Explaining the Implicit Neural Canvas: Connecting Pixels to Neurons by Tracing their Contributions

Supplementary Material

The main purpose of this appendix is to expand our results from the main paper by showing analysis for frames from additional videos, to demonstrate that our analysis and findings hold across a variety of different videos.

6. Implementation Details

6.1. INR Architecture Settings

Neither the INR or NeRV have established out-of-the-box settings for arbitrary, in-the-wild videos. Thus, we have to carefully explore and select settings that allow for meaningful comparisons. We try to ensure the FFN and NeRV have similar compression ratios; however, this is not fully practical for one key reason. NeRV is able to leverage the temporal redundancies in addition to the spatial redundancies. We also try to ensure that both networks achieve similar reconstruction quality. For the videos we choose, the NeRVs have a mean PSNR of 33.50, and for the frames, the FFNs have a mean PSNR of 34.29.

For all the NeRVs, we ensure they have 978,557 parameters; for all FFNs, we ensure they have 32,971; for a 30 frame video (for Cityscapes-VPS all are 30 frames, for VIPSeg we only use videos between 30 and 45 frames), where the NeRV represents all 30 frames, and a given FFN represents a single frame, this gives roughly equivalent bits-per-pixel. For the FFN, we feed the 208-dimension Fourier position encoding to a network with 3 layers, with hidden sizes of 104 for the inner layers, ReLU activations, and output as 3-channel rgb predictions. For the NeRV, we use 4 layers, with the upsampling layers having strides 4, 2, and 2, respectively. We borrow the parameter reduction trick from HNeRV to balance parameters between layers, and set this to $r = 1.2$ [9], with a minimum width of 6, and the model size hyperparameter set to 1 million. The inner layers use GELU activations [20]. For other settings, we use the original NeRV defaults [7]. Both are trained for 1000 epochs with L2 loss.

6.2. Gabor Features

We already have mechanisms for addressing some types of low-level and high-level features. Since we have dense instance and background segmentation masks, we are able to identify how neuron contributions correspond with instances and background. We can also cluster pixels by color and space. However, we notice low-level patterns in the contribution maps, such as a tendency to focus on edges, that are not captured fully by instance masks, color, or space. So, we use an additional mechanism for clustering pixel locations – Gabor filters [16, 33].

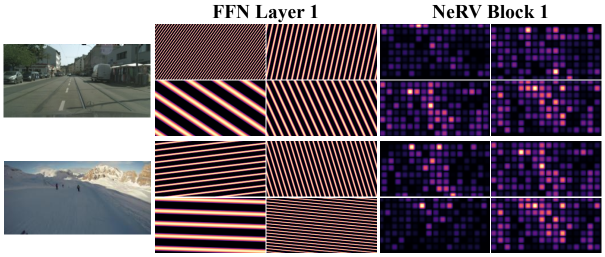

Figure 11. **The implicit neural canvas.** We show the contribution maps for sample first layer neurons of FFN and NeRV.

Gabor filters offer a robust method for texture analysis, detecting patterns across various orientations and scales. In our experiments, we utilize Gabor filters with four orientations and three scales, ensuring a comprehensive analysis of diverse textures within the contribution maps. The filters are applied to the maps, producing a distinct feature map for each filter. These feature maps are then stacked. For each pixel coordinate, a feature vector is constructed from values across all filters. These vectors are utilized to group contributions with similar traits into Gabor clusters (see Section 4.6). Please refer to the code for precise settings and implementation.

7. Further Results

7.1. Contribution Maps

We show the missing layers from Figure 3 in Figure 11. Note that the earliest layers have patterns that heavily correlate with the fourier features (FFN) and positional encoding (NeRV). Also, due to the nature of the PixelShuffle, for the first layer of NeRV, some neurons cannot possibly represent certain pixels. In that sense, the representation of each neuron is forced to be somewhat sparse.

Using Gabor filter features, we cluster neurons at each layer of MLP-based and CNN-based INRs and plot them using UMAP. For a few of these clusters, Figures 12 and 13 show the contribution maps of four neurons sampled from each cluster. We see that early FFN layers learn Fourier patterns, and the last layer of NeRV tends to resemble the image. Further, note how neurons belonging to the same cluster tend to learn similar contribution maps as compared to neurons belonging to different clusters.

