HybridMQA: Exploring Geometry-Texture Interactions for Colored Mesh Quality Assessment

Supplementary Material

1. Experimental Setup Details

1.1. Details of Datasets

To validate the performance of our proposed method, we conduct experiments on four publicly available color MQA datasets: Nehmé *et al*. [6], SJTU-TMQA [1], TSMD [9], and CMDM [7]. The Nehmé *et al*. dataset is the largest public dataset of 3D textured meshes, containing 55 source meshes distorted by a mixture of geometric and color distortions to obtain 3000 distorted meshes. The SJTU-TMQA dataset consists of 21 reference and 945 distorted textured meshes. Distorted meshes were generated through geometric or color distortions or a combination of both. The TSMD dataset includes 39 source 3D textured meshes (excluding 3 source meshes as they were not publicly available: "Mitch", "Nathalie", and "Thomas"), each distorted at five levels with a combination of geometric and color distortions, resulting in a total of 195 distorted meshes. Finally, the CMDM dataset consists of vertex-color meshes, with 5 source meshes each subjected to geometric or color distortions, resulting in 80 distorted meshes. Mean opinion scores (MOS) were computed and reported as ground truth quality labels for all distorted models across the four datasets, based on subjective evaluations from 4513, 73, 74, and 72 study participants, respectively. In total, the four datasets encompass a wide variety and strength levels of geometric and color distortions. We note that the TSMD and SJTU-TMQA datasets have overlapping source meshes which were excluded from the training set (TSMD dataset) in our generalization test.

1.2. Implementation Details

We use Adam optimizer [3] with the default $1e^{-5}$ weight decay and $1e^{-4}$ initial learning rate that is gradually reduced to $1e^{-5}$ with cosine annealing scheduler [4]. The default batch size is set to 8, and the model is trained for 15 epochs by default. The loss balance term λ is set to 1. During training and testing on the CMDM dataset, we skip the base encoder and directly initialize the feature graph with raw vertex color, normal, and position values as vertex-color meshes lack 2D texture maps and UV mapping data. To allow for faster training and larger batch sizes given the limitations of our GPU (NVIDIA V100 32GB), we implement viewpoint dropout, where we randomly select two out of six camera viewpoints in each training iteration and only render those two projections.

Data Augmentation. We use camera angle augmentation in training to enhance the model's robustness and generalization capabilities. Specifically, we set the original azimuth and elevation angles as the mean of a normal distribution with a standard deviation of 22.5° and sample new azimuth and elevation angles in each training iteration. We also employ flip augmentation on patches extracted from 3D feature and colored projections.

1.3. Details of Evaluation Metrics

To compare the performance of different MQA methods, we employ two mainstream evaluation criteria: the Spearman rank-order correlation coefficient (SRCC) and the Pearson linear correlation coefficient (PLCC). SRCC measures prediction monotonicity, while PLCC evaluates prediction accuracy [2]. The PLCC score is calculated by using a logistic non-linear fitting method to align the predicted scores with the ground truth scale [2]. Higher SRCC and PLCC absolute values signal a higher correlation between MOS and predicted quality scores and hence a better performance.

2. Further Ablation Studies

We perform additional ablation experiments on Nehmé *et al*. dataset [6].

2.1. Cross-attention Mechanism

We perform further ablation studies to highlight the impact of the cross-attention mechanism. Specifically, given the encoded 3D surface representation $\boldsymbol{f}_m$ and the textural representation $\boldsymbol{f}_t$, we replace the proposed cross-attention mechanism with: (1) addition; (2) weighted addition of $\boldsymbol{f}_m$ and $\boldsymbol{f}_t$, where we learn the weights using a convolutional block that takes the two representations as input; (3) concatenation; (4) elementwise multiplication; and (5) self-attention of $\boldsymbol{f}_m$ and $\boldsymbol{f}_t$ followed by concatenation. Table 1 presents the results. We can observe that all replacements result in significant drops in performance. This highlights the effectiveness of the proposed cross-attention mechanism in capturing interactions between 3D geometry and textural representations of the mesh, emphasizing the importance of these texture-geometry interactions for achieving accurate MQA.

