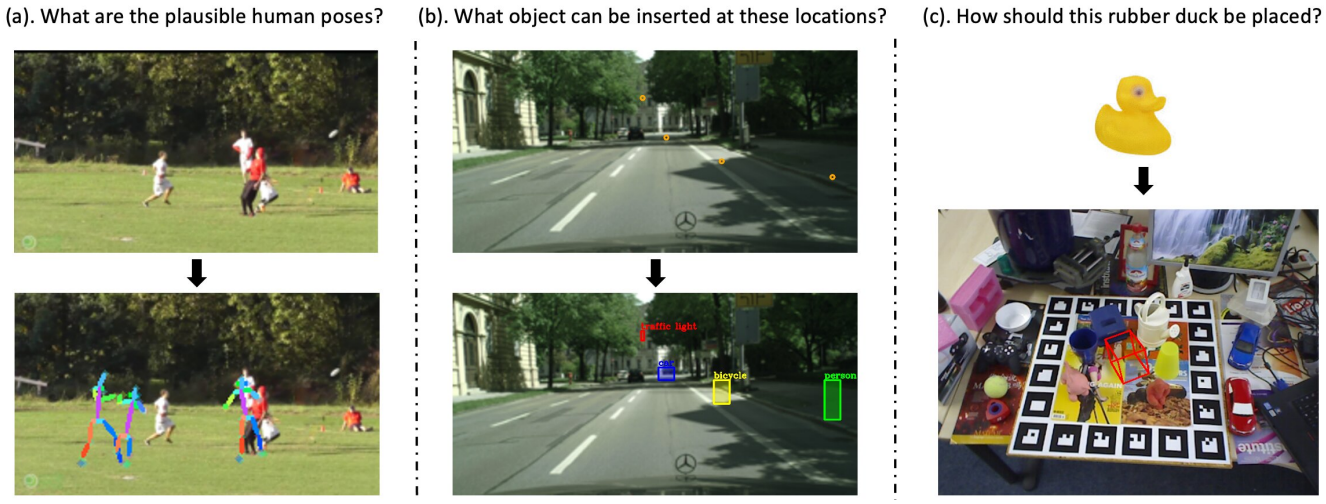

Figure 1: Our proposed Inpaint2Learn framework allows us to train models that can predict affordances in various scenarios. In panel **a**, our model can predict plausible and diverse human poses in an in-the-wild image. In panel **b**, our model can predict what type of objects to insert at a given location and how large it should be. In panel **c**, our model can predict the 6D pose (3D translation + 3D rotation) required to place an object into a target scene.

Abstract

Perceiving affordances – the opportunities of interaction in a scene, is a fundamental ability of humans. It is an equally important skill for AI agents and robots to better understand and interact with the world. However, labeling affordances in the environment is not a trivial task. To address this issue, we propose a task-agnostic framework, named Inpaint2Learn, that generates affordance labels in a fully automatic manner and opens the door for affordance learning in the wild. To demonstrate its effectiveness, we apply it to three different tasks: human affordance prediction, Location2Object and 6D object pose hallucination. Our experiments and user studies show that our models, trained with the Inpaint2Learn scaffold, are able to generate diverse and visually plausible results in all three scenarios.

1. Introduction

Affordance[11], a concept originated from psychology, refers to the opportunities of interaction in a scene. Understanding affordances is a fundamental ability of humans and an important part of socialization[11]. It is an equally important skill for AI agents and robots to better understand and interact with the world. In fact, predicting affordances from the environment has received more and more attention over the years, with applications in domestic robotics, content analysis, context-aware scene understanding[12].

A key challenge in affordance learning is the lack of large-scale, well-labeled data. Existing approaches find different ways to address this issue. In human affordance prediction, for example, Wang et al.[27] leverage large-scale sitcom videos, which contain interactions in repeated indoor scenes to create a 2D affordance dataset. Li et al.[22] generate 3D affordance labels by learning a pose synthesizer from the sitcom dataset, mapping the generated poses to 3D using camera parameters and finally adjusting for physical correctness. While these approaches have proved their strength in generating plausible human poses given in-

* indicates equal contribution.
† W. Du is currently affiliated with Nuro, Inc.

