Supplemental Material: Visualization of Convolutional Neural Networks for Monocular Depth Estimation

Junjie Hu[1,2] Yan Zhang[2] Takayuki Okatani[1,2]
[1] Graduate School of Information Sciences, Tohoku University, Japan
[2] Center for Advanced Intelligence Project, RIKEN, Japan
{junjie.hu, zhang, okatani}@vision.is.tohoku.ac.jp

In this supplementary document, we provide more experimental results and analyses of the proposed method.

1. More Analyses

1.1. Pixels Invalidating Mask Application

The proposed method predicts a mask and multiplies it with an input image, masking out irrelevant pixels. However, the application of the mask will be ineffective for pixels with nearly zero values $(0, 0, 0)$, as they will not change their values before and after the mask application. Note that the input images here are normalized by z-score normalization. Thus, pixels with values $(0, 0, 0)$ have colors of the mean image (i.e., gray color) in the original, pre-normalization images, since the normalization is done by subtraction of the mean image of ImageNet.

To examine the effects of such pixels, we count their population in all the images of NYU-v2. To be specific, we count the pixels of the post-normalization images that satisfy $\max(|r|, |g|, |b|) < \tau$ for small τ. The results show that they occupy only 0.19% and 1.67% of all the pixels of all images for $\tau = 0.1$ and 0.2, respectively. Thus, the population of pixels that could invalidate the mask application is very small, and we may think that their impacts on the mask prediction and analyses based on them will be limited.

1.2. Visualization by Multiplication with Input Images

In the main paper, we use only predicted masks for visualization, which indicate which pixels are (ire)relevant for depth estimation. We show here another type of visualization, input images multiplied with the predicted masks for them, in Fig. 1. To be specific, for each input image,we multiply its normalized version with the predicted mask, and then 'unnormailized' the multiplied image. For the latter step, we scale the pixel values and then add the mean image, so that their pixel values lie in the range $[0, 255]$.

In Fig. 1, from left to right, the input I and I multiplied with a) the mask M, b) its binarized version M', and c) edge map, respectively; (a) and (c) have the same sparseness level. An interesting observation is that it is a lot easier for our visual system to infer depth from (a) or (b) than from (c). As seen from these, there are stark differences between M (or M') and the edge map, validating the discussion given in the main paper.

Table 1. The RMSE error for the mask prediction network (G) with different depth layers.

Encoder	Params	RMSE(GT)
DRN-D-22	25.3M	0.740
DRN-D-38	35.4M	0.656
DRN-D-54	44.7M	0.647
DRN-D-105	63.7M	0.639

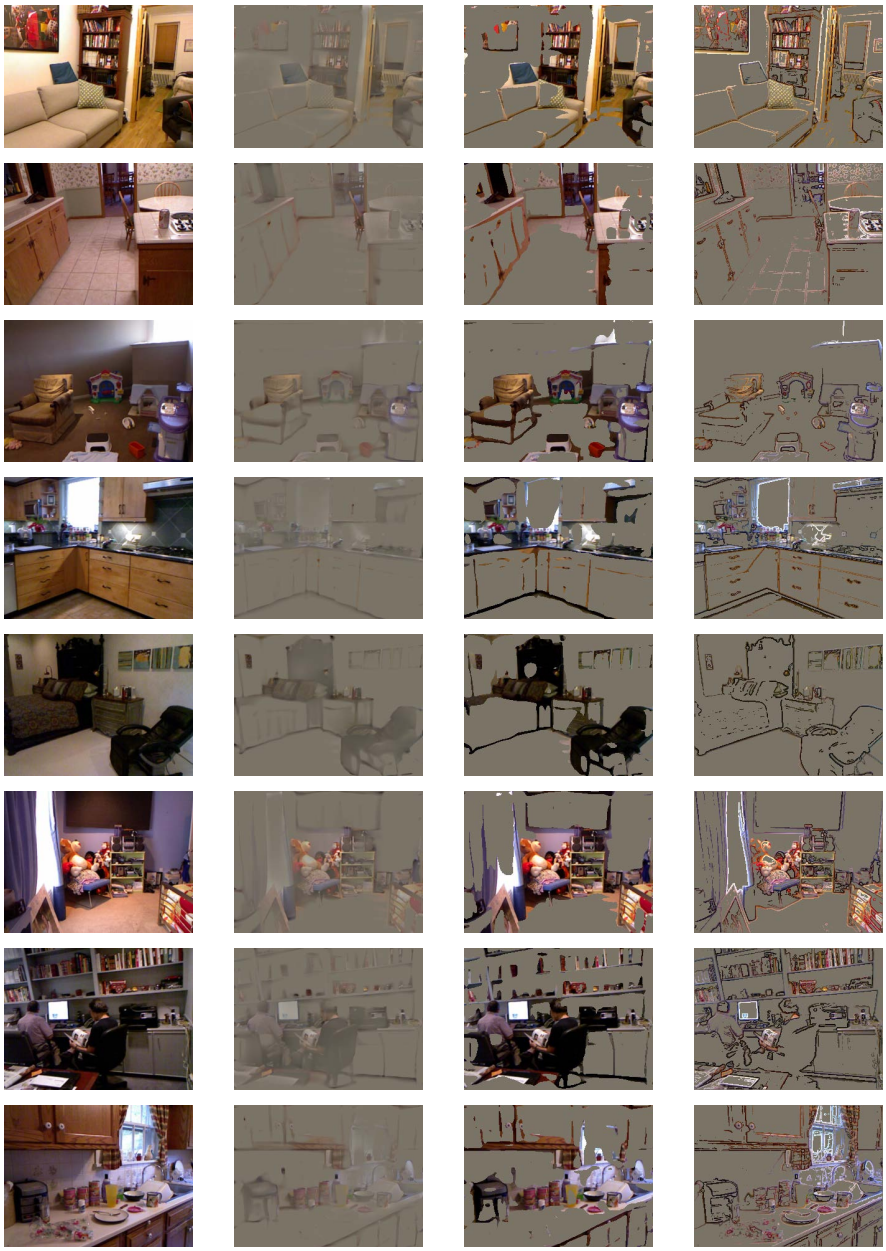
1.3. Influence of Capacity of the Mask Prediction Network