In Figure 14, we provide a supplement to Figure 10 by

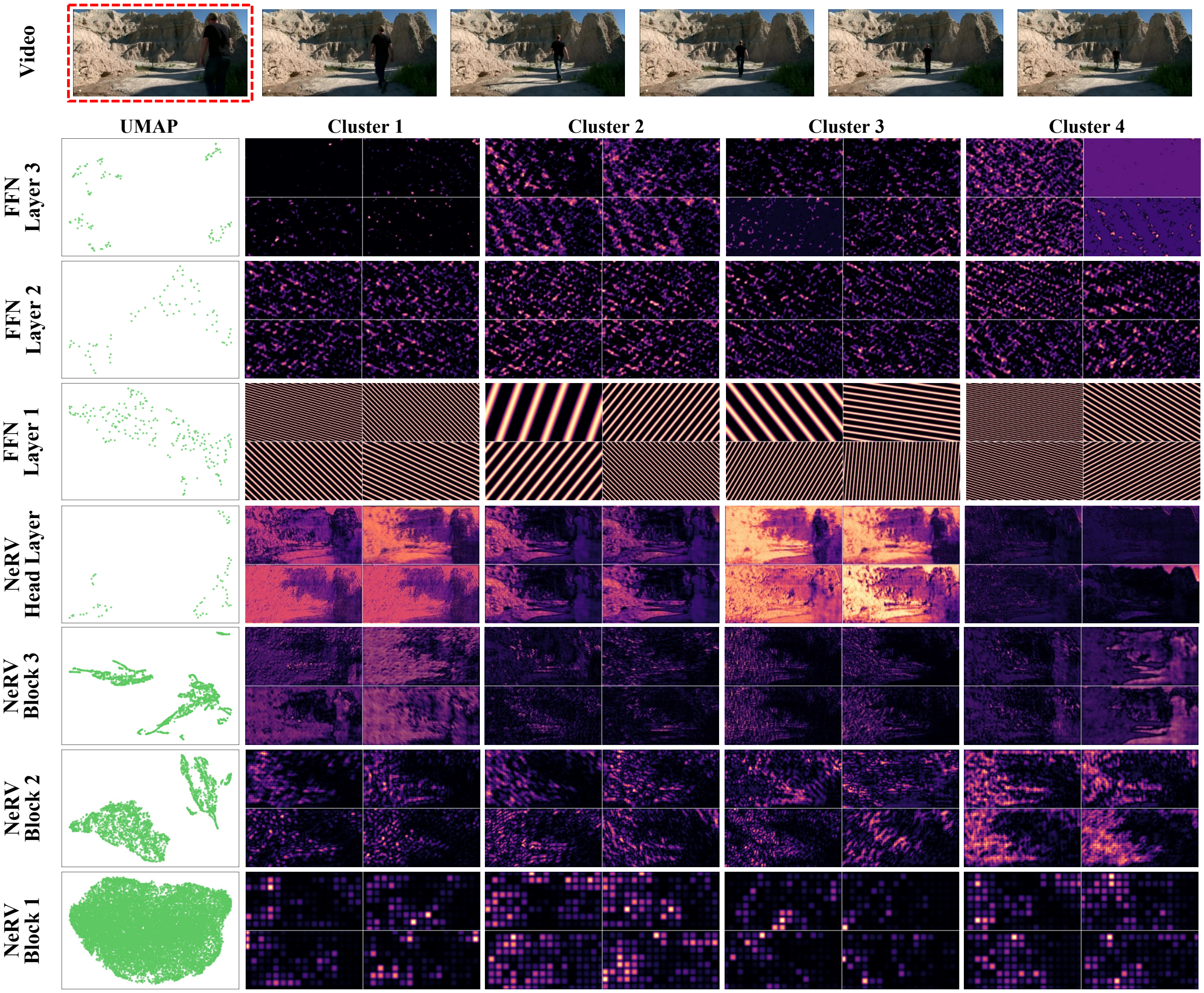

Figure 12. **Neuron Clusters.** We cluster neurons for each layer/block into 4 clusters, and then sample 4 neurons from each cluster. We show contribution maps of sample clustered neurons for the first (red-bordered) frame for the indicated video (top). See another video in Figure 13.

clustering neurons from models with five different seeds, for layers of both MLP-based and CNN-based INRs on frames from five videos.

7.2. Grouping Contributions

In Figure 16, we provide a supplement to Figure 4 by showing results for frames from 5 additional videos. As expected, the major trends hold. We see far higher neuron contribution difference variances when using the instances, RGB clusters, and Gabor clusters, compared to space clusters (gridcells). We see that in general Gabor and RGB clusters have equal explanatory power for neuron contributions compared to instances, reinforcing our hypothesis that these networks have low-level, rather than high-level, object semantics. Interestingly, the trends are consistent across all layers, except for the trends for space are weaker with the middle layers of NeRV than with the FFN, perhaps due to the spatial bias from NeRV's convolutional kernels.

