

CONTROLLABLE DATA GENERATION WITH HIERARCHICAL NEURAL REPRESENTATIONS – APPENDIX

Anonymous authors

Paper under double-blind review

A DETAILS ON EXPERIMENTAL SETUP

Implementation details. The LoE structure can be configured with the number of layers L , the number of experts at each layer K , the channel dimension of each expert C , and the dimension of the latent at each layer H , denoted as a tuple (L, K, C, H) . We train LoEs of $(7, 384, 128, 128)$, $(5, 256, 64, 256)$, $(6, 256, 64, 64)$, and $(5, 64, 64, 64)$ in CelebA-HQ (Karras, 2017), ShapeNet (Chang et al., 2015), SRN-Cars (Sitzmann et al., 2019), and AMASS (Mahmood et al., 2019) datasets, respectively. We follow mNIF (You et al., 2024) on the data processing protocols for CelebA-HQ, ShapeNet, and SRN-Cars datasets. Details about the AMASS dataset are provided in Sec. B.3.

Training details. In Stage-1, we train LoEs via meta-learning on CelebA-HQ, ShapeNet, and AMASS, and with auto-decoding on SRN-Cars. We use a batch size of 32, an outer learning rate of $1e-4$, an inner learning rate of 1 with 3 steps, and train the LoE for 800 epochs in the meta-learning setting. For auto-decoding experiments on SRN-Cars, we use a batch size of 8, a learning rate of $1e-4$, and train the LoE for 1000 epochs. In both settings, we use the AdamW (Loshchilov, 2017) optimizer without weight decay. In Stage-2, we set the training batch size to be 32, learning rate $1e-4$, and cosine scheduler with minimum learning rate 0.0. We train the HCDM for 1000 epochs with the AdamW optimizer.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 GENERALIZABILITY ANALYSIS THROUGH RETRIEVAL

We use retrieval to compare the generalizability of CHINR and mNIF on the CelebA-HQ dataset. Specifically, we generate samples and retrieve the closest images from the training set. As shown in Fig. 1, mNIF generates samples that are very similar to the training images, suggesting a higher chance of “memorization”. In contrast, CHINR demonstrates better generalization by producing “new” samples that differ more noticeably from the training data.

B.2 MORE GENERATED SAMPLES

Fig. 2 shows more generated samples on CelebA-Net, ShapeNet, and SRN-Cars datasets.

B.3 AMASS EXPERIMENTS

We apply our proposed CHINR model to the AMASS dataset of 3D human motions. For each motion sequence, we use 200 frames, with each frame represented by 165 values corresponding to the locations and rotations of body joints. As a result, each data instance is formatted as a grid with size 200×165 . In Stage-1, the LoE is employed to fit the motion instances. In Stage-2, we set the binary lengths to 8 to avoid memorizing conditions.

Reconstruction and generation results. The reconstruction performance is shown in Table. 1. The randomly generated motions are shown in Fig. 3.

Semantic-level Interpolation. Since the LoE successfully learns the consistent latent space, we can perform semantic-level interpolation for motions. As shown in Fig. 4, given two fitted sequential

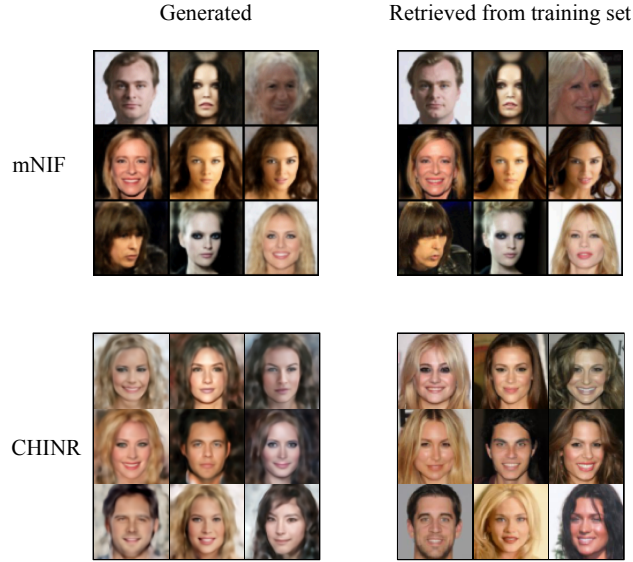


Figure 1: Retrieval on CelebA-HQ: mNIF retrieves images closely resembling those from the training set, while CHINR demonstrates better generalization by producing distinct new images.

