BRIDGE: Building plan Repository for Image Description Generation, and Evaluation

Shreya Goyal*, Vishesh Mistry†, Chiranjoy Chattopadhyay‡ and Gaurav Bhatnagar§
Indian Institute of Technology Jodhpur, India
Email: *goyal.3@iitj.ac.in, †mistry.1@iitj.ac.in, ‡chiranjoy@iitj.ac.in, §goravb@iitj.ac.in

***Abstract*—In this paper, a large scale public dataset containing floor plan images and their annotations is presented. BRIDGE (Building plan Repository for Image Description Generation, and Evaluation) dataset contains more than 13000 images of the floor plan and annotations collected from various websites, as well as publicly available floor plan images in the research domain. The images in BRIDGE also has annotations for symbols, region graphs, and paragraph descriptions. The BRIDGE dataset will be useful for symbol spotting, caption and description generation, scene graph synthesis, retrieval and many other tasks involving building plan parsing. In this paper, we also present an extensive experimental study for tasks like furniture localization in a floor plan, caption and description generation, on the proposed dataset showing the utility of BRIDGE.**

***Index Terms*—Floor Plan; Dataset; Evaluation; Captioning**

I. Introduction

In the field of document image analysis, a floor plan is a document containing drawing of the house, apartment or any other building. These are documents which aid architects to show the interior of a building along with components. Floor plan image analysis involves semantic segmentation, symbol spotting and identifying a relationship between them. Tasks such as symbol spotting, thick and thin wall classification, doors and window detection, room and sub-room detection are various aspects in floor plan image analysis. Describing a floor plan in natural language is a task which has applications in areas such as obstacle avoidance, navigation of a visually impaired or robot, spotting a specific type of room or object in the plan. Detection and localization of symbols become mandatory when it comes to describing a floor plan image in words. Automatic synthesis of floor plans from their respective descriptions is another task which may be required by an architect to give a customer rough idea of his/her requirement.

In the literature, the number of floor plan datasets, which are publicly available for research is four. They are: ROBIN [1], CVC-FP [2], SESYD [3], and FPLAN-POLY [4]. However, these datasets contain very less number of sample images, various symbols for furniture, doors and windows and do not contain annotations for objects, and their descriptions. These datasets are primarily constructed to find the solution for problems such as image segmentation, retrieval, and layout analysis. These datasets are not suitable for the purpose of caption generation and description synthesis. Now a days with the advent of deep neural networks, tasks such as symbol spotting, caption generation, retrieval, semantic segmentation are getting more accurate and robust. However, there is a requirement of a large number of samples and corresponding annotations specific to each task for training these models. There are many large scale datasets publicly available in the literature in the context of natural images. Examples include visual genome [5], MS COCO [6], which has a large number of natural images along with their descriptions or captions, object annotations, region graphs, and other metadata. Since the publicly available floor plan datasets are not having a large number of image samples, the deep learning models perform poorly when trained and tested on them.

In order to perform all the above-mentioned tasks efficiently with deep neural networks, a large scale dataset of floor plan images along with their task-specific annotations is the need of the hour. In this work, we propose the BRIDGE (Building plan Repository for Image Description Generation, Evaluation and other purposes) dataset [1]. The rationale behind the nomenclature is that the proposed dataset bridges the two modalities, i.e., image and text for floor plan images. In the domain of floor plan datasets, BRIDGE contains a collection of 13000+ floor plan images. Annotations for symbols, region descriptions, and paragraph descriptions are part of the dataset. The construction of BRIDGE is inspired by the visual genome dataset which includes the annotations for similar tasks in the context of natural images. In this paper, along with the features of the dataset, we also present experimental results on tasks like symbol spotting, region wise caption generation, and paragraph generation, using deep learning models.

The rest of the paper is organized as follows. Section II describes the various state of the art floor plan datasets and experiments performed on them. Section III describes the methods in which dataset was constructed and its contents. In Sec. IV, we present the results of various experiments performed on BRIDGE dataset. Section V concludes the paper.

II. Literature survey

Various floor plan datasets have been proposed in past for purposes such as symbol spotting, retrieval, semantic and layout segmentation. Table I lists out the details of the publicly available datasets, number of samples present in them, and a brief description. There are several techniques in the literature [1], which have used one or more of these four datasets.

