A. Implementation Details

We describe the implementation details of our framework for generalizable INRs via instance pattern composers. In all experiments of this study, our transformer-based hypernetwork with 768 hidden dimensions consists of six self-attention blocks with 12 attention heads, where each attention head has 64 dimensions. We use Adam [3] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the constant learning rate of 0.0001 to train our transformers. The training epochs are different in experiments, and we describe the details below. Given a rank r of $\mathbf{U}$ and $\mathbf{V}^{(n)}$, and the hidden dimension d of the MLP, we initialize $\mathbf{U} \sim \mathcal{N}(0, 1/\sqrt{rd})$ for the stability of training, since we intend to scale the initialization of $\mathbf{W}_{\mathrm{m}}^{(n)}$ as $\mathbf{W}_{\mathrm{m}}^{(n)} \sim \mathcal{N}(0, 1/\sqrt{d})$ while $\mathbf{V}^{(n)} \sim \mathcal{N}(0, 1)$ at initialization. We also use weight standardization [5] for the weights of coordinate-based MLPs.

A.1. Audio Reconstruction

We train our framework on the train split of LibriSpeech-clean [4] for the audio reconstruction, while evaluating on its test split. We use 200 sizes of non-overlapping patches to unfold and tokenize each audio instance, which is sampled by 16kHz, and then a second of audio is expressed as a sequence of 80 data tokens. Since we train our generalizable INRs to represent one or three seconds of audios, we randomly crop training audio. For evaluation, we trim a test audio instance into one or three seconds for audio. We train both our framework and TransINR for 1,000 epochs. For a fair comparison with TransINR, which predicts 257 weight tokens, our transformer-based hypernetwork predicts $r = 256$ weight tokens for instance pattern composers. A coordinate-based MLP has five layers with $d = 256$, where $d_{\text{in}} = 1$ and $d_{\text{out}} = 1$.

A.2. Image Reconstruction

We evaluate our generalizable INRs on image reconstruction on facial images such as CelebA [8] and FFHQ [2], and natural images of ImageNette [1, 6]. We use a zero-padding for 178×178 images to convert them into the 180×180 resolution, and use non-overlapping 9×9 patches to represent an image as the sequence of 400 data tokens. For 256×256 and 512×512 images, we use 16×16 and 32×32 size of non-overlapping patches, respectively. A coordinate-based MLP has five layers with $d = 256$, where $d_{\text{in}} = 3$ and $d_{\text{out}} = 3$. We train our framework with $r = 256$ and TransINR on 178×178 CelebA, FFHQ, and ImageNette during 300, 1000, and 4000 epochs, respectively, until the training converges.

For 256×256 and 512×512 FFHQ, a model is trained during 400 epochs due to the limited computational resources, but the performance consistently improves as we train the model longer. In addition, considering the balance of computational costs of transformers and MLP for high-resolution images, we subsample 10% of coordinates to compute the mean-squared error.

A.3. Novel View Synthesis

We evaluate our framework with $r = 256$ on novel view synthesis of a 3D object based on the ShapeNet Chairs, Cars, and Lamps datasets. We follow the experimental settings of previous studies, TransINR [1, 6], except for the manual decay of learning rate in TransINR [1], but use a constant learning rate until the training converges. We train our framework and TransINR for 1000 epochs for Chairs and Cars until the training converges. However, we use 400 epochs for Lamps, since TransINR starts overfitting after 400 epochs, although our framework consistently improves the performance. Before we tokenize each image, we concatenate the starting point and direction of each emitted ray from every pixel into the RGB channels of each pixel, and then each spatial coordinate has nine channels of features. Given 128×128 resolution of images per view, we use 8×8 non-overlapping patches to represent each image as the sequence of 256 data tokens. When multiple support views are used, we concatenate the data tokens of each view as one sequence. A coordinate-based MLP has six layers with $d = 256$, $d_{\text{in}} = 3$, $d_{\text{out}} = 4$. We use adaptive random sampling [1] during the first epoch to stabilize the training. We subsample 128 rays during training.

B. Examples of Novel View Synthesis

Figure B, we attach more examples of novel view synthesis on ShapeNet Chairs, Cars, and Lamps by our framework, where the number of support views is increased from one to five. The quality of synthesized images increases as the number of support views increases, while our framework modulates only one weight matrix of the instance pattern composer.

C. Comparison with overfitted INRs

Figure A shows the efficiency to represent data as INRs with or without individual training of MLPs. Since our framework exploits FFNets for INRs, we perform test-time optimization (TTO) on FFNets initialized 1) randomly, 2) by our meta-learning, and 3) by our transformer-based hypernetwork. Since the inference time for weight modulation is shorter than one optimization step, the time is negligible, while providing meaningful representations without individual training of FFNets. When the number of trainable parameters is equivalent, the PSNRs converge to similar values after TTO, but our TTO w/ $\mathbf{V}^{(n)}$ can maintain interpretable structures in Figure D. When we optimize the entire weights of INRs, our framework shows better performance during training than random initialization.

