VSCHH 2023: A Benchmark for the View Synthesis Challenge of Human Heads

Youngkyoon Jang[§†1], Jiali Zheng[§†1], Jifei Song[†2], Helisa Dhamo[†2], Eduardo Pérez-Pellitero[†2], Thomas Tanay[†2], Matteo Maggioni[†2], Richard Shaw[†2], Sibi Catley-Chandar[†2,3], Yiren Zhou[†2], Jiankang Deng[†1], Ruijie Zhu, Jiahao Chang, Ziyang Song, Jiahuan Yu, Tianzhu Zhang, Khanh-Binh Nguyen, Joon-Sung Yang, Andreea Dogaru, Bernhard Egger, Heng Yu, Aarush Gupta, Joel Julin, László A. Jeni, Hyeseong Kim, Jungbin Cho, Dosik Hwang, Deukhee Lee, Doyeon Kim, Dongseong Seo, SeungJin Jeon, YoungDon Choi, Jun Seok Kang, Ahmet Cagatay Seker, Sang Chul Ahn, Aleš Leonardis[†4], and Stefanos Zafeiriou[‡†1]

[1]Imperial College London [2]Huawei Noah's Ark Lab [3]Queen Mary University of London [4]University of Birmingham

Abstract

This manuscript presents the results of the "A View Synthesis Challenge for Humans Heads (VSCHH)", which was part of the ICCV 2023 workshops. This paper describes the competition setup and provides details on replicating our initial baseline, TensoRF. Additionally, we provide a summary of the participants' methods and their results in our benchmark table. The challenge aimed to synthesize novel camera views of human heads using a given set of sparse training view images. The proposed solutions of the participants were evaluated and ranked based on objective fidelity metrics, such as PSNR and SSIM, computed against unseen validation and test sets. In the supplementary material, we detailed the methods used by all participants in the VSCHH challenge, which opened on May 15th, 2023, and concluded on July 24th, 2023.

1. Introduction

Recent advances in novel view synthesis using Neural Radiance Fields (NeRF) [18] have unlocked diverse scenarios, such as reconstructing 3D scenes with only a few images [19] or using unstructured collections of photographs [14], editing scenes [33, 10], rendering city-scale scenes [25], and novel high dynamic range (HDR) view synthesis [16]. These scenarios have been made possible by resolving technical challenges such as noisy [12, 26] or unknown [35] camera poses, sparse views [23], motion-blurred images [29], and unbounded scenes [1]. As methods in novel view synthesis have advanced, attention has also increased towards creating realistic human head avatars. To address the technical challenges and benefits of targeting human heads, several novel datasets [2, 36, 40, 32, 8, 37] and methods [21, 30, 6, 5, 15, 27] have been proposed.

	Registration	Submission		
		DevPhase	ChaPhase	Report
Record Date (July)	12th	12th	19th	24th
Count	97	20	12	9
% from Prev.Phase	100%	20.62%	60.00%	75.00%

Table 1: From interest in the ILSH dataset to active participation in the VSCHH challenge.

Although recent advances in targeting human heads have continuously resolved issues such as the dynamic movement of talking heads [39, 31] and the need for a generalized model [15], they are still far from achieving real-time training and rendering speeds while achieving high fidelity output, in part due to their high computational complexity and data requirements, e.g. the need for dense and accurate camera poses, which remain a necessity when aiming at high-quality outputs. Due to the challenges of implementing realistic human avatars that run in real-time, recent research approaches [13, 9] that consider common commercialization constraints often use conventional 3D vision techniques rather than neural rendering methods. For example, Project Starline [9], which targets remote communication, uses an image-based formulation of geometry fusion to merge multiple depth and color images. It also combines

[§]These authors contributed equally to this work.

[‡]S.Zafeiriou (s.zafeiriou@imperial.ac.uk) is the corresponding author.

