Humans are still the best lossy image compressors

Ashutosh Bhown[1,*], Soham Mukherjee[2,*], Sean Yang[3,*],
Shubham Chandak[4], Irena Fischer-Hwang[4], Kedar Tatwawadi[4], Tsachy Weissman[4]

[1]Palo Alto High School   [2]Monta Vista High School
[3]Saint Francis High School   [4]Stanford University
schandak@stanford.edu

Abstract

Lossy image compression has been studied extensively in the context of typical loss functions such as RMSE, MS-SSIM, etc. However, it is not well understood what loss function might be most appropriate for human perception. Furthermore, the availability of massive public image datasets appears to have hardly been exploited in image compression. In this work, we perform compression experiments in which one human describes images to another, using publicly available images and text instructions. These image reconstructions are rated by human scorers on the Amazon Mechanical Turk platform and compared to reconstructions obtained by existing image compressors. In our experiments, the humans outperform the state of the art compressor WebP in the MTurk survey on most images, which shows that there is significant room for improvement in image compression for human perception.
Data: The images, results and additional data is available at `https://compression.stanford.edu/human-compression`.

Introduction

Since the advent of electronic media, image compression has been studied extensively, leading to multiple image formats and compression techniques such as PNG [1], JPEG [2], JPEG2000 [3], JPEG XR [4], BPG [5] and WebP [6]. In order to achieve significant reduction in image size, most compression techniques allow some loss while compressing images. However, the loss functions used do not correspond to human perception, and the resulting images may be blurry and unnatural at high loss levels. The left two panels of Figure 1 show an example in which compression and reconstruction using WebP [6] results in a severely blurred image.

It seems natural to posit that better compression results can be achieved using a loss function optimized for human perception. We refer to such a loss function as "human-centric." The rightmost panel of Figure 1 shows an example of a possible human-centric reconstruction which prioritizes image content over pixel-by-pixel fidelity of grass texture. Indeed, there has been a large body of work in the computer vision community [7][8][9] in order to better understand human perception, and hence a loss function governing human vision. Some compression methods, for example, take advantage of the fact that human vision is more susceptible to differences in intensity than in color, and quantize color space more crudely than intensity space in order to achieve better compression performance.

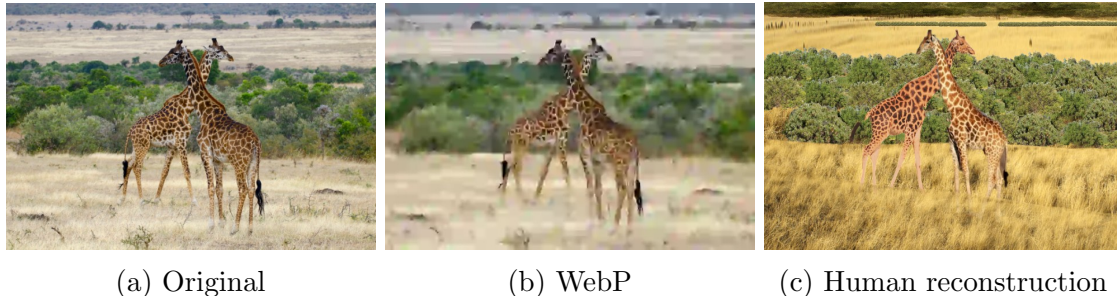| (a) Original | (b) WebP | (c) Human reconstruction |

Figure 1: Giraffe image along with WebP and human reconstructions of similar sizes.

Unfortunately, despite these efforts we still lack a single metric that accurately summarizes the human perception loss. To assess the importance of a human-centric loss function, we present the results of image compression experiments performed by *humans*. In our human compression experiments, two humans communicate through a text chat system in which one "describer" human describes an image to a "reconstructor" human via text instructions. In order to emulate the human ability to perceive and recognize scenes based on prior memory or knowledge of locations or objects, our human compression scheme allows the describer to share URL links to reference images in the text chats from contextually similar images that are publicly available on the web. The describer may also send the reconstructor text instructions for manipulating the images in order to better approximate the describer's image. By employing the growing repertoire of publicly available image databases, our experiment is designed to understand the limits of human-centric compression in the context of side information.

