Supplementary Material for "FaceMap: Distortion-Driven Perceptual Facial Saliency Maps"

Fig. 13. The relation between JOD value and probability of selection. If a stimulus is 1 JOD away from the reference, this can be interpreted as the reference being chosen 75% of the time in a 2AFC task in our experiment.

A APPLICATIONS

A.1 Re-meshing Validation Study Details

Our re-mesh application and user study are described in Sec. 6.1. Pre-rendered images were used as stimuli, shown with a slight rotation off-center. As in the main study, a timer was present but not enforced. The labeled reference face was always shown on the right side of the screen, and test faces on the left.

After a brief training session allowing the participants to familiarize themselves with a number of conditions, including all rendering modes and face models, participants were tasked with rating the test face in terms of quality on a scale of 1-10. Qualitative descriptions were placed under the scale ("bad", "poor", "fair", "good", "excellent"), as shown by Madhusudana et al. [2021]. The study consisted of 4 unique face models, each of which was rendered at 6 different mesh resolutions (30, 20, 10, 8.3, 6.6, and 5% or the original), and using one of 3 resampling allocation methods (FaceMap, uniform, or the automatic saliency map of Song et al. [2014]). As FaceMap is only defined in the face region, the neck, ears, and hair of the model were rendered identically (as in the reference), and only the face area had modifications. Each mesh was presented 3 times throughout the experiment. Including 12 training trials, the total number of trials was: $4 \times 6 \times 3 \times 3 + 12 = 228$. Study duration averaged 40 minutes.

Figure 19 shows an example of the automatic saliency map and the process of creating adaptive UV. Figures 16 and 17 show all models used in this study. Note that as the overall density of the mesh increases, the quality of faces improves independently of the re-meshing method used, which reduces the difference between the two rendering modes.

A.2 Gaussian Splatting Validation Study Details

Our 3DGS application and validation are described in Sec. 6.2. Pre-rendered videos were used as stimuli, showing a rotating face, similar to the main experiment, but interactive rotation was not present. As in the main study, a timer was present but not enforced.

Participants were tasked with answering which face has better quality - left or right in the 2AFC task. The study consisted of 10 unique face models, each of which was rendered at 5 different allocation densities. As FaceMap is only defined in the face region, the neck, ears, and hair of the model were rendered identically (uniform allocation), and only the face area was modified between allocation conditions. Each pair was presented twice - uniform on the left and FaceMap on the right and vice-versa. The total number of trials was: $10 \times 5 \times 2 = 100$, presented in random order. The study lasted for an average of 34 minutes.

Figure 14 shows all models used in this study. Note that as the overall number of Gaussian increases, the quality of faces improves independently of allocation, which reduces the difference between the two rendering modes. As shown in Fig. 12, FaceMap is strongly preferred for low-bandwidth models (98.6% preference for $1k$, 91.8% preference for $4k$, 75.4% preference for $16k$), but this effect is reduced for the higher quality renders (54.1% preference for $65k$, 57.7% preference for $262k$). An additional qualitative comparison using an initialization derived from [Song et al. 2014] on a template mesh (as described in Supplementary A.1) is provided in Figure 15.

A.3 Validation Conclusions

As shown in Figs. 8 and 12, FaceMap is strongly preferred for low-bandwidth models. As the model quality increases, the gain in quality due to FaceMap is less obvious for both applications. This leads us to conclude that perceptual priors like FaceMap are especially important to improve visual quality of bandwidth and memory limited applications, such as mobile rendering.

A.4 Texture Compression

Similarly to our geometry simplification application (Sec. 6.1), the texture of a face model can also be compressed by employing a quad-tree-based image compression to efficiently reduce the memory requirement for storage [Shusterman and Feder 1994] or to speed up machine learning algorithms [Jewsbury et al. 2021]. A quad-tree is built from a given image patch recursively by analyzing the detail metric on the patch and subdividing it by four if a threshold is not met until it reaches a maximum depth specified by the user.