[†]These authors are the "To NeRF or not to NeRF: VSCHH 2023" organizers, while the other authors are participants in the VSCHH challenge. See Appendix A for the affiliations of the participants.

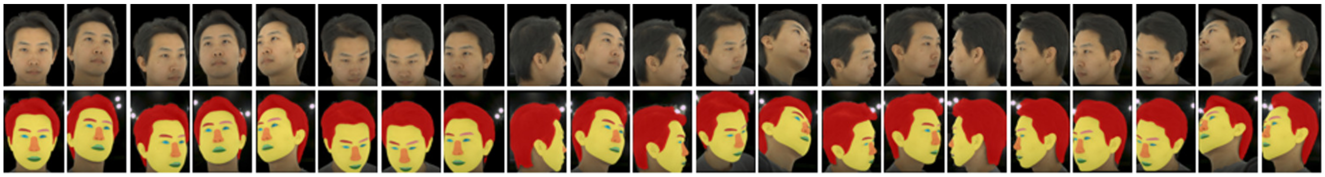

Figure 1: Example images and face masks, excluding three backside views where the face detection failed and masks could not be generated.

2D facial landmark estimation, 3D triangulation, and double exponential filtering to track 3D facial features.

Both NeRF and non-NeRF conventional 3D vision approaches have their own advantages and disadvantages. The theoretical basis for neural rendering builds upon both conventional 3D vision (e.g. multi-view geometry) and computer graphics (i.e. volume rendering, ray marching). Since both NeRF and non-NeRF approaches are rooted in similar theoretical foundations, inviting both communities and approaches by offering a unified benchmark to explore potential methods may create an opportunity to understand their potential bridging points. With this idea in mind, we organized the VSCHH challenge in conjunction with the "To NeRF or Not to NeRF" workshop to invite participants from diverse communities to submit competitive methods using the newly released, publicly available ILSH dataset [37] for the task of novel view synthesis for human heads.

This paper introduces the VSCHH challenge and our baseline approach (Sec. 2), which serves as an initial starting point. In addition, this paper summarizes and discusses the overview and achievements of the participants' methods in a generic manner using our benchmark table (Sec. 3). Detailed approaches of all participants are presented in the supplementary material. The main contribution of this report is the provision of comprehensive benchmarks derived from the results of all participants in the VSCHH challenge. By enabling participants to explore any potential methods without restrictions, their selection of baselines and additional approaches to address the challenges represent valuable investigations using the novel light-stage head dataset [37].

2. The VSCHH Challenge

We organized a challenge called "To NeRF or not to NeRF[*]: A View Synthesis Challenge for Human Heads (VSCHH)" based on the publicly available ILSH dataset [37]. The VSCHH challenge comprises a novel view synthesis task, which aims to test the capability of algorithms to generate new views for human head images given a set of relatively sparse views in training that also have visually disturbing light blooms serving as noise (as they are occluded from different viewpoints and do not always appear). The challenge consists of two phases: the Development Phase and the Challenge Phase. During the Development Phase, participants have the opportunity to test their ideas using our hidden validation set, which can only be validated through our CodaLab[¶] submission platform. During the final Challenge Phase, participants are required to submit their results produced using the test pose inputs[‖]. A feedback for the final Challenge Phase was not provided until after the challenge had ended.

The ILSH dataset contains light-stage captured human head images from 52 subjects, captured using 24 cameras under uniform illumination conditions. This results in a total of 1,248 close-up head images, border masks, and camera pose pairs. Along with the dataset, we released a codebase that includes scripts for restructuring downloaded subdatasets, loading data, checking submission files, visualizing camera poses, and evaluating results. Please refer to the ILSH paper [37] for details on how the dataset was collected and prepared to support a view synthesis challenge.