In the main paper, we consider only a single artchitecture for the mask prediction network G. We conducted an additional experiment to examine how large influence architectural differences of G will have. To be specific, we increase the layers of the encoder part of G from 22 (the original) to 38, 54, and 105. As seen in Table 1, the results (RMSE(M) w/ $\lambda = 5$ in the same setting are 0.740, 0.656, 0.647, and 0.639, respectively, indicating that the more layers G has, the better mask M will be obtained.

2. More Results on KITTI

2.1. More Scenes

Figures 2 - 3 show visualization results for different networks and input images from the test split of the KITTI dataset in the same setting of the paper. The same observations as those in the main paper apply to these results. Firstly, as with the above indoor images, there are clear differences between the edge maps and the predicted masks. For example, almost all shadow boundaries do not appear in the masks despite their strong presence in the edge maps,



(a) I (b) $I \otimes M$ (c) $I \otimes M'$ (d) $I \otimes Edge$

Figure 1. Another visualization using the predicted mask for the NYU-v2 dataset. An input image is multiplied with the predicted mask for it. The multiplication is performed with normalized input images, which gives the actual input to the depth estimator N; they are then 'unnormalized' for the purpose of visualization. The images shown in (b)-(d) show the 'unnormalized' version.

e.g., Figs. 2(2), 2(3), 2(6), 3(2)-(6), *etc*. This is also the case with the white lines on the road surface in Fig. 2(1), 2(3), *etc*. The borders between the roadway and sidewalks are often highlighted in the predicted masks but are not clearly seen in edge maps, such as those in Figs. 2(3), 3(1) and 3(6), *etc*. Secondly, the vanishing points and distant regions are highlighted for all cases. As is seen also in the indoor images, small to medium size objects are highlighted in the mask not only with their boundaries but with their internal regions.

