Single Image Furniture Placement by Indoor Commonality Relationships

Yating Luo
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, China
13531665402@sjtu.edu.cn

Xiaohan Mao
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, China
mxh1999@sjtu.edu.cn

Yifan Xu
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, China
xuyifan_frank@sjtu.edu.cn

ABSTRACT

Where to place a piece of furniture in a room is a very common problem. Previous works usually put the problem in 3D space since 3D images carry more information than 2D images. However, 3D images are relatively rare compared with 2D images, which limits the application of those 3D methods. Solving the problem in 2D space can be more practical in the real world. Therefore, we propose a new method for 2D furniture arrangement. Given an image of the furniture and another image of the indoor background, our method can generate a heatmap on the background image, indicating how suitable each position is for the furniture arrangement. For each position of furniture arrangement, we detect relationships between the furniture and other indoor objects and use statistical methods to predict the confidence scores of these relationships. We judge how well a position fits based on both confidence and commonality of these relationships. We demonstrate that our method fully considers relationships between indoor objects and can infer reasonable areas for placing the furniture.

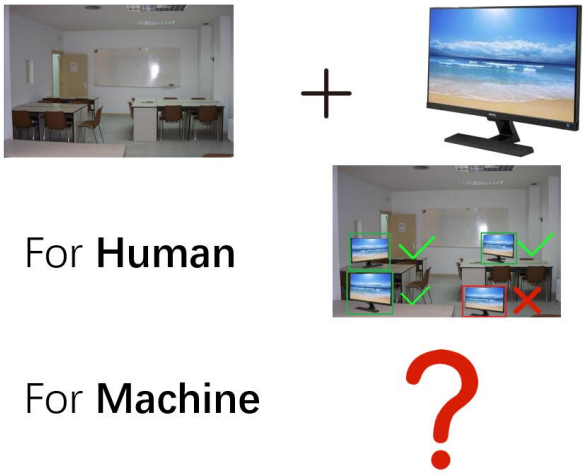


Figure 1: The overview of the task

1. CCS CONCEPTS

•**Computing methodologies** → **Computer vision**
•**Computing methodologies** → **Computer graphics**

Keywords

furniture arrangement; visual relationship detection; Computer Vision

2. INTRODUCTION

Given a piece of the furniture and an indoor scene, where should the furniture be placed in the indoor scene? Which areas are suitable for the placement of the furniture? The problem is very common and practical. In the current situation, people usually solve this problem manually with their prior knowledge and personal preference. But doing furniture arrangement manually in large amounts of data is not a efficient and economic way, So a model should be designed to make these decisions. Previous works [8] usually try to address furniture placement problem from the perspective of 3D modeling. With the help of sufficient information extracting from 3D modeling, the problem can be well solved in 3D space.However, these methods are focusing on 3D images and thus have their own limitations. In the real world, most indoor scenes are 2D images. The number of 3D images are relatively small and it is not easy to convert 2D scenes to 3D. Therefore we want to address this problem on the 2D level with only a single scene image as input.

On the 2D level. People also have some ideas similar to furniture placement. One of them is using an empty house plan to generate an house floor layout by adding furniture one by one. [6]. Given the top-down view of an indoor scene, they build a model to extract free space information from the current view and then select a suitable objects from the furniture list to place in a proper position. This method is very beneficial to make a overall furniture arrangement. In the real world, Only using the features extracting from the top-down view is not available for small objects placement.To some extent top-down view of an indoor scene is also difficult to obtain, so we want to address this problem

.

with front view scene images.

Another idea comes from using the consistency of the image. In this paper [1], for a specific object on the image, they consider whether there are other possible positions to place it. By comparing the background appearance of the target object with the background appearance of other positions, they generate a heatmap to demonstrate the consistency of the image, from which it's easy to judge which position is suitable. Why didn't we migrate this method to our task? The reason is that in our task the furniture we need to place may be not exist in the image before. we have no background appearance information. If we use the furniture's background appearance information extracted from other images, it will introduce high bias.