In the VSCHH challenge, submissions were limited to a maximum of 200 per individual participant, with a daily limit of 20 submissions for the Development Phase. For the final Challenge Phase, we only allowed a maximum of 3 submissions per day and 20 in total. After the challenge opened on May 15th, the Imperial College London team shared the ILSH dataset, following a careful process of receiving an End User License Agreement (EULA) document with the academic faculty (or line manager)'s signature and collecting the identity of individual researchers. This was done to track and confirm that only guaranteed research teams had access, as advised by the Imperial College London Ethics Committee. After releasing the dataset, a total of 97 teams registered to express their interest in downloading it by July 12th. In addition, during the Development Phase, 20 teams tested their methods on the ILSH dataset using the validation set. Finally, by the end of the final Challenge Phase, 12 of these teams submitted their test results using the test set. A total of 9 teams ultimately decided to submit their complete results, including final output images, technical reports, and train/test codes for validation of their development, as shown in Table 1.

[*] Website: https://sites.google.com/view/vschh/

[¶] CodaLab: https://codalab.lisn.upsaclay.fr/competitions/13273

[‖] The validation and test sets of the ILSH dataset [37] are not shared publicly. Instead, they are kept within the Codalab platform for evaluation purposes only, along with face masks as shown in Fig. 1.

Evaluation Region	Full Region		Masked Region		Time (Sec.)	NVIDIA GPU	Details
Evaluation Metric	PSNR	SSIM	**PSNR**	SSIM			
C1:MPFER-H	28.05	0.84	28.90	0.83	1.50	V100	Supp. Material Sec. 1
C2:DINER-SR	22.37	0.72	28.50	0.83	87.25	V100	Supp. Material Sec. 2
MPFER [24]_*C1	26.28	0.81	*27.82*	0.82	0.75	V100	Supp. Material Sec. 1
T1:OpenSpaceAI	21.66	0.68	**27.02**	0.83	76.88	RTX 3090	Supp. Material Sec. 3
T2:NoNeRF	20.37	0.69	**26.43**	0.82	175.58	RTX 3090	Supp. Material Sec. 4
T3:CogCoVi	21.49	0.70	**26.33**	0.82	806.00	A40	Supp. Material Sec. 5
TensoRF [3]_*C0	20.54	0.71	*26.17*	0.82	94.02	V100	Sec. 2.1
T4:CUBE	21.07	0.66	25.72	0.81	95.00	A100 & H100	Supp. Material Sec. 6
T5:Y-KIST-NeRF	20.73	0.71	25.54	0.82	15.10	RTX 6000	Supp. Material Sec. 7
*TensoRF [3]_T6	20.09	0.65	25.30	0.81	758.00	A10	Supp. Material Sec. 8
T6:xoft	20.01	0.64	25.02	0.80	727.00	NVIDIA A10	Supp. Material Sec. 8
*TensoRF [3]_T1	20.28	0.70	24.70	0.81	31.13	RTX 3090	Supp. Material Sec. 3
*TensoRF [3]_T2	20.13	0.70	24.37	0.81	72.55	RTX 3090	Supp. Material Sec. 4
*NeuS [28]_T3	21.02	0.72	24.33	0.80	944.00	A40	Supp. Material Sec. 5
*Mip-NeRF360 [1]_T4	20.59	0.71	24.13	0.80	78.00	A100 & H100	Supp. Material Sec. 6
*TensoRF [3]_T5	19.87	0.69	24.07	0.81	14.50	RTX 6000	Supp. Material Sec. 7
T7:KHAG	22.14	0.64	23.39	0.79	2.58	RTX A6000	Supp. Material Sec. 9
*Nvdiffrec [20]_T7	20.03	0.62	22.96	0.78	0.22	RTX A6000	Supp. Material Sec. 9
DINER [22]_*C2	14.81	0.58	*22.72*	0.78	86.37	V100	Supp. Material Sec. 2

Table 2: Results of all participants' methods, as well as the baselines, obtained using the ILSH dataset. The asterisk symbol * represents a baseline method that each team has chosen and tested. T# represents a participant team ID, while C# represents a challenge organizing team ID.