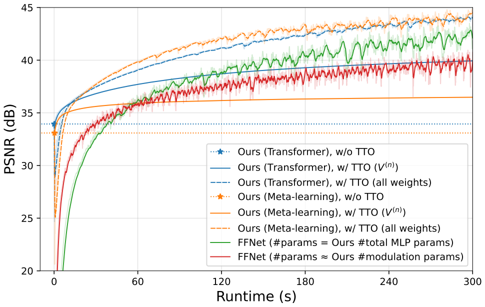


Figure A. Time/PSNR trade-off during training of randomly initialized FFNets and our generalizable INRs. Each FFNet is trained per a sample in randomly selected 10 images in FFHQ 256×256.

D. Visualization Analysis of MLP Activations

After we separately train a coordinate-based MLP, denoted as FFNet [7], on each image, we visualize the activation patterns of each neuron in a layer over all coordinate inputs. Since the two FFNets memorize their training sample separately, the activation maps neither capture the common representations across instances nor be easily interpreted. On the other hand, Figure D shows that our generalizable INRs can exploit the instance-agnostic pattern composition rule, which enables each neuron to capture common and interpretable structures across instances. The visualization of activation maps validates that our framework enables the learned representations and pattern composition rule of coordinate-based MLP to be effectively generalized to unseen data instances.

References

[1] Yinbo Chen and Xiaolong Wang. Transformers as meta-learners for implicit neural representations. In *European Conference on Computer Vision*, pages 170–187. Springer, 2022. 1

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1

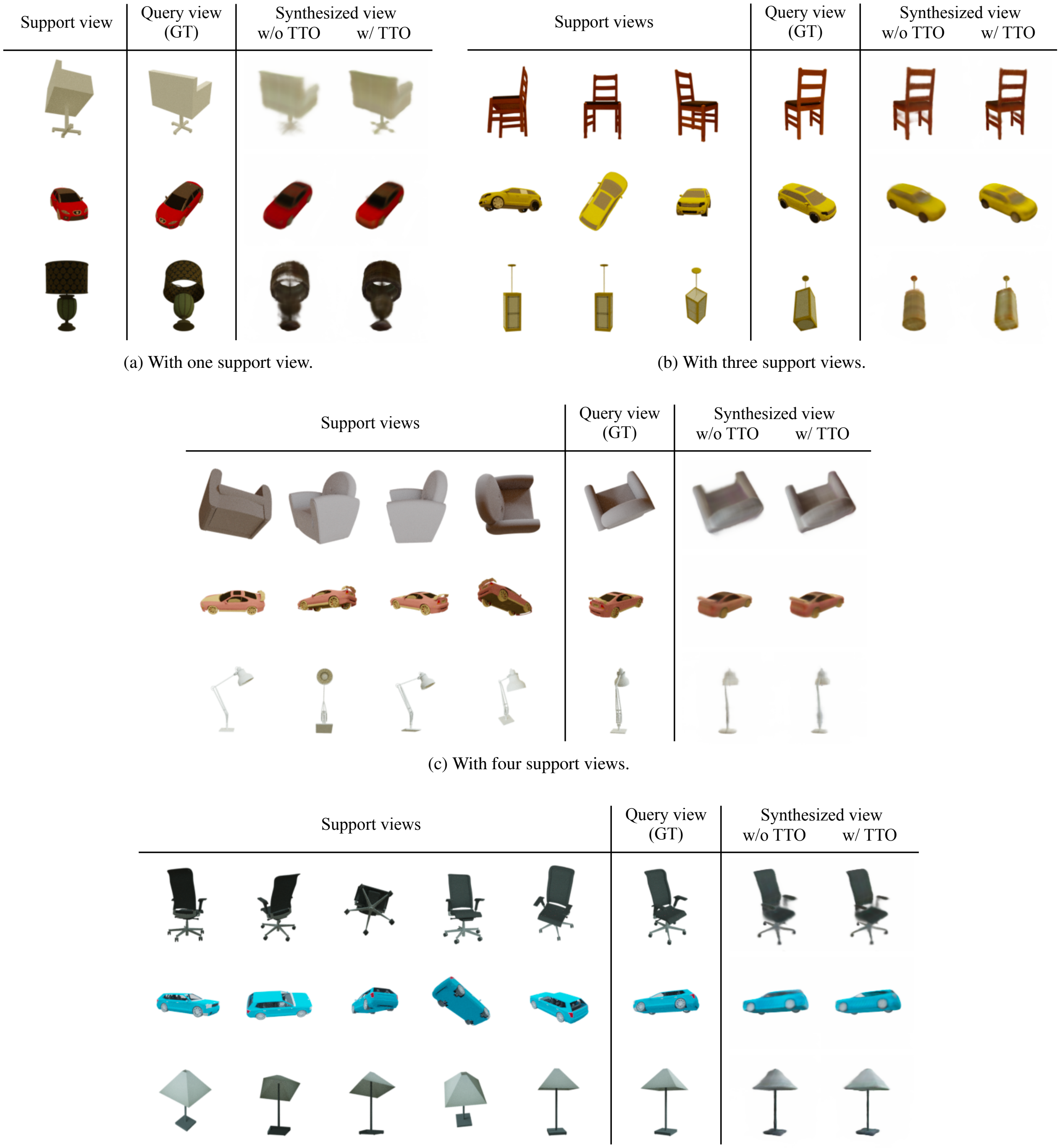
[4] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. 1