Our proposed method solved furniture placement problem on the 2D level with only a single front-view image as scene input. We place target furniture in each possible position and output a heatmap to demonstrate which areas are better for placement. Our method is based on the relationship detection technique. But only using relationship information will cause high bias because relationship detection models more focus on the relationship prediction correctness of two objects without ability to detect whether the current arrangement is reasonable. The key idea of our method is using some prior distance vector distribution and commonality information to give each position a confidence score and commonality score. Combining the features above makes our method perform well in this problem. The experiment result shows that our method can generate a heatmap which is very close to human expectation.

3. RELATED WORK

Furniture Arrangement in 3D Scenes. Many previous work tried to address furniture arrangement problem based on 3D model. To arrange furniture properly, previous work [8] first represents an object with features extracting from the 3D model, such as bounding surfaces, center, orientation and so on, which are not easy to detect in 2D level. With these features and some layout information, they design a model to put furniture into suitable space in the scene. It's the fundamental work of building virtual world.

Indoor Scene Synthesis Given an empty room image, in what order should we place the furniture and where should the furniture be placed? This paper [6] propose a convolution network to address these problems. They generate an indoor scene base on features extracted from top-down view.

Image Consistency. The basic goal of this paper [1] is boosting instance segmentation but they proposed a location probability map based method to explore the feasible locations that objects can be placed based on image consistency.They define a descriptor to represent the background of an object and a function to measure the distance between two background appearance. But this method is only useful for such object that already exists in the image.

Visual Relationship Detection. The visual relationship detection task means that given an image, we not only care about how many objects are there in the image and their corresponding positions, but also want to detect the relationship between the objects using some global and local information. They usually output a probability distribution to demonstrate the relationship.Relationship is expressed as a tuple(object1, prediction, object2). Many previous work have tackled this task from different angles.The paper [4] separates the task into two parts: training object detection and training relationship prediction for the difficulty in predicting a relationship tuple directly. Meanwhile it combines semantic features to solve zero-shot problem. People have also attempted to map the features of two objects to a vector space and use the translation vector between them to represent the relationship [9]. In this work [7], people tried to construct a scene graph with objects relationship involved.Its advantage is to predict each relationship using global contextual information. Visual relation detection is the basis for understanding a image, so it has many meaningful applications, such as Image Captioning, VQA (Visual Question Answer), etc. Our visual relationship detection module is based on the model proposed by this paper [3], which is a follow-on work of this paper [4].

4. METHOD

In general, the arrangement of the furniture should consider its relationships and consistency with other objects in the room. A good arrangement of the furniture usually means that the furniture meets common indoor relationships with other objects. Based on the basic idea, we propose the following pipeline to generate heatmap indicating which areas are suitable for the furniture. First, we use faster-RCNN [5] to detect all of objects in the indoor scene. Then we search different positions for the target furniture on the scene image and predict a score for each position. For a given position, we input several features of indoor objects and the furniture to a visual relationship detection model. The model can give out the relationships between the furniture and other objects along with their confidence score. Comparing these relationships with statistical information from the indoor image dataset, we quantify whether the position is appropriate for the placement of the furniture. Figure 2 shows our pipeline.

We adopts Deep Structural Ranking (DSR) [3] for visual relationship detection. It uses multiple features to represent the object instance including visual appearance, spatial location and semantic embedding.

1. **visual appearance**. For a relation instance (s, p, o), let $\boldsymbol{b}_s = (x_s, y_s, w_s, h_s), \boldsymbol{b}_o = (x_o, y_o, w_o, h_o)$ be the bounding boxes of the subject and the object. Let $\boldsymbol{b}_u = (x_u, y_u, w_u, h_u)$ be the union of $\boldsymbol{b}_s$ and $\boldsymbol{b}_o$. Using the convolutional neural network, the model extracts the RoI Pooling features of $\boldsymbol{b}_s, \boldsymbol{b}_o$ and $\boldsymbol{b}_u$ from the last convolutional layer as visual appearance features.