Our facial saliency map can be integrated to guide this compression. The texture image is weighted using the *overall* saliency map. As a result, the computed detail metric will be low for non-salient regions, prompting fewer subdivisions in those areas. We reconstruct the original image using a built tree, where each tree leaf is replaced with a square of the mean color. A result of this method is shown in Figure 18, contrasted against a mesh obtained via the default detail metric based on a weighted histogram difference. Note that in this example, our map is applied on a mesh with different connectivity and UV map, and we use the same set of landmark locations to define and interpolate the saliency values.

B RANDOMIZED LANDMARK VALIDATIONS

To study the impact of our choice of landmark locations (Section 3.2) and interpolation scheme used to obtain a continuous importance

map of the face (Section 6), we designed an additional validation study with randomized landmark locations.

We randomly selected 8 additional landmark locations by placing a box on the UV map of the face as seen from a frontal angle, and performed random sampling of 2D points (seed=6, shown in Figure 20, left). Next, the artifact generation algorithm was ran as described for our main study (Section 3.3) on the same three base face models. We then proceeded to repeat our main study's steps for this new randomized dataset following the same protocols and using the same hardware for an additional 10 participants.

The same scaling procedure was employed to compute JOD scores for the new randomized landmarks (no outliers were detected). These newly measured values were then compared against the interpolated result from the original model at the same locations (shown in Figure 20, right). Datapoints showed good consistency, with root mean squared error between the subjective and interpolated data points at 0.242 JOD, comparable to the 0.209 JOD standard deviation of the bootstraped baseline data (see Figure 22). Further, the two groups of values present a Pearson Linear Correlation Coefficient of 0.83 ($p \ll 0.05$), and a Spearman's Rank Correlation Coefficient of 0.74 ($p \ll 0.05$). These values can be interpreted as a strong correlation [Schober et al. 2018], demonstrating that our original semantic anchor choice and interpolation scheme does not significantly distort the results for other points on the face.

C CORRELATION ANALYSIS WITH AUTOMATIC METHODS

Many automatic methods to compute saliency exist, some of which target 3D models and could be applied to meshes representing faces, like the ones used in this study. We set out to evaluate the accuracy of automatic saliency estimators and metrics in predicting the importance map obtained from our experimental data.

We selected two representative methods for analysis: the no-reference saliency estimator of Song et al. [2014], and the reference-based textured mesh Graphics-LPIPS distance metric developed by Nehmé et al. [2023] (both discussed in Section 2).

The saliency map for the method of Song et al. [2014] was obtained as detailed in Section 6 and shown in Figure 19.

The method of Nehmé et al. [2023] targets the perception of 3D models, but requires 2D images as input. To accommodate this requirement, we employed a setup mimicking that described by the authors, including directional lights coming from the top right, and choose 3 relevant views to render images at a resolution of 650×550. For each of the distorted meshes in our study (see Section 3.3), we produced 3 renderings as seen from views directly in front, 45° to the left and 45° to the right. We then decompose the rendering into overlapping 64×64 patches, following the paper. Using the pre-trained model, we computed the averaged predicted perceptual loss with respect to the reference image of the same model, generating values for each of the conditions measured in our experiment, then average the values for corresponding subject and level.

Finally, we analyze the predictive power of the obtained results for both methods via correlation analysis (illustrated in Figure 21). The saliency map of Song et al. [2014] obtains a SROCC of 0.306 ($p \ll 0.05$) and a PLCC of 0.234 ($p = 0.0225$). The metric evaluations of Nehmé et al. [2023] result in a SROCC of 0.190 ($p \ll 0.05$) and a PLCC of 0.234 ($p \ll 0.01$). These results can be classified as having weak correlation with the study data [Schober et al. 2018], indicating they are not capable of accurately predicting facial importance. This is not surprising, as neither method models the unique aspects of perception of human faces, as done in our study.

D N-WAY ANOVA ANALYSIS

Table 1 shows the p-values associated with condition of our studies.