To determine the quality of the reconstruction, we solicit human opinion on the reconstructed image using the Amazon Mechanical Turk (MTurk) platform [10]. The compressed size of the text chat in our framework represents the size of the compressed image, and the MTurk score is considered the (negative) "loss" associated with human compression. We present the results of human compression on 13 high-resolution images of different types. The results show that our human compression scheme performs better than the WebP compressor on 10 out of 13 images.

Related Works

There has been significant work on human-centric compression, and attempts to quantify human perception. Many commonly used compressors such as JPEG, JPEG2000 and WebP already attempt to implicitly capture properties of human perception. For example, the human psychovisual system is prone to discarding sharp edges in images, so JPEG quantizes high frequency components heavily. The MS-SSIM metric was developed in order to exploit image similarity, and is used by [11] and [12] for optimizing image compression. The compressor Guetzli [13] includes a perceptual JPEG encoder optimized for a new image similarity metric dubbled "butteraugli"[14]. More recently, [15] trained a neural network to predict human perceptual quality scores on a large dataset of human-scored images.

Another interesting line of work attempts to capture the effects of human perception by using generative models for lossy compression, which implicitly capture distributions of natural images. Then, discriminator models are used to train the generative models instead of image similarity metrics like RMSE or MS-SSIM. Furthermore, the discriminator models are themselves trained to distinguish between natural and synthetically generated images. For example, [16] uses GANs (Generative Adversarial Networks) to obtain visually pleasing images at low bitrates.

Video encoders such as MPEG [17] attempt to exploit extreme structural similarity (i.e., translational similarity) between adjacent video frames. However, apart from video data, not much work has been done on utilizing semantically similar (i.e., images containing similar high-level features such as objects, persons, etc.) or structurally similar images for compression.

Methods

Here, we describe our human compression scheme. Our setup involves two distinct humans, referred to as the describer and reconstructor, as presented in the introduction.
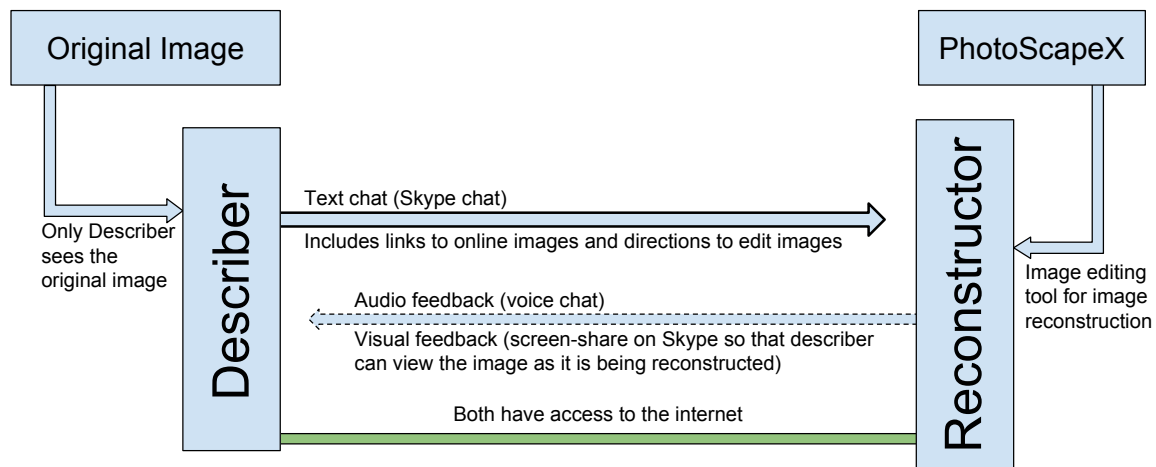

Figure 2: Block diagram showing the human compression process. The describer attempts to describe the image with URL links and text instructions. The describer can see the image as it is being reconstructed and can also hear the reconstructor's voice.

For every input target image, the roles of the describer and reconstructor are as follows:

- **Describer**: Analyzes/recognizes the input image and informs the reconstructor of the necessary steps to best recreate the target image. The describer communicates with the reconstructor only via a real-time text chat and may view the reconstructed image in progress, as well as hear verbal communications from the reconstructor.

https://www.worldwildlife.org/habitats/grasslands

try transformations
elongate the fence bit
only focus on the vertical…

there's a line of shrubbery that goes across the middle third of the image…
that's the largest bush in the pic
so keep the others sizes equal to or smaller that that and make it look continuous
and make sure to make the bushes smaller as you work your way up so that there's a sense of depth…

there will be a line of tiny shrubs along that line…
the line itself starts about a quarter from the left…

