VisualEchoes: Spatial Image Representation Learning through Echolocation (Supplementary Materials)

Ruohan Gao[1,3], Changan Chen[1,3], Ziad Al-Halah[1],
Carl Schissler[2], Kristen Grauman[1,3]

[1]The University of Texas at Austin, [2]Facebook Reality Lab, [3]Facebook AI Research
{rhgao,changan,ziad,grauman}@cs.utexas.edu, carl.schissler@fb.com

The supplementary materials for [3] consist of:

A. Qualitative results on monocular depth prediction.
B. Qualitative results on surface normal estimation.
C. More qualitative results on visual navigation.
D. More qualitative results from the case study of spatial cues in echoes.
E. t-SNE embedding of echoes
F. Ablation study.
G. Low-shot experiments varying the amount of training data.
H. Comparing to state-of-the-art on monocular depth prediction.
I. Network/Dataset/Implementation details.

A Qualitative results on monocular depth prediction

Fig. 1 shows example results on monocular depth prediction as described in Sec. 3.4 in the main paper. Using our pre-trained VisualEchoes network as initialization leads to more accurate depth prediction results compared to no pre-training, demonstrating the usefulness of the learned spatial features. Although the features are learned in the Replica environments, they transfer well to do depth predictions in these photos from NYU-V2. For example, while the model trained from scratch cannot well predict the depth values for the side walls of the first and the last examples, our model with enhanced spatial knowledge can make predictions that better match the ground-truth.

B Qualitative results on surface normal estimation

The surface normal is a unit norm 3-dimensional vector (x, y, z) at each pixel location. Following [6], the surface normals are clustered into 40 clusters through k-means and surface normal estimation is formulated as a 40-way classification task. Fig. 2 shows example results on surface normal estimation of color-coded surface normals of the 40 classes. Using our pre-trained VisualEchoes network as initialization leads to more accurate surface normal estimation compared to

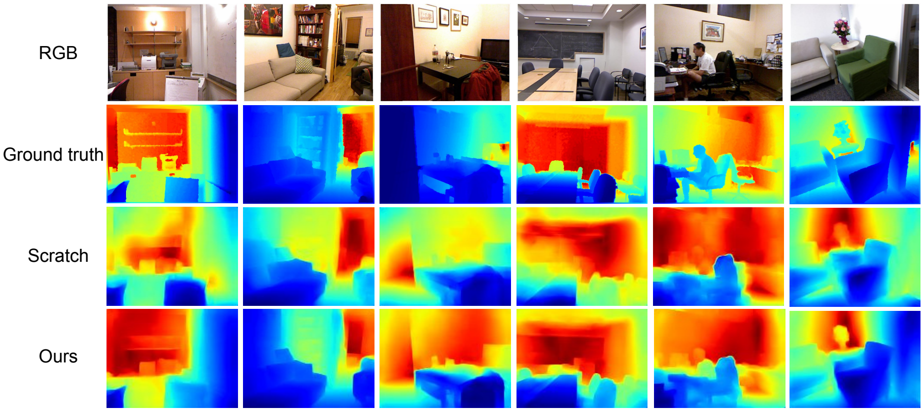

Fig. 1: Qualitative results of monocular depth prediction on the NYU-V2 dataset.

no pre-training, again demonstrating that the learned spatial features are useful for downstream vision tasks that require spatial reasoning. For example, while the model trained from scratch makes a noisy prediction, our model makes more consistent predictions, especially on large surfaces and the predicted surface normal better matches the ground-truth.