Main Study	p-value
distortion strength	1.8×10^{-11}
distortion type	7.3×10^{-8}
distortion location	3.3×10^{-4}
strength:type	0.97
strength:location	0.13
type:location	0.06
strength:type:location	0.99

Gaussian Study	p-value
method	1.5×10^{-21}
model	0.24
participant	3.8×10^{-3}
method:model	0.08
method:participant	0.92
model:participant	0.79
method:model:participant	0.43

Re-meshing Study	p-value
method	7.1×10^{-65}
model	0.3
participant	3.3×10^{-6}
level	9.6×10^{-206}
method:model	0.031
method:participant	0.1
method:level	0.71
model:participant	0.8
model:level	0.25
level:participant	0.0058
method:model:participant	0.71
method:model:level	0.55
method:participant:level	0.29
model:participant:level	0.62
method:model:participant:level	0.074

Table 1. N-way ANOVA results for the Main Study (Section 5, top), Gaussian Splatting application (Section 6.2, mid), and Re-Meshing application (Section 6.1, bot).

Fig. 14. This image shows all the stimuli presented in the Gaussian Splatting validation study, described in Sec 6.2. Rows show different base faces, and columns show conditions with an increasing number of Gaussians from left to right. Each face is split vertically showing the FaceMap allocation on the left, and uniform allocation on the right. Please zoom in to see details.

