Supplemental Material: Mononizing Binocular Videos

WENBO HU, MENGHAN XIA, CHI-WING FU, and TIEN-TSIN WONG, The Chinese University of Hong Kong

ACM Reference Format:
Wenbo Hu, Menghan Xia, Chi-Wing Fu, and Tien-Tsin Wong. 2020. Supplemental Material: Mononizing Binocular Videos. *ACM Trans. Graph.* 39, 6, Article 228 (December 2020), 7 pages. https://doi.org/10.1145/3414685.3417764

This supplemental material contains five sections: Section 1 presents the network architecture details; Section 2 gives details on the compiled 3D movie dataset; Section 3 shows an evaluation on the quantization layer; Section 4 shows more qualitative results; and Section 5 provides details on the user study.

1 DETAILS ON THE NETWORK ARCHITECTURE

There are three levels in our feature extractor. Its details is shown in Figure 1 below. The three feature extractors employed in our framework share the same architecture but with different channel numbers, since they process different amount of information.

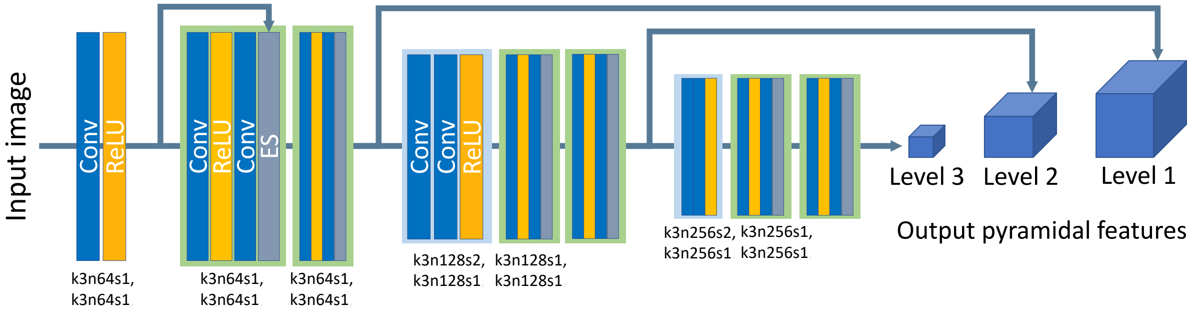


Fig. 1. Architecture of our feature extractor. Here, "ES" denotes element-wise summation, and the "k3n64s1" notation means that the convolution kernel size is three, the number of channels is 64, and the stride is one.

The reconstructor is symmetric with the feature extractor, as shown in Figure 2. Similarly, the three reconstructors used in our framework also share the same architecture but with different channel numbers. There are no Tanh or Sigmoid activation layer for the output convolution, since we empirically found doing so performs slightly better. Source code for the model will be released upon the publication of this work.

Authors' address: Wenbo Hu; Menghan Xia; Chi-Wing Fu; Tien-Tsin Wong, The Chinese University of Hong Kong, Hong Kong, [wbhu,mhxia, cwfu,ttwong]@cse.cuhk.edu.hk.

0730-0301/2020/12-ART228 $15.00
https://doi.org/10.1145/3414685.3417764

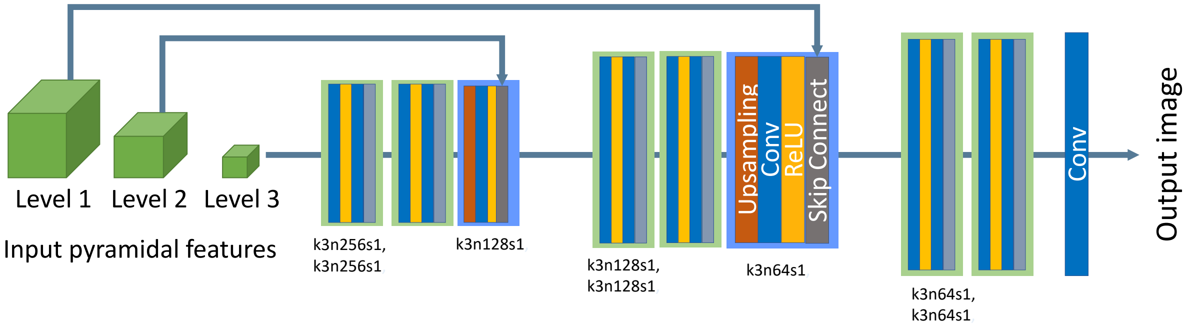


Fig. 2. Architecture of our reconstructor. Here, "upsampling" is achieved using a bilinear interpolation, the "skip connection" first concatenates the two source features and then uses a 1 × 1 convolution to learn the linear combination of the weights.