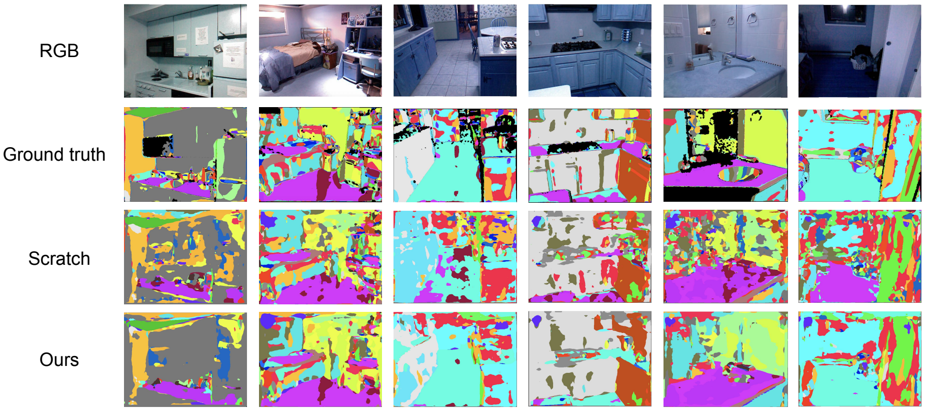

Fig. 2: Qualitative results of surface normal estimation on the NYU-V2 dataset.

C More qualitative results on visual navigation

Fig. 3 shows more qualitative examples of navigation trajectories on top-down maps, supplementing Fig. 6 in the main paper. Our visual-echo consistency pre-training task allows the agent to better interpret the room's spatial layout to find the goal more quickly than the baselines.

Fig. 4 shows an example of the training curves for the model using our pre-trained VisualEchoes network as initialization and a model trained from scratch. VisualEchoes equips the embodied agents with a better sense of room geometry and allows them to learn faster.

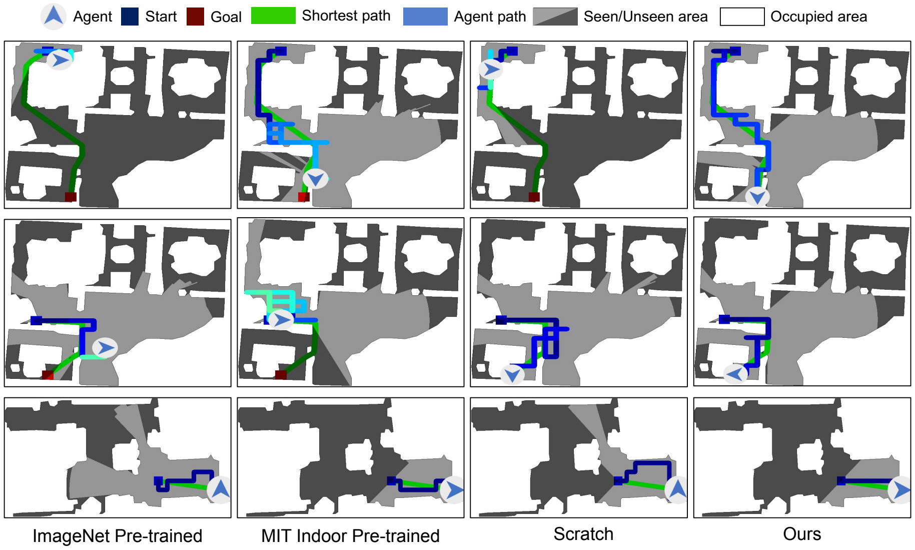

Fig. 3: More qualitative examples of visual navigation trajectories on top-down maps. Blue square and arrow denote agent's starting and ending positions, respectively. The green path indicates the shortest geodesic path to the goal, and the agent's path is in dark blue. Agent path color fades from dark blue to light blue as time goes by. Note, the agent sees a sequence of egocentric views, not the map. We can see that the agents for the baseline methods often get stuck in corners or roam around to look for the target, while the agent of our model takes fewer turns and finds the target more smoothly due to its better sense of the spatial structure of the room.

D More qualitative results for the case study

In addition to the qualitative examples shown in Fig. 3 of the main paper, in Fig. 5 we show more results of our case study on monocular depth estimation in unseen environments using echoes. It is clear that echo responses indeed contain cues of the spatial layout; the depth map captures the rough room layout, especially its large surfaces. When combined with RGB, the predictions are more accurate. The last row shows a typical failure case, where the echoes alone cannot capture the depth as well maybe due to sound absorbing and far away surfaces with weaker received echo signals.