2. **spatial location**. The model takes spatial masks (which sets 1 in the bounding box and 0 outside) of the subject and the object as spatial location features and uses convolutional neural networks to convert these features to low-dimensional features. The spatial mask of the union bounding box is also used to suggest spatial relationships.

3. **semantic embedding**. Leveraging language priors from semantic embedding can boost the performance of visual relationship prediction. Thus the models uses semantic the embedding layer to map objects to embedding vectors.

DSR model uses these three kinds of features to predict visual relationships between objects along with their ranking

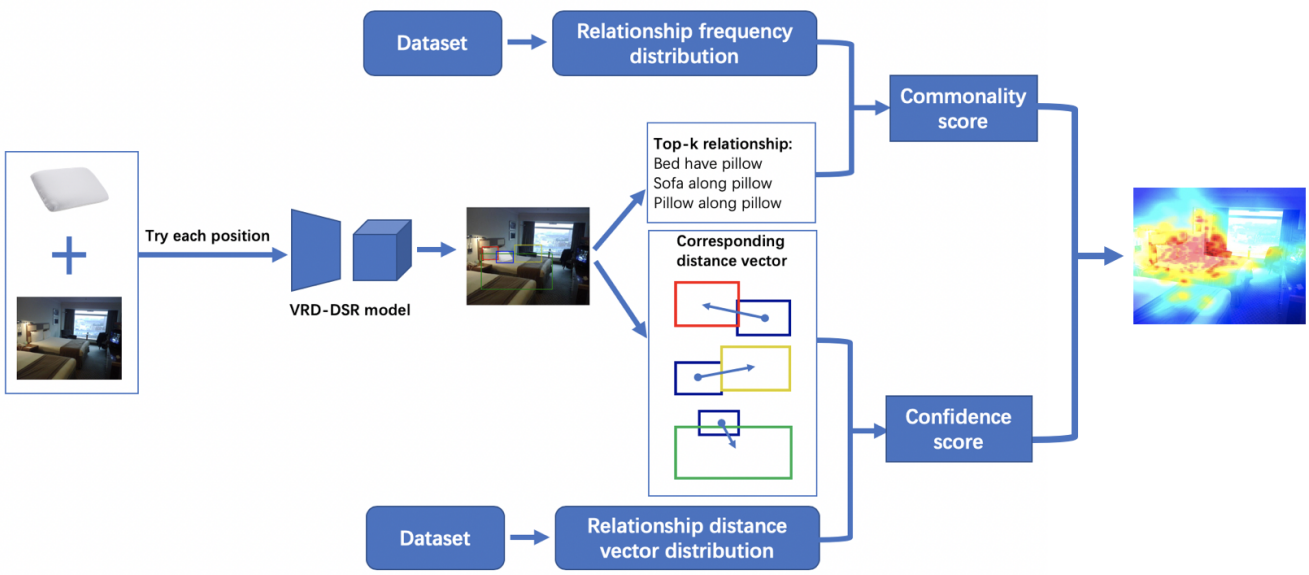


Figure 2: Overall network of the proposed algorithm

scores. The higher the ranking score is, the more likely the corresponding relationship is. In the end the model lists the most likely relationships based on their ranking scores.

We can filter out the relationships involving the target furniture from the results of DSR and leverage these relationships and their ranking scores to infer whether the position is suitable for the target furniture. However, we find that DSR is not sensitive to the variation of the spatial location. We observe that given the subject and the object, the predicate that DSR predicts is usually right. However, even though DSR takes spatial locations as one of the features, it has difficulty telling whether the relative location of the subject and the object is reasonable or not. DSR predicts that bed-have-pillow is the most likely relationship when pillow is on bed, which is reasonable. However, DSR still predicts bed-have-pillow with high ranking score when pillow is far away from bed. If we use the ranking scores provided by DSR to infer the fitness of the position, we may suffer from DSR's insensibility of the relative spatial location of the subject and the object.