[5] Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019. 1

[6] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021. 1

[7] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2, 4

[8] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE international conference on computer vision*, pages 3676–3684, 2015. 1



(a) With one support view.

(b) With three support views.

(c) With four support views.

(d) With five support views.

Figure B. The examples of Novel view synthesis on Chairs, Cars, and Lamps by our framework with one, three, four, and five support views (a-d).

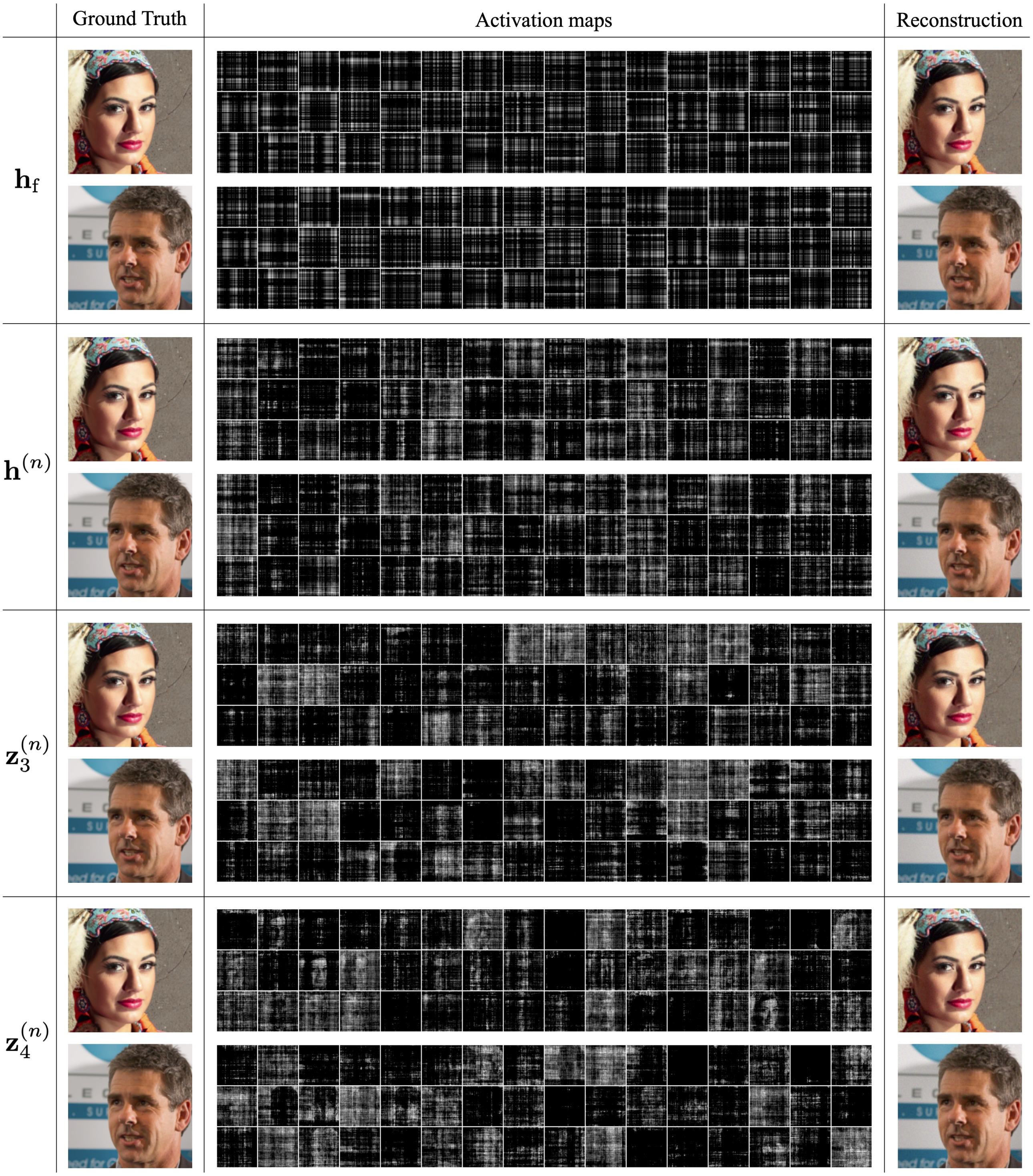


Figure C. Activation maps of two FFNets [7], which are separately trained to memorize each of the two images.