2.2. Data Augmentations

We also conduct experiments to measure the importance of camera angle and flip augmentations in the method's performance. Table 2 presents the results of excluding each of

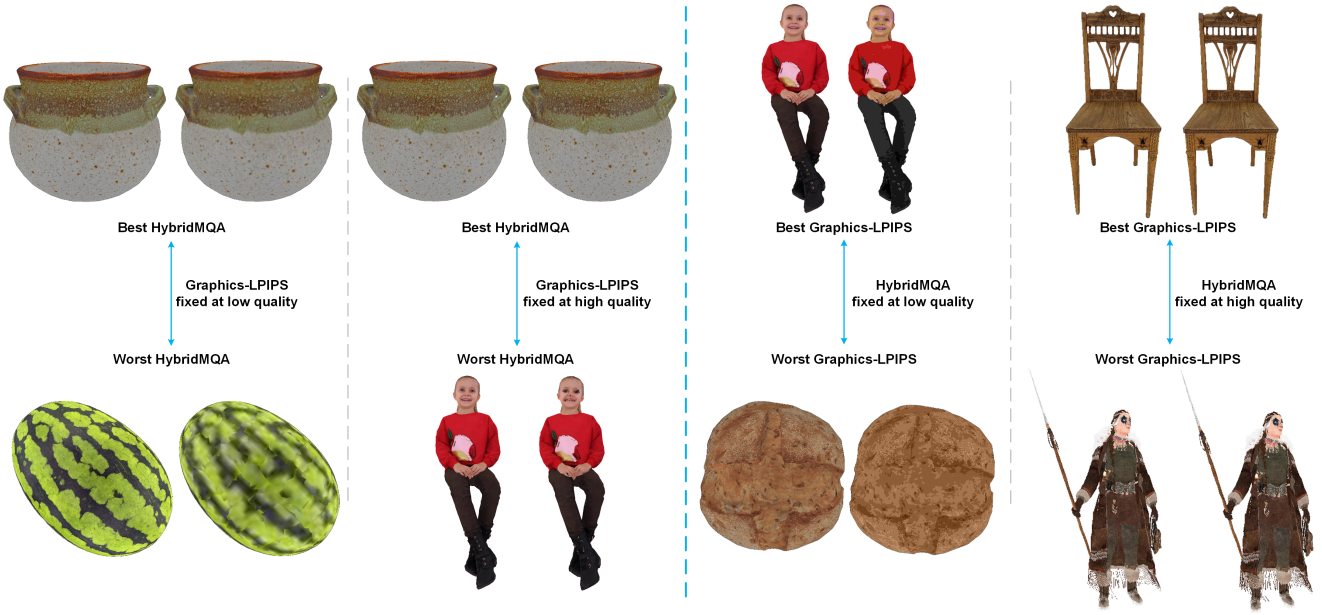

Figure 1. HybridMQA clearly outperforms Graphics-LPIPS [6] in gMAD competition [5]. Columns one and two showcase results with Graphics-LPIPS fixed at low and high quality, respectively, while columns three and four display results with HybridMQA fixed at low and high quality. In each column, the left objects are the references, while the right ones are the distorted meshes. The most perceptually important viewpoint of each object is selected for visualization.

Configurations	SRCC	PLCC
addition: $\boldsymbol{f}_m + \boldsymbol{f}_t$	0.842	0.842
weighted addition: $\boldsymbol{f}_m + \boldsymbol{w} \odot \boldsymbol{f}_t$	0.846	0.861
concat.: $\boldsymbol{f}_m \oplus \boldsymbol{f}_t$	0.845	0.857
multiplication: $\boldsymbol{f}_m \odot \boldsymbol{f}_t$	0.848	0.849
self-att. + concat.: $\mathrm{SA}(\boldsymbol{f}_m) \oplus \mathrm{SA}(\boldsymbol{f}_t)$	0.852	0.857
cross-attention (proposed)	**0.892**	**0.897**

Table 1. Ablation on cross-attention mechanism on Nehmé *et al*.

the two data augmentations. We observe that both data augmentations improve performance, with camera angle augmentation having a more pronounced effect.

Angle Aug.	Flip Aug.	SRCC	PLCC
✓	–	0.876	0.883
–	✓	0.857	0.857
✓	✓	**0.892**	**0.897**

Table 2. Ablation on data augmentations on Nehmé *et al*.