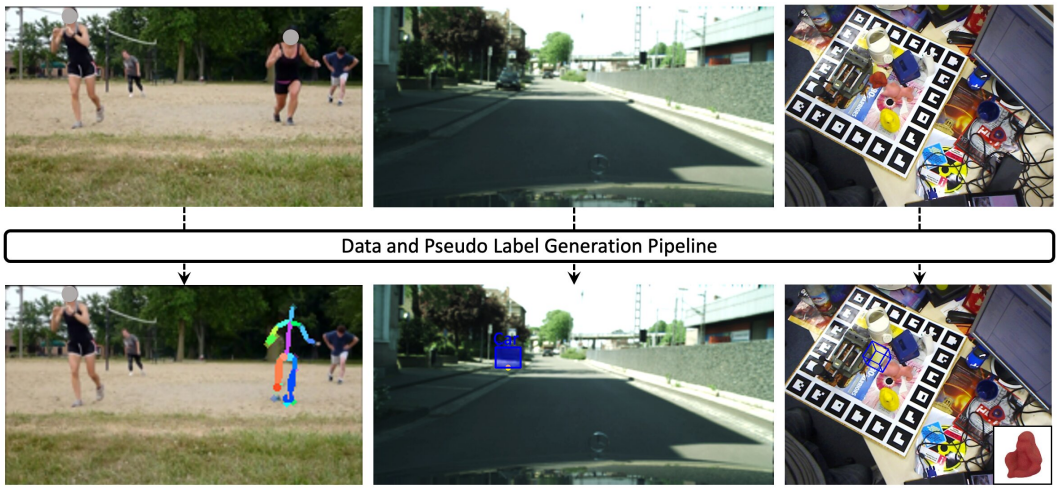

Figure 2: In our Inpaint2Learn data generation pipeline, we first segment the object of interest, cut it out, and generate the pseudo labels paired with the image without object. The first row shows the raw images, and the second raw shows the processed image with visualized pseudo labels.

door contexts, they suffer from limited generalization abilities from a less diverse dataset. People live in and interact with varied scenes and environment everyday, ranging from outdoor lawns to indoor gyms. The diverse scenarios and interaction types cannot be fully captured through a limited number of sets in TV shows.

To this end, we introduce a framework, *Inpaint2Learn*, that allows affordance learning in the wild. We propose a data preprocessing pipeline in which we remove an object/person of interest from the image by obtaining its instance segmentation mask, cutting it out and filling the hole with image inpainting technique. While the now object-absent image serves as input to the network, the cut out instance provides pseudo affordance label. Following this scheme, we are able to generate ground truth affordances for any RGB image dataset in the wild. To demonstrate the effectiveness of our framework, we apply Inpaint2Learn to three specific tasks in affordance learning: 1) human affordance prediction in the wild, 2) Location2Object in driving scenarios, and 3) 6D object pose hallucination. Our experiments show diverse and visually plausible results in all three cases, which are then verified through user studies.

In the task of human affordance prediction, we follow previous works [19][22] in designing a two-stage *where* and *what* training pipeline, with a new semantics-and-geometry-aware adversarial learning strategy. Our experiments reveal that the semantic segmentation and depth information are crucial for the network to generate meaningful predictions that obey the rules of projective geometry.

Our main contributions are as follows:

1. We present an easy-to-use, task-agnostic framework for generating large-scale affordance labels on images in the wild. We plan to release three datasets with pseudo affordance labels on the task of human affordance prediction, Location2Object and 6D object pose hallucination.

2. To the best of our knowledge, we are the first to study human affordances in the wild. We propose a semantics-and-geometry-aware adversarial learning strategy, which proves helpful to generating meaningful and plausible results.

3. We demonstrate the effectiveness of our self-supervised affordance framework in two additional tasks: Location2Object and 6D object pose hallucination. Our proposed models outperform the baselines and provide strong benchmarks for these tasks.

2. Related Work

Affordance[11], described as what the environment provides or furnishes the animal, is a term coined by psychologist James Gibson in the late 70s. Understanding affordances is an innate ability of humans. According to Gibson, people not only perceive affordances, but also modify their environment to change its affordances to better suit them. A further study[1] reveals that merely looking at an object primes the human brain to perform the action the object affords.

In the computer vision community, predicting affordances is often defined as predicting opportunities of interaction in a scene. Many previous works [36][5][24][8] have tackled the task in a data-driven manner. Wang et al.[27] predict plausible 2D human poses given context by scaling and deforming a clustered pose from sitcom videos. Li et

al.[22] synthesize 3D poses from a scene image by factorizing pose distributions into where and what. Tan et al.[25] propose a branching CNN to jointly learn location and size of a person given a scene, and retrieve plausible human instances to insert. Corona et al. [7] generate hand shapes and poses for grasping objects from RGB images. Given 3D scene meshes, Zhang et al.[35] predict 3D human bodies with various shapes and poses to interact with the environment; Zhang et al.[34] generate human-scene proximal relationships and infer a plausible human body.

Besides learning plausible human poses from context, previous research has also studied modeling interaction based on objects. Lin et al.[23] proposed a Spatial Transformer GAN to learn a realistic image compositional distribution in order to insert indoor furniture in a simulated environment. Azadi et al.[3] propose a cycle-consistent adversarial training to compose and decompose a pair of objects, such as a chair and a table. Zhan et al. [32] propose a Spatial Fusion GAN to fuse an instance into a background using a geometry synthesizer and an appearance synthesizer. Fang et al.[10] learn the interaction region of objects and action labels from demonstration videos.

Learning to place objects in a scene is another way to model affordances in an environment. Lee et al.[19] insert object instance masks into a driving scene image by predicting where and what with two generative modules. Zhang et al.[33] predict a distribution of common sense locations given a foreground object and a background scene.