This work is partially supported by Science and Engineering Research Board, India under the project id ECR/2016/000953.

[1] Available at http://home.iitj.ac.in/~chiranjoy/research/dataset/BRIDGE.zip

TABLE I
DETAILS OF PUBLICLY AVAILABLE EXISTING FLOOR PLAN DATASETS

Dataset	Details
ROBIN [1]	• 510 samples, 3 classes, 170 plans/ class • Plans with various global layout and room count • Purpose: Symbol spotting and retrieval
CVC-FP [2]	• 122 samples in 4 Sub-Categories, with varying wall textures • Purpose: Semantic segmentation
SESYD [3]	• 1000 samples, 100 layouts/ class • Differ in arrangement of symbols and layout • Purpose: Retrieval, symbol spotting
FPLAN-POLY [4]	• 42 samples, used for symbol spotting

However, none of the above-mentioned datasets connects the images with text modality. If the task is to generate a textual description from the floor plan images, then it is a mandatory requirement to learn the relationship between the various decors and architectural artifacts. Moreover, to have human-like description synthesis, the textual annotation is also necessary. In the context of natural images, there are datasets, for example, visual genome which connects $108,077$ images with textual annotation, 5.4 Million region descriptions and other annotations for various tasks such as caption generations, visual question answering. MS-COCO and MS-COCO captions [7] are examples of datasets which contain over 330000 images and over one and a half million captions (5 captions per image). The dataset is currently being used for caption generation, object segmentation tasks.

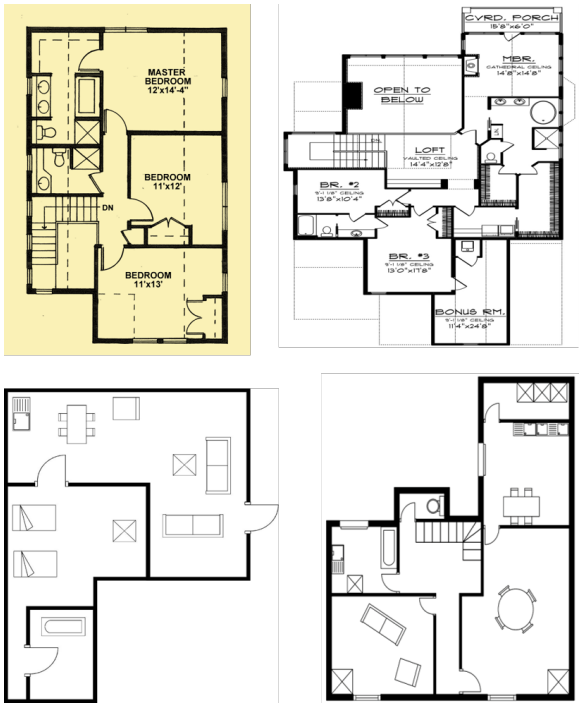


Fig. 1. Sample images in the BRIDGE dataset.

Along with dataset creation, researchers have also proposed algorithms that use these datasets. In order to perform tasks like object detection, several schemes have been proposed before for example YOLO [8], Fast-RCNN [9], Faster-RCNN [10] in the context of natural images. YOLO presents a fast approach for object detection which presents object detection as a regression problem to spatially separate bounding boxes and class probabilities. In Fast-RCNN, authors have proposed a single stage training algorithm that jointly learns to classify object proposals and refine their spatial locations. In the context of caption generation for natural images Densecap [11] has proposed an algorithm which generates region wise captions in images. Hierarchical recurrent network [12] based paragraph generation technique produces a paragraph like description for the entire image. From the above discussion it is clear that in the domain of document images, specifically floor plan, there is lack of a dataset which has a large number of sample images as well as annotations for those images. BRIDGE is the first dataset that caters both the requirements and can be used for several tasks.

III. CONSTRUCTION OF BRIDGE

To construct the BRIDGE dataset we have followed two approaches. First, we have collected floor plan images from the publicly available datasets (i.e., ROBIN, SESYD etc.). In the second approach, we have collected the remaining floor plan images from the internet. In total, we have over 13000 floor plan images in this dataset. Along with the images BRIDGE also has object annotations, region descriptions, and paragraph description for the floor plans. Till date, this is the largest annotated floor plan dataset created for the document analysis and research (DAR) community. For creating annotations we asked volunteers for marking bounding boxes around each decor items. We used LableImg graphical annotation tool [13] for marking the bounding boxes in the images. For generating region descriptions also we used the same tool and later converted them in the JSON format.