Try and make the grass look less tall on the bottom…

when you're done with that take a look at these
https://public-media.smithsonianmag.com/filer/32/f2/32f24473-b380-43f5-94df-da0e58644439/16301090250_acf80be87f_o.jpg
https://img.purch.com/w/192/aHR0cDovL3d3dy5saXZlc2NpZW5jZS5jb20vaW1hZ2VzL2kvMDAwLzA2OC8wOTQvaTMwMC9naXJhZmZlLmpwZwZz8xNDA1MDA4NDQy
sure
while you're editing that giraffe
its spots are too dark
make it look like the other giragge…

make the right one bigger than the left
make the heads level
wait back
put the left one where it was before
good
now move the right giraffe to the left so that their necks cross
good
move them both to the center
make them both taller as well
their heads should be above the middle line of shrubs…

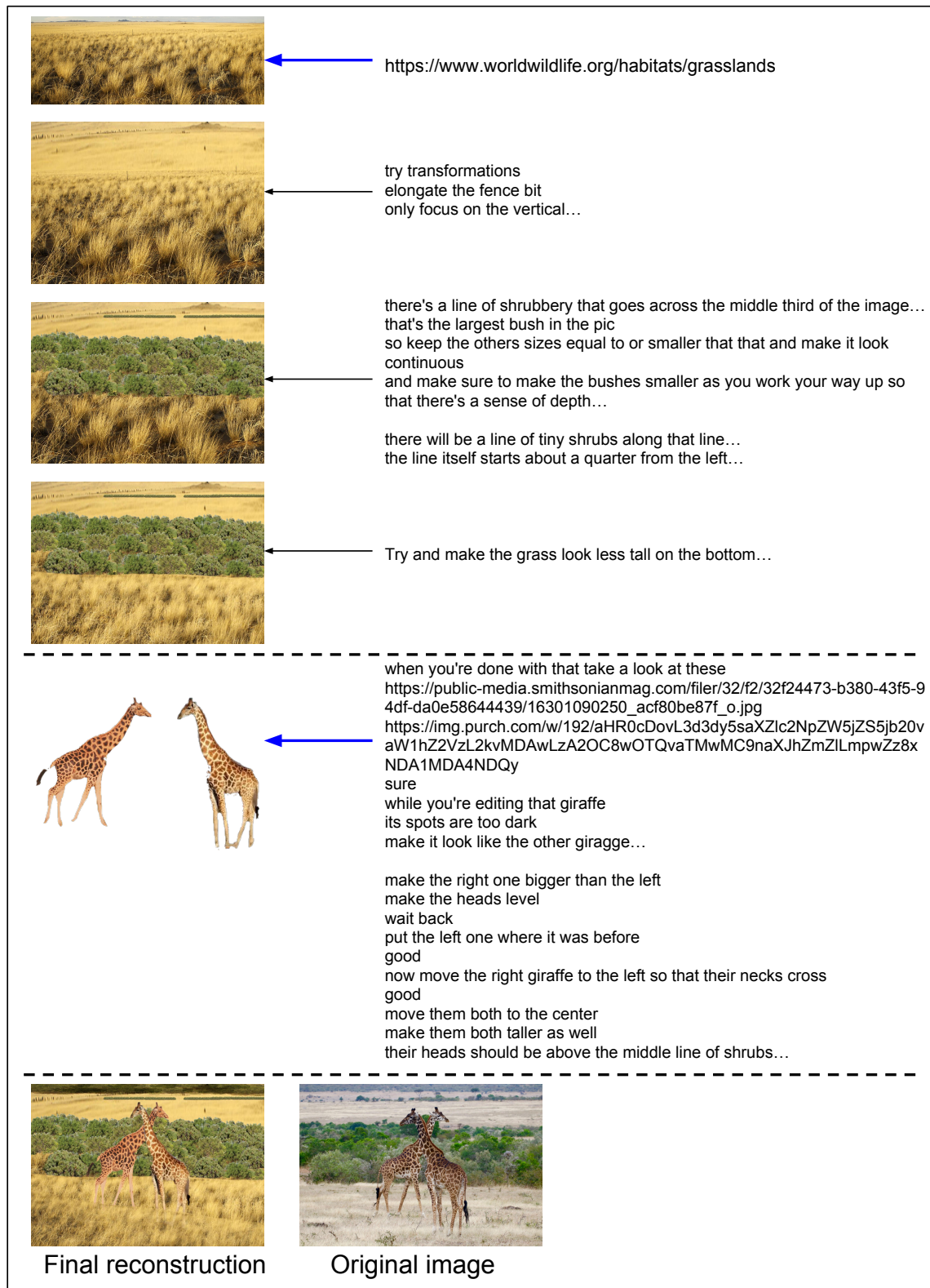Final reconstruction   Original image

Figure 3: Human compression process for the giraffe image (simplified). The text on the right shows text communications sent from the describer to the reconstructor. Various stages of reconstruction are shown, describing (i) the background grass and bush and (ii) the giraffes. Internet links (blue arrows) to a publicly available images are may also be included.

- **Reconstructor**: Interprets text instructions from the describer in order to produce a reconstruction of the original image. The reconstructor is not permitted to view the original image until the reconstruction is complete, but may communicate with the describer.

Unlike machine-based compression schemes, our human compression scheme involves two streams of one-way communication, one text-based from the describer to the reconstructor, and one in any format from the reconstructor back to the describer. In our scheme, only the text transcript from the describer to the reconstructor is considered to be the "compressed" representation of the input image; any communication from the reconstructor to the describer is not counted towards the size of the compressed image.

To understand why only the text communication from describer to reconstructor is included in the final compressed representation of the image, we compare our human compression scheme to a machine implementation of the same. In our experiments, the compression process involves communication between a describer and reconstructor in order to produce a reconstructed image, as well as the text transcript. In a machine implementation, the compression process again involves describer and reconstructor parties, but only generates the text transcript. A reconstructed image would instead be produced by a decompression process that executes the instructions in the text transcript and simulates the communication from the reconstructor to the describer. The decompression process, by definition, is such that the reconstructor behaves identically as it did during the compression process, meaning that only the describer's instructions are needed for decompression, and is the justification for counting only descriptions flowing to the reconstructor in the overall size of compressed files (cf. e.g., [18]). As a result, in our human-based implementation of a compression scheme, only the communication from describer to reconstructor is saved.