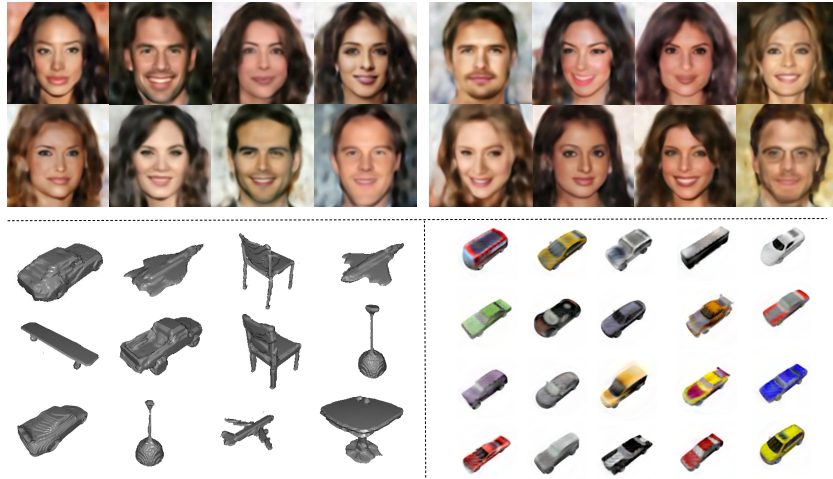


Figure 2: More generated samples of CelebA-HQ, ShapeNet, and SRN-Cars data.

Table 1: Quantitative results on AMASS.

Model	MSE↓
mNIF (You et al., 2024)	0.015
CHINR	0.011

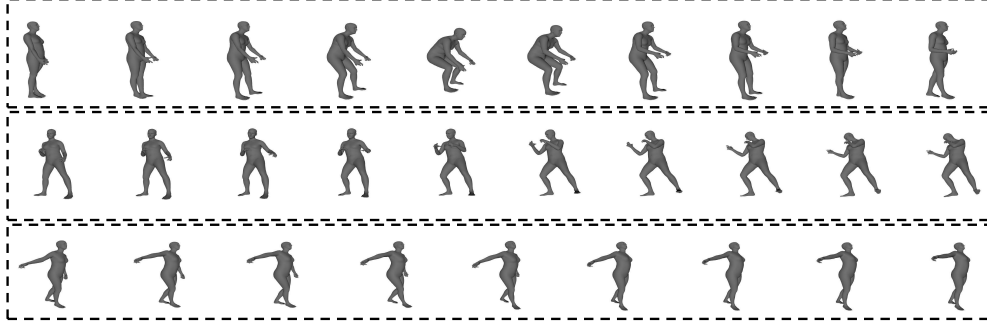


Figure 3: Generated motions with HCDM. each row denotes one sampled data.