7.3. Representation is Distributed

We extend the results from Section 4.4 for frames from some additional videos.

Figure 15 reveals the same trends as Figure 6, for frames from 5 additional videos. This is consistent whether the frame is street or open domain, and whether it is dim or more brightly lit.

The same trends from the video in Figure 7 are consistent for 5 additional video frames in Figure 17. Pixels for the

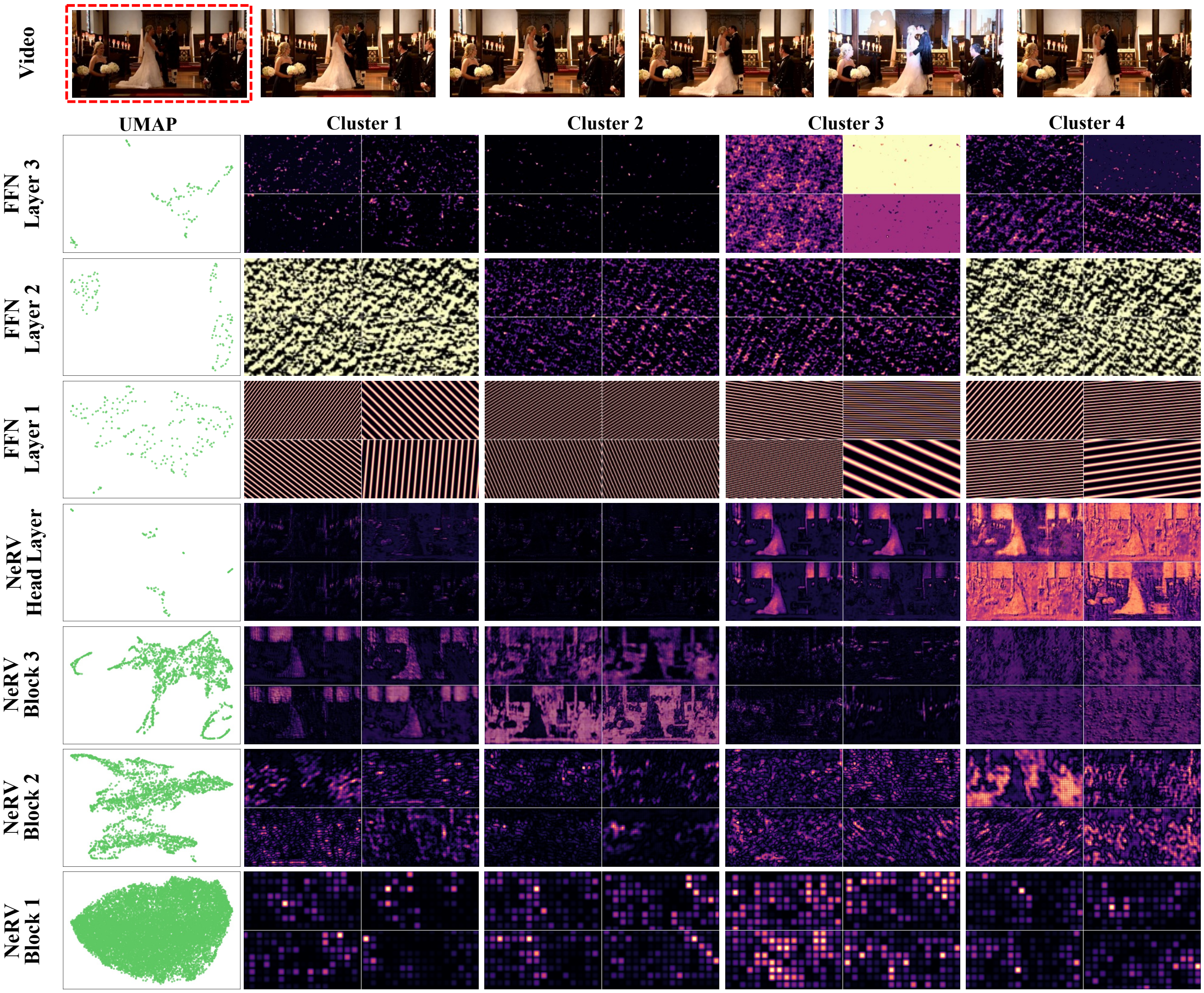

Figure 13. **Neuron Clusters.** Following the same procedure as in Figure 12, we cluster neurons for each layer/block into 4 clusters, and then sample 4 neurons from each cluster. We show contribution maps of sample clustered neurons for the first (red-bordered) frame for the indicated video (top).