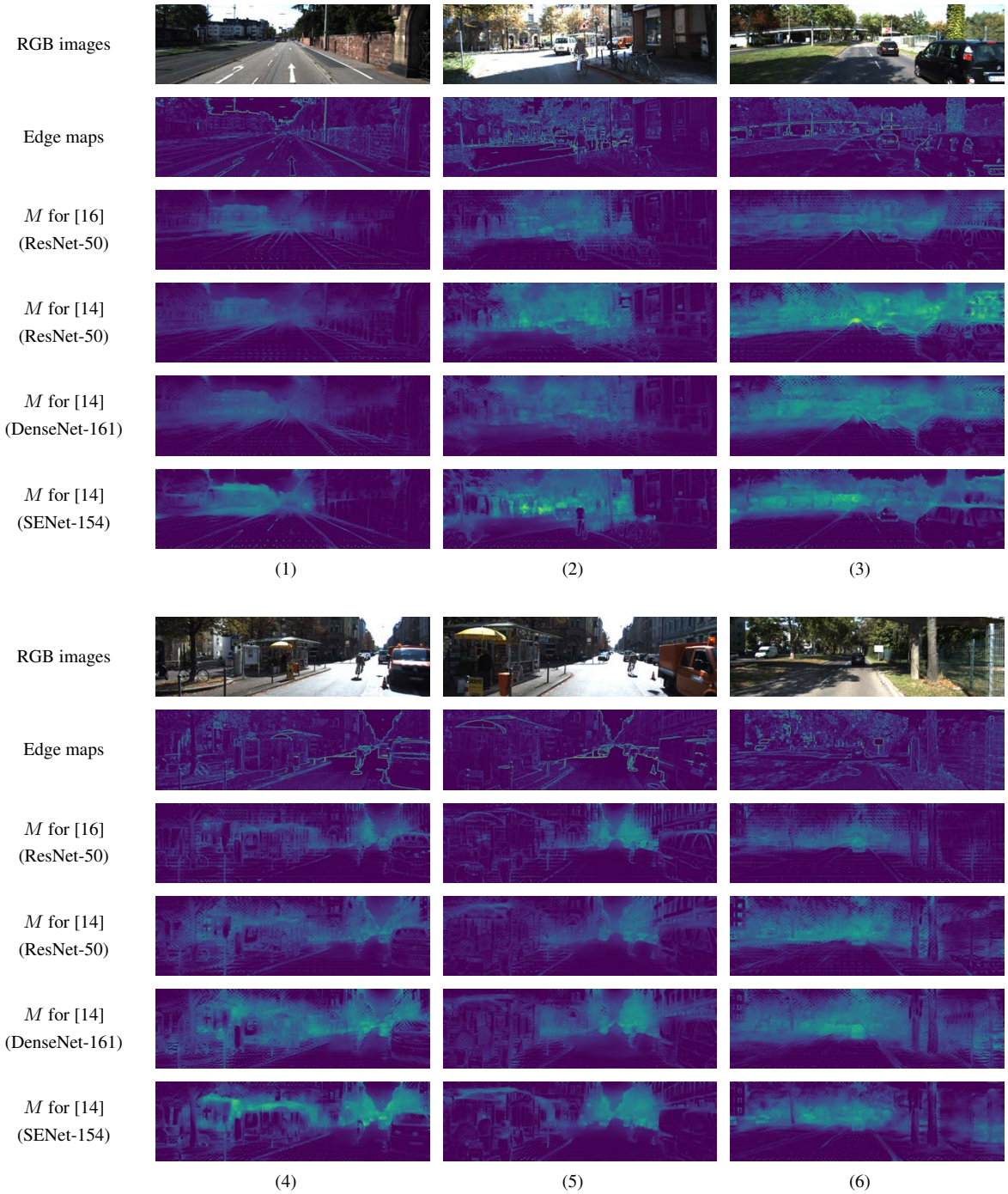


Figure 2. Predicted masks for different networks trained on the KITTI dataset for different input images from the test split.