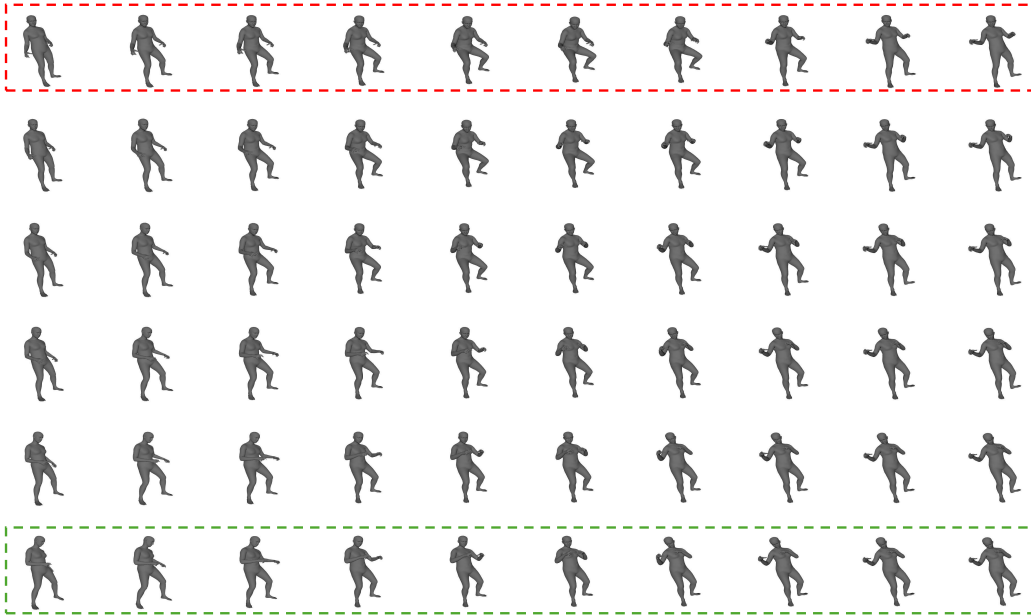


Figure 4: Semantic interpolation for AMASS data. Anchor sequential motions (indicated by the red and green dashed boxes) are first fitted with LoE to obtain latents. Then semantic-level interpolation is performed by interpolating the latents. The red dashed box denotes the start motion, and the green dashed box denotes the end motion.

motions with LoE, each corresponds to a latent, we can interpolate the latent from the start motion (indicated by the red dashed box) to the end motion (indicated by the green dashed box) linearly with ratio $[0.2, 0.4, 0.6, 0.8]$. We can see that the interpolated motions change smoothly from the start to the end. Semantic-level interpolation can be useful in the gaming industry, and 3D-digital content generation.

Temporal-level interpolation. Since the INR can generate data instances in any resolution, we can easily enlarge the input coordinates’ resolution in the time dimension to achieve temporal-level interpolation. We set the length of the time dimension to be 200 and 400, then get motions with LoE. The interpolated results are submitted as videos named “motion_short.mp4” and “motion_long.mp4”.

B.4 HIERARCHICAL CONTROLLABLE GENERATION

More examples of hierarchical controllable data generation are presented in Fig 5.

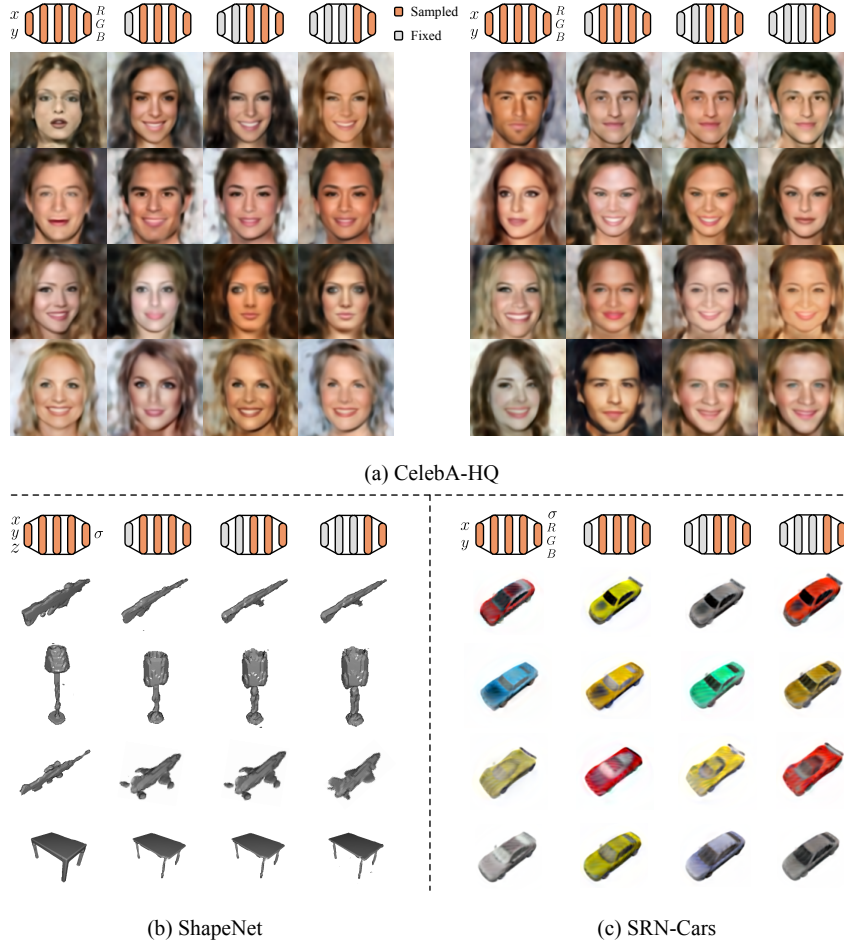


Figure 5: More examples of hierarchical controllable generation on CelebA-HQ, ShapeNet, and SRN-Cars data.