Since the confidence scores that DSR gives are unreliable, we predict confidence scores based on statistical analysis of the indoor image dataset. In another word, given a relationship $r = (s, p, o)$, we evaluate a new instance of r based on the existing instances from the dataset. Let $b_s = (x_s, y_s, w_s, h_s)$ and $b_o = (x_o, y_o, w_o, h_o)$ be the bounding boxes of the subject s and the object o, respectively. We define the relative distance vector from s to o

$$d = \left(\frac{x_{oc} - x_{sc}}{w_s}, \frac{y_{oc} - y_{sc}}{h_s} \right) \tag{1}$$

where $(x_{sc}, y_{sc}) = (x_s + \frac{w_s}{2}, y_s + \frac{h_s}{2})$, $(x_{oc}, y_{oc}) = (x_o + \frac{w_o}{2}, y_o + \frac{h_o}{2})$ are centers of b_s and b_o respectively. Suppose that there are n instances of r in the dataset. We can get n relative distance vectors $d_1, d_2, \ldots, d_n$. Assume the relative distance vector of the new instance of r is d'. We want to evaluate the confidence score of the new instance by comparing d' with $d_1, d_2, \ldots, d_n$. The problem is actually using n samples to infer the probability distribution p and p may be complex. Intuitively, if d' is close to some d_i, the confidence score of the new instance is high. Otherwise, if d' is far away from any other d_i, the confidence score is low. Here we take a method which is kind of similar to k-nearest-neighbor. Assume $d_{a_1}, d_{a_2}, \ldots, d_{a_k}$ be the closet k vectors to d' in $\{d_1, d_2, \ldots, d_n\}$, where k is a hyperparameter. The confidence score of the new instance of r is

$$s_r = \frac{1}{\sum_{i=1}^{k} \|d_{a_i} - d'\|_2 + 1}. \tag{2}$$

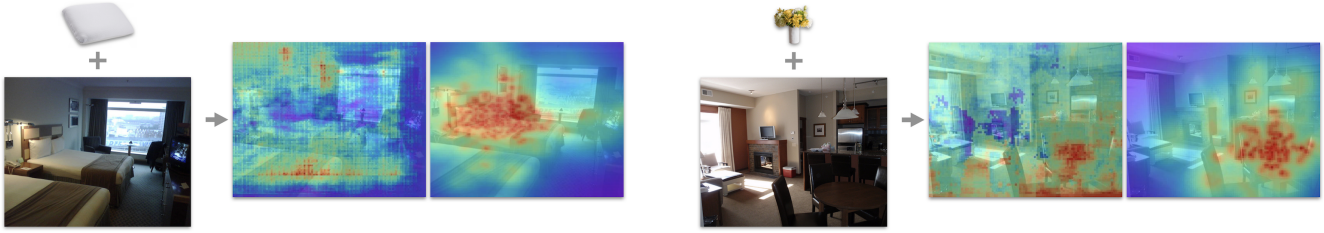
Given the positon (x, y), we use DSR to gain the relationships between the target furniture and other objects in the background and predict the confidence scores of these relationships. Now we decide $h_{x,y}$, the final score of the position (x, y), which indicates whether position (x, y) is suitable for placing the target furniture. Besides considering the confidence scores of relationships, whether the relationship is common indoor should be considered, too. In general, the more common the relationship is, the more suitable the position is. Let t be the target furniture. $r_1, \ldots, r_m$ are m relationships detected by DSR, which all involve t. For any $1 \le i \le m$, p_i is the confidence of the other object besides t involving in the relationship r_i, which is given by the object detector. s_i is confidence score of r_i. And c_i is the number of instances of r_i that appear in the dataset. Then

$$h_{x,y} = \sum_{i=1}^{m} p_i s_i c_i. \tag{3}$$

For every position (x, y) in the background image, we calculate $h_{x,y}$ and then we generate the heatmap of the suitable positions for the target furniture based on h.