Implementation Details

The describer is provided an input image for compression, and a Skype call is initiated between the describer and reconstructor with the following restrictions. First, the describer may only communicate to the reconstructor through the inbuilt Skype text chat. The describer turns off their outgoing audio/video to avoid inadvertently leaking information to the reconstructor. Now, the reconstructor may communicate verbally with the describer through audio/video/text chat. Finally, the reconstructor may share partial, in-progress reconstructions with the describer in real time using Skype's screen share feature.

With these restrictions in place, the describer begins to send a series of instructions for the reconstructor to attempt image reconstruction. Generally, the describer may send URL links to images that already exist on the internet, as well as specific text instructions for altering the image. A variety of image editing tasks may be sent, including: spatially translating image elements, performing affine or perspective transformations, erasing or adding certain objects in the image, enlarging a portion of the image, compositing multiple images, etc. Figure 3 shows parts of the

reconstruction process for the giraffe image.

When reconstruction has been completed to the level of the describer's satisfaction, the experiment is stopped. The Skype text transcript containing all instructions from the describer to the reconstructor is saved. Finally, the transcript is processed by removing timestamps and compressing it using the bzip2 [19] compressor. The bzip2 encoded Skype transcript represents the final compressed representation of the input image. The quality of image reconstruction can now be compared to that of a standard lossy image compressor, as described in the next section.

Experiments

Data Collection

We first created a data set of original images that are not publicly available on the web. The creation of original images prevents trivial encoding via an exact copy of a non-original picture. Original images were captured with a digital camera or smartphone camera at high resolution. A wide variety of images (e.g., faces, landscapes, sketches, etc.) unknown to the describers and reconstructors were captured for the experiments. From these, we selected 13 diverse high-resolution images for our comparison experiments. The images and additional details are available in the appendix and at `https://compression.stanford.edu/human-compression`.

Experimental Setup

We describe the experimental procedure for evaluating the quality of reconstructions by human compressors and WebP:

1. **Human compression:** The input image is compressed and reconstructed by the human compression system using the previously described procedure. The size (in bytes) of the compressed text instructions is recorded.

2. **WebP compression:** The WebP compressor is used to lossily compress the input image into a size similar to that of the compressed human text instructions.