B.5 LATENT-BASED RETRIEVAL

We show an application of data retrieval by latents, since they already embed rich semantic meanings. We first obtain the latents for the target data by fitting it to the LoE through a few gradient steps. Once the latents are optimized, they can be used to retrieve similar data by comparing their latent representations to the searched set, allowing us to search for semantically similar examples within the latent space. Fig. 6 shows this process by using images from the test-split of CelebA-HQ as the targets, and train-split images as the searched set. We demonstrate two approaches for retrieval: (1) using the flattened \mathbf{h} for all layers, and (2) layer-wise retrieval using each layer’s latent \mathbf{h}^l . As shown in Fig. 6, retrieval by the flattened \mathbf{h} will retrieve samples that are broadly similar, while layer-wise retrieval retrieves samples with specific semantic similarities. For example, \mathbf{h}^2 retrieves faces with similar orientations, while \mathbf{h}^3 retrieves faces with similar facial features such as eye shape and expressions.

C ANALYSIS

In this section, we provide more analysis of the latent space and the functionalities of binary conditions.

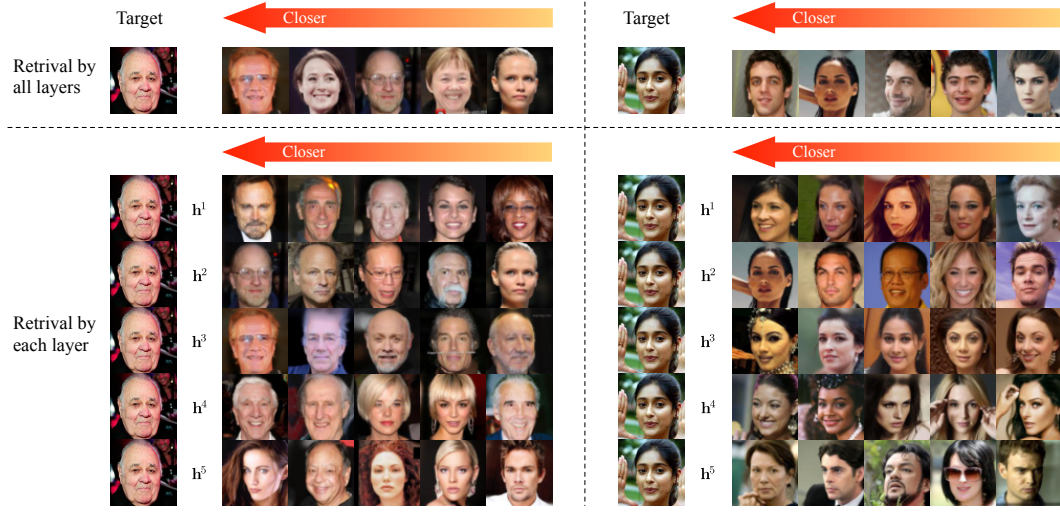


Figure 6: Latent-based retrieval via two approaches: retrieval by all layers and retrieval by each layer.

C.1 LATENT SPACE ANALYSIS

Here, we analyze the latent space further, focusing on its interpolation capabilities and providing additional results of correlation analysis.

C.1.1 LATENT INTERPOLATION

To illustrate that our model learns a consistent and metric latent space, following definitions in Du et al. (2021), we perform latent space interpolation in two ways: complete interpolation, and layer-wise interpolation.

