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, Jiaohao Chang, Ziyang Song, Jiahan 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.

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...
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 sub-datasets, 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.
<table>
<thead>
<tr>
<th>Evaluation Region</th>
<th>Full Region</th>
<th>Masked Region</th>
<th>Time (Sec.)</th>
<th>NVIDIA GPU</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
<td>PSNR</td>
<td>SSIM</td>
<td></td>
</tr>
<tr>
<td>C1:MPFER-H</td>
<td>28.05</td>
<td>0.84</td>
<td>28.90</td>
<td>0.83</td>
<td>1.50</td>
</tr>
<tr>
<td>C2:DINER-SR</td>
<td>22.37</td>
<td>0.72</td>
<td>28.50</td>
<td>0.83</td>
<td>87.25</td>
</tr>
<tr>
<td>*MPFER [24], CI</td>
<td>26.28</td>
<td>0.81</td>
<td>27.82</td>
<td>0.82</td>
<td>0.75</td>
</tr>
<tr>
<td>T1:OpenSpaceAI</td>
<td>21.66</td>
<td>0.68</td>
<td>27.02</td>
<td>0.83</td>
<td>76.88</td>
</tr>
<tr>
<td>T2:NeRF</td>
<td>20.37</td>
<td>0.69</td>
<td>26.43</td>
<td>0.82</td>
<td>175.58</td>
</tr>
<tr>
<td>T3:CogCoVi</td>
<td>21.49</td>
<td>0.70</td>
<td>26.33</td>
<td>0.82</td>
<td>806.00</td>
</tr>
<tr>
<td>*TensoRF [3], C0</td>
<td>20.54</td>
<td>0.71</td>
<td>26.17</td>
<td>0.82</td>
<td>94.02</td>
</tr>
<tr>
<td>T4:CUBE</td>
<td>21.07</td>
<td>0.66</td>
<td>25.72</td>
<td>0.81</td>
<td>95.00</td>
</tr>
<tr>
<td>T5:Y-KIST-NeRF</td>
<td>20.73</td>
<td>0.71</td>
<td>25.14</td>
<td>0.81</td>
<td>15.10</td>
</tr>
<tr>
<td>*TensoRF [3], T6</td>
<td>20.09</td>
<td>0.65</td>
<td>25.30</td>
<td>0.81</td>
<td>758.00</td>
</tr>
<tr>
<td>T6:xoft</td>
<td>20.01</td>
<td>0.64</td>
<td>25.02</td>
<td>0.80</td>
<td>727.00</td>
</tr>
<tr>
<td>*TensoRF [3], T1</td>
<td>20.28</td>
<td>0.70</td>
<td>24.70</td>
<td>0.81</td>
<td>31.13</td>
</tr>
<tr>
<td>*TensoRF [3], T2</td>
<td>20.13</td>
<td>0.70</td>
<td>24.37</td>
<td>0.81</td>
<td>72.55</td>
</tr>
<tr>
<td>*NeuS [28], T3</td>
<td>21.02</td>
<td>0.72</td>
<td>24.33</td>
<td>0.80</td>
<td>944.00</td>
</tr>
<tr>
<td>*Mip-NeRF360 [1], T4</td>
<td>20.59</td>
<td>0.71</td>
<td>24.13</td>
<td>0.80</td>
<td>78.00</td>
</tr>
<tr>
<td>*TensoRF [3], T5</td>
<td>19.87</td>
<td>0.69</td>
<td>24.07</td>
<td>0.81</td>
<td>14.50</td>
</tr>
<tr>
<td>T7:KHAG</td>
<td>22.14</td>
<td>0.64</td>
<td>23.39</td>
<td>0.79</td>
<td>2.58</td>
</tr>
<tr>
<td>*NvidiaFrec [20], T7</td>
<td>20.03</td>
<td>0.62</td>
<td>22.96</td>
<td>0.78</td>
<td>0.22</td>
</tr>
<tr>
<td>*DINER [22], C2</td>
<td>14.81</td>
<td>0.58</td>
<td>22.72</td>
<td>0.78</td>
<td>86.37</td>
</tr>
</tbody>
</table>

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 challenging 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 = ilff, downsample_train = 1.0, ndc_ray = 0, n_iters = 50000, n_lamb_sigma = [16,4,4], n_lamb_sh = [48,12,12], shading_Mode = 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
Having discussed external and not released, as previously evaluated the result within the face region using face masks, there are two groups of results calculated: one group results over all processed images, as shown in Table 2. The rankings in Table 2 reflect the official resolution of the input image and saved it as a reference. 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.

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 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 Table 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 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 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 downsampling the dataset to a resolution of 300 pixels in width to detect the upsampled 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 upsampled 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 Nvdiffr [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.
VSCCHH 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 VSCCHH 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 Shaw¹ (richard.shaw@huawei.com), Sibi Catley-Chandar¹,²

Affiliations. ¹Huawei Noah’s Ark Lab, ²Queen Mary University of London

Team-1. OpenSpaceAI (Supp. Material Sec. 3)

Members. Ruijie Zhu, Jiahao Chang, Ziyang Song, Jiahuan Yu, Tianzhao Zhang (tzzhang@ustc.edu.cn)

Affiliations. University of Science and Technology of China

Team-2. NoNeRF (Supp. Material Sec. 4)

Members. Khanh-Binh Nguyen¹ (binhnhk@skku.edu), Joon-Sung Yang²

Affiliations. ¹Sungkyunkwan University, ²Yonsei 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 Kim¹,² (hyeseongkim@yonsei.ac.kr), Jungbin Cho³, Dosek Hwang¹, Deukhee Lee²

Affiliations. ¹Yonsei University, ²Korea Institute of Science and Technology

Team-6. xoft (Supp. Material Sec. 8)

Members. Doyeon Kim¹ (doyooo.kim@lge.com), Dongseong Seo², SeungJin Jeon³, YoungDon Choi⁴

Affiliations. ¹LG Electronics, ²Seoul National University, ³Dongguk University, ⁴Korea Water Resources Corporation

Team-7. KHAG: KIST-Head Avatar Generator (Supp. Material Sec. 9)

Members. Jun Seok Kang¹, Ahmet Cagatay Seker², Sang Chul Ahn² (asc@kist.re.kr)

Affiliations. ¹University of Science and Technology, ²Korea Institute of Science and Technology

1125
Figure 4: Example of synthesized test results produced by each team for the same test subject and viewpoint.
References