Instead of focusing on one single task, we propose a general, self-supervised framework for all three different aspects of affordance learning. Our data generation pipeline, inspired by [33], is task agnostic and produces training labels with no additional information other than RGB images.

3. Approach

3.1. Data Generation for Affordance Learning

A key challenge to learning affordances is the lack of training data, which should ideally contain an object, a context image without the object, and their corresponding spatial or geometric relationship as a pair. As shown in Figure 2, we build a data preprocessing pipeline to generate the above labels from a single RGB image: we obtain instance segmentation mask for the object of interest, cut it out, and then inpaint the hole in the context image. Specifically, we use Mask R-CNN [13] for instance segmentation and ProFill[31] for image inpatining. In the task of human affordance prediction, we take advantage of Alphapose[9][21][29] to extract human pose labels. Furthermore, as auxiliary information to aid prediction, we compute semantic segmentation map with HRNet [26] and depth map with [28]. Since our pipeline does not assume any constraint on the input images and uses no additional information other than off-the-shelf prediction results, it opens the door to learning affordance in the wild, giving rise to large-scale, diverse training scenarios and object-context relationships. We demonstrate its wide applications in a number of affordance prediction tasks, described in the following sections.

3.2. Human Affordance Prediction in the Wild

Given a context image, a human may appear at many locations, interacting with the environment in a number of ways. Thus, hallucinating plausible human poses is a multimodal problem in nature. As shown in Figure 3, we use a two-stage training pipeline with conditional VAE models [18] to capture a diverse yet reasonable distribution for human poses given context. First, we predict a set of bounding boxes, encoding plausible spatial locations and shapes for the person; then, we predict a set of possible poses based on the bounding boxes.

3.2.1 Predicting Human Bounding Box

The first stage of our model is a conditional VAE that encodes bounding boxes (parameterized as affine transformation matrices θ), conditioned on RGB image, semantic segmentation and depth map. At the bottleneck, we tile the latent embeddings z_θ of the affine matrices to the same spatial dimension as the image. Specifically, the 4-dimensional latent code is tiled into a dimension of $H \times W \times 4$, where H and W are image height and width, and then concatenated to the image, segmentation and depth map channel-wise as input for the decoder. Following the cVAE convention, our loss is formulated as $\mathcal{L}_{kld\_bbox} + \mathcal{L}_{1\_bbox} = KL(z_\theta||\mathcal{N}(0,1)) + ||\theta - \hat{\theta}||_1$, where $\mathcal{L}_{kld\_bbox}$ is Kullback-Leibler divergence ensuring the encoded distribution of θ follows a Gaussian distribution and $\mathcal{L}_{1\_bbox}$ is a L1 reconstruction loss between the predicted affine matrix $\hat{\theta}$ and ground truth θ regularizing the one-to-one mapping between latent space and prediction space. During inference, we draw from Gaussian distribution $\mathcal{N}(0,1)$ to produce a set of bounding boxes for plausible human locations and shapes.

In order for the bounding box generator to produce diverse results that satisfy both semantic plausibility and geometric correctness, we use a semantics-and-geometry-aware discriminator as an additional module. Leveraging Spatial Transformer Network (STN)[16], we transform a $H \times W$ canonical mask with all ones to binary masks $\hat{B}$ and B with predicted and ground truth affine matrices $\hat{\theta}$ and θ. $\hat{B}$ and B represent predicted and ground truth bounding boxes and are spatially aligned with the image. We then concatenate them with the image, semantics and depth map, respectively as fake and real examples for the discriminator.