2 DETAILS ON THE 3D MOVIE DATASET

It contains 122 3D movie sequences with 720p resolution (1280 × 720), 5876 frames in total, collected from the *Inria* dataset[1] and *YouTube*[2]. Overall, it covers eight different types of scenes: *Animal*, *Indoor*, *Outdoor*, *Architecture*, *Vehicle*, *Scenery*, *Nightscape*, and *Cartoon*. Some sample frames in the dataset are shown in Figure 3 below. We randomly selected 69 sequences from it as the training set and used the remaining 53 sequences as the test set.

3 EVALUATION ON QUANTIZATION LAYER

To evaluate the effectiveness of the quantization layer, we removed the quantization layer from our full framework, re-trained our framework on the 3D movie training set, and tested it on the 3D movie test dataset. Moreover, we tried to replace our quantization layer with the universal quantization technique, which formulates the quantization error as uniformly-distributed noise. The results are reported in Table 1 below.

Table 1. Comparing the visual quality of our results (mononized videos and restored binocular videos) without the quantization layer and with universal quantization in our framework.

Method variants	Mono-frame		L. Bino-frame		R. Bino-frame	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Our full framework	39.0	0.98	39.8	0.99	38.7	0.98
without quantization layer	39.4	0.98	38.7	0.97	25.5	0.79
with universal quantization	39.2	0.98	40.1	0.99	38.2	0.98

We can see that if we remove the quantization layer from our framework, the quality of the restored binocular frames (particularly the right frames) will drop substantially. This is because the stereo information encoded in the mononized view is not quantization-friendly if the quantization layer is not adopted, and the quantization error on the mononized view in the testing phase will distort the stereo information. Also, we can see that universal quantization has similar performance with our quantization layer in the framework. Since the quantization layer is more straightforward, we adopt it in our full framework.