A. Floor Plan images

Along with the images obtained from available public datasets, images were collected from two websites, www.architecturalhouseplans.com and www.houseplans.com. These websites contain multiple floor plan images for a single house design for both single storied and multi-storied buildings. The similarity between the images taken from both websites is that the floor plans belong to real homes, available for a customer to use and they are not generated for any specific task for example retrieval or segmentation. They are similar in the symbols of objects used in them. There is a diversity between the layouts and complexity of the plans. Each image in the dataset has a unique ID (image name) which depicts the source where it has been obtained from along with the floor plan images of the same house for the different floor. Figure

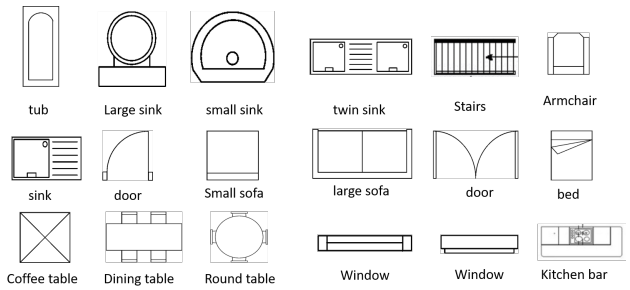


Fig. 2. Various classes of decor items available in the dataset.

1 shows some of the sample floor plan images available in BRIDGE. There is a lot of variability in the formation of the images in the dataset. Floor plan images from SESYD dataset are synthetically generated, while images from ROBIN are handcrafted. Floor plan images collected from the web are standard plans designed by the architects. All the images taken did not have any ground truth data or annotations available which are required of any machine learning algorithm to train. Hence, we further proceeded to create annotation for various tasks and evaluation purposes.

B. Symbol Annotations

Detection of several decor items is an important step when for parsing a floor plan image and information extraction. Object detection schemes have been used in the context of objects in natural images. In the line of architectural drawings, techniques involving handcrafted features is used multiple times in the literature. However, techniques using deep neural networks are still needed to be explored in the context of document images. For the task of symbol detection and localization, symbol annotations were required. Figure 2 shows various decor symbols for 16 classes of objects in the proposed dataset. Each annotation file is in XML format with information such as image name, path, the bounding box for each decor symbol with their names. Figure 3 (a) shows the sample annotation XML for objects. All of the symbol annotations were generated using LabelImg tool [13].

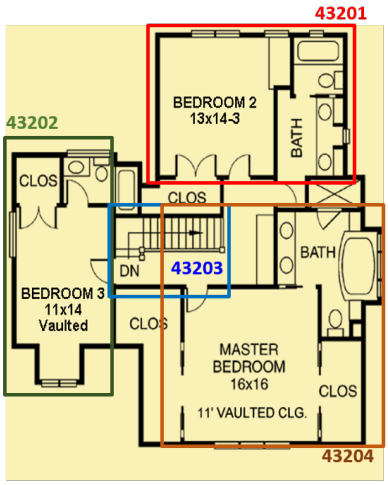
C. Caption Annotations

In the literature, there are image datasets with image captions (MS-COCO) and region wise captions (visual genome). For a floor plan, region wise caption generation is an important step. In [14], [15], authors have used handcrafted features for identifying decor symbol, room information and generating region wise caption generation. Deep neural networks, like CNN, RNN, and LSTM have shown superior performance for natural images for the same task. However, the same can not be translated for floor plans due to the lack of data and pre-trained models. To compensate this issue we annotated image regions by describing them in the form of a dictionary. The region descriptions are inspired by the region descriptions provided in the visual genome dataset. The regions in the floor plan images are taken as different rooms, for example, bedrooms, kitchen, living room etc. For each image, region description contains a region ID which is unique over the entire dataset. Additionally, coordinates of the bounding box for each region as x,y coordinate of the top left point, height and width along with a describing caption with the field "phrase" is also given. Figure 3(c) shows the sample region descriptions created for one of the floor plan images in Fig. 3(b)