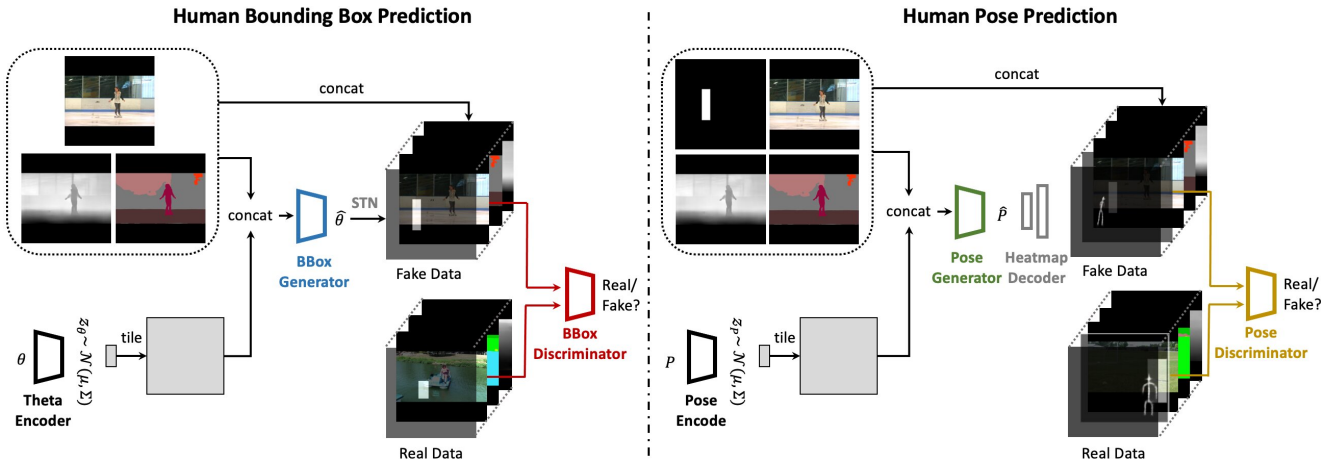

Figure 3: Human Affordance Prediction Model. **Left**: Our human bounding box prediction model first encodes the ground truth affine matrices into a Gaussian distribution, reparameterizes the code, and then tiles the code into the same spatial dimension as the image. The concatenated tiled code, image, segmentation and depth are fed into a bounding box generator to predict the human location, which is checked by a conditional discriminator. **Right**: Our human pose prediction model follows a similar pipeline as the bounding box prediction model. Note that the pretrained heatmap decoder maps pose coordinates into a heatmap in order to spatially align the prediction with the image.

The adversarial loss is formulated as follows:

$$\begin{aligned}\mathcal{L}_{adv\_bbox} = \mathbb{E}_{data}\left[log(D(B|I,S,D))\right] \\ + \mathbb{E}_z\left[log(1 - D(\hat{B}|I,S,D))\right]\end{aligned} \tag{1}$$

where I, S, D refers to the RGB image, semantic segmentation and depth map.

Our insight is that this explicit conditioning of the discriminator is crucial for the network to pick up semantics and depth cues in generating realistic results. For example, the height of the predicted bounding boxes should be similar to those at approximately the same depth level. Moreover, the bounding box size should be roughly proportional to the depth in a scene. The overall objective of the bounding box predictor can be written as:

$$\mathcal{L}_{bbox} = \mathcal{L}_{kld\_bbox} + \mathcal{L}_{1\_bbox} + \mathcal{L}_{adv\_bbox} \tag{2}$$

3.2.2 Synthesizing Human Pose

Our human pose prediction model follows a similar pipeline. In this case, the conditional VAE model encodes 2D human joint locations P into latent embeddings z_P. Its loss is written as $\mathcal{L}_{kld\_pose} + \mathcal{L}_{1\_pose} = KL(z_P||\mathcal{N}(0,1)) + ||P - \hat{P}||_1$. To disentangle location from pose, we normalize all ground truth poses to $[0, 1]$ within their bounding box labels as a preprocessing step.

In addition to RGB-D and semantics information, we also feed in predicted bounding boxes $\hat{B}$ from the first stage to the cVAE decoder and discriminator, so that the pose synthesis is conditioned on a specific location. To spatially align the predicted 2D joint locations with the image, we use a pretrained pose heatmap renderer $\phi(\cdot)$, consisted of a few transposed convolutional layers, and map the normalized pose joints and the bounding box to a heatmap same size as the image. Then both predicted and ground truth pose heatmaps are fed into the discriminator as the fifth input (other four being RGB-D, semantics and bounding box mask) for adversarial learning. Our heatmap renderer ϕ is fully differentiable and frozen during training, allowing gradients to pass to the cVAE decoder. The adversarial loss is formulated as follows:

$$\begin{aligned}\mathcal{L}_{adv\_pose} = \mathbb{E}_{data}\left[log(D(\phi(P)|B,I,S,D))\right] \\ + \mathbb{E}_z\left[log(1 - D(\phi(\hat{P})|B,I,S,D))\right]\end{aligned} \tag{3}$$

The overall objective for the pose predictor can be written as:

$$\mathcal{L}_{pose} = \mathcal{L}_{kld\_pose} + \mathcal{L}_{1\_pose} + \mathcal{L}_{adv\_pose} \tag{4}$$

We first pretrain the bounding box and the pose module respectively, and then jointly train them with a lower learning rate. More implementation details are described in the supplementary materials.

3.3. Location2Object in the Driving Scene

In the Location2Object task, our goal is to predict what to put at a user-specified location in an image and how large it should be. This could prove useful in interactive content generation, where the user can enrich image content easily by clicking on a place to insert objects.

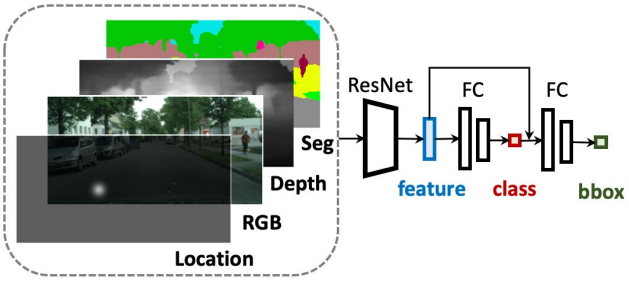

Figure 4: The Location2Object model takes location Gaussian mask, image, depth map, and semantic segmentation map as inputs, and feed them into a two-stage network to predict semantic class and bounding box shape sequentially.