[1] Inria: https://www.di.ens.fr/willow/research/stereoseg/

[2] YouTube: https://www.youtube.com/

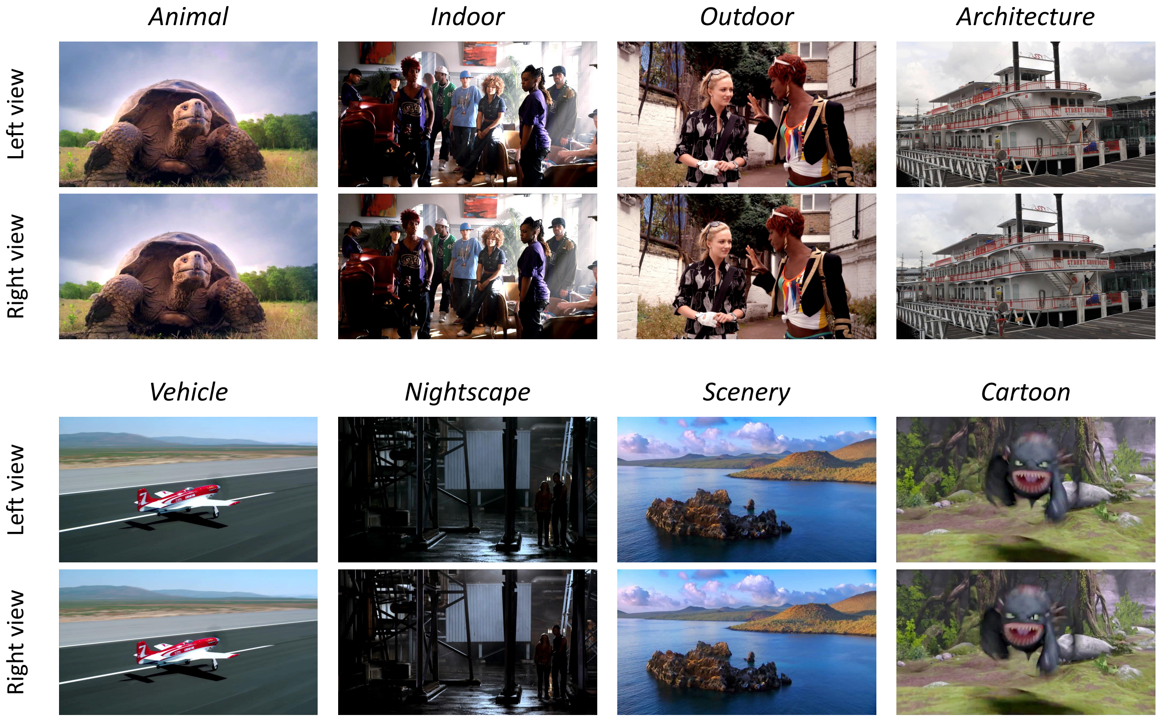


Fig. 3. Sample frames from the eight types of scenes in the 3D movie dataset. For each type, we show the left view on top and the right view on bottom.

4 QUALITATIVE EVALUATION RESULTS

Besides the results shown in the paper, we additionally show eight sets of results (every pair of rows) produced by our method in Figures 4 and 5 on the next two pages. In each result, we show the input left & right views (1st column), generated mononized view and its difference map from the input left view (2nd column), restored binocular views (3rd column), and their difference maps from the inputs (4th column). In each result, the numbers show PSNR and SSIM, while in each error map, the number shows the mean absolute difference (scale of [0,255]). From the figures, we can see that the PSNR values of our results are well above 35 dB and the pixel color difference from the ground truths is typically very small, as revealed in the error maps.