FFN layers tend to be represented by similar amounts of neurons, whereas for the NeRV, these numbers vary widely. Some neurons are represented by many pixels, others by very few (relatively). The contrast is sharper in general for last layers (NeRV head layer, FFN layer 3).

Figure 18 shows how a large portion of the raw contributions are relatively smaller in magnitude for each layer. The thresholds for Figure 17 are selected using the contribution value at the 10th and 50th percentile of these curves.

7.4. Objects and Categories

We offer results for the first two NeRV blocks corresponding to Figure 8 in Figure 19, reserved for this appendix due to space constraints. Overall, the representations for objects are relatively constant over time for Block 2, with some notable exceptions, such as the representation of one of the persons for one of the four sampled neurons, which increases dramatically at the final frame of the video. We also note that Block 1 seems less structured than the other blocks. Specifically, whereas certain object types might dominate for the other blocks, in Block 1 the contributions are more evenly distributed across objects overall.

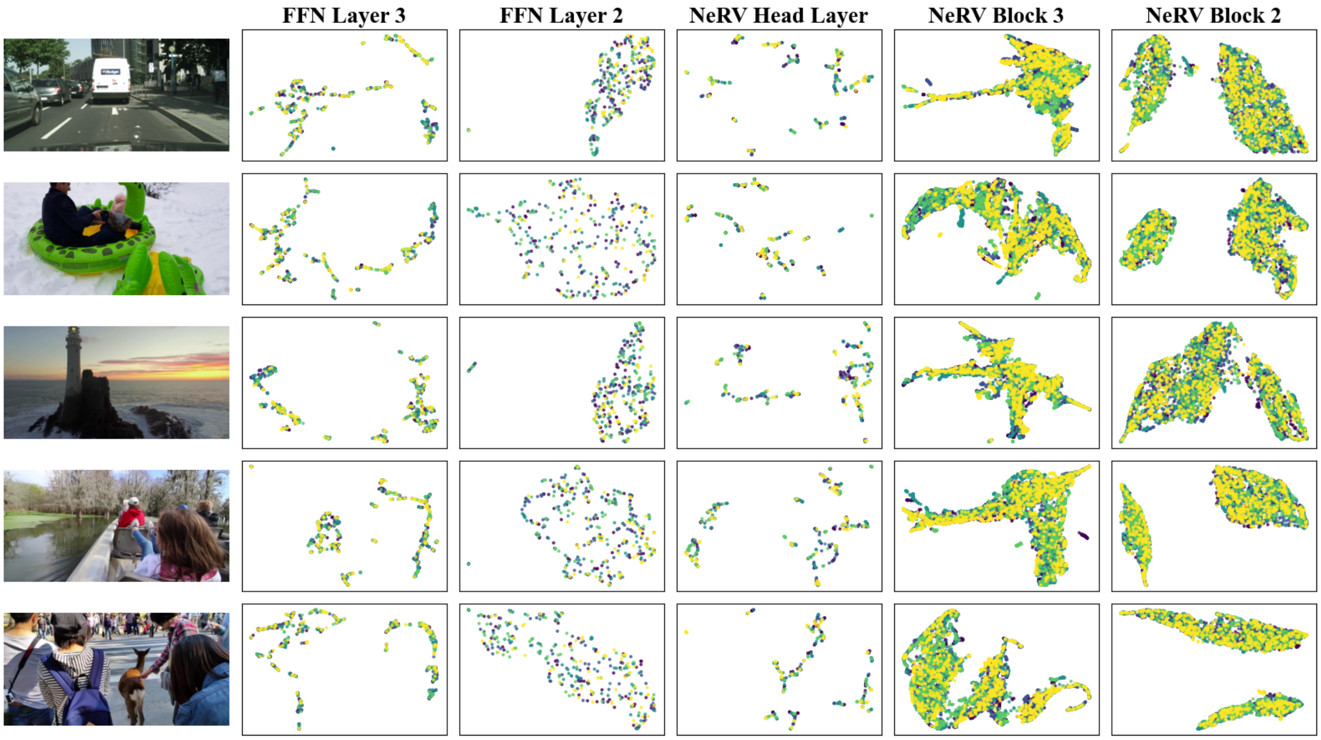

Figure 14. **Layerwise UMAP for INRs trained with different seeds**. We show results of Gabor filter based clustering neurons in each layer of MLP- and CNN-based models with different seeds. As seen in Figure 10, each cluster has neurons belonging to different seeds, indicating that models of different seeds learn a set of neuron "types".