We propose a simple yet effective two-stage network that first predicts the semantic class of the object to be inserted, and then generates its bounding box based on the predicted class. The overall model pipeline is shown in Figure 4. As a preprocessing step, we convert the user-specified location into a Gaussian mask so that it is spatially aligned with the image. We then concatenate the pre-computed semantic segmentation and depth map to the image and the Gaussian mask as input to our network. We use ResNet-18 [14] as backbone to extract visual features and two sets of fully-connected layers in sequential order to predict the class label and the bounding box shape (height and width). We use a weighted cross-entropy loss as training objective to relieve imbalanced class distribution in the dataset, which is represented as follows:

$$\mathcal{L}_{class} = \frac{\sum_{i=1}^{N} -weight[label[i]] \log \frac{x[label[i]]}{\sum_{j=1}^{C} \exp(x[j])}}{\sum_{i=1}^{N} weight[label[i]]} \quad (5)$$

We also use a L1 loss to minimize the difference between predicted and ground truth bounding box shapes $\hat{B}$ and B, written as $L_{bbox} = ||\hat{B} - B||_1$. The overall objective is $L = L_{class} + L_{bbox}$.

3.4. 6D Object Pose Hallucination

In this task, we wish to find the translation and rotation required to place an object into a scene, provided with the class label of the object and a context image. Essentially, the network is asked to hallucinate the 6D pose of the object to be inserted in the target scene.

Similar to the Location2Object model, we design a two-stage regression network that first predicts the translation, then estimates the rotation based on the predicted translation. We use Resnet-18 [14] as backbone for feature extraction, as shown in Figure 5 below.

We use (x, y, z) coordinates to represent translation and quaternions to represent rotation. Compared to Euler angles, which suffer from a singular case known as the gimbal lock, quaternion is more compact and numerically stable. We denote predicted and ground truth translation as $\hat{T}$ and T, and predicted and ground truth rotation as $\hat{R}$ and R. Our overall objective can be written as $\mathcal{L}_{total} = \mathcal{L}_{translation} + \mathcal{L}_{rotation}$, where $\mathcal{L}_{translation} = ||T - \hat{T}||_2$ and $\mathcal{L}_{rotation} = ||R - \hat{R}||_2$.

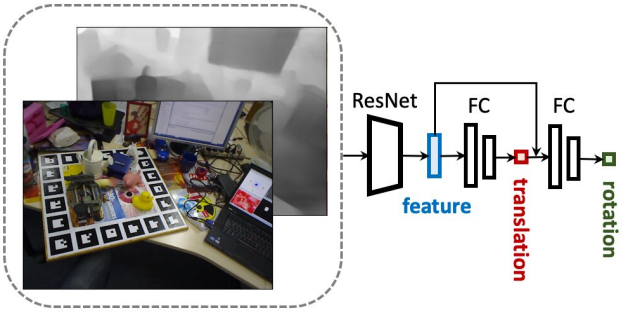

Figure 5: We use ResNet [14] to first extract features, and sequentially predict the object 3D translation and 3D rotation.

Since this task requires understanding of 3D geometry of the scene, we concatenate the pre-computed depth map to the RGB context image as input to our network. Our experiments in the next section demonstrate the superior performance of the model using quaternions and depth information.

4. Experimental Results

4.1. Preliminaries

For human affordance prediction task, we experiment on the MPII dataset[2], which includes 25K images containing 40K people with 410 human activities. We run the Inpaint2Learn pipeline to generate data pairs consisted of ground truth pose labels and inpainted context images.

For Location2Object task, we use the Cityscape dataset [6] and our Inpaint2Learn pipeline to generate a ground truth object class label, a bounding box shape and a context image free of the object as a data pair, yielding 222,844 training data points and 10,000 test data points.

For 6D object pose hallucination task, we use the Linemod dataset[15], which contains 15 different types of objects with ground truth 6D pose labels and roughly 1.4k images per object. We run our Inpaint2Learn pipeline to generate context images free of the objects and use a train-test split of 9:1.

We plan to release the generated affordance labels in all three datasets to the research community in the near future. Due to space limit, please see the supplementary materials for more training and implementation details. In the following subsections, we will discuss and analyze the experimental results for each task.