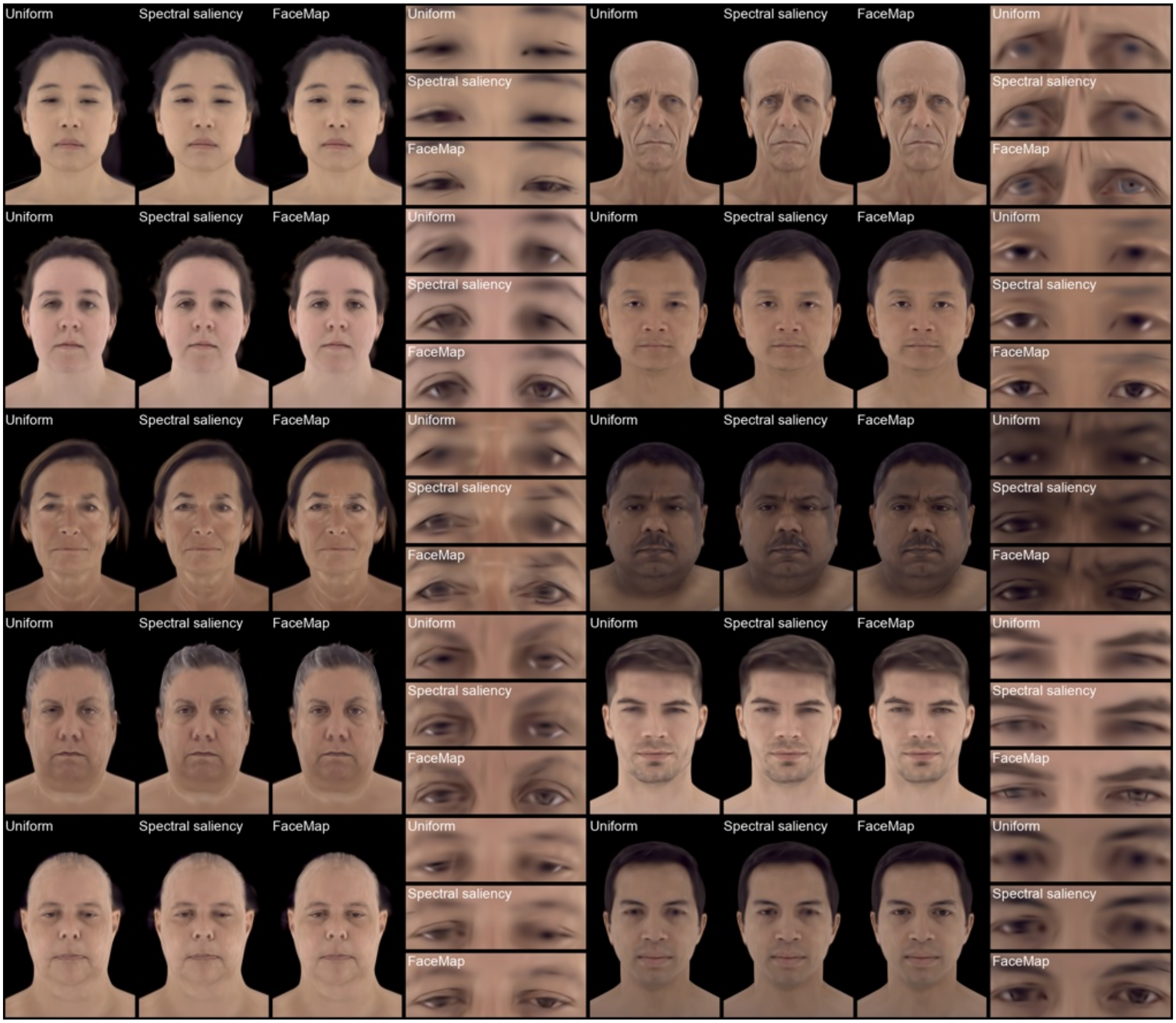

Fig. 15. A qualitative comparison for Gaussian Splatting with 1K primitives. Initialization with uniform weight, a spectral saliency weight, and our facemap weight is shown (as described in Supplementary 6.2). Note that Facemap's results consistently allocate higher quality for the eye and mouth regions, which were found to have greater perceptual importance in our study. Please zoom in to see the details.

