Line2depth: Indoor Depth Estimation from Line Drawings

Pavlov Sergey\textsuperscript{a}, Kanamori Yoshihiro\textsuperscript{b} and Endo Yuki\textsuperscript{c}

Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan

Keywords: Depth Estimation, Line Drawing, Convolutional Neural Network, Conditional GAN.

Abstract: Depth estimation from scenery line drawings has a number of applications, such as in painting software and 3D modeling. However, it has not received much attention because of the inherent ambiguity of line drawings. This paper proposes the first CNN-based method for estimating depth from single line drawings of indoor scenes. First, to combat the ambiguity of line drawings, we enrich the input line drawings by hallucinating colors, rough depth, and normal maps using a conditional GAN. Next, we obtain the final depth maps from the hallucinated data and input line drawings using a CNN for depth estimation. Our qualitative and quantitative evaluations demonstrate that our method works significantly better than conventional photo-aimed methods trained only with line drawings. Additionally, we confirmed that our results with hand-drawn indoor scenes are promising for use in practical applications.

1 INTRODUCTION

Depth estimation from a single image has traditionally been one of the major challenges in computer vision. Properly estimated scene depth has many applications in various areas such as robotics, augmented reality, and 3D modeling. Although this task has been quite difficult with traditional methods, the recent rise of deep learning (DL) has allowed researchers to achieve substantial progress in depth estimation (Liu et al., 2018; Liu et al., 2019). While the main targets of research have mostly been photos and videos, to the best of our knowledge, no previous study has investigated for line drawings as inputs, yet it may also have practical applications in many areas, such as painting software, 3D modeling, and manga creation.

Depth estimation only from a single line drawing is quite challenging due to a number of reasons. The main issue is the underlying ambiguity of line drawings. In contrast to color images, line drawings usually lack textures and shading, leaving only contours with white insides to work with. Even for a human observer, it might be complicated to decide whether the shape is convex, concave, or flat. Another challenge is the lack of datasets containing both line drawings and depth maps, which makes the task of training an effective convolutional neural network (CNN) substantially difficult.

However, despite the difficulty, line drawings might have enough cues to estimate depth maps. Although line drawings may lack certain information, they do show the shape of objects. For example, object shape can hint at its spatial orientation and whether the object is planar or non-planar. This observation holds particularly for indoor scenes, which are filled with many planar objects such as walls, floors, and furniture. We thus tackle the following research question in this study: Can we estimate the depth map from a single line drawing of an indoor scene?

This paper proposes the first method for estimating depth from single line drawings of indoor scenes based on supervised learning using CNNs. First, to combat the ambiguity of line drawings, we enrich the input line drawings by hallucinating colors, rough depth and normal maps and using them as additional inputs. For this hallucination, we integrate the conditional generative adversarial network (GAN), pix2pix (Isola et al., 2017). Next, we obtain the final depth maps from the combined inputs (i.e., line drawings, hallucinated colors, depth maps, and normal maps) using PlaneNet (Liu et al., 2018), one of the recent CNNs for depth estimation, which explicitly handles planar regions in the scene. Our qualitative and quantitative evaluations demonstrate that our method works significantly better than conventional methods trained only with line drawings. Also, we confirmed that our results with hand-drawn indoor scenes are promising for use in practical applications.
scenes are promising for use in practical applications.

2 RELATED WORK

Here we explain two major groups of prior studies relevant to ours, i.e., single-photo depth estimation and 3D mesh reconstruction from line drawings.

**Single-photo Depth Estimation.** Previous depth-estimation studies have mostly focused on photos, i.e., RGB images. Because both photos and our targets, i.e., line drawings, represent a scene rather than a single object, the research literature in single-photo depth estimation is essentially valuable for our study.

Most of the modern methods use DL because, in contrast to traditional methods, they can automatically extract appropriate features and are thus more robust. Roy and Todorovic (Roy and Todorovic, 2016) introduced the neural regression forest for single-image depth estimation. Liu et al. proposed using additional modules to classify images into planar and non-planar regions and regressing plane equations (Liu et al., 2018; Liu et al., 2019). Ramamonjisoa and Lepetit (Ramamonjisoa and Lepetit, 2019) used a classic network architecture (Ronneberger et al., 2015) and improved the depth estimation quality by applying a novel edge-preserving loss function. However, when naively applied to depth estimation for line drawings, these methods suffer from the severe lack of visual information in line drawings, as explained in Section 1.