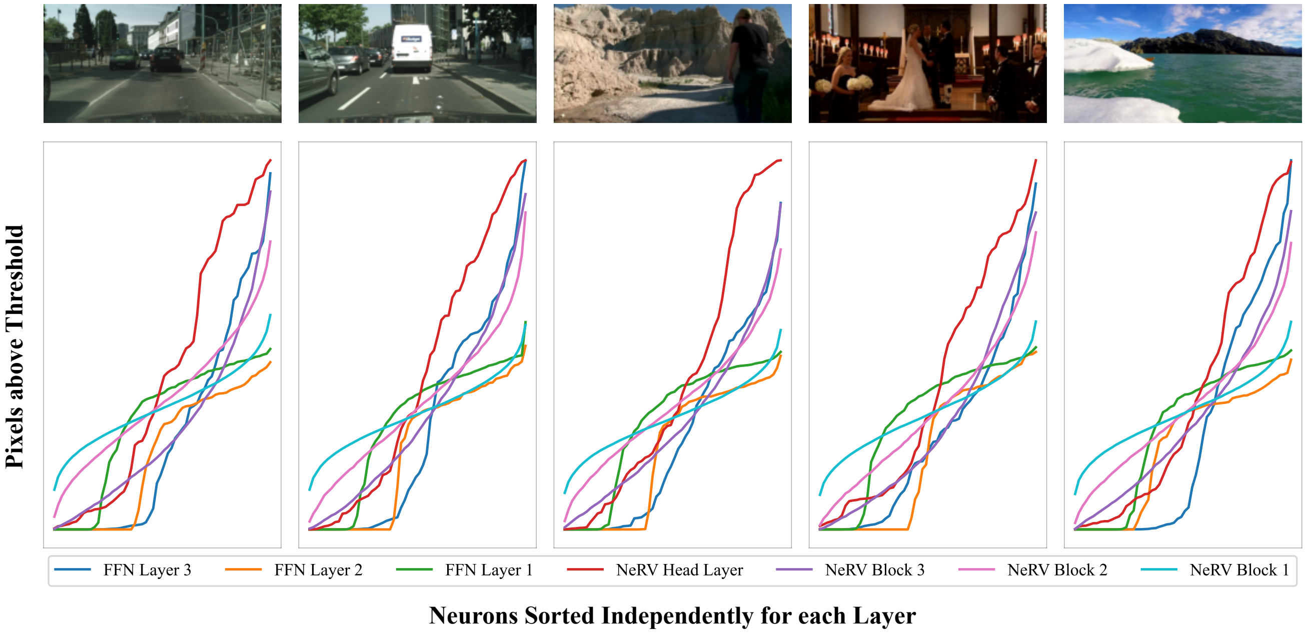

Figure 15. **Pixels per neuron.** We supplement Figure 6 by plotting the pixels activated per neuron for frames from 5 additional videos. These results reveal similar trends as seen in Figure 6.