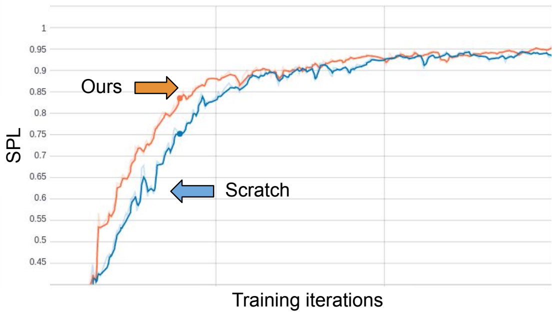

Fig. 4: Training curves for the model using our pre-trained VisualEchoes network as initialization and that trained from scratch. Training is done on the Replica training environments.

E T-SNE Embedding of Echoes

To visualize that echo responses indeed contain spatial cues and how our VisualEchoes representation learning framework leverages them, Fig. 6 displays a t-SNE [12] embedding of the echo responses at different spatial locations for the training scenes. We use the audio features extracted from the audio branch (EchoNet) of our representation learning network (Fig. 4 in the main paper). Echos from locations of similar spatial layout (e.g., open areas, blocking walls, corners, etc.) tend to cluster together, confirming the spatial cues contained in echoes and also demonstrating that our representation learning framework successfully make use of these spatial signals in order to make the right prediction of the agent's orientation. Note that the shape of the embedding space is in the form of a coherent curve, because the agent moves coherently in the environments to get echo responses.

F Ablation study

In this section, we perform some ablation studies to demonstrate that the design of our spatial representation learning framework introduced in Sec. 3.3 in the main paper is essential and effective. We compare with the following two variants:

- SimpleVisualEchoes: This baseline is the same as our method except that we simplify it to only two classes: The echo is either received from the same orientation as the agent's current view or the opposite direction.
- BinaryMatching: For this baseline, we train a network to discriminate whether the echoes come from the same environment as the RGB image. The positive pairs are matched RGB-Echo pairs where the echoes come from one of the four directions, while negative pairs are unmatched RGB-Echo pairs where the RGB views and echoes are from unrelated environments.

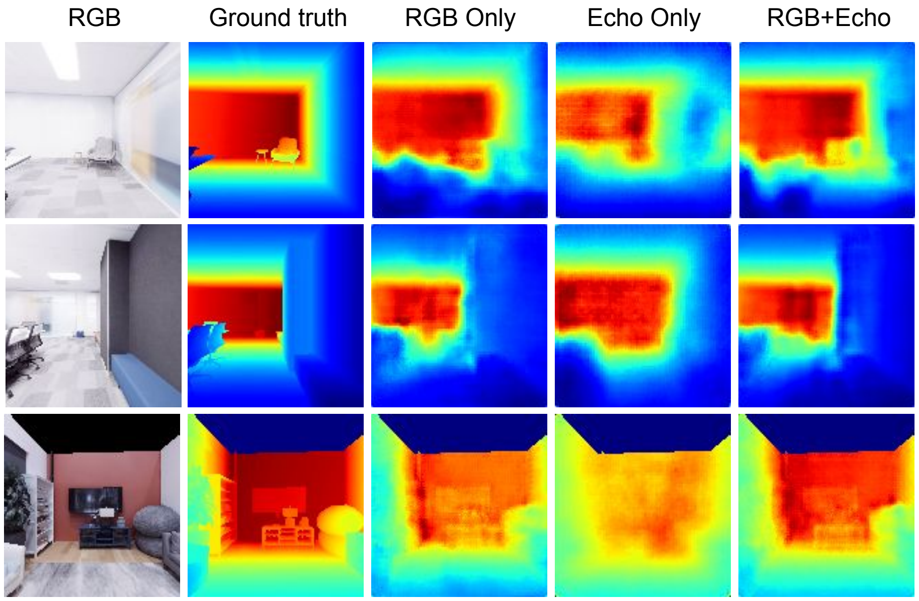

Fig. 5: Additional qualitative results of our case study on monocular depth estimation in unseen environments using echoes. The last row shows a typical failure case, where the echoes alone cannot capture the depth as well maybe due to sound absorbing and far away surfaces with weaker received echo signals.

Table 1 shows the results. Our method performs much better than the BinaryMatching baseline where the views and echoes come from unrelated environments. These "easy" examples are not as useful for learning spatial features as our method. The audio-visual data at offset views used in the design of our framework will naturally be related to one another, thus forcing the network to reason about the spatial structures of different orientations. The SimpleVisualEchoes variant also performs worse than our original method, showing the gain of using more fine-grained spatial orientations.