**3D Object Reconstruction from Line Drawings.** There exist several methods for reconstructing 3D meshes from single line drawings. Due to the inherent ambiguity of 3D shape in line drawings, some methods require different types of user annotations to specify 3D shapes, e.g., (Li et al., 2017). Our method learns to work with grayscale line drawings without any additional user input.

Recent methods adopt CNNs. Lun et al. (Lun et al., 2017) proposed a method to reconstruct a 3D model from line drawings in two object views. However, the network requires an object class as an additional input, which constrains the number of possible object classes and drastically limits free-form modeling. To address free-form modeling, Li et al. (Li et al., 2018) proposed smoothing ground-truth (GT) 3D meshes, thus, making the CNN independent from shape features specific to exact 3D models. However, these approaches require contours to be explicitly specified. Zheng et al. (Zheng et al., 2020) proposed a shading GAN which implicitly infers 3D information, but such information cannot be used directly and requires further processing. Our approach does not require to specify object contours or classes and infers final depth maps of whole scenes.

3 OUR METHOD

Our preliminary experiment revealed that our baseline method (Liu et al., 2018) failed to estimate depths solely from line drawings. This might be caused by the lack of information, which leads to our key idea: data enrichment. To enrich the input line drawings, our method integrates three streams of networks for coloring, initial depth, and normal estimation. Next, our method obtains the final depth map by refining the intermediate data. Figure 1 shows our depth estimation pipeline. Our depth estimation pipeline requires various data for training. Line drawings are required as an input to all the modules. Data enrichment modules require original RGB images, depth and normal maps as the ground truth. The refinement module requires ground truth depth maps, planar segmentations, and planar equations.

3.1 Pix2pix Modules for Data Enrichment

To tackle the detail shortage problem in line drawings, we integrate three branches of conditional GAN for colorization, initial depth, and normal estimation. Namely, we adopt the pix2pix architecture (Isola et al., 2017) for all of them. We train the first branch, edge2pix, to hallucinate original RGB images. The second and the third branches, edge2depth and edge2norm, are trained to estimate rough depth and normal maps, respectively.

After processing the input line drawing with pix2pix modules, we concatenate initial line drawings, hallucinated RGB images, initially estimated depth and normal maps. Next, we feed the concatenated result to the PlaneNet module.

3.2 PlaneNet Module

To obtain final depth maps from the initial depth maps and intermediate data, we use the PlaneNet (Liu et al., 2018) module. This module is based on a dilated version of the ResNet network (He et al., 2016; Yu et al., 2017) and has three branches following it. The first branch regresses plane parameters represented as three-dimensional vectors $\mathbf{n}$, where $d$ are offsets and $\mathbf{n}$ are unit normal vectors that define plane equations. The second branch segments an image into planar regions and a single non-planar mask. The third branch

---

This is a summary of the content from a page in a document. The page contains text related to related work in depth estimation, particularly focusing on single-photo depth estimation and 3D object reconstruction from line drawings. It discusses the challenges and approaches used in these areas, highlighting the need for data enrichment to improve depth estimation from line drawings. The page concludes by introducing a method that integrates three streams of networks for data enrichment and a PlaneNet module for refining depth maps.
estimates the non-planar depth map. Finally, the outputs of all three branches are merged into a single depth map output.

Training each module of the pipeline with the whole dataset will result in pix2pix modules overfit to the dataset, making PlaneNet insensitive to the intermediate results, which will severely affect the testing results. To avoid this issue, we divide the training dataset into two subsets. The first subset is used to train all of the pix2pix modules because they are independent of each other. The second one is used to train the PlaneNet module.