2.1. The Baseline: TensoRF [3]

The ILSH dataset presents its own challenges, such as relatively sparse camera views available for training and small object bounding boxes compared to the actual scene box. In addition to these generic challenges, the limited timeframe of the challenge (about two months) and the number of subjects to be tested within that timeframe are additional constraints that participants must consider. Developing ideas based on conventional (slow) methods such as vanilla NeRF [17] may not be suitable for training and advancing within the given timeframe of the Development Phases. Given these constraints, we tested non-face-specific but known-to-be-fast models for training and testing, and decided to provide a baseline using TensoRF [3].

TensoRF [3] is a neural radiance field method that models 3D scenes as a 4D tensor, i.e., a 3D voxel grid with a per-voxel feature channel. Its core idea is to decompose this 4D tensor into low-rank tensors, resulting in improved performance and run-time compared to vanilla NeRF. We found that TensoRF is relatively sensitive to scene-bound specification and camera normalization, such as pose scaling and centering. This property is associated with voxel-based neural radiance models in general, as they define a bounded scene box where the rendered subject is expected to be located close to the box center. Furthermore, due to the ambiguity caused by view sparsity, we found that tight scene bounds help reduce floaters.

To make TensoRF compatible with the dataset, we applied the following changes: **1.** Created a configuration file for the dataset with these specifications: (dataset_name = llff, downsample_train = 1.0, ndc_ray = 0, n_iters = 50000, n_lamb_sigma = [16,4,4], n_lamb_sh = [48,12,12], shadingMode = MLP_Fea, fea2denseAct = relu, view_pe = 0, fea_pe = 0, TV_weight_density = 1.0, TV_weight_app = 1.0). **2.** Set the scene bound (near far=[3.5, 7.0]) and object_bound (near far=[0.4, 2.8]). **3.** Disabled all functionalities related to NDC, as it is intended for forward-facing scenes, unlike ours. **4.** Used the provided border masks to train only in valid image regions (where mask = 1). **5.** Disabled pose centering in the data loader.

We shared these replication steps with the participants to enable them to quickly join the Challenge and focus on developing their algorithms. The results tested using this baseline method are shown as *TensoRF [3]_C0 in Table 2.

2.2. Evaluation Metrics

In addition to the baseline method, we explained our evaluation metrics on the challenge website (CodaLab). These metrics are also demonstrated using the starting kit in the dataset. The following descriptions provide a detailed

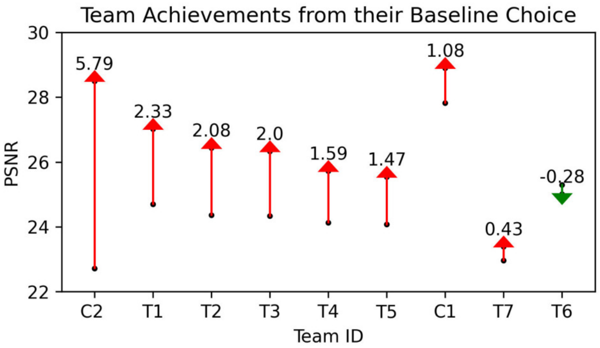

Figure 2: Scores differences between baseline and improvement from each team. The team IDs are arranged in the order of their final score improvement compared to their own baseline scores. This leads to a mismatch with the actual ranking in the benchmark as shown in Table 2.

explanation of our evaluation metrics, specifically designed for the VSCHH challenge.

Description of the evaluation metrics. Submissions are evaluated using quantitative metrics, such as the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). These metrics are commonly used to measure the quality of novel view synthesis images. Evaluations were conducted using the CodaLab platform, which is designed for organizing competitions and submitting results. There are two groups of results calculated: one group evaluates the result within the face region using face masks, which are external [4, 38] and not released, as previously discussed. Another group evaluates the result over the full region, without using face masks. However, the official ranking is based on the results calculated within the face region. For the VSCHH benchmarks, we report the average results over all processed images, as shown in Table 2.