2.3. Viewpoint Dropout & Batch Size

We conduct further experiments to evaluate different configurations of viewpoint dropout and batch size, as introduced in Sec. 1.2. Specifically, we evaluate three configurations: randomly selecting two or four viewpoints in each training iteration or using all six viewpoints (no dropout). These configurations are tested across batch sizes of 2, 4, and 8. We note that the largest possible batch size varies depending on the number of viewpoints: 8 for two viewpoints, 4 for four viewpoints, and 2 for six viewpoints. Table 3 presents the results. We can see that performance improves as the batch size increases for each viewpoint configuration. Notably, the best performance is achieved with two viewpoints, which allows for a batch size of 8—the largest among the tested configurations. This demonstrates the effectiveness of the viewpoint dropout mechanism.

$N_v \backslash N_b$	2	4	8
2 Views	0.837/0.844	0.864/0.873	**0.892/0.897**
4 Views	0.859/0.867	0.866/0.873	OOM
6 Views	0.838/0.846	OOM	OOM

Table 3. SRCC/PLCC results of the ablation on the number of viewpoints and batch sizes in training on Nehmé *et al*. N_v and N_b denote the number of viewpoints and batch size, respectively. OOM stands for out of memory.

3. Further Qualitative Results

3.1. gMAD Competition

We perform gMAD competition [5] to qualitatively compare the performance of HybridMQA with Graphics-LPIPS

[6]. gMAD competition identifies 3D meshes that one method estimates to be of similar quality, while the other method rates them as having significantly different quality. Through this competition, at least one of the methods will be discredited due to producing quality judgments that do not correlate with human opinions. We perform the gMAD competition on the SJTU-TMQA dataset [1], where we gather quality judgments of the two methods on all validation sets of the 5-fold cross-validation test.

Figure 1 presents the results of the competition, where HybridMQA clearly outperforms Graphics-LPIPS. As we can see, Graphics-LPIPS judges the 3D meshes in the first column (*pottery vessel* and *watermelon*) to be of similarly low quality. This is clearly in contradiction with human judgments as well as HybridMQA predictions. The second column shows a similar trend: HybridMQA predictions align with human judgments, while Graphics-LPIPS incorrectly rates the *girl* 3D mesh as having high quality. We then switch the roles of the two methods in the third and fourth columns. In column three, Graphics-LPIPS assigns higher quality prediction to the *girl* compared to the *bread*. However, both 3D meshes are severely contaminated by JPEG compression [1] and judged by human viewers to be of similarly low quality. HybridMQA successfully rates the two meshes as having poor perceptual quality. Similar conclusions can be made in the fourth column, where HybridMQA accurately assigns high quality scores to both meshes. These results demonstrate the clear superiority of HybridMQA over Graphics-LPIPS in colored MQA.

3.2. GradCAM on meshes

Figures 2 and 3 provide additional examples of GradCAM [8] applied to graph features in the model branch. The highlighted regions successfully identify noticeable geometrical artifacts that align well with human perception. This showcases the model branch's effectiveness in capturing geometry-aware quality representations.

3.3. GradCAM on Cross-attention

Figure 4 provides additional examples of GradCAM [8] applied before and after cross-attention. The two branches concentrate on distinct regions, with the model branch emphasizing geometric artifacts. Through cross-attention, the framework effectively identifies and focuses on perceptually important regions by exploring interactions between geometry and texture. This demonstrates the effectiveness of our hybrid method in exploiting interactions between representations learned in texture and model branches.