We have also performed an additional ablation study that predicts depth from only the left/right ear spectrogram. The RMS results for LeftEcho2Depth, RightEcho2Depth and Echo2Depth on are 0.801, 0.815 and 0.713, respectively (see Table 1 in the main paper). This confirms the importance of binaural spatial perception and the rationale of our task design that equips the embodied agent with two ears. To demonstrate the robostness of our system to noise, we add Gaussian white noise to the echoes to simulate the noise from an average quality microphone. The RMS results are 0.733 and 0.359 for Echo2Depth and RGBEcho2Depth, respectively—only slightly worse than the counterparts without noise (Table 1 in the main paper). This suggests the network has some robustness to non-idealities in the real world.

	RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Scratch	0.360	0.214	0.078	0.747	0.879	0.940
SimpleVisualEchoes	0.340	0.198	0.073	0.763	0.892	0.948
BinaryMatching	0.345	0.199	0.074	0.760	0.889	0.944
VisualEchoes (Ours)	**0.332**	**0.195**	**0.070**	**0.773**	**0.899**	**0.951**

(a) Replica

	RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Scratch	0.818	0.252	0.103	0.586	0.853	0.950
SimpleVisualEchoes	0.803	0.248	0.101	0.595	0.859	0.954
BinaryMatching	0.813	0.250	0.103	0.589	0.854	0.953
VisualEchoes (Ours)	**0.797**	**0.246**	**0.100**	**0.600**	**0.863**	**0.956**

(b) NYU-V2

	RMS ↓	REL ↓	log 10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Scratch	2.352	0.481	0.214	0.321	0.581	0.742
SimpleVisualEchoes	0.278	0.448	0.203	0.333	0.604	0.760
BinaryMatching	2.351	0.479	0.211	0.323	0.583	0.745
VisualEchoes (Ours)	**2.223**	**0.430**	**0.198**	**0.340**	**0.610**	**0.769**

(c) DIODE

Table 1: Ablation study results. ↓ lower better, ↑ higher better.

G Low-shot experiments

Fig. 7 shows the results of low-shot experiments varying the amount of training data on the three datasets (as referenced in Sec. 4 of the main paper). We can see that models initialized with our pre-trained VisualEchoes network have consistent gains across varied percentages of training data, further demonstrating the usefulness of the learned spatial features.

H Comparing to State-of-the-art on Monocular Depth Prediction

We show the power of feature learning from echoes as a pre-training mechanism for spatial tasks, which is orthogonal to advances on architectures for each individual task. Therefore, we compare our method with the supervised pre-trained baselines in an apples-to-apples manner, and our method even outperforms them on two tasks (see Table 3 in the main paper).

As a reference, we compare to the state-of-the-art methods for monocular depth prediction on NYU-V2. The RMS results for Eigen et al. 2014 [2], Liu et al. 2015 [11], Cao et al. 2017 [1], Li et al. 2017 [9], Lee et al. 2018 [8], Qi et al. 2018 [14] and Ours are 0.907 / 0.824 / 0.819 / 0.635 / 0.569 / 0.683, respectively (lower is better). Lee et al. and Qi et al. use more advanced network architectures, loss functions and ImageNet supervised pre-trained networks, leading to their better performance.

I Network/Dataset/Implementation Details

Network Details: In our case study, the RGB+Echo2Depth network consists of a visual branch and an audio branch. For the visual branch, we adopt a UNet [15] style network due to its effectiveness in dense prediction tasks [10] and multi-modal feature fusion [13,4,5]. The network takes the agent's view of dimension 128 × 128 × 3 as input, and extracts a feature map of dimension 4 × 4 × 512 through a 5-layer encoder. The audio branch Echo-Net consists of three convolution layers of kernel size 8 × 8, 4 × 4, and 3 × 3, respectively. Batchnorm and ReLU are applied after each convolution layer. A final linear layer is used to reduce the feature dimension to 512. We replicate the audio feature vector 4×4 times, tile them to match the visual feature dimension, and then concatenate the echo and visual feature maps along the channel dimension. The final audio-visual feature map is of dimension 4 × 4 × 1024. Five layers of up-convolution layers followed by a `Sigmoid` layer are applied to the feature map to predict the depth map. The predicted depth maps are multiplied by a scalar, which is the maximum depth value for the dataset. The RGB2Depth variant directly upsamples from the visual feature map without considering audio. For Echo2Depth, we directly upsample the $512 \times 1 \times 1$ audio feature map through 7 upconvolution layers to predict the 128×128 depth map. We use the following loss function to train the network:

$$L_{\text{depth}} = \frac{1}{W \times H} \textstyle\sum_{i=1}^{W \times H} \ln(1 + \|d_i - g_i\|_1), \tag{1}$$

where d_i is the predicted depth and g_i is the ground-truth depth at pixel i. We take the logarithm of depth errors to encourage correct predictions for pixels of small depth values, following [7].

Our VisualEchoes network used for representation learning also consists of two branches. The audio branch uses the same Echo-Net as used in the case study and extracts an audio visual feature of dimension 512. The visual branch VisualEchoes-net uses different models depending on the corresponding downstream task, and we use a conv1x1 layer in the end to reduce the channel dimension and then flatten the feature map as the visual feature. The audio-visual fusion layer is a fully-connected layer followed by ReLU to reduce the concatenated audio-visual feature to dimension $D = 128$. A final fully-connected layer is used to make the prediction among the four classes.

Dataset Details: 1) **Replica** [17]: contains 18 3D scenes having 1,740 navigable locations × 4 orientations $= 6,960$ agent states in total. We use 15 scenes for pre-training, and the rest (apartment 2, frl apartment 5, and office 4) are held out for evaluation. For downstream tasks, the training data for our method and the baseline methods are the same 15 scenes, and the held-out three scenes are split in two halves for validation and testing. 2) **NYU-V2** [16]: consists of a variety of indoor scenes. For monocular depth prediction, we use the standard splits of 464 scenes, and use 249 scenes for training and 215 for testing following [7]. For surface normal estimation, we use the dataset split as formulated in [6]. 3) **DIODE** [18]: the first public dataset that includes RGB-D images of both indoor and outdoor scenes. We only use the indoor scenes (as the Replica training environments are limited to indoors). We resize all images to 128 × 128 and use the dataset for RGB2Depth as described in Sec.4.1 in the main paper. We use the official train/val split: 8,574 images are used for training, and the official validation split of 325 images are divided half by half for validation and testing in our experiments.

Implementation Details: 1) For our case study in Sec. 3.2 and RGB2Depth experiment in Sec. 4.1, we use images of resolution 128×128. We use batchsize of 128, starting learning rate of 0.0001, and Adam optimizer with weight decay of 0.0005. 2) For our representation learning framework, we crop 128×128 regions from images images of resolution 140×140 along with color/contrast/brightness jittering as data augmentation. We also jitter the volume of echoes by 0% - 5% as audio data augmentation. We use batchsize of 256, starting learning rate of 0.0001, and Adam optimizer with weight decay of 0.0005. 3) For monocular depth prediction in Sec. 4.2, see [7] for details. 4) For surface normal estimation in Sec. 4.2, see [6] for details. 5) For visual navigation, we also use images of resolution 128×128 and Adam optimizer with a learning rate of 2.5e-4. We discount rewards with a decay of 0.99 and we train the RL policy for 15M agent steps.