To create the face masks used in the evaluation, we downsampled the dataset to a resolution of 300 pixels in width to detect [4] and parse [38] faces in the input image. We then upsampled the parsed face mask output to the original resolution of the input image and saved it as a reference mask, as shown in Fig. 1. Although the upscaled face masks are not pixel-perfectly watertight for the face region in the original high-resolution images, we internally agreed to use them as they seem to reasonably cover the overall face region with just a few pixels of difference, which is acceptable for the main purpose of using the face masks, i.e., face masked-region evaluation.

3. VSCHH Benchmark and Technical Report

Challenge organizing teams. As part of the organizational effort, two teams explored the use of MPFER [24] and DINER [22] as additional baselines, in addition to our initial baseline TensoRF [3], as described in Sec. 2.1. These teams participated in the VSCHH challenge independently and under the same conditions as other teams, following the same timeline and using the same CodaLab evaluation platform. However, their entries were not included in the final ranking for prizes and awards. These teams are referred to as Challenge organizing team-# and C#, e.g., C1 in Table 2.

Summary of benchmarks. During the VSCHH challenge, various neural rendering models, including TensoRF [3], MPFER [24], DINER [22], MIPNeRF [1], NeuS [28], and Nvdiffrec [20], were applied. TensoRF [3] was the most commonly chosen baseline among participants. In addition to the baselines, participants identified and reported various artifacts using the ILSH dataset. These included floaters, which are dark occlusions in front of the heads; texture issues, where texture mapping on surfaces appeared inconsistent or distorted; gridding, characterized by grid-like patterns in the rendered images; color shift, indicating subtle differences in color representation; geometric inconsistency, where object or surface shapes were distorted or misaligned; and blurriness, indicating poor image sharpness.

To address these challenges and improve the final output, participants introduced several novel ideas into their models. Common approaches include using a face mask, such as SAM [7], to exclude the image background, and employing a ray selection method, such as FreeNeRF [34] and NerfAcc [11], to help the proposed neural rendering model initiate the sampling process efficiently and reliably. Please refer to Fig. 4 for examples of the results submitted by all the finalists of the participants. For details of the methods proposed and used by all participants in the VSCHH challenge, please refer to the supplementary material.

Team achievements made by various ideas, independent of baseline choice. The rankings in Table 2 reflect the overall performance, including the selection of appropriate baseline methods and the incorporation of additional ideas to produce the final results. However, we also wanted to emphasize the unique ideas of each participant, which they independently applied to improve their results, starting from the scores generated by their chosen baseline method, as shown in Fig. 2. Fig. 2 shows that C1's baseline score was anchored at the best starting point, and also illustrates that C2's method achieved the most significant performance boost on top of their initial baseline score. In addition, we can confirm that the efforts to try additional ideas on top of the baseline score helped participants achieve a better ranking. However, choosing the best baseline still had a significant impact on achieving the first position.

In general, except for the first position, from the second (C2) to the seventh (T5) ranking, the effort put into improving their scores was successful, regardless of whether their initial baselines performed better or worse. Although all participants' improvements appear to be marginal, the

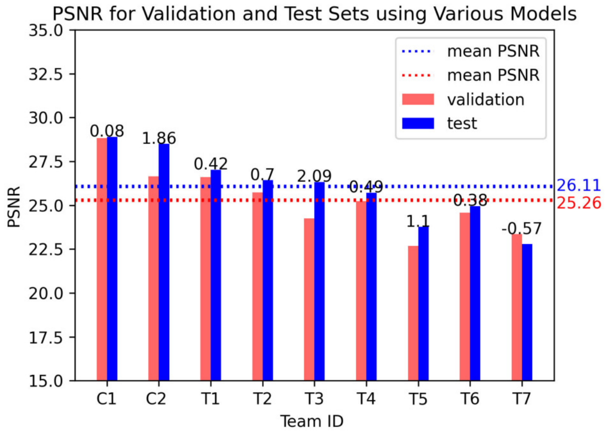

Figure 3: Score differences between the validation and the test sets from each team. Although each team used different baseline methods, the result scores tested on the validation and test sets remained consistent across the teams using different baselines. Note that the scores for T5, T6 and T7 are based on a subset, while the scores for all other teams are based on the full set. See text for more details.