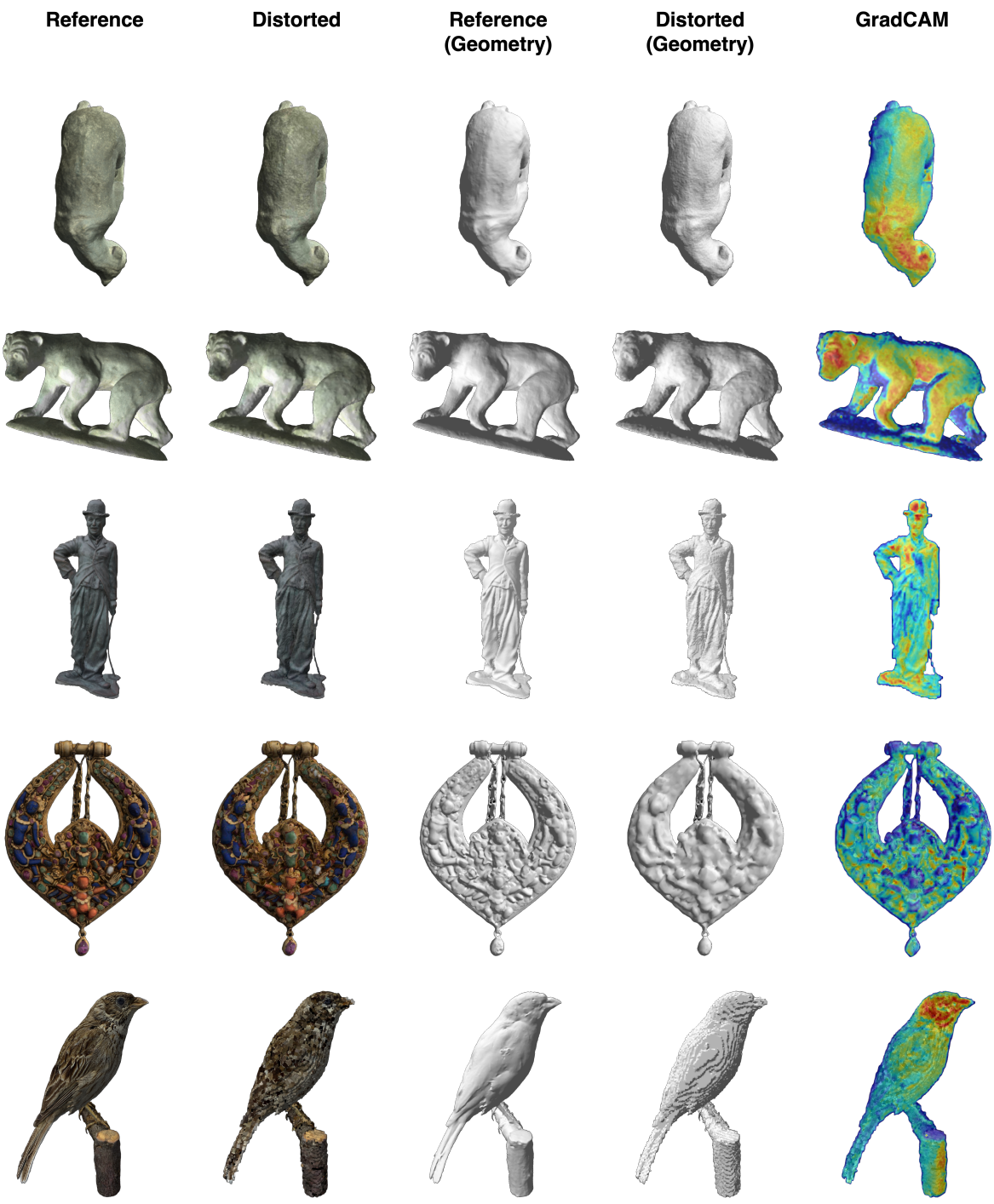

Figure 2. More GradCAM [8] results on meshes.

Figure 3. More GradCAM [8] results on meshes.