3. **Quality evaluation:** We compared the quality of WebP and human compressed images using human scorers on the MTurk platform.

WebP Compression

WebP [6] is a relatively recent image compressor released by Google. We chose WebP as the reference compressor for comparing the human reconstruction quality since WebP outperforms JPEG and JPEG2000 at the high compression levels achieved by the human compression scheme. This is illustrated in Figure 4.

Even when compressing images using WebP at the lowest allowed quality level (quality parameter set to 0), the compressed files were much larger than those of the human compressors. As a result, we first reduced the resolution of the images before compressing with WebP with quality parameter 0 in order to attain the target size, always erring on the side of the WebP file being larger than that associated with the human compressor.

(a) JPEG   (b) JPEG2000   (c) WebP

Figure 4: Comparison of JPEG, JPEG2000 and WebP at a high compression level.
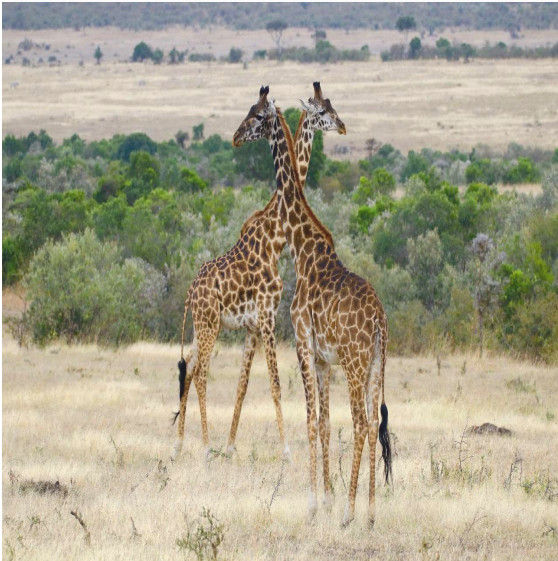
Quality Evaluation using MTurk

Figure 5: A screenshot of the Mturk survey.
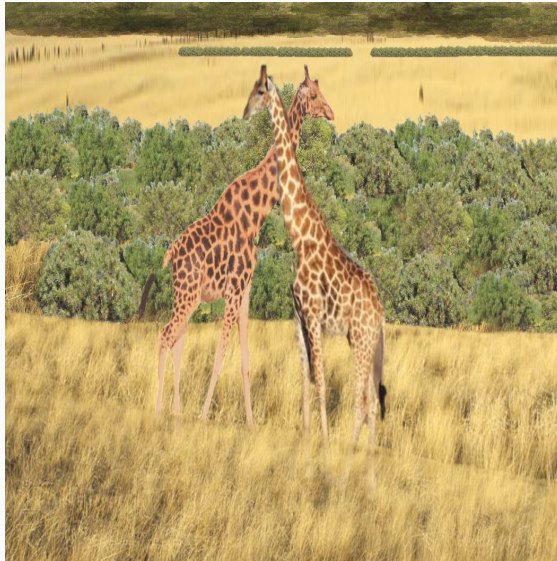
We compare the quality of compressed images using human scorers (workers) on Amazon Mechanical Turk (MTurk) [10], a platform for conducting large scale surveys. For each image, we display the original image and a reconstructed image, and ask workers to rate the reconstruction on a scale of 1 to 10. Since human perception is not yet well understood nor defined, we defined the scale on a general "level of satisfaction" with the reconstruction, rather than a specific metric like accuracy. For every experiment and for both types of reconstruction (human compression and

Image	Original size (KB)	Compressed chat size (KB)	WebP size (KB)	Mean score Human	Mean score WebP	Median score Human	Median score WebP
arch	1119	3.805	3.840	4.04	**5.1**	3	**5**
balloon	92	1.951	2.036	**6.22**	5.45	**7**	6
beachbridge	3263	4.604	4.676	**4.34**	3.92	4	4
eiffeltower	2245	4.363	4.394	**5.98**	5.77	6	6
face	1885	2.649	2.762	2.95	**5.47**	3	**6**
fire	4270	2.407	2.454	**6.74**	5.09	**7**	5
giraffe	5256	3.107	3.144	**6.28**	4.48	**7**	4
guitarman	1648	2.713	2.730	**4.88**	4.07	**5**	4
intersection	3751	3.157	3.238	**6.8**	4.15	**7**	4
rockwall	4205	6.613	6.674	4.41	**4.85**	4	**5**
sunsetlake	1505	4.077	4.088	**5.08**	4.82	5	5
train	3445	1.948	2.024	**6.85**	3.62	**7**	3
wolfsketch	1914	0.869	0.922	**8.25**	3.46	**9**	3

Table 1: Original image size and compressed sizes along with mean and median Mturk scores for human and WebP reconstructions. Best results are boldfaced.