VSCHH challenge successfully attracted a diverse range of approaches, including generic / scene-specific models, improved losses, leveraging geometric meshes, incorporating super-resolution, and background removal methods. Thus, we recommend referring to the supplementary material for details on the first set of approaches proposed and used by the initial group of participants who downloaded and experimented with the ILSH dataset through the VSCHH challenge.

Additional experiments for dataset fairness. As the ILSH dataset [37] emphasizes its careful design for fair evaluation using the chosen baseline method TensoRF [3], our challenge organizers double-checked to ensure that consistency was still maintained across participant teams using diverse methods other than the tested TensoRF. The final scores from all participants' results once again confirmed that the well-designed data split of the ILSH dataset supports fair comparison, demonstrating a close gap (0.85 PSNR gap) between validation and test scores across various methods independently, as shown in Fig. 3. However, compared to its original discussion in [37], our experiments, as shown in Fig. 3, reveal a larger gap than expected. We found that this is due to the fact that participants further developed their methods during the final Challenge Phase. Although no additional validation and test results were provided, participants were still able to improve and test potential improvements using the toy example, which includes two subjects along with full-camera viewpoint images.

Note that in Fig. 3, the scores for teams T5, T6 and T7 are based on a subset (the first three images) of the ILSH dataset, for both the validation and test sets. In contrast, the scores for all other teams are based on the full set of images from the validation and test sets. This is because teams T5, T6 and T7 only submitted results for the subset during the Development Phase (using the validation set). To ensure a fair comparison, we used the same number of subset to compare their results with those submitted for the test set during the Challenge Phase, even though the subjects and viewpoints being evaluated differ between the validation and test sets. Fig. 3 is intended to show consistent performance between the validation and test sets within each team, rather than to compare performance between different teams.

A. Teams and Affiliations

Challenge organizing team-1. MPFER-H: MPFER for Heads (Supp. Material Sec. 1)
Members. Thomas Tanay (thomas.tanay@huawei.com), Matteo Maggioni
Affiliations. Huawei Noah's Ark Lab

Challenge organizing team-2. DINER-SR (Supp. Material Sec. 2)
Members. Richard Shaw[1] (richard.shaw@huawei.com), Sibi Catley-Chandar[1,2]
Affiliations. [1]Huawei Noah's Ark Lab, [2]Queen Mary University of London

Team-1. OpenSpaceAI (Supp. Material Sec. 3)
Members. Ruijie Zhu, Jiahao Chang, Ziyang Song, Jiahuan Yu, Tianzhu Zhang (tzzhang@ustc.edu.cn)
Affiliations. University of Science and Technology of China

Team-2. NoNeRF (Supp. Material Sec. 4)
Members. Khanh-Binh Nguyen[1] (binhnk@skku.edu), Joon-Sung Yang[2]
Affiliations. [1]Sungkyunkwan University, [2]Yonsei University

Team-3. CogCoVi (Supp. Material Sec. 5)
Members. Andreea Dogaru (Andreea.Dogaru@fau.de), Bernhard Egger
Affiliations. Friedrich-Alexander-Universität Erlangen-Nürnberg

Team-4. CUBE (Supp. Material Sec. 6)
Members. Heng Yu, Aarush Gupta, Joel Julin, László A. Jeni (laszlojeni@cmu.edu)
Affiliations. Carnegie Mellon University