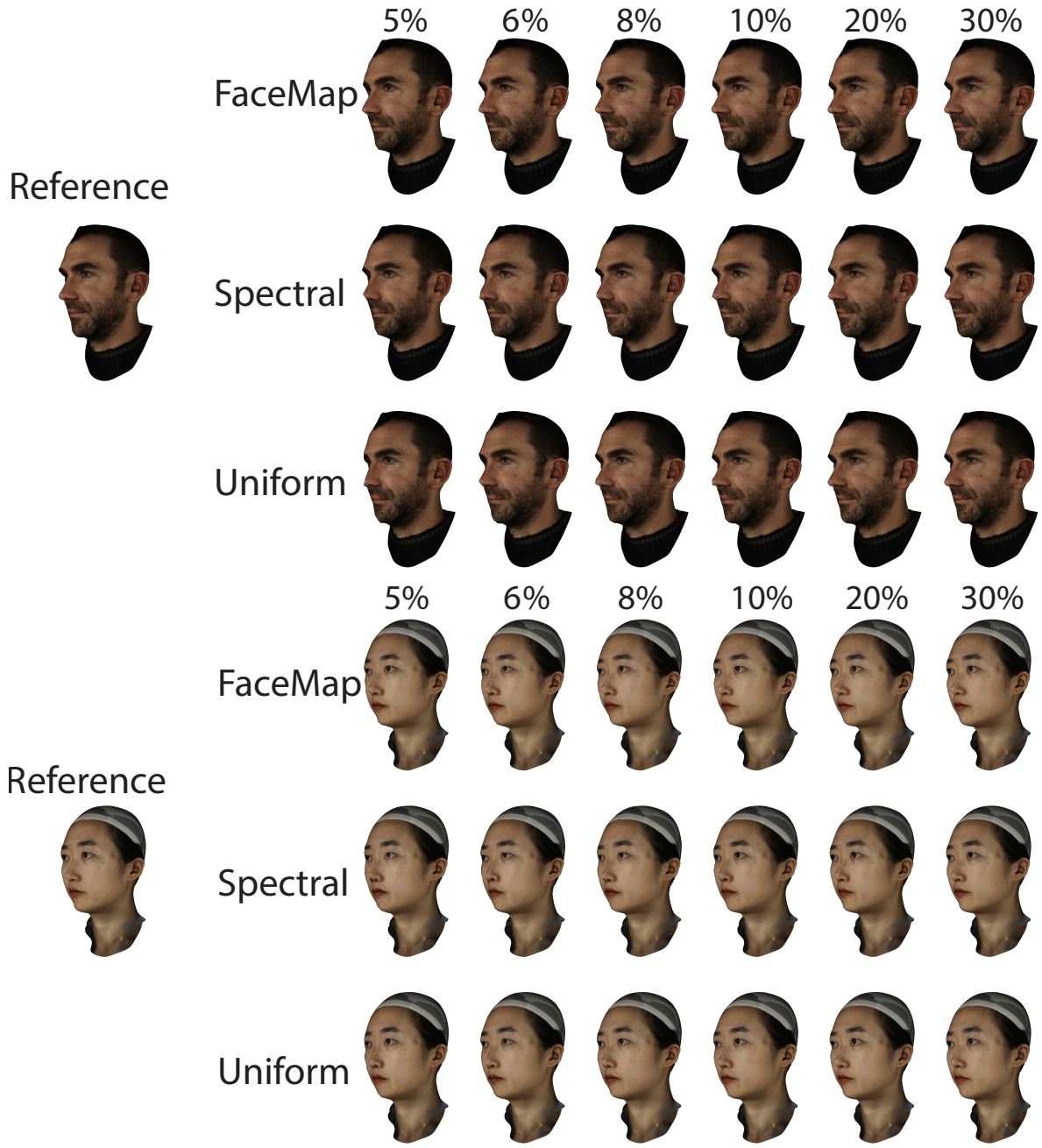

Fig. 16. This image shows all the levels of compression used in the geometry compression validation study (Sec. 6.1) for identities 1 and 2. Please zoom in to see the details.