D. Paragraph based description annotations

Going further in the line of describing images with captions as well as in the form of paragraphs, there was a requirement of paragraph based textual descriptions which described images in free flow with variability. Previous works [14] [15]

(a) Symbol annotation



(b) Regions in floor plan

[
{"image_id": 432,
"regions":[{"region_id": **43201**, "width": 220,
"height": 160, "image_id":432, "x": 5, "y": 227,
"phrase": "the bedroom is with a private bathroom with dual sink, tub and toilet space"},
{"region_id": **43202**, "width": 140, "height": 270,
"image_id": 432, "x": 150, "y": 5, "phrase": "another bedroom with private bathroom having tub, sink and walk in closet"},
{"region_id": **43203**, "width": 250 "height": 90,
"image_id":432, "x": 255, "y": 170, "phrase": "There are stairs to the other floors"},
{"region_id": **43204**, "width": 190 "height":310,
"image_id":432, "x": 256, "y": 240, "phrase": "master bedroom has a master bathroom with dual sink, tub, toilet and closet"}]
}
]

(c) Region Description

Fig. 3. Illustrations of the several steps of annotation process.

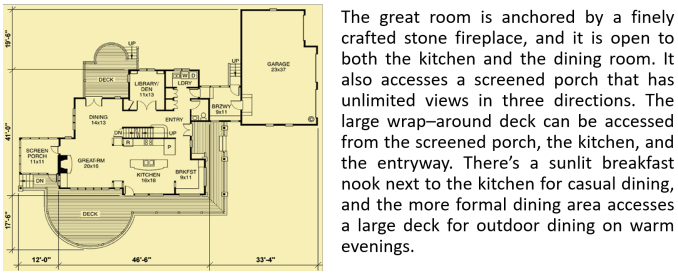


Fig. 4. Paragraph annotation for an example floor plan image.

have generated template based paragraph descriptions which was generated by concatenating region based captions. To have more variability in description there was a requirement of using RNN based deep learning models which require a bulk amount of data for training. Hence we annotated floor plan images with paragraph based annotations. Along with images, we have also collected the descriptions given on the web pages for each floor plan image. The images which did not have descriptions with them were written by the volunteers. The paragraph descriptions were collected in the form of raw text and cleaning was done by removing extra white spaces, alpha-numeric characters, new line characters, and non-ASCII characters by using NLTK tool. All the paragraph annotations are converted into a JSON file for free flow description synthesis of floor plan images. Figure 4 shows paragraph annotation for the image shown. In the next section, we demonstrate experimental results for the various state of the art techniques to prove the usability of our proposed dataset.

IV. Experiments

All the experiments on the proposed dataset were performed on a system with NVIDIA GPU Quadro $P6000$, with 24 GB GPU memory, 256 GB RAM.

A. Symbol Spotting

The symbol spotting algorithms are needed when it comes to identifying the decor and other symbols in the floor plan images. In the context of floor plan images, symbol spotting techniques using handcrafted features have been used widely in the literature. With the application of deep neural networks, authors in [16] have explored Faster-RCNN technique for symbol spotting on a very small scale dataset. However, aforementioned task with deep neural networks, trained on a large scale dataset, is hardly explored. We experimented with YOLO and Faster RCNN methods by fine tuning the pre-trained network with the proposed dataset. Fine tuning was done using 2500 annotations and testing was done with another 500 annotations. Figure 5 shows the distribution of various symbols over the training dataset. Results of the symbol spotting on BRIDGE are described next.

1) YOLO: YOLO is a single Convolutional network, which simultaneously predicts multiple bounding boxes and class probabilities (confidence value) of those boxes. It defines confidence score as $Prob(object) * IoU$, where IoU is the intersection of union between predicted bounding box and the ground truth bounding box (Eq. 1) and $Prob(object)$ is the probability of detecting the object in that bounding box. We fine-tuned the pre-trained tiny YOLO network with BRIDGE for 16 classes of objects. The original network has 9 convolutional layers, having max-pooling layers in between, where the final layer had 105 filters (for our dataset) and linear activation function. Figure 6 (a) shows the average precision observed for each object after testing. Figure 7 shows the detected and localized objects using YOLO with their respective confidence score.