Team-5. Y-KIST-NeRF: Yonsei-KIST NeRF (Supp. Material Sec. 7)
Members. Hyeseong Kim[1,2] (hyeseongkim@yonsei.ac.kr), Jungbin Cho[1], Dosik Hwang[1], Deukhee Lee[2]
Affiliations. [1]Yonsei University, [2]Korea Institute of Science and Technology

Team-6. xoft (Supp. Material Sec. 8)
Members. Doyeon Kim[1] (doyooo.kim@lge.com), Dongseong Seo[2], SeungJin Jeon[3], YoungDon Choi[4]
Affiliations. [1]LG Electronics, [2]Seoul National University, [3]Dongguk University, [4]Korea Water Resources Corporation

Team-7. KHAG: KIST-Head Avatar Generator (Supp. Material Sec. 9)
Members. Jun Seok Kang[1], Ahmet Cagatay Seker[2], Sang Chul Ahn[2] (asc@kist.re.kr)
Affiliations. [1]University of Science and Technology, [2]Korea Institute of Science and Technology

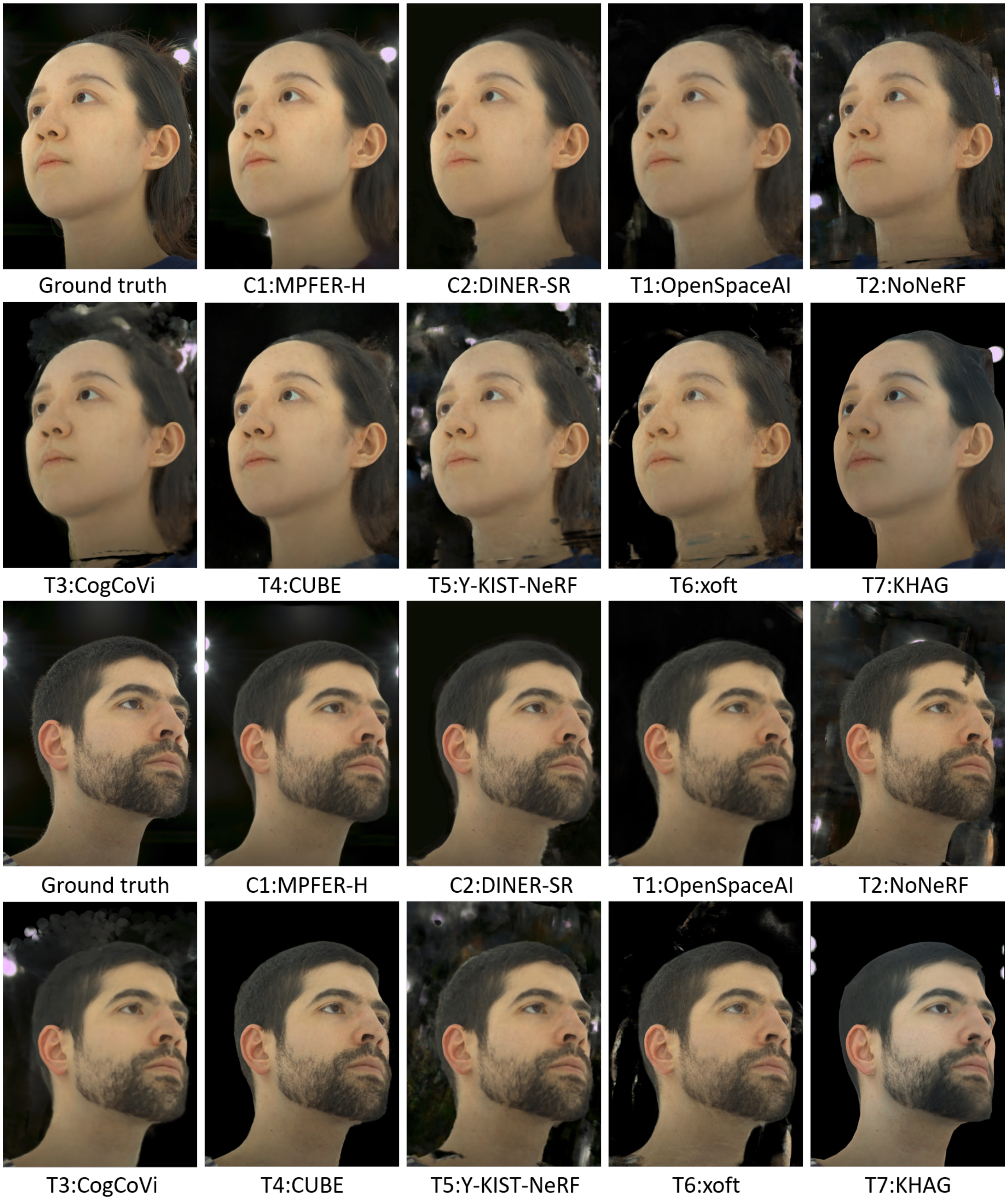

Figure 4: Example of synthesized test results produced by each team for the same test subject and viewpoint.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 1, 3, 4

[2] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, mar 2014. 1

[3] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 4, 5

[4] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020. 4

[5] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 1

[6] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4

[8] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), Jul 2023. 1

[9] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. Project starline: A high-fidelity telepresence system. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6), 2021. 1

[10] Verica Lazova, Vladimir Guzov, Kyle Olszewski, Sergey Tulyakov, and Gerard Pons-Moll. Control-nerf: Editable feature volumes for scene rendering and manipulation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4329–4339, 2023. 1

[11] Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: Efficient sampling accelerates nerfs. *arXiv preprint arXiv:2305.04966*, 2023. 4

[12] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[13] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4), jul 2019. 1

[14] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 1

[15] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*, 2022. 1

[16] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. *CVPR*, 2022. 1

[17] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[19] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 1

[20] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8280–8290, June 2022. 3, 4