### 4 DATASET

For the dataset, we use the PlaneNet (Liu et al., 2018) version of the ScanNet (Dai et al., 2017) dataset. We discard extremely bright and blurry samples by examining the image edge strength. Next, we extract line drawings using one of the recent CNNs for edge extraction (He et al., 2019), preceded by the contrast limited adaptive histogram equalization (CLAHE) (Zuiderveld and Heckbert, 1994). Whilst we found such line drawings not plausible, they worked surprisingly well for the training of our pipeline.

### 5 EXPERIMENTAL RESULTS

We trained each of the pix2pix modules up to 200 epochs and the PlaneNet module up to 50 epochs. For comparison, we chose the PlaneNet (Liu et al., 2018) network as the baseline method and trained it with the same dataset up to 50 epochs. Our dataset has approximately 45,000 training and 1,000 testing samples. We use 15,000 samples to train the pix2pix modules and 30,000 samples to train the PlaneNet module. In total, the training took 34 hours using NVIDIA GeForce RTX 2080 Ti GPU.

Figure 2 shows some of the depth reconstruction results produced by our method and some of the recent approaches. As can be seen, PlaneNet (Liu et al., 2018) produces globally plausible but locally inconsistent outputs. In contrast, SARPN (Chen et al., 2019) produces locally consistent results, but globally they are significantly different from the ground truth. Our method is mostly consistent both locally and globally. It indicates that our method has successfully learned how to deal with the ambiguity in indoor-scene line drawings and to estimate their depth maps. This hypothesis is also supported by the 3D representations shown in Figure 3.

Table 1 provides a quantitative evaluation of the recent methods (Liu et al., 2018; Chen et al., 2019) and our pipeline on various metrics used in a prior
work (Eigen et al., 2014). The four metrics on the left
represent various error statistics such as rooted-mean-
square-error (RMSE) and relative difference (Rel).
The three metrics on the right show the percentage of
pixels, for which the relative difference between the
GT and the predicted depths is within a certain thresh-
old. As can be observed, our method works better not
only visually but also quantitatively. Also, as an abla-
tion study, we trained our pipeline alternately without
the edge2pix, edge2norm, and both edge2depth and
edge2norm modules. While some of the omitted ver-
sions are statistically better than the baseline methods,
we found them to perform qualitatively worse than the
full version. It is also clear that the edge2depth mod-
ule plays the key role in the pipeline.

Figure 2 additionally shows some of the plane
segmentation results. Note that the random col-
ors in segmentations are used just for distinguishing
planes and do not correspond among different meth-
ods. For some images plane segmentations are esti-

ted wrongly, merging perpendicular planes (e.g.,
“Ours” for Scene 2, where the green segment in-
cludes three different planes). However, for others,
the segmentation seems to work properly, and even
in the cases where a single plane is separated into
various planar instances (e.g., “Ours” for Scene 1,
where the red and green segments are mixed), these
instances appear to have nearly identical plane param-
eters, which means that even improper plane segmen-
tation usually does not strongly affect the final accu-

cacy of depth maps.

5.1 Evaluation with Hand-drawn Line
Drawings

Toward practical applications, we also evaluated our
method using hand-drawn line drawings of three
scenes. As can be seen from Figure 4, the resulting
depth maps are of slightly worse quality than the re-
sults using the evaluation dataset (Figure 2). How-
ever, the results are still visually plausible and account
for the overall object layout in the scenes.

6 CONCLUSION AND FUTURE WORK

This paper has proposed the first pipeline to estimate
depth of indoor-scene line drawings. To combat the
problem of ambiguity, our method integrates three
streams of conditional GAN. Next, to obtain the final
depth, our pipeline integrates the PlaneNet module, a
recent depth estimation method. Our method handles
indoor scenes including the hand-drawn line drawings
effectively.

Future work includes training and testing with a
high-resolution dataset. It might also include intro-
ducing a neural module to classify lines to texture-
based and geometry-based ones, thus solving this
complicated task for the main pipeline. Another
promising yet challenging direction might be to inte-
egrate vanishing points detection into the loss function or as an additional input to the network. Line drawings in our dataset might be further improved by employing apparent ridges for line extraction (Judd et al., 2007).

REFERENCES


Figure 4: Our results from hand-drawn line drawings. From left to right: line drawings, estimated segmentations, estimated depth, and ground truth depths of the corresponding images.