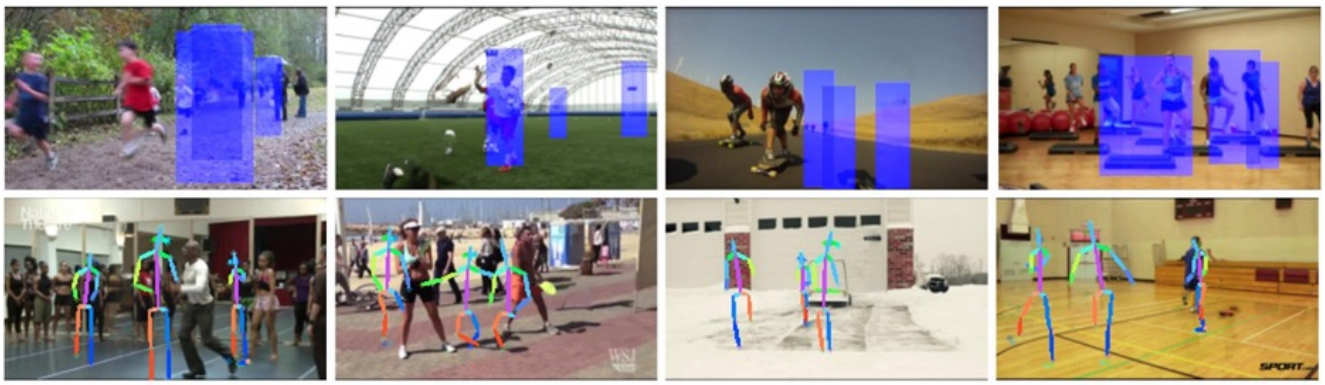

Figure 6: Qualitative results of human location/shape (first row) and pose (second row) predictions in the wild. The results show that our model can predict bounding boxes with various sizes at different depth levels. In addition, the predicted human poses also have consistent sizes with respect to the other existing people in the image.

Methods	BBox Score (AMT) ↑	Pose Score (AMT) ↑	Diversity ↑
Heuristic	54.13	58.70	-
RGB Only	63.39	68.14	0.0977
+ Depth	67.97	71.54	0.0989
+ Segmentation	66.36	74.14	0.0779
+ Depth & Segmentation	70.75	75.94	0.1022

Table 1: Quantitative results of human affordance prediction model. The BBox Score and Pose Score are the averaged user preference rate among 30 AMT workers over 50 testing images. The diversity is the averaged pairwise distance between ten sampled predictions across 1,000 testing images.

4.2. Human Affordance Prediction Results

We evaluate our human affordance prediction results from two aspects: 1) plausibility of the predicted bounding boxes in terms of locations and shapes, and 2) correctness of the synthesized human poses. We conduct extensive user studies using Amazon Mechanical Turk (AMT). Our survey contains 50 test questions, where we provide an image overlaid with a predicted bounding box and the same image visualized with a synthesized human pose, and ask the subjects whether the box and the pose look reasonable to them in two separate sub-questions. We recruit 30 subjects for the survey and average the evaluation scores among all subjects. The survey description and user interface can be found in the supplementary materials.

4.2.1 Comparison with Heuristic Baselines

We design a heuristic baseline to compare with our learned model. We empirically define the bottommost visible joint as the contact point of the person with the environment. We manually decide if a semantic segmentation category is stand-able or sittable. For all training data, we group contact point locations and poses by semantic class, obtaining location/pose banks for each class that we later sample from. We also compute the height of each person and group it by the disparity value at the contact point. We then compute the average for each disparity bucket and use it as the recommended height for that disparity.

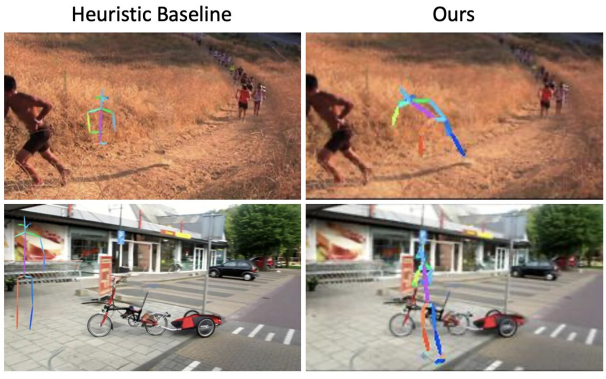

Figure 7: Visual comparison between heuristic baseline and ours.

During inference, given an RGB image, we look at its segmentation map to see if any stand-able or sittable surface is present. If yes, we choose to randomly place the person within the lower part of the largest stand-able or sittable surface, because we observe empirically that the con-

tact points, usually being the person's feet or pelvis, fall in the bottom part of the image. Then the pose is randomly selected from the bank of poses of that class. If no such surface is present, we randomly select a location from all known contact points and obtain its corresponding semantic class, which determines which group of poses we then randomly sample from. Finally, we scale the selected pose to the recommended height according to the disparity value of the contact point.

As shown in Table 1, all of our learned models perform better than the heuristic baseline. In Figure 8, we see that our learned model is able to infer interactions with objects or people in the environment, while the heuristic baseline, though provided with the same segmentation and depth info, fails to capture such intricacies.