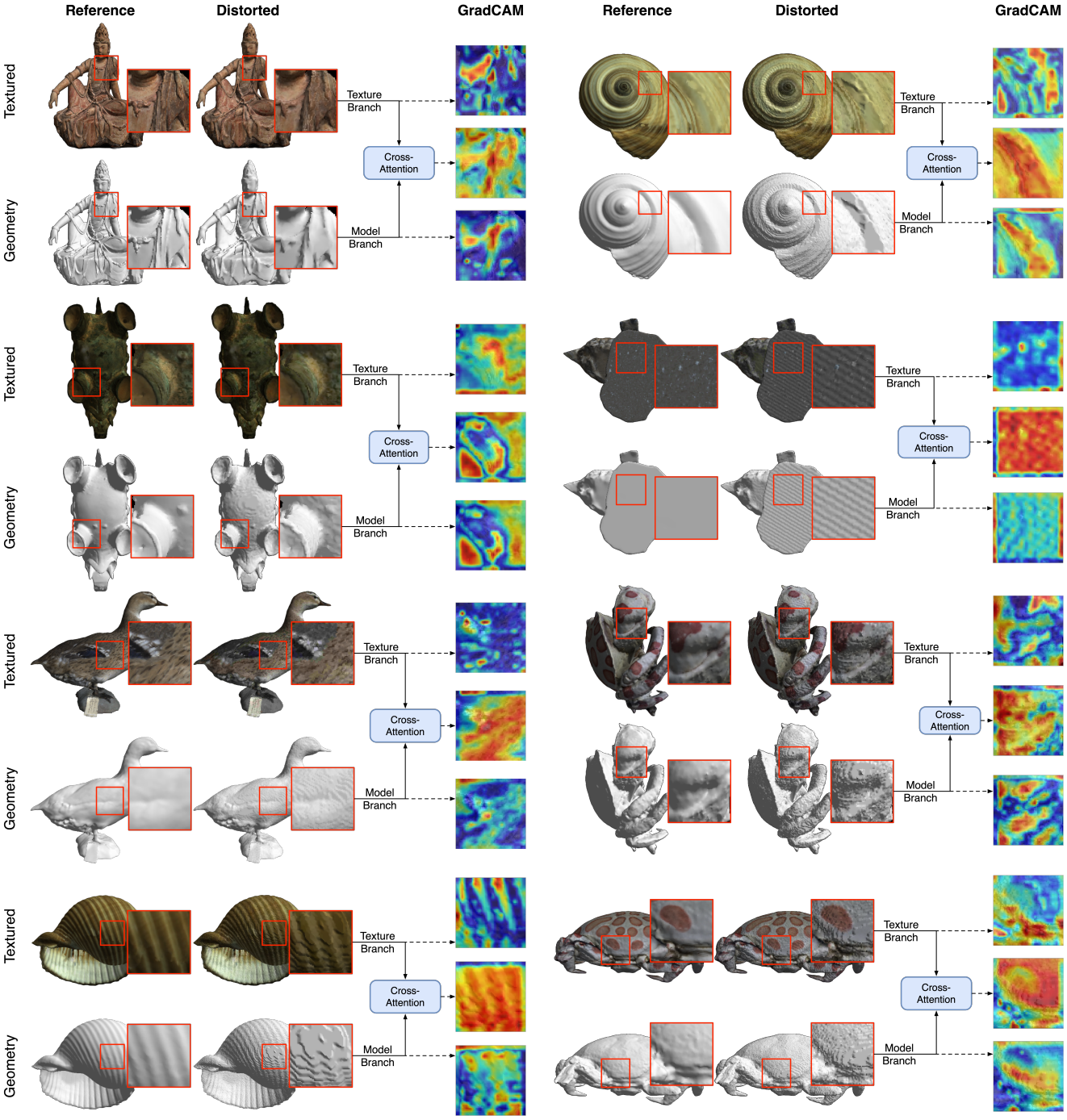

Figure 4. More GradCAM [8] results on cross-attention.

References

[1] Bingyang Cui, Qi Yang, Kaifa Yang, Yiling Xu, Xiaozhong Xu, and Shan Liu. Sjtu-tmqa: A quality assessment database for static mesh with texture map. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7875–7879, 2024. 1, 3

[2] Video Quality Experts Group et al. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II. *VQEG*, 2003. 1

[3] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 1

[4] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[5] Kede Ma, Zhengfang Duanmu, Zhou Wang, Qingbo Wu, Wentao Liu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):851–864, 2020. 2

[6] Yana Nehmé, Johanna Delanoy, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. Textured mesh quality assessment: Large-scale dataset and deep learning-based quality metric. *ACM Trans. Graph.*, 42 (3), 2023. 1, 2, 3

[7] Yana Nehmé, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué. Visual quality of 3d meshes with diffuse colors in virtual reality: Subjective and objective evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2202–2219, 2021. 1

[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3, 4, 5, 6

[9] Qi Yang, Joel Jung, Haiqiang Wang, Xiaozhong Xu, and Shan Liu. Tsmd: A database for static color mesh quality assessment study. In *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2023. 1