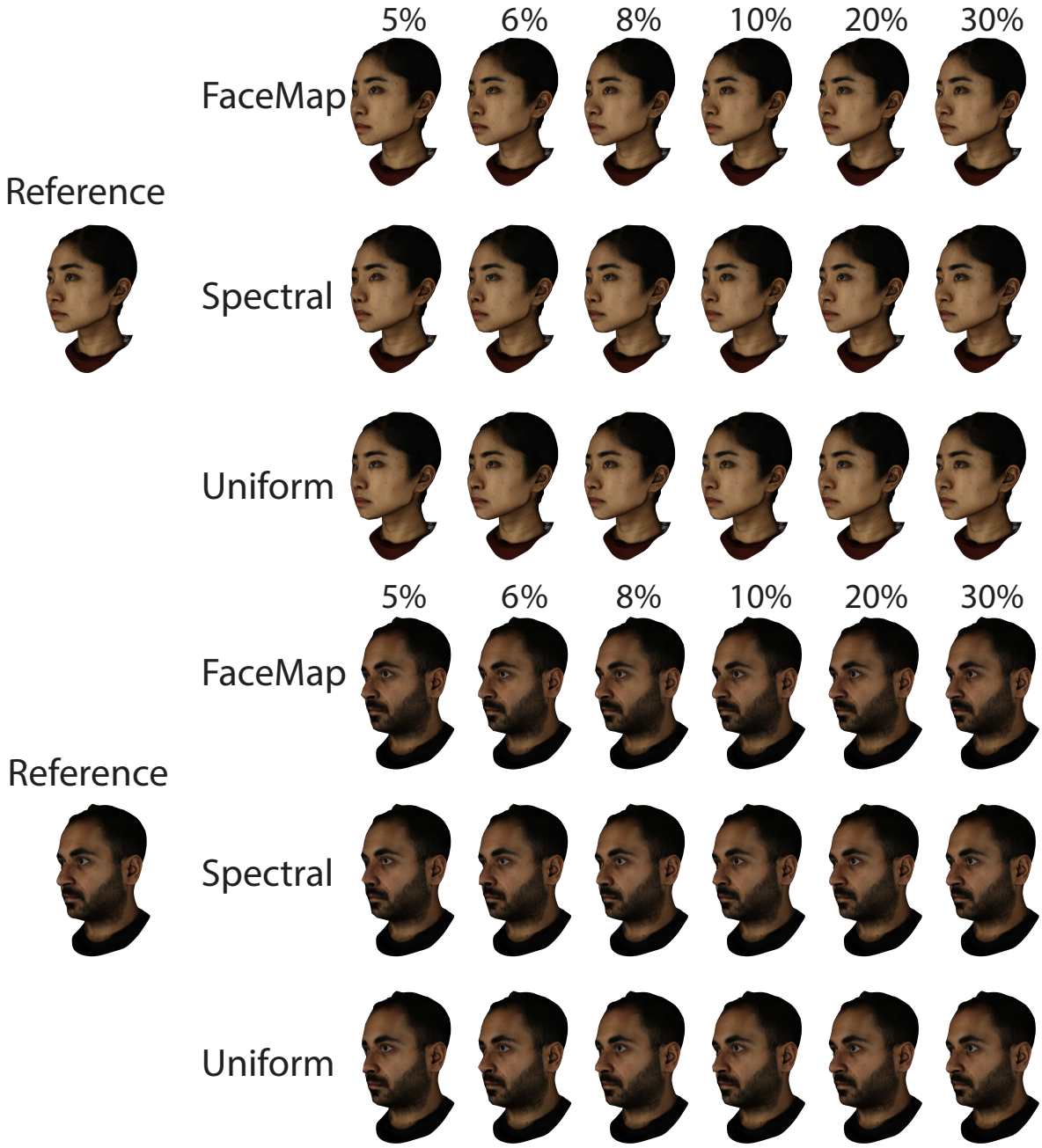

Fig. 17. This image shows all the levels of compression used in the geometry compression validation study (Sec. 6.1) for identities 3 and 4. Please zoom in to see the details.

Fig. 18. We show texture image compression with a quad-tree-based approach. (a) Original mesh and texture. (b) using the default metric, we obtain a texture with 11098 color nodes. (c) using a saliency-guided metric, we obtain texture with 11290 color nodes, with a higher quality in the eye and mouth regions. Note that our method generalizes well despite a different topology and texture map for the base asset.

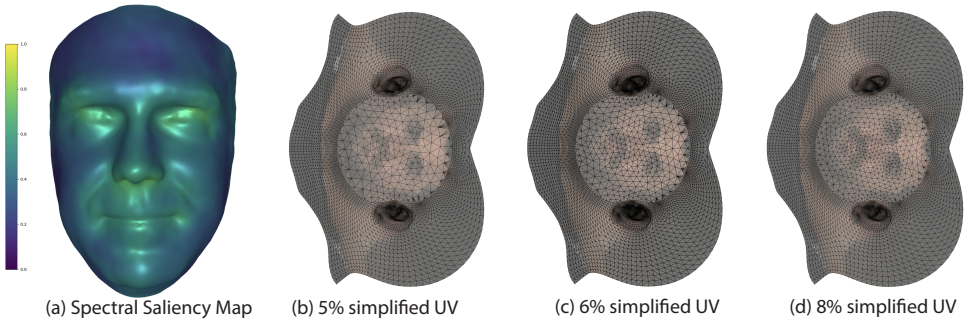

Fig. 19. (a) shows the spectral saliency map from the method of [Song et al. 2014]. The map place emphasize on features with varying curvature, including eyes and region between nose and mouth. (b,c,d) And the UV domain of 3 different simplification levels (5%, 6.6%, 8.3%) are shown.

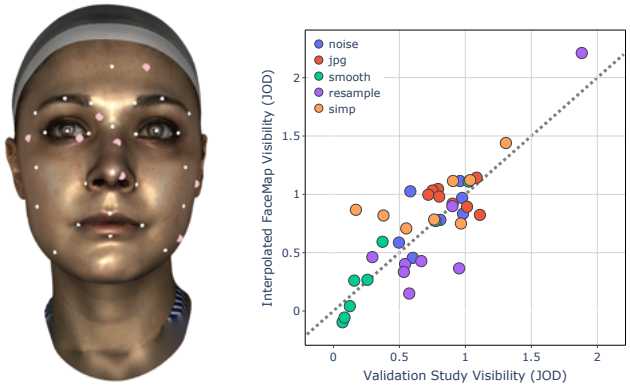

Fig. 20. (Left) In pink, the randomized landmarks used in the validation study described in Section B. White dots show the original landmarks from the main study for reference. (Right) A comparison of the subjective scores of the randomized landmarks along the x-axis (see Section B) against the interpolated results for the same points from the main study along the y-axis. The dashed line represents identity.