4.2.2 Useful Auxiliary Information

We conduct experiments on our full model and three ablation baselines. They each take the following inputs: 1) RGB image, 2) RGB image + depth, 3) RGB image + segmentation, 4) RGB image + depth + segmentation (full model). As shown in Table 1, our user study indicates that our full model generates the most plausible results among the four. To quantify the diversity of the affordance prediction models, we also compute the averaged pairwise distance between ten samples of poses in the validation set. As indicated by the scores, our full model achieves on-par or slightly better diversity compared to other baselines. This shows that adding segmentation and depth information as input improves the plausibility of the predictions without sacrificing their diversity.

Figure 6 demonstrates some qualitative results of our model. We observe that our predictions are diverse and plausible: the synthesized poses occupy various locations in the image and tend to interact with the environment and other existing humans. Moreover, the bounding boxes shrink as they go further away from the camera, leading to the vanishing point. Figure 8 shows a qualitative comparison between baseline model with RGB input and our full model with additional semantics and depth input. Compared to baseline, the full model generate more diverse and realistic results in terms of semantics and geometry. The baseline model suffers from a small variation in locations and shapes and tends to generate bounding boxes of the same size.

4.3. Location2Object Results

In the Location2Object task, our model predicts a categorical label and a bounding box shape for an object to be inserted at a user-specified location. Following the same setup as before, we conduct experiments on our full model and three baselines with the following inputs: 1) RGB, 2) RGB + depth, 3) RGB + segmentation, 4) RGB + depth + segmentation (full model). As evaluation metrics, we compute object classification accuracy and the mean squared error of the bounding boxes using the Inpaint2Learn pseudo ground truth labels in the test set. As shown in Table 2, our full model outperforms the other baselines under both metrics. In addition, we conduct an AMT user study of 100 test images for each model among 20 subjects. In the survey, the subjects are asked whether they think the predicted object categorical labels and bounding boxes look reasonable. The user study results are consistent with our quantitative evaluation, where the full model is either on-par or better than the baselines. Both studies demonstrate the importance of adding semantics and depth information when learning object affordance in the Cityscape driving scene[6].

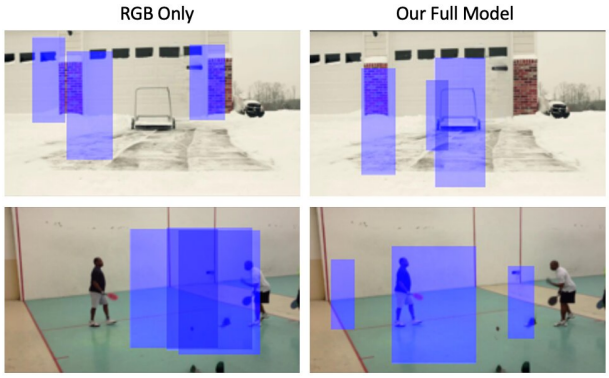

Figure 8: Comparison between the baseline model trained with RGB image only (left) and our full model trained with both segmentation and depth map (right). We can see that our full model can produce bounding boxes with more diverse shapes and sizes, which are also more geometrically plausible with respect to the depth.

4.4. 6D Object Pose Hallucination Results

In the 6D object pose hallucination task, our model predicts 3D rotation and translation parameters to insert an object into a scene. We compare our full model (using quaternion representation and with depth input) with two baselines without depth input, one using Euler angles and the other using quaternions. In evaluation, we convert quaternion results to Euler angles and compute the mean squared error of rotation and translation bewteen prediction and ground truth labels in the test set. As shown in Table 10, using quaternion representation significantly increases the prediction accuracy, and the depth input further improves the performance. In addition, we conduct an AMT user study consisted of 60 questions and recruit 20 subjects to evaluate whether the predicted 3D bounding boxes are plausible placement of the objects. The results confirm that using quaternion representation aids prediction, as satisfaction rate is boosted from 13.6170% (Euler angles) to 52.3404%

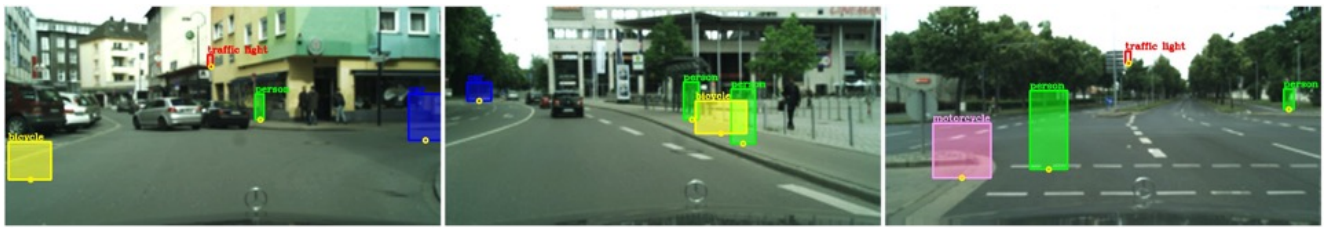

Figure 9: Qualitative results of Location2Object model prediction. The results show that our model can reliably predict not only the common objects, such as person and car, but also relatively rare categories, such as stop sign, motorcycle, etc.