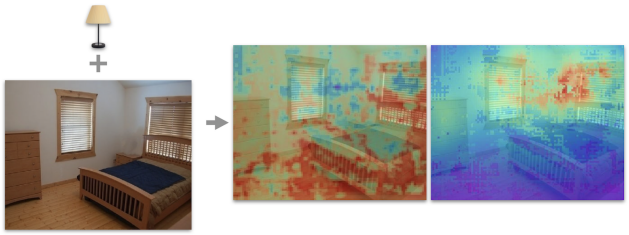
5. EXPERIMENTS

We use Visual Genome (VG) [2] as our dataset, which provides visual relationship annotations for each image. We select 5386 indoor images from VG by checking the keywords

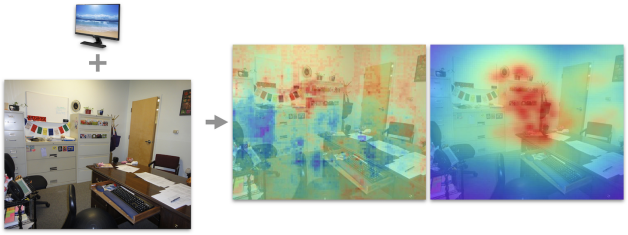


(a) Place a pillow in the bedroom

(b) Place a vase in the living room



(c) Place a lamp in the bedroom



(d) Place a screen in the study

Figure 3: Experimental results of our method. The left heatmap uses ranking scores to predict position scores while the right heatmap uses confidence scores.

in annotations of images. We divide these indoor images into two parts, 4315 for training and 1071 for testing. We choose the most frequent 202 objects and 229 predicates and discard the relationships involving other objects and predicates. In this way we ensure that for most relationships there are relatively sufficient instances.

We train DSR on the indoor dataset for 10 epochs. DSR model can correctly predict most relationships between objects in indoor images. However, the model is not sensitive to different spatial locations of the subject and the object and thus the ranking scores it gives are not very reliable. Why DSR model gives high ranking score to the relationship bed-have-pillow even if the spatial locations of bed and pillow obviously don't match? Our explanation is that the high frequency of the appearance of bed-have-pillow overwhelms spatial irrationality. Hence we use the statistical method to predict the confidence score, which is more aware of spatial locations of the relationship. We set hyperparameter $k = 3$ in equation(2). Figure 3 shows some results. For each figure, the left heatmap is the result of using ranking scores that DSR provides to predict position scores while the right heatmap is the result of using confidence scores. As we can see, Right heatmaps are more reasonable than left ones and suggest some suitable areas for placing the target furniture.

6. CONCLUSIONS

We propose a new method to place a piece of furniture in an indoor background, from a pure 2D perspective. Where to place the furniture mainly depends on its relationships with other objects in the background. Our method first detects objects in the background and chooses a candidate position for the furniture by detecting its relationships with other objects. In order to make up the visual relationship detector's insensitivity of spatial locations, we predict confidence scores of these relationships by statistical analysis. Then we predict how well the position fits based on those relationships and their confidence scores. Our method can infer which areas are good for the arrangement of the furniture reasonably.

7. REFERENCES

[1] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li, and C. Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. *arXiv preprint arXiv:1908.07801*, 2019.

[2] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[3] K. Liang, Y. Guo, H. Chang, and X. Chen. Visual relationship detection with deep structural ranking. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[4] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[5] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[6] K. Wang, M. Savva, A. X. Chang, and D. Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)*, 37(4):70, 2018.

[7] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017.

[8] L.-F. Yu, S. K. Yeung, C.-K. Tang, D. Terzopoulos, T. F. Chan, and S. Osher. Make it home: automatic

optimization of furniture arrangement. *ACM Trans. Graph.*, 30(4):86, 2011.

[9] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017.

Authors' background

Your Name	Title	Research Field	Personal website
Yating Luo	undergraduate student	NLP, active learning	
Xiaohan Mao	undergraduate student	Computer Vison	
Yifan Xu	undergraduate student	Computer Vison,	