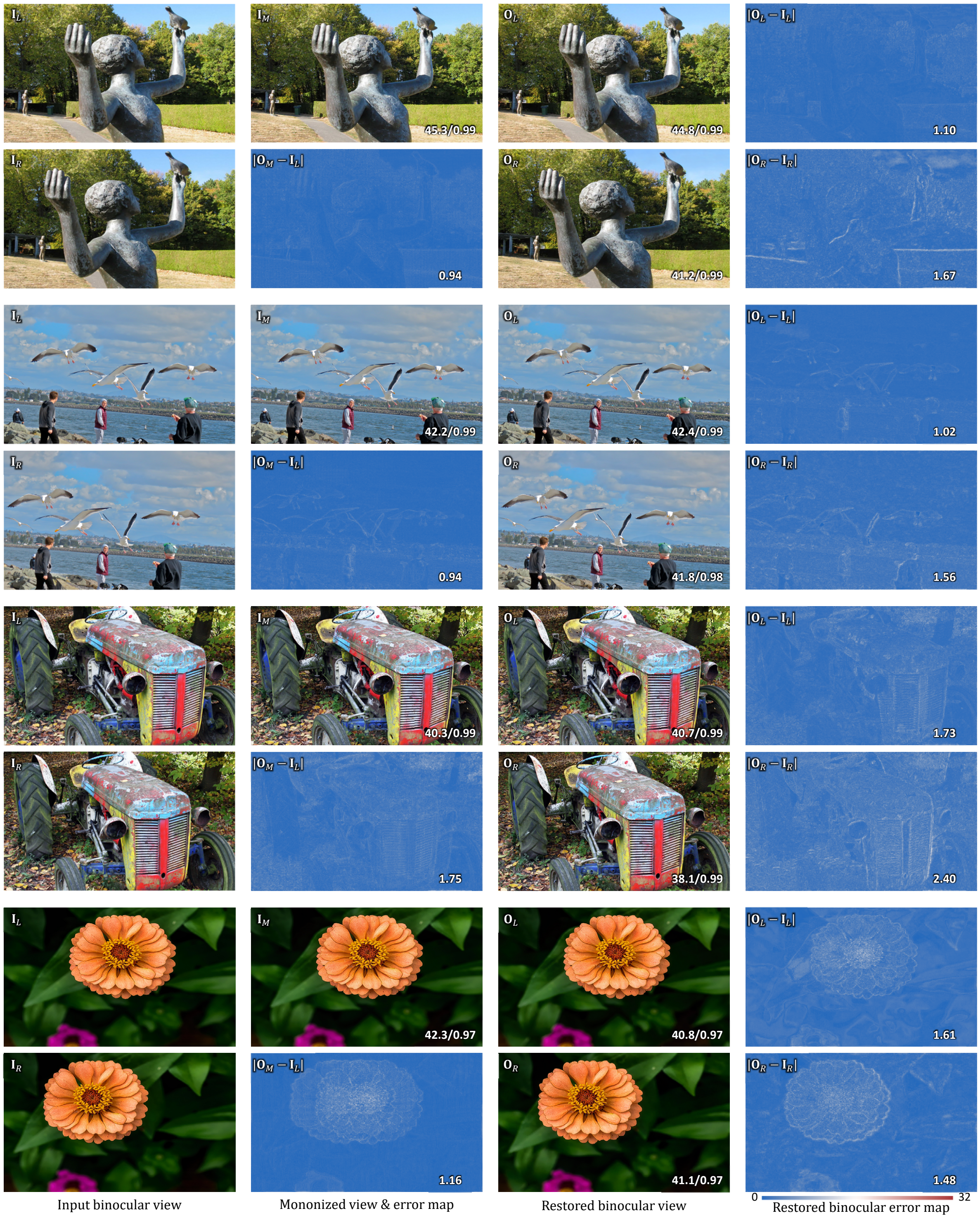


Fig. 4. Example results (every pair of rows) produced by our method.