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{1}$$

2) Faster RCNN: There are two modules in Faster RCNN; (i) A deep fully convolutional region proposal network, which proposes regions, (ii) a Fast-RCNN detector. The proposed regions are used for their classification. Figure. 6(b) shows the average precision observed for each object using faster-RCNN technique. Figure 8 shows the detected and localized objects using faster-RCNN and respective confidence score.

B. Caption generation

Captioning an entire image is a task which has been explored widely on natural images. A caption is a single line sentence consisting of information of the entire image. However, in case of floor plan images, generating a single line caption for the entire plan is insufficient for accurate representation. Hence it is required that captions to be generated region wise. Dense captioning is one such task which generates a caption for regions over the images. In this context, symbol spotting is a special case of dense captioning where the target labels generate one word. Along with fully convolutional network, the network has a fully convolutional localization layer which proposes a region of interest and respective confidence score. Recognition network and RNN language model succeed the previous networks where regions are refined by the earlier one and captions are generated by the later one. Figure 9(a) depicts the requirement of a trained dense captioning model on document images. The available data and pre-trained networks do not work on floor plan images because they are trained

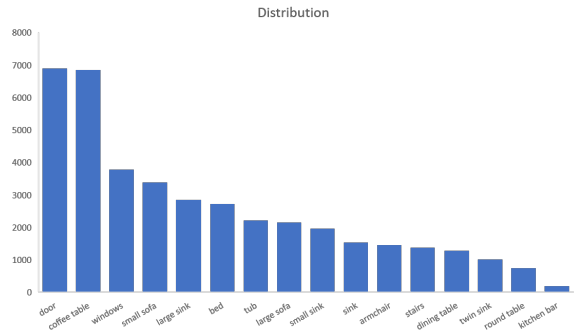
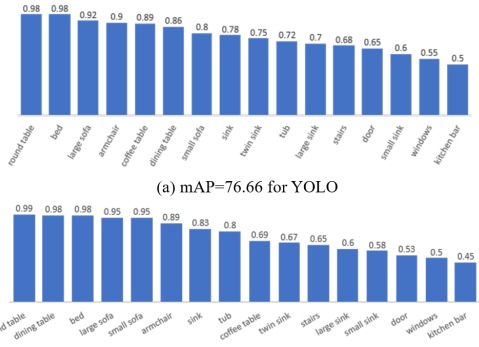


Fig. 5. Distribution of various objects in the training dataset.



(a) mAP=76.66 for YOLO

(b) mAP=75.25 for Faster-RCNN

Fig. 6. Quantitative analysis of the symbol spotting task.

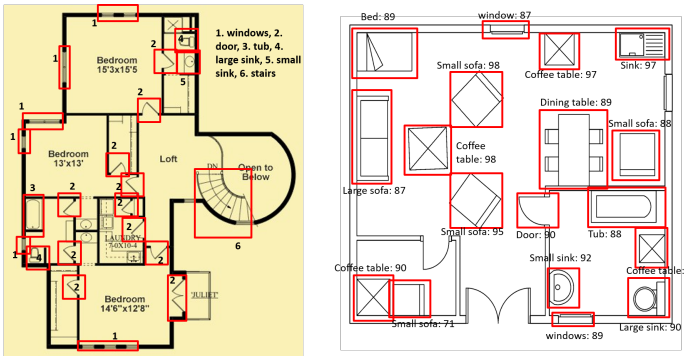


Fig. 7. Symbol spotting using YOLO.

on natural images. Hence there was a need of dataset having floor plan images along with their region descriptions as shown in Fig. 3(b), (c). Figure 9(b), (c) shows the results of dense captioning task on the floor plan images from BRIDGE. The generated caption for a specific region is depicted in the same color as the bounding box for the region.

In the publicly available models for natural images, maximum number of tokens/caption is 15. For floor plans, we released the limit since these images require more words for a caption to make sense. The word which appears less than 15 times is given a special token ⟨UNK⟩. Figure 9(d) depicts

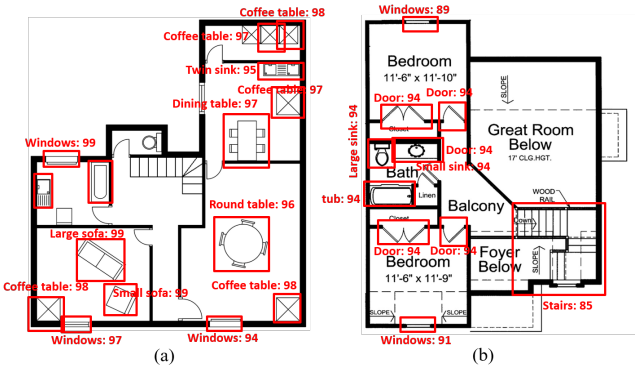


Fig. 8. Symbol spotting using faster-RCNN.