Methods	Classification ↑	BBox Error ↓	Semantic Score (AMT) ↑	Shape Score (AMT) ↑
RGB Only	92.56	7.9379	80.13	67.83
+ Depth	94.87	7.5095	85.61	77.25
+ Segmentation	95.18	7.7919	89.02	73.16
+ Depth & Segmentation	95.51	7.3762	88.71	79.78

Table 2: Quantitative results of Location2Object model prediction. The classification accuracy and bounding box error are computed with respect to the ground truth in the test set. The semantic score and shape score are the averaged preference rate from 20 AMT workers over 100 testing images.

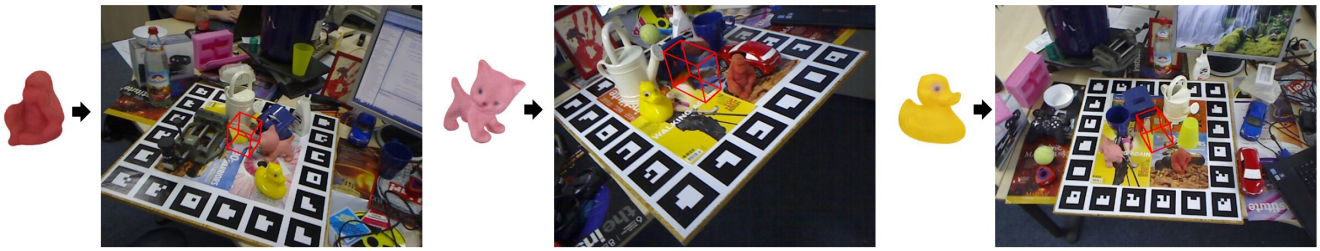

Figure 10: Qualitative results of 6D object pose hallucination in context.

Methods	Translation Error ↓	Rotation Error ↓	User Study (AMT) ↑
Baseline	0.0351	0.1675	13.6170
Use Quaternion	0.0274	0.0787	52.3404
Use Quaternion & + Depth	0.0265	0.0569	59.5745

Table 3: Quantitative results of 6D object pose hallucination in context. The translation error and rotation error are computed with respect to the ground truth in the test set. The user study scores are the averaged preference rates from 20 AMT workers over 60 testing images.

(Quaternions) and that the additional input of depth is helpful to the task.

5. Conclusion

In this work, we propose a task-agnostic framework to generate ground truth labels for affordance learning. To demonstrate its effectiveness, we apply it to three different aspects of affordance prediction: human affordance prediction in the wild, Location2Object in driving scenes and 6D object pose hallucination. In human affordance learning, we adopt a two-stage where and what training pipeline with a new semantics-and-geometry-aware adversarial learning strategy. Our experiments show that our proposed models in all three tasks are able to generate diverse and visually plausible predictions, which are verified through our user studies. Finally, we plan to release all three affordance datasets with strong benchmark to the research community in the near future.

As one potential extension, we plan to expand human affordance learning to the video domain. We will leverage video inpainting techniques [17, 30, 20, 4] and study plausible human environment interactions in motion, which should yield better understanding of the scene and its affordances.

References

[1] Stephen J. Anderson, Noriko Yamagishi, and Vivian Karavia. Attentional processes link perception and action. *Proceeding of the Royal Society: Series B*, 269(1497):1225–1232, June 2002. 2

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 5

[3] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning conditional image composition. 2018. 3

[4] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9066–9075, 2019. 8

[5] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5, 7

[7] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Gregory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[8] Thanh-Toan Do, Anh Nguyen, Ian Reid, D. Caldwell, and N. Tsagarakis. Affordancenet: An end-to-end deep learning approach for object affordance detection. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5, 2018. 2

[9] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 3

[10] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[11] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979. 1, 2

[12] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *CoRR*, abs/1807.06775, 2018. 1

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 5

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 3

[17] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019. 8

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[19] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 10414–10424, Red Hook, NY, USA, 2018. Curran Associates Inc. 2, 3

[20] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4413–4421, 2019. 8

[21] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 3

[22] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *CVPR*, 2019. 1, 2, 3

[23] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 3

[24] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 186–201, Cham, 2016. Springer International Publishing. 2

[25] Fuwen Tan, Crispin Bernier, Benjamin Cohen, Vicente Ordonez, and Connelly Barnes. Where and who? automatic semantic-aware person composition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1519–1528. IEEE, 2018. 3

[26] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[27] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 1, 2

[28] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 611–620, 2020. 3

[29] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 3

[30] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2019. 8

[31] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 3

[32] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3653–3662, 2019. 3

[33] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *Computer Vision – ECCV 2020*, 2020. 3

[34] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, Nov. 2020. 3

[35] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[36] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 408–424, Cham, 2014. Springer International Publishing. 2