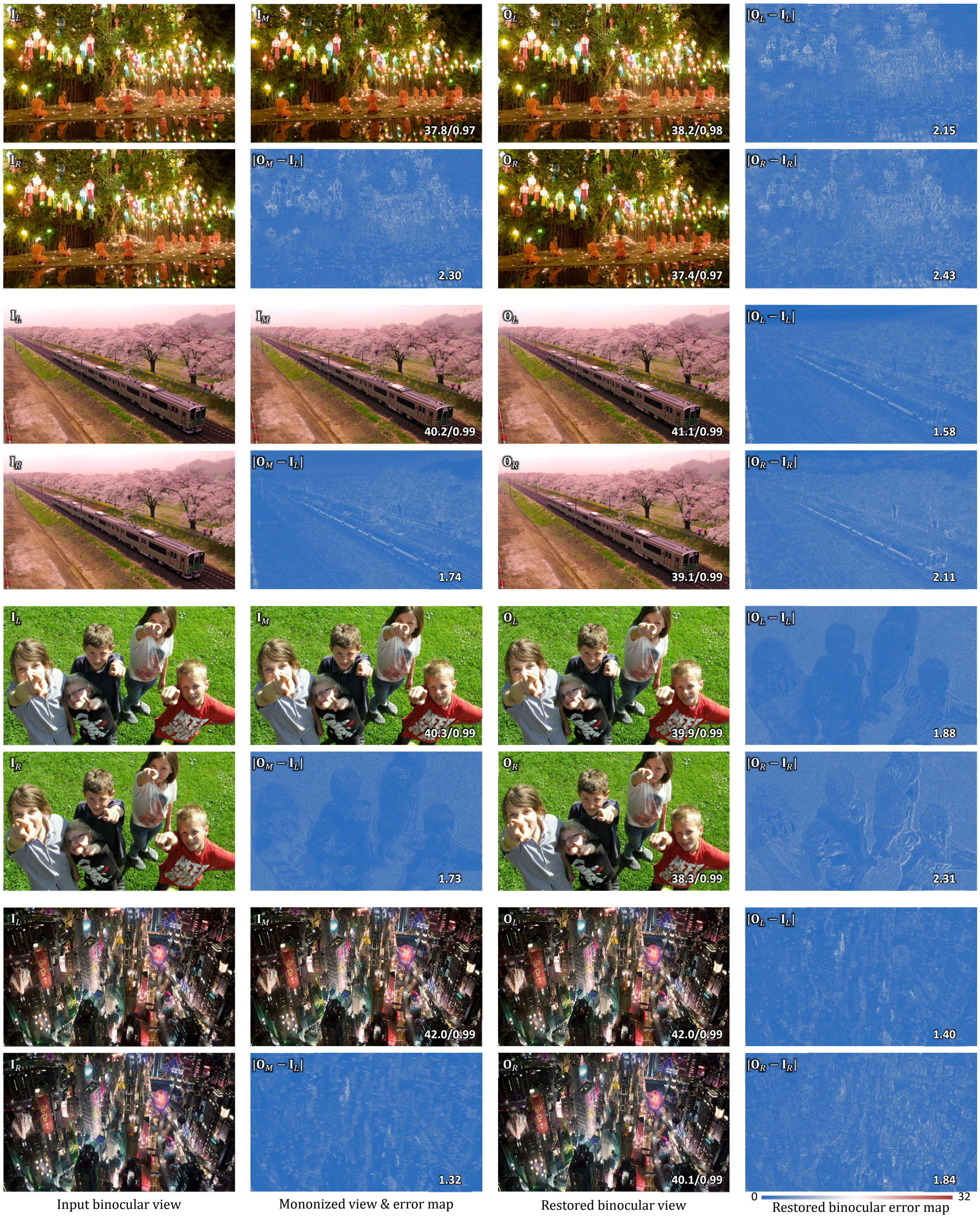
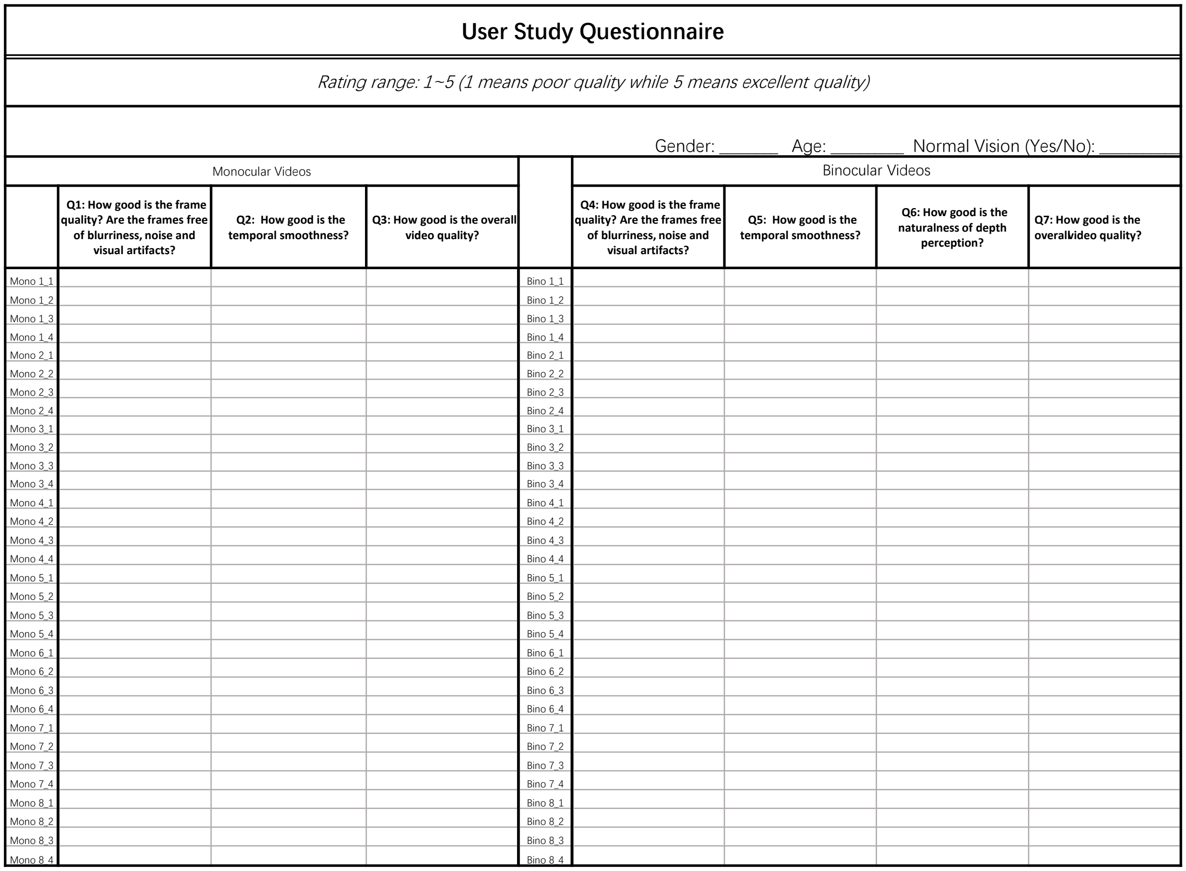


Fig. 5. Example results (every pair of rows) produced by our method.