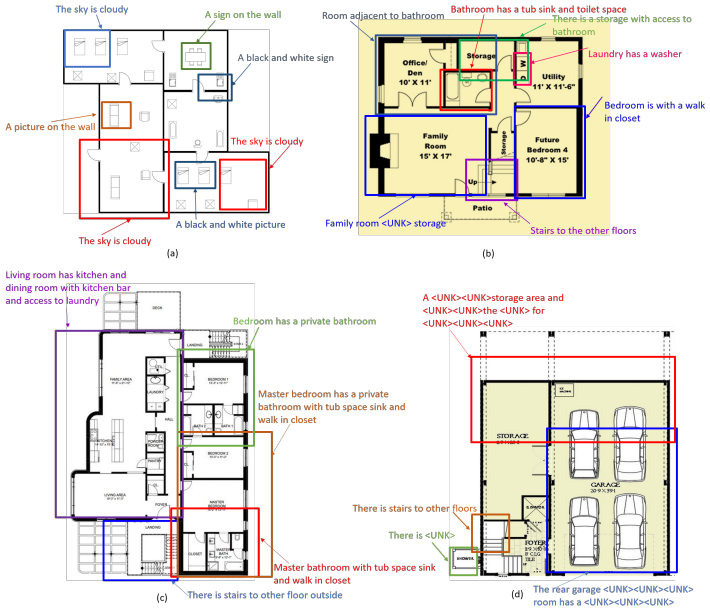


Fig. 9. Results of dense captioning with the proposed dataset.

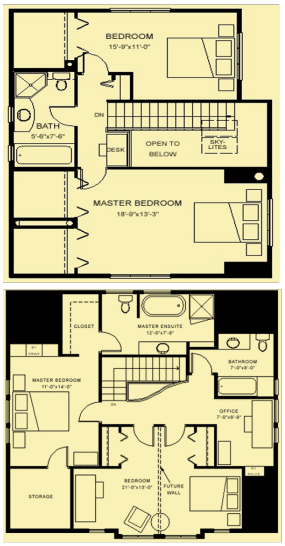
a failure case, where the caption is incorrect, because the detected region is not very common in training dataset. Most of the failure cases are due to less variability in the text data for the floor plan images. The limitation of annotating a floor plan image with region wise description is less variability in the design of floor plans as compared to natural images. Hence it affects the efficiency of the model.

C. Description synthesis

It is insufficient to describe an image by a single caption. Hence we need a system which could generate paragraph based descriptions and has variability. There are many state-of-the-art techniques which have generated paragraphs in the context of natural images. However, these models need annotated images to be used to train the deep neural networks and further test and evaluate the models. There is no such publicly available floor plan dataset to support this task. We have generated paragraphs by using two techniques and evaluated them with the annotations in BRIDGE.

1) Template based: Paragraph based descriptions are generated by using technique proposed in [14]. In this method rooms are detected by semantic segmentation of the floor plan and symbols are detected by symbol detection techniques using handcrafted features. Further, a feature Bag of Decor feature is formed and the classifier is trained using that feature for each image. Paragraph based description is generated by using a manually specified template where words for rooms and decor symbols are proposed by above-mentioned methods. Figure 10 depicts comparison between descriptions generated using template based, and Denscap model (described next), along with the ground truth (collected from web).

2) Densecap-concat: In this method, the paragraphs were synthesized by concatenating top 5 captions generated by



Densecap Concatenation

Bedroom is with bed. Bedroom is with bathroom which has tub, sink and toilet space. Master bedroom is near stairs and has a bathroom with tub shower and toilet space. Bathroom has a separate shower and sink space. There are stairs to the other floors.

The bathroom has space for tub sink toilet and separate shower. Bathroom has tub and sink. There are stairs to upper floors. Master bedroom has a bed and closet. Bedroom has a bed.