WebP), we collect 100 survey responses and obtain summary statistics. Figure 5 shows a screenshot of the MTurk survey as seen by the workers.

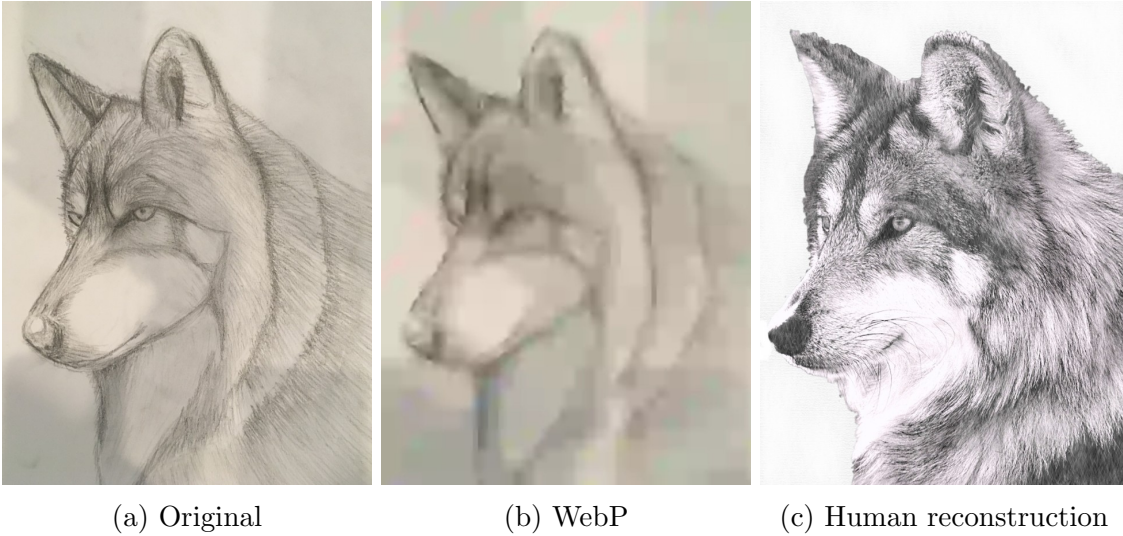(a) Original (b) WebP (c) Human reconstruction

Figure 6: Original wolfsketch image along with the WebP and human reconstructions.

Results

Table 1 shows the results of our human compression scheme and MTurk evaluation on 13 diverse high-resolution images. The human compressor was ranked higher than WebP on 10 out of the 13 images from the dataset. Qualitatively, the human reconstructions seem more natural and sharper to the MTurk workers, as compared to the WebP compressed images (see Figures 1, 6, 7) while still achieving high compression ratios, ranging from around 100x to 1000x.

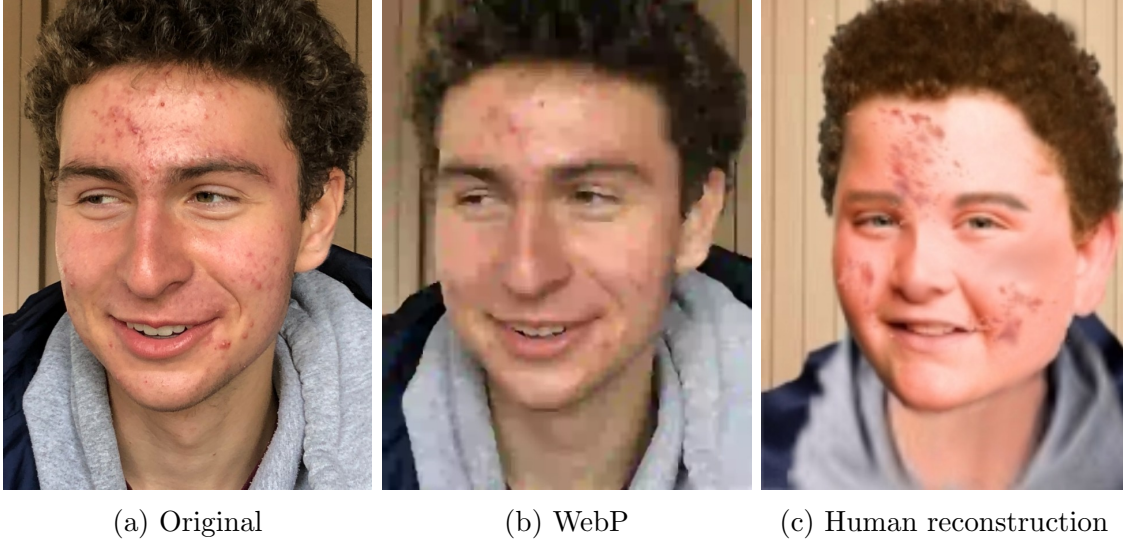(a) Original   (b) WebP   (c) Human reconstruction