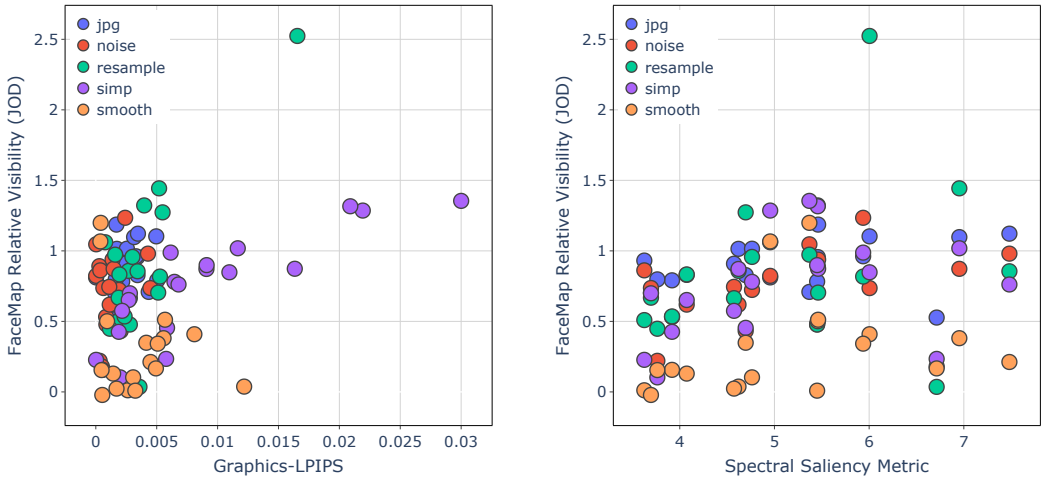

Fig. 21. This figure illustrates the relationship between our study's results (y-axes) and two automatic methods (x-axes), as described in Section C: Nehmé et al. [2023] (Top) and Song et al. [2014] (Bottom).

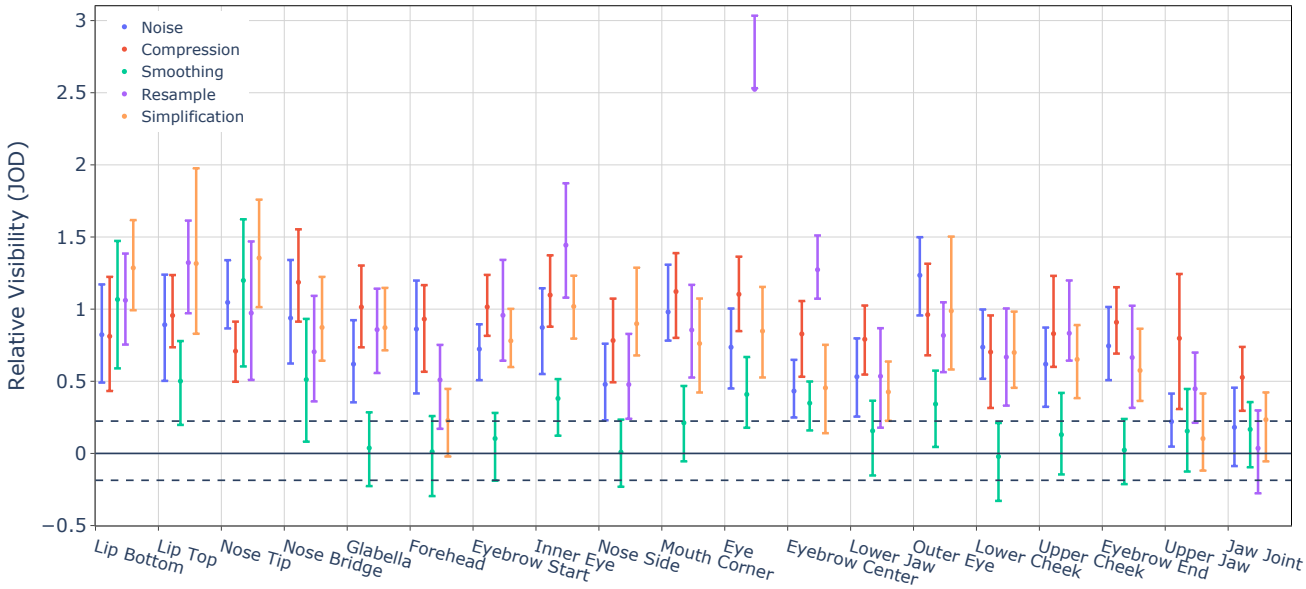

Fig. 22. This figure shows the results of our main study across all studied locations on the face and different artifact types, aggregated across 2 levels. Vertical lines represent 95% confidence intervals. The dashed lines represents the 95% confidence interval for the reference undistorted model.