5 DETAILS OF THE USER STUDY

This section gives details on the questionnaire and control-group B video frames employed in the user study.

Questionnaire. Figure 6 below presents the questionnaire we prepared for the user study. After completing the first session (tutorial) of the study, we gave this questionnaire to each participant for him/her to fill in the background information on top, then we asked each participant to fill the ratings on each video sample per question: first for the monocular videos (the second session), then for the binocular videos (the third session).



User Study Questionnaire

Rating range: 1~5 (1 means poor quality while 5 means excellent quality)

Gender: _______ Age: ________ Normal Vision (Yes/No): ________

	Monocular Videos				Binocular Videos			
	Q1: How good is the frame quality? Are the frames free of blurriness, noise and visual artifacts?	**Q2: How good is the temporal smoothness?**	**Q3: How good is the overall video quality?**		**Q4: How good is the frame quality? Are the frames free of blurriness, noise and visual artifacts?**	**Q5: How good is the temporal smoothness?**	**Q6: How good is the naturalness of depth perception?**	**Q7: How good is the overallvideo quality?**
Mono 1_1				Bino 1_1				
Mono 1_2				Bino 1_2				
Mono 1_3				Bino 1_3				
Mono 1_4				Bino 1_4				
Mono 2_1				Bino 2_1				
Mono 2_2				Bino 2_2				
Mono 2_3				Bino 2_3				
Mono 2_4				Bino 2_4				
Mono 3_1				Bino 3_1				
Mono 3_2				Bino 3_2				
Mono 3_3				Bino 3_3				
Mono 3_4				Bino 3_4				
Mono 4_1				Bino 4_1				
Mono 4_2				Bino 4_2				
Mono 4_3				Bino 4_3				
Mono 4_4				Bino 4_4				
Mono 5_1				Bino 5_1				
Mono 5_2				Bino 5_2				
Mono 5_3				Bino 5_3				
Mono 5_4				Bino 5_4				
Mono 6_1				Bino 6_1				
Mono 6_2				Bino 6_2				
Mono 6_3				Bino 6_3				
Mono 6_4				Bino 6_4				
Mono 7_1				Bino 7_1				
Mono 7_2				Bino 7_2				
Mono 7_3				Bino 7_3				
Mono 7_4				Bino 7_4				
Mono 8_1				Bino 8_1				
Mono 8_2				Bino 8_2				
Mono 8_3				Bino 8_3				
Mono 8_4				Bino 8_4				

Fig. 6. Questionnaire employed in our user study.

Videos in control-group B. Figures 7 and 8 on the next page show example frames in some of the control-group B monocular videos and binocular ones, respectively. For the monocular videos, we compressed the original ones using the H.264 video codec at around 1 Mpbs, to create visual artifacts in the video frames. Also, we randomly removed some frames (50% probability to drop each frame) in the videos to create temporal flickering. For the binocular videos, we replaced the right view by the left view shifted by 15 pixel units. Therefore, the disparity of the control-group B binocular videos is a constant, no matter how far the object is, while for real binocular videos, the disparity should vary according to depth.