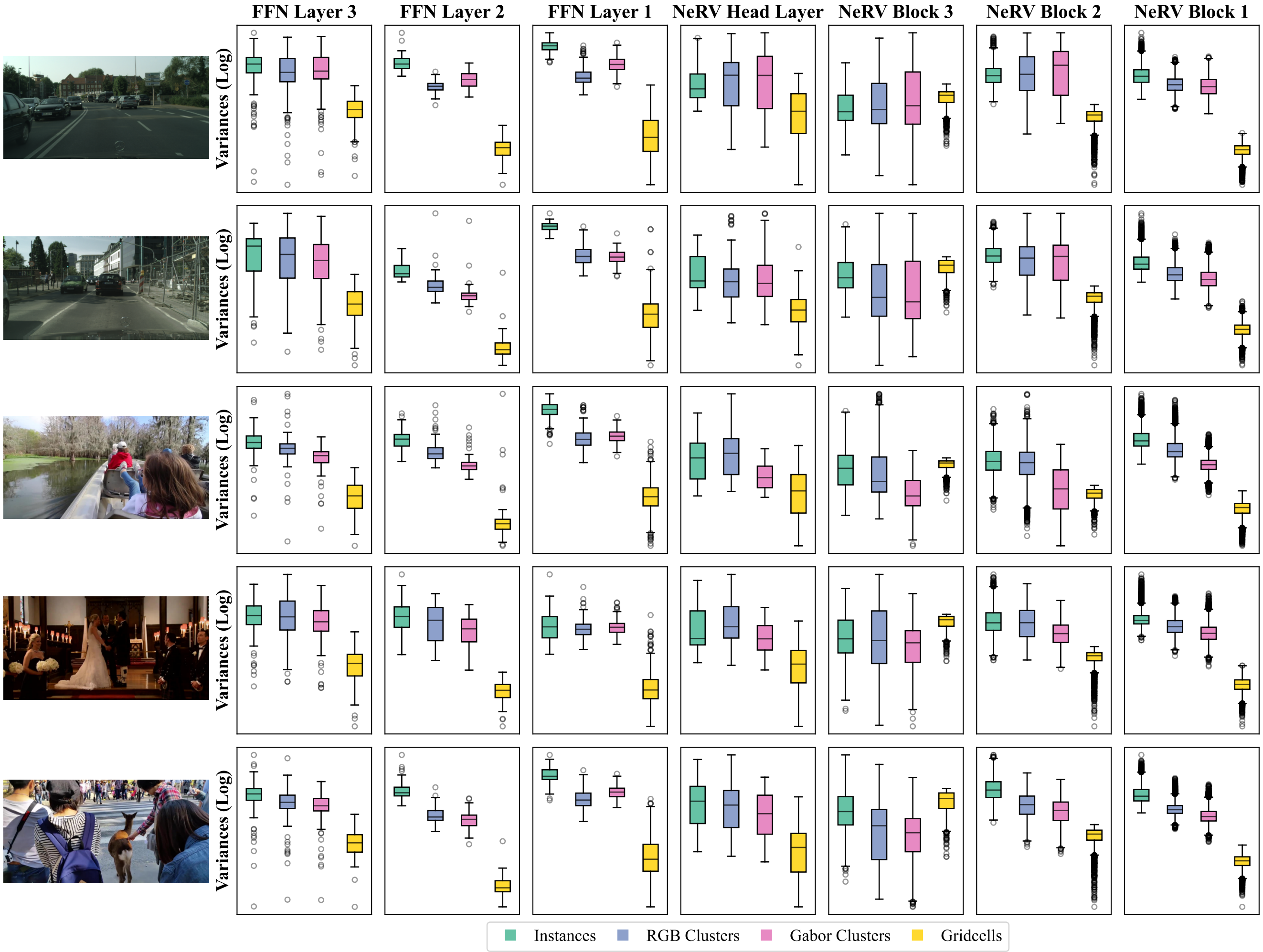

Figure 16. **Grouping contributions.** Similar to Figure 4, we observe higher variances in neuron contribution differences when using instances, RGB clusters, and Gabor clusters in contrast to space clusters (gridcells). These observations lend support to our hypothesis that INRs prefer low-level object semantics while demonstrating a tendency to disregard space.