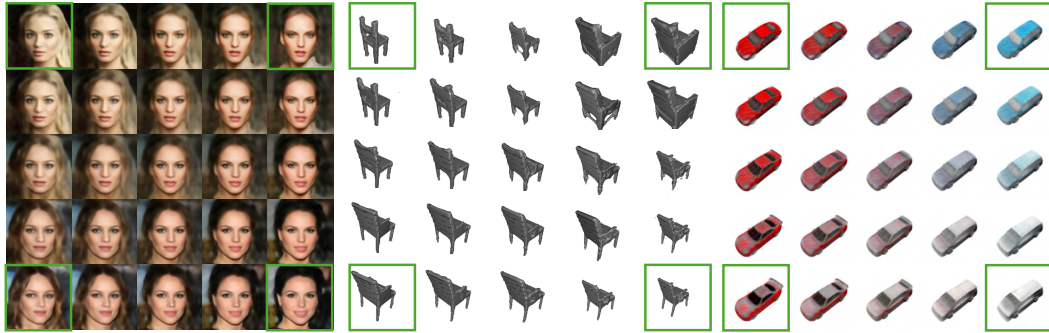


Figure 7: Latent space interpolation is performed for LoE, with four corner points representing the anchor examples rendered in stage 1. The intermediary points are generated through the bilinear interpolation of the latents associated with these four anchors. The interpolation is evaluated on datasets CelebA-HQ, ShapeNet, and SRN-Cars.

Complete Interpolation is shown in Fig.7. Four corners present the signals with latent generated from Stage-1. The intermediary signals are bilinearly interpolated from four corners in latent space. The results demonstrate that the learned latent is metric and consistent with human perception.

Layer-wise Interpolation. Since our LoE embeds semantics hierarchically in different parts of the latent, we can interpolate each part to control specific semantics. As shown in Fig. 8, we interpolate the second, third, and fourth parts of the latent associated with red-boxed signals, with the corresponding parts of the right side latent. For CelebA-HQ samples, we find that the facial orientation,

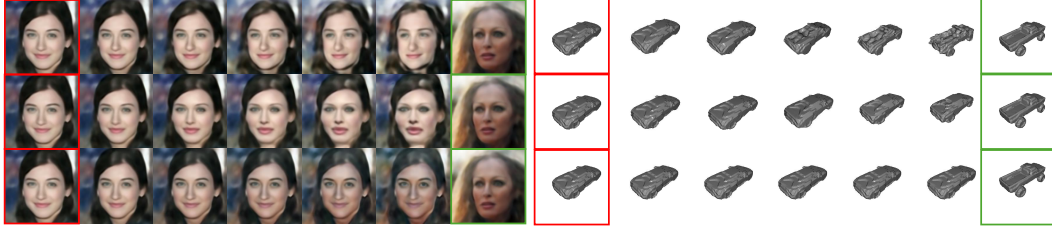


Figure 8: Layerwise interpolation. The red boxes denote the start and the green boxes denote the end. For the CelebA-HQ, the layers 2 \rightarrow 4 are interpolated respectively while other layers are fixed. For the ShapeNet, the layers 1 \rightarrow 3 are interpolated respectively.

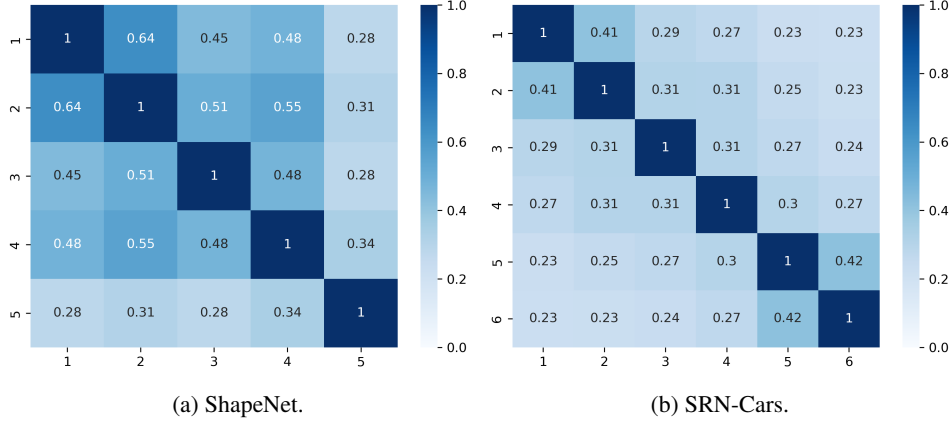


Figure 9: Correlation between the learned latents across layers, trained on ShapeNet (Chang et al., 2015) and SRN-Cars (Sitzmann et al., 2019). The non-negligible correlation between adjacent layers (e.g., h^1 and h^2) reveals the necessity of conditional distribution learning.

facial features, and skin tone can be interpolated independently. This demonstrates that each part of the latent also constructs a metric and consistent manifold.