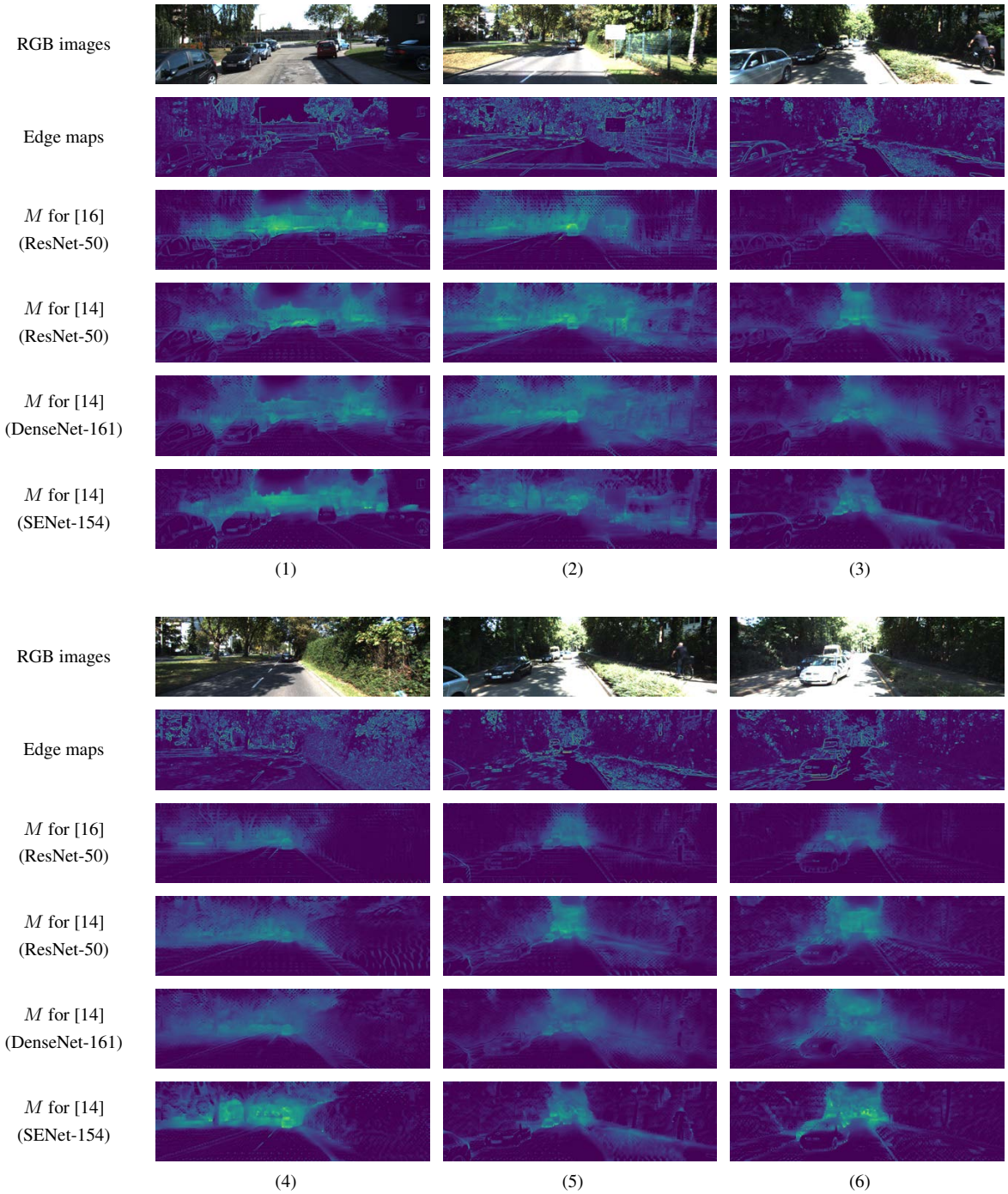


Figure 3. Predicted masks for different networks trained on the KITTI dataset for different input images from the test split.

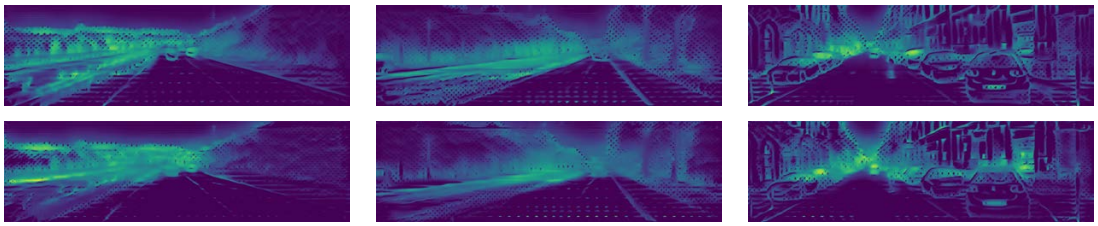


Figure 4. Predicted masks for the three images of Fig.7 in the main paper with two different losses, *i.e*. mean REL (1st row) and mean $\log 10$ error (2st row).

2.2. Predicted Masks with Other Loss Functions

In the predicted masks for images of the KITTI dataset, distant scene pixels are almost always highlighted, which we think indicates the importance of vanishing points for depth inference with outdoor scenes. However, one may wonder if it is because of the loss employed for training in the experiments described in our main paper (i.e., the absolute difference in depths), which tends to give more weights on distant scene pixels.

To clarify this, we conduct experiments with different loss functions. To be specific, we tested two loss functions, *i.e*., mean relative error (REL) and mean $\log 10$ error, both of which are scale-invariant and thus do not have the above tendency (i.e., more weights on more distance pixels). The results with the ResNet-50-based model are shown in Fig. 4 for the images of Fig.7 of the main paper; REL and $\log 10$ are in the 1st and 2nd rows of Fig. 4, respectively. Although there are slight differences, the same observation as reported in the main paper holds true, *e.g*., more saliency on distant points. The same applies to other images than these three.