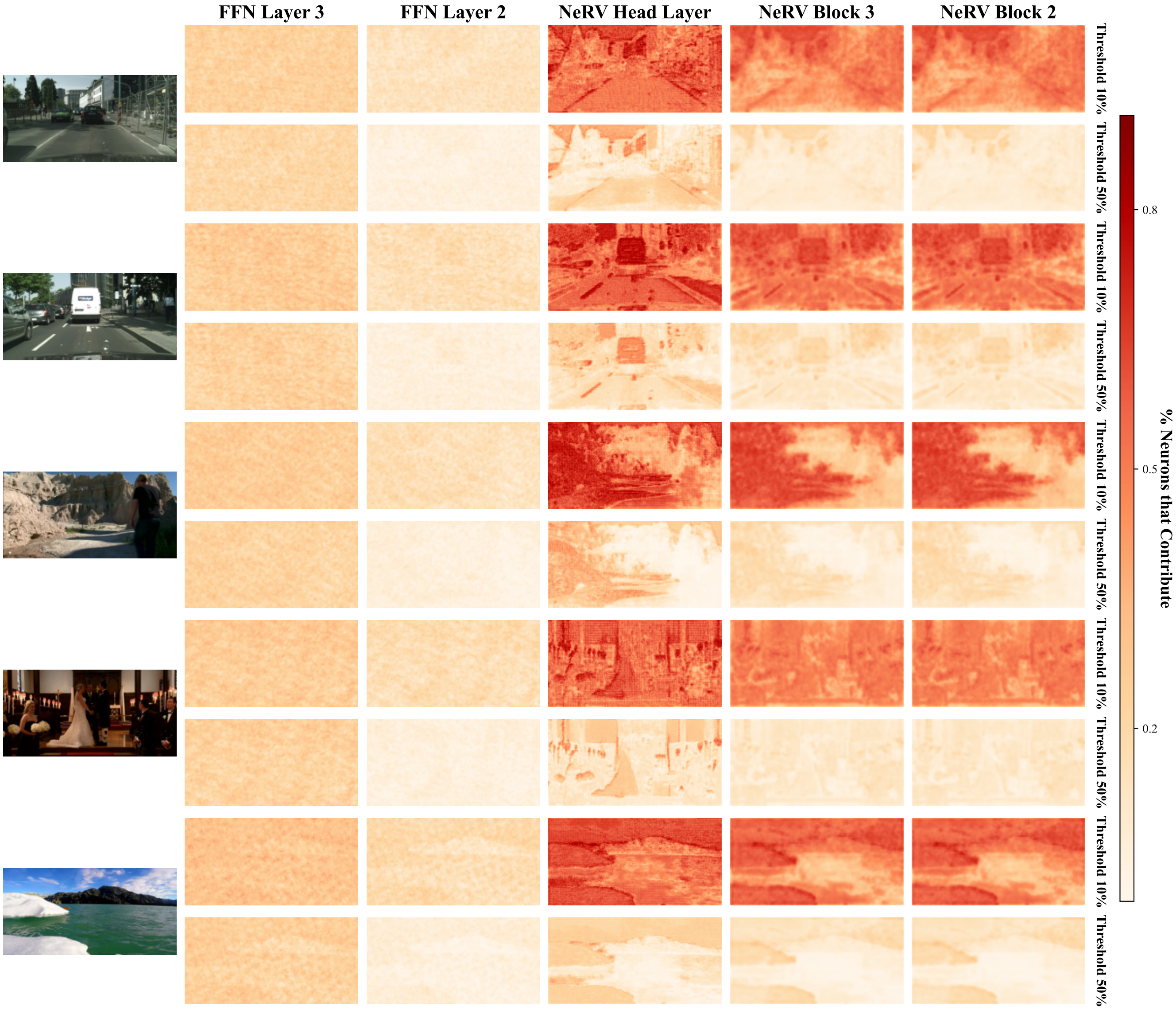

Figure 17. **Neurons per pixel.** We show the percentage of neurons in each layer, that represent significant portions of each pixel, at two different thresholds. This figure reveals properties consistent with Figure 7 for 5 additional video frames.