Figure 7: Original face image along with the WebP and human reconstructions.

To better understand the results, we discuss some specific examples. For the giraffe image (Figure 1), we suspect that human compression achieved a better rating than WebP because human scorers give more priority to image sharpness over accuracy. In contrast, for the face image (Figure 7) human compression achieved a significantly lower quality score than WebP. We hypothesize that this is because the identity of the human being is more important than other semantic features of the image (such as facial blemishes). On the other hand, humans achieve a much better score than WebP for the wolfsketch image (Figure 6), likely because differences in the facial features of an animal are not significant to human scorers. We also observed that human compression achieved better compression ratios and MTurk scores when similar images were publicly available. This was the case for images of famous monuments such as the Eiffeltower image, and for the intersection image where Google Street View provided images of similar road intersections.

Discussion & Conclusion

We designed an experiment to better understand the potential for improving lossy image compression based on a human-centric loss. Using two humans playing the roles of describer and reconstructor, we compressed 13 diverse images of landscapes, portraits, animals and urban settings. We evaluated the quality of human compression by comparing their reconstructions with WebP compressed images. On most of our natural images, the human reconstructions are preferred over the WebP reconstructions. These results suggest that the human compression process was better at identifying and preserving image properties that are relevant to human scorers. This highlights the fact that there is significant room for improving image compression using human-centric loss functions in the current era of increasingly comprehensive public image databases.

The human compression scheme is able to exploit semantically similar images quite

effectively during compression. However, most popular compressors do not appear to take advantage of this rich public resource. Our experiment suggests that effective utilization of semantically and structurally similar images (or parts of images) can dramatically improve compression ratios. This is particularly relevant today, when images can be easily found using image search tools such as the one offered freely by Google.

While the human compression framework is useful as an exploratory tool, it is clearly not practical due to its labor-intensive nature. We did not strive to optimize our protocols in any way, and we could have undoubtedly achieved substantially better compression and reconstruction scores had we done so. Notably, each of the image reconstructions took a few hours to complete. Furthermore, redundancies in English language resulted in sub-optimal compression, even though this is partly resolved by the use of bzip2. Our drawing skills, use of rudimentary software for image editing, inefficiencies due to occasional misunderstandings of describer instructions by the reconstructor, and the difficulty of manually searching for similar images all contributed to transcript size. Improvements on any of these fronts would further result in improved image reconstruction quality.

We plan to use the insights obtained from this work to build an image compressor that is both optimized for human perception loss and able to utilize side information in the form of publicly available databases. We look to the work in [15], which trains a neural network to predict human scores, as a strategy for training machine-based compressors for the human perception loss. We also expect to take advantage of reverse image search tools in order to better utilize side information. We believe these techniques will be key to significantly improved lossy image compression.

Our work was inspired in part by Claude Shannon's 1951 paper [20], where humans were used to establish an upper bound on the fundamental limit of English language compression. At the time, humans were better compressors than any practically implementable algorithm, and the paper motivated subsequent developments in text compression to match and eventually surpass the 2.3 bits/symbol shown to be achievable by human compressors. Similarly, we hope that the performance we showed for human-based image compression will motivate and guide construction of lossy image compression algorithms that will achieve and eventually surpass human performance.

Acknowledgement

References

[1] Greg Roelofs, *PNG: The Definitive Guide*, O'Reilly & Associates, Inc., Sebastopol, CA, USA, 1999.