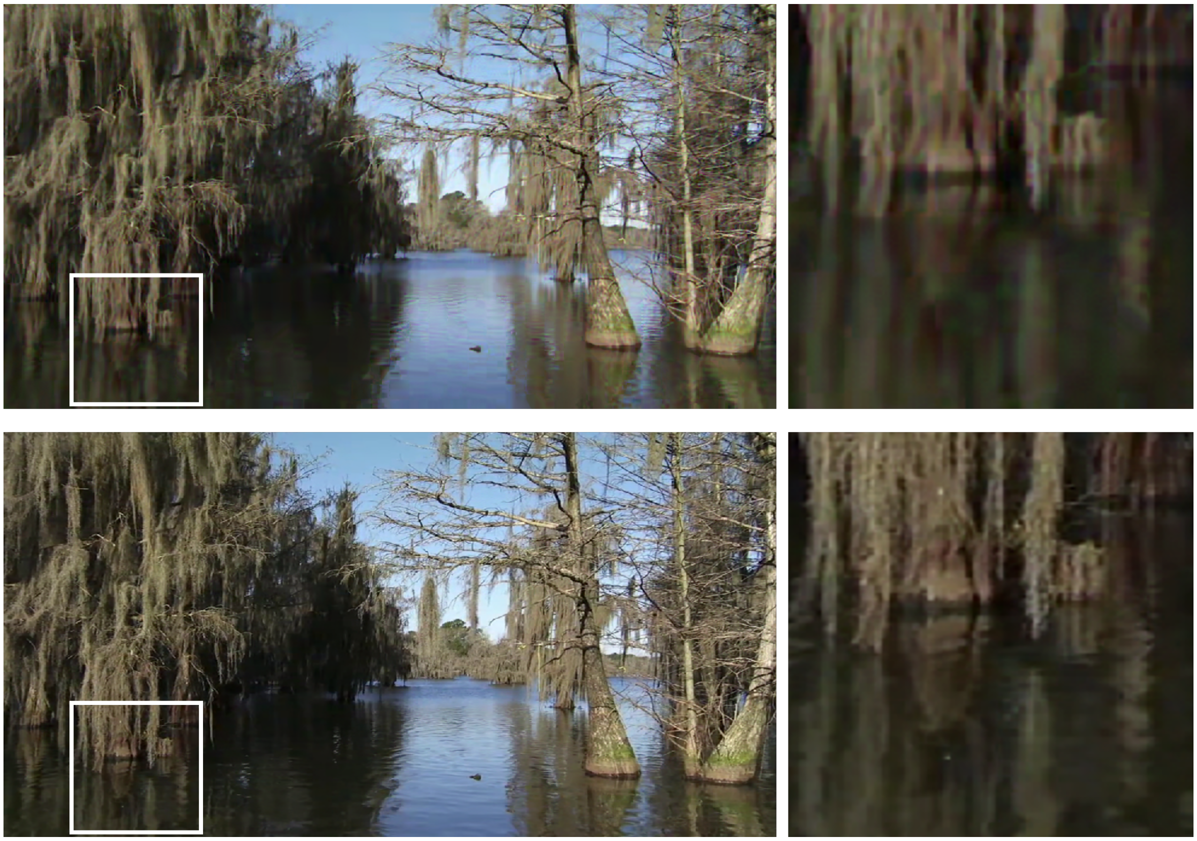


Fig. 7. Comparing an example frame in a control-group B monocular video (top) vs. the corresponding original video (bottom). Top left shows the control-group B video frame, with the blown-up view on its right to highlight the visual artifacts (blurriness), while bottom left shows the original video frame with the corresponding blown-up view for comparisons.

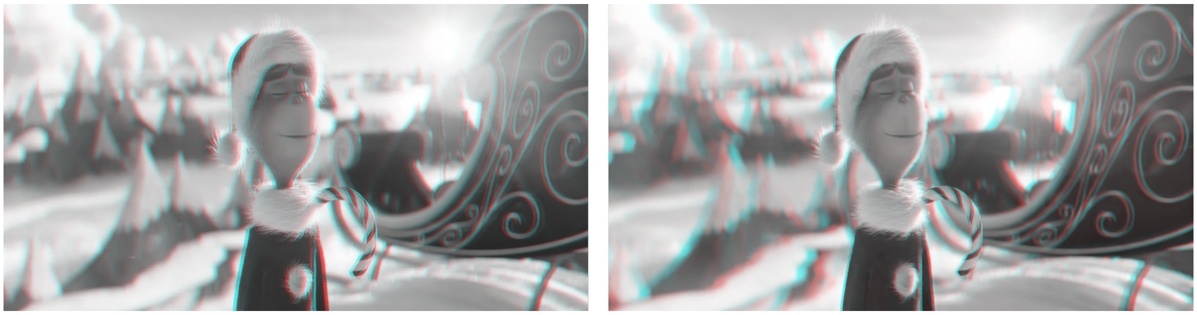


Fig. 8. Comparing an example frame in a control-group B binocular video (left) vs. the corresponding original binocular video (bottom). Note that we present the binocular frames using a red-cyan format to reveal the disparity information, but in the user study, we employed a polarized 3D display to show these binocular videos. Here, since the disparity in the control-group B binocular videos is intentionally set to a constant, the red-cyan format is not revealed in the video frame (left), while the example frame (right) for the original binocular video clearly shows the disparity information.