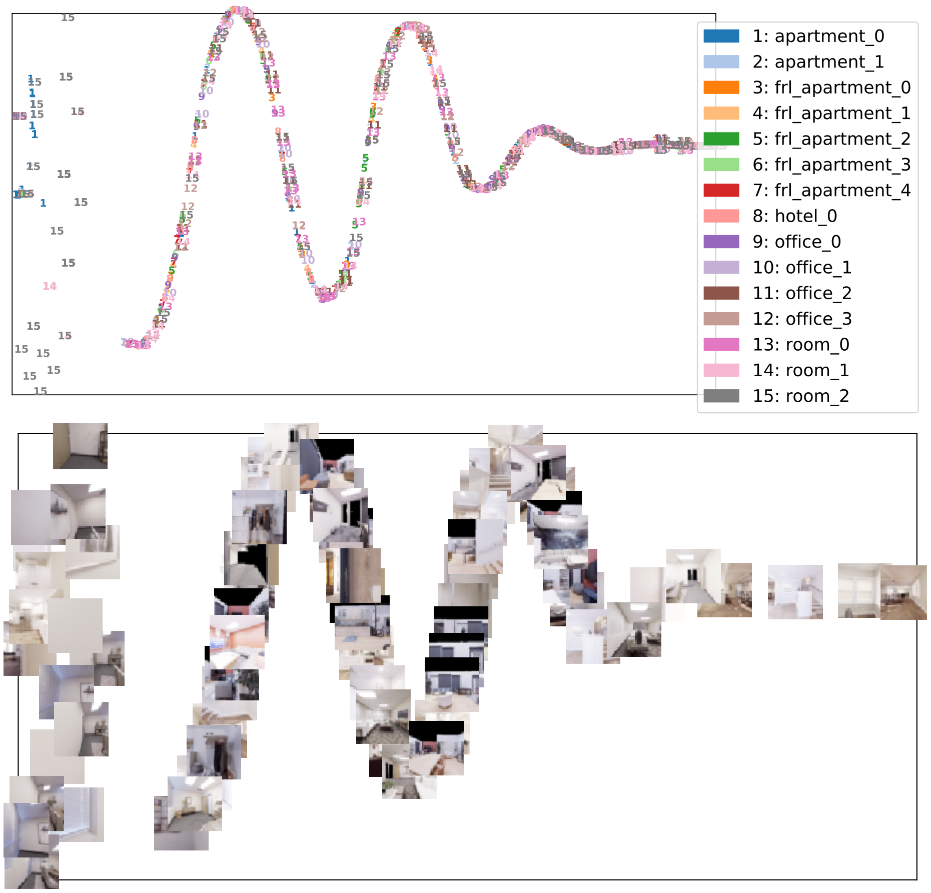

Fig. 6: Echoes embeddings from 15 scenes used for training, visualized with t-SNE in two ways: (top) scene categories are color-coded, and (bottom) the agent's views are shown at the t-SNE embedding the corresponding echo features. Best viewed in pdf with zoom.

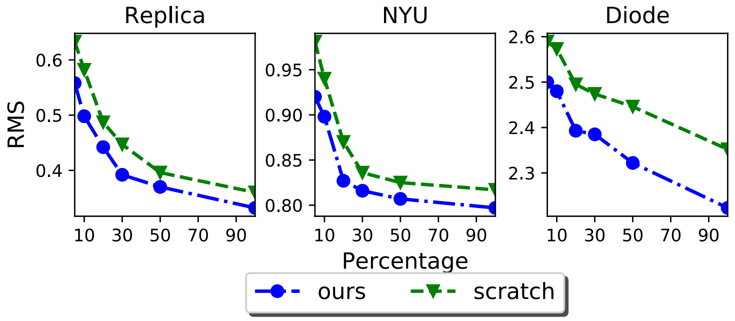

Fig. 7: Low-shot experiments varying the amount of training data.

References

1. Cao, Y., Wu, Z., Shen, C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. TCSVT (2017)
2. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NeurIPS (2014)
3. Gao, R., Chen, C., Al-Halah, Z., Schissler, C., Grauman, K.: Visualechoes: Spatial image representation learning through echolocation. In: ECCV (2020)
4. Gao, R., Grauman, K.: 2.5d visual sound. In: CVPR (2019)
5. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: ICCV (2019)
6. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: ICCV (2019)
7. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: WACV (2019)
8. Lee, J.H., Heo, M., Kim, K.R., Kim, C.S.: Single-image depth estimation based on fourier domain analysis. In: CVPR (2018)
9. Li, J., Klein, R., Yao, A.: A two-streamed network for estimating fine-scaled depth maps from single rgb images. In: ICCV (2017)
10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis (2017)
11. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: CVPR (2015)
12. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. JMLR (2008)
13. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)
14. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: CVPR (2018)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
16. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
17. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
18. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G.: DIODE: A Dense Indoor and Outdoor DEpth Dataset. arXiv preprint arXiv:1908.00463 (2019)