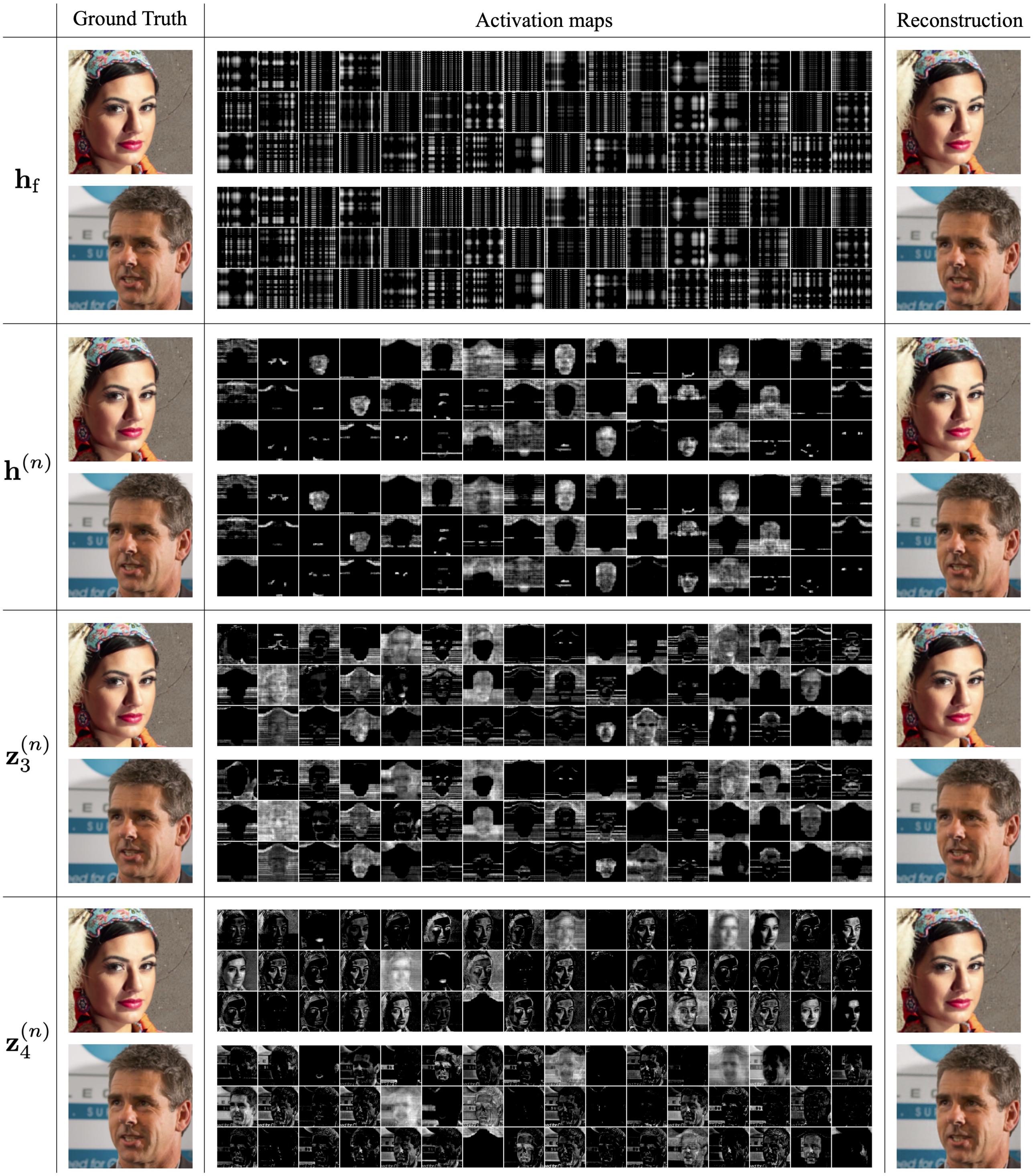


Figure D. Activation maps of two INRs predicted by our framework of generalizable INRs for each of the two images.