Template based

In this architectural floor plan there are 3 rooms. There is one bedroom. Bedroom has a bed in the east side of the room. There is one bathroom. Bathroom has 1 tub in south side of the room, 1 large sink in the north side of the room, 1 small sink in the west side of the room. There is one bedroom. Bedroom has a bed in the east side of the room

In this architectural floor plan there are 4 rooms. There is a bathroom. Bathroom has 1 tub in the north side of the room, 1 large sink in the north west of the room, 1 small sink in the south side of the room. There is a bathroom. Bathroom has 1 tub in the south side of the room, 1 large sink in the north side of the room, 1 small sink in the north west side of the room. There is one bedroom. Bedroom has 1 bed in the west side of the room. There is one bedroom. Bedroom has 1 bed in the east side of the room.

Collected Description

On the second floor, the balcony is open to the entrance foyer below, and has a nook for a desk. There is a master bedroom, a third bedroom, and a full bath to share. The bedrooms have a 9' flat ceiling that slope with the roof at the end walls. Closet spaces are tucked under the sloping roofs.

The upper level has a flexible floor plan which works well for growing families. The master bedroom offers a walk–in closet, spacious master bath, and a bonus storage room. The second bedroom is extra large, and can be easily divided by adding a center wall to create two smaller bedrooms. There is also an extra nook which can be used as an office or play area. The curved handrail in the hallway creates a small opening to view those on the ground floor.

Fig. 10. Paragraphs generated using Densecap-concat, Template based method and evaluation.

TABLE II
EVALUATION OF GENERATED PARAGRAPHS WITH DIFFERENT METRICES (METEOR (M), BLEU (B), ROUGE (R)).

Method	B-1	B-2	B-3	B-4	M	R_L
Densecap-concat	0.117	0.054	0.0195	0.003	0.189	0.122
Template	0.199	0.044	0.025	0.002	0.133	0.126

Densecap. This method yields the bounding boxes of the regions along with the respective captions. Difference between Densecap-concat and template based method is that the former predicts the whole sentence without having a predefined structure within. Table II shows the evaluation of generated paragraphs with the paragraph annotations available in BRIDGE using METEOR [17], ROUGE [18], and BLEU-{1,2,3,4} [19]. Results show that used schemes for paragraph generation are not performing well on BLEU-3 and BLEU-4 because generated descriptions do not contain flexibility in terms of positioning of words. Hence, they contains similarity in terms of words chosen to describe (BLEU-1, BLEU-2) but not in structuring of sentences. Hence a robust paragraph generation method for floor plan images is required.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented, for the first time, a novel large scale (13000+ images) floor plan dataset BRIDGE, which has images and metadata. This dataset could be used for various tasks on floor plan analysis using deep learning model. We are also planning to release the trained deep learning models along with the dataset for the community.

REFERENCES

[1] D. Sharma, N. Gupta, C. Chattopadhyay, and S. Mehta, "Daniel: A deep architecture for automatic analysis and retrieval of building floor plans," in *ICDAR*, 2017.

[2] L.-P. de las Heras, O. R. Terrades, S. Robles, and G. Sánchez, "Cvc-fp and sgt: a new database for structural floor plan analysis and its groundtruthing tool," *IJDAR*, vol. 18, no. 1, pp. 15–30, 2015.

[3] M. Delalandre, E. Valveny, T. Pridmore, and D. Karatzas, "Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems," *IJDAR*, vol. 13, no. 3, pp. 187–207, 2010.

[4] A. Barducci and S. Marinai, "Object recognition in floor plans by graphs of white connected components," in *ICPR*, 2012.

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[7] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[9] R. Girshick, "Fast r-cnn," in *ICCVn*, 2015, pp. 1440–1448.

[10] 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*, 2015, pp. 91–99.

[11] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *CVPR*, 2016.

[12] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *CVPR*, 2017.

[13] "Labelimg: Graphical image annotation tool," https://github.com/tzutalin/labelImg.

[14] S. Goyal, C. Chattopadhyay, and G. Bhatnagar, "Asysst: A framework for synopsis synthesis empowering visually impaired," in *MAHCI*, 2018, pp. 17–24.

[15] ——, "Plan2text: A framework for describing building floor plan images from first person perspective," in *CSPA*, 2018.

[16] Z. Ziran and S. Marinai, "Object detection in floor plan images," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, 2018.

[17] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.