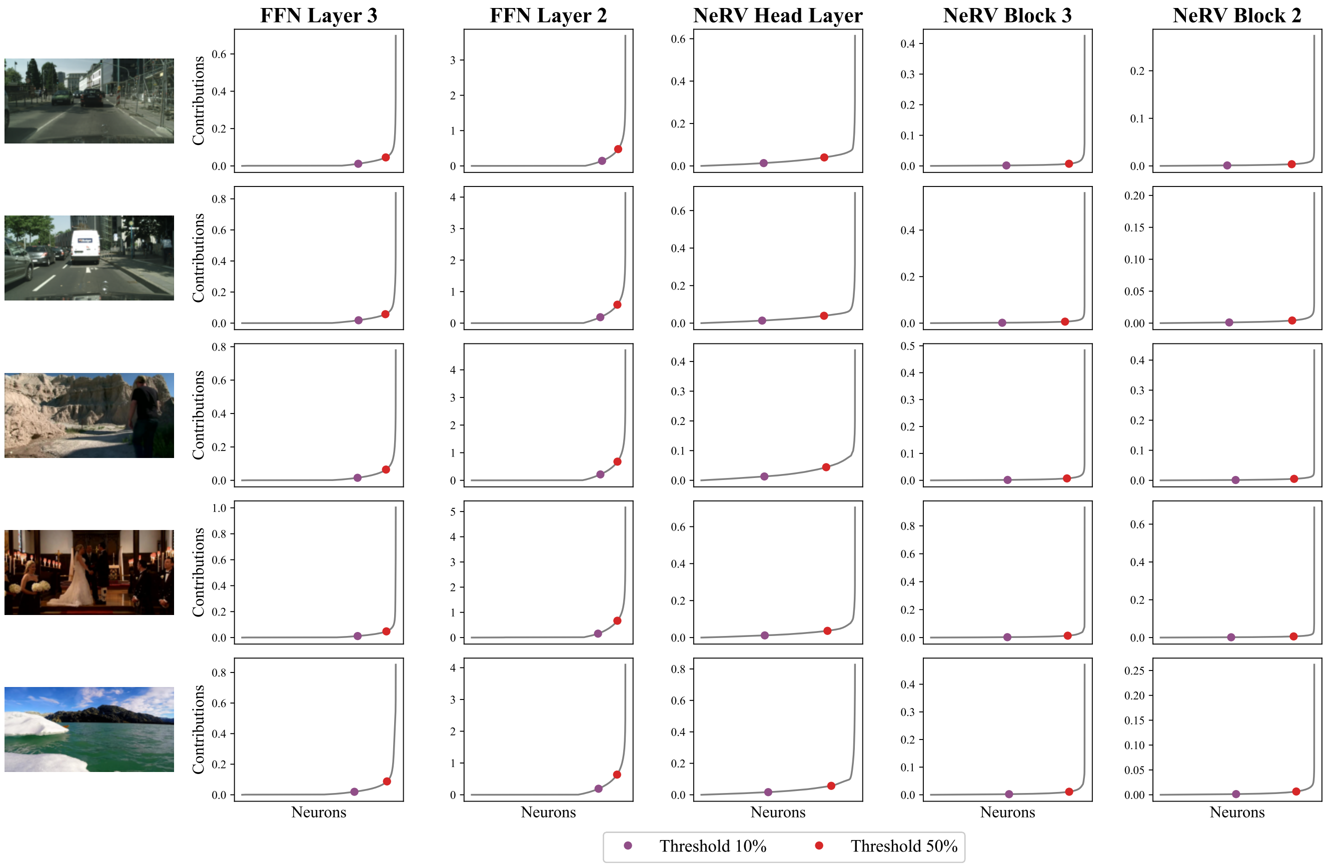

Figure 18. **Distribution of Neuron Contributions.** We plot the distribution over all neuron contributions in each layer. This shows that a large fraction of neuron contributions tend to be relatively smaller in magnitude. The contribution values corresponding to the 10th and 50th percentile of each distribution are used in selecting thresholds for Figures 7 and 17.

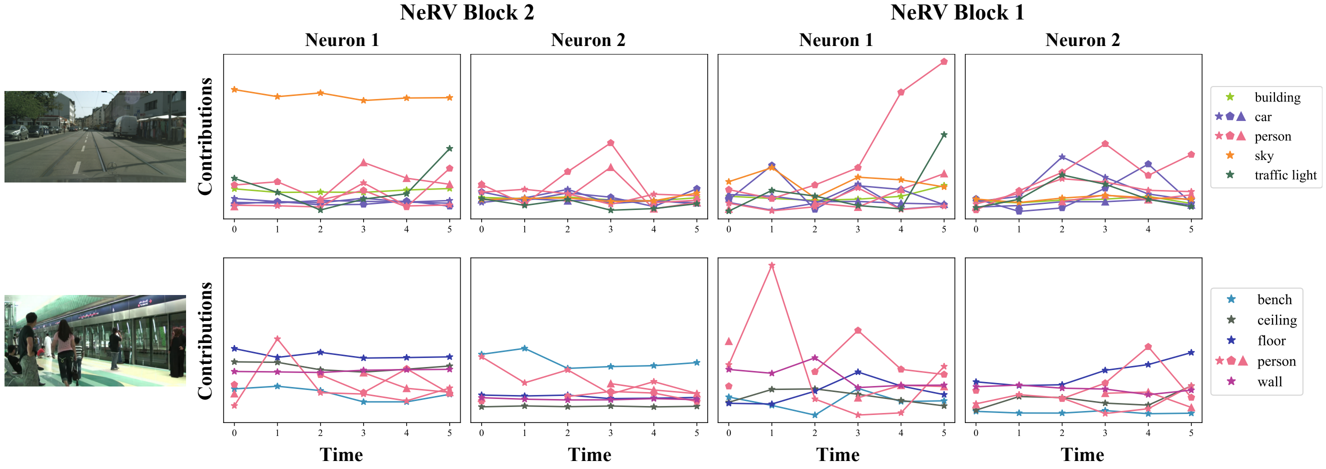

Figure 19. **Neuron contributions to things and stuff** for the first two blocks of NeRV. See Figure 8 for other blocks/layers.