C.1.2 LAYER-WISE CORRELATION ANALYSIS

We provide correlation analysis on additional datasets in Fig. 9 and Fig. 10. This highlights the significance of conditional modeling in the hierarchical generation process.

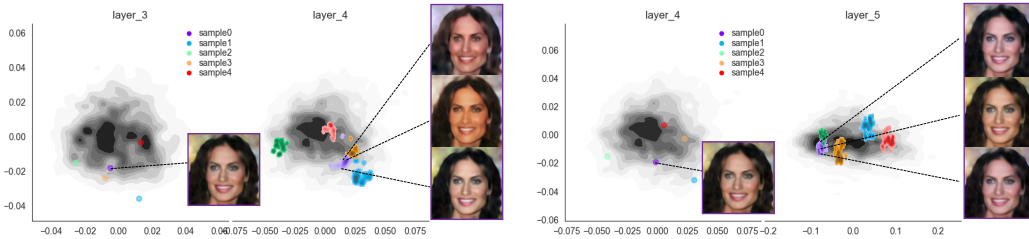


Figure 10: Visualization of conditional distributions across layers 3, 4, 5. The gray regions present the distribution of latents from Stage-1, while the colored regions represent the sampled latents from Stage-2.

C.2 BINARY CONDITION ANALYSIS

We analyze the clustering of latents and binary conditions on CelebA-HQ dataset, as shown in Fig. 11. Firstly, we use the KMeans algorithm to get 10 clusters of latents, shown as the dots in

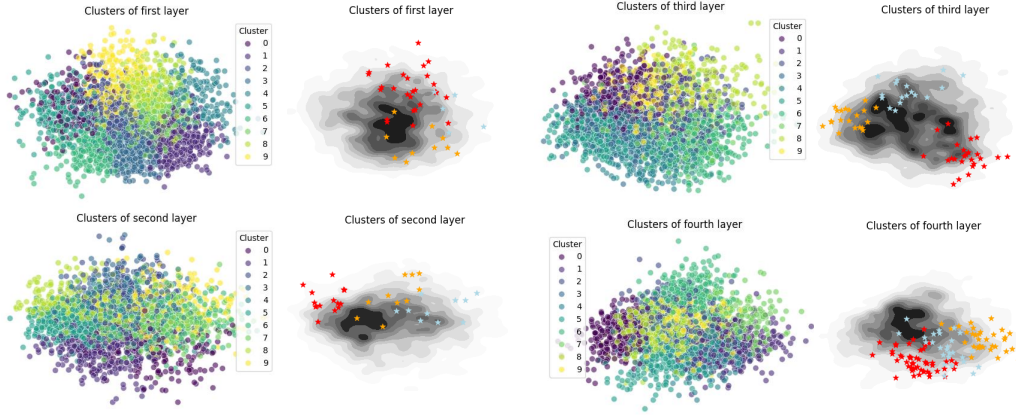


Figure 11: Clusters of each part of latent and binary conditions. The dotted plot presents clusters of each part of latents trained on ClebA-HQ. The gray distribution plot presents the distribution of each part of latents, and starred scatter plot presents clusters of latents with similar binary conditions.

the figure. Then we select three anchor latents, generate three binary conditions with HCDM, and search the nearest binary-corresponded latents. The nearest neighbors are represented by the stars. We can observe that the binary conditions embed the latents’ information and form a consistent binary condition space. This binary condition space corresponds to the latent space.

REFERENCES

- Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.
- Yilun Du, Katie Collins, Josh Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields. *Advances in Neural Information Processing Systems*, 34:8320–8331, 2021.
- Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.
- Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.
- Tackgeun You, Mijeong Kim, Jungtaek Kim, and Bohyung Han. Generative neural fields by mixtures of neural implicit functions. *Advances in Neural Information Processing Systems*, 36, 2024.