[21] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 1

[22] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware Image-based NEural Radiance fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4

[23] Samarth Sinha, Jason Y. Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B. Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21349–21359, June 2023. 1

[24] Thomas Tanay, Aleš Leonardis, and Matteo Maggioni. Efficient view synthesis and 3d-based multi-frame denoising with multiplane feature representations. *CVPR*, 2023. 3, 4

[25] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. *arXiv*, 2022. 1

[26] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 4190–4200, June 2023. 1

[27] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, 2022. 1

[28] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 3, 4

[29] Peng Wang, Lingzhe Zhao, Ruijie Ma, and Peidong Liu. Bad-nerf: Bundle adjusted deblur neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4170–4179, June 2023. 1

[30] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5704–5713, June 2021. 1

[31] Dong Wang Bin Zhao Zhigang Wang Mulin Chen Bang Zhang Zhongjian Wang Liefeng Bo Xuelong Li Weichuang Li, Longhao Zhang. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2023. 1

[32] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022. 1

[33] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, October 2021. 1

[34] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[35] Christopher Choy Animashree Anandkumar Minsu Cho Yoonwoo Jeong, Seokjun Ahn and Jaesik Park. Self-calibrating neural radiance fields. In *ICCV*, 2021. 1

[36] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2987–2997, 2020. 1

[37] Jiali Zheng, Youngkyoon Jang, Athanasios Papaioannou, Christos Kampouris, Rolandos Alexandros Potamias, Foivos Paraperas Papantoniou, Efstathios Galanakis, Aleš Leonardis, and Stefanos Zafeiriou. ILSH: The imperial light-stage head dataset for human head view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2023. 1, 2, 5

[38] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 4

[39] Guojun Qi Zhen Lei Lei Zhang Zhiyuan Ma, Xiangyu Zhu. Otavatar : One-shot talking face avatar with controllable tri-plane rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2023. 1

[40] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, et al. FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction. *arXiv preprint arXiv:2111.01082*, 2021. 1