[2] Gregory K Wallace, "The JPEG still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[3] David Taubman and Michael Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice*, Springer Publishing Company, Incorporated, 2013.

[4] "JPEG XR," `https://jpeg.org/jpegxr/`, Accessed: 2018-10-22.

[5] "BPG," `https://bellard.org/bpg/`, Accessed: 2018-10-22.

[6] "Webp," `https://developers.google.com/speed/webp/`, Accessed: 2018-10-16.

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Ieee, 2003, vol. 2, pp. 1398–1402.

[9] Troy Chinen, Johannes Ballé, Chunhui Gu, Sung Jin Hwang, Sergey Ioffe, Nick Johnston, Thomas Leung, David Minnen, Sean O'Malley, Charles Rosenberg, et al., "Towards a semantic perceptual image metric," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 624–628.

[10] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?," *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.

[11] Thomas Richter and Kil Joong Kim, "A ms-ssim optimal jpeg 2000 encoder," in *Data Compression Conference, 2009. DCC'09*. IEEE, 2009, pp. 401–410.

[12] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[13] Giaime Ginesu, Maurizio Pintus, and Daniele D Giusto, "Objective assessment of the webp image coding algorithm," *Signal Processing: Image Communication*, vol. 27, no. 8, pp. 867–874, 2012.

[14] "butteraugli," `https://github.com/google/butteraugli/`, Accessed: 2018-10-16.

[15] Troy Chinen, Johannes Ballé, Chunhui Gu, Sung Jin Hwang, Sergey Ioffe, Nick Johnston, Thomas Leung, David Minnen, Sean O'Malley, Charles Rosenberg, et al., "Towards a semantic perceptual image metric," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 624–628.

[16] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool, "Generative adversarial networks for extreme learned image compression," *arXiv preprint arXiv:1804.02958*, 2018.

[17] Didier Le Gall, "MPEG: A video compression standard for multimedia applications," *Communications of the ACM*, vol. 34, no. 4, pp. 46–58, 1991.

[18] A. No and T. Weissman, "Rateless lossy compression via the extremes," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5484–5495, Oct 2016.

[19] "bzip2," `http://www.bzip.org/`, Accessed: 2018-10-16.

[20] Claude E Shannon, "Prediction and entropy of printed english," *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

Appendix

Additional details

Tables 2 and 3 contain additional details about the images and the mechanical turk experiments.

Image	Original resolution	WebP resolution	Original size (KB)	Compressed chat size (KB)	WebP size (KB)
arch	1762 × 2286	506 × 656	1119	3.805	3.840
balloon	1024 × 683	630 × 420	92	1.951	2.036
beachbridge	4032 × 3024	500 × 375	3263	4.604	4.676
eiffeltower	2448 × 3264	492 × 656	2245	4.363	4.394
face	3024 × 4032	435 × 580	1885	2.649	2.762
fire	3036 × 4048	375 × 500	4270	2.407	2.454
giraffe	5472 × 3648	528 × 352	5256	3.107	3.144
guitarman	1136 × 640	550 × 310	1648	2.713	2.730
intersection	3024 × 4032	450 × 600	3751	3.157	3.238
rockwall	3036 × 4048	531 × 708	4205	6.613	6.674
sunsetlake	3264 × 2448	1148 × 861	1505	4.077	4.088
train	4032 × 3024	340 × 255	3445	1.948	2.024
wolfsketch	2698 × 3539	290 × 380	1914	0.869	0.922

Table 2: Resolution and original/compressed size for the images. Chat transcripts were compressed with bzip2. WebP resolution was reduced till the file size just exceeded the compressed chat transcript size, keeping quality parameter 0 and aspect ratio fixed.

Image	Mean score		Median score		Standard deviation	
	Human	WebP	Human	WebP	Human	WebP
arch	4.04	**5.1**	3	**5**	2.27	2.11
balloon	**6.22**	5.45	**7**	6	2.3	2.54
beachbridge	**4.34**	3.92	4	4	2.27	2.17
eiffeltower	**5.98**	5.77	6	6	2.2	2.15
face	2.95	**5.47**	3	**6**	1.87	2.01
fire	**6.74**	5.09	**7**	5	2.31	2.25
giraffe	**6.28**	4.48	**7**	4	2.37	2.08
guitarman	**4.88**	4.07	**5**	4	2.55	2.03
intersection	**6.8**	4.15	**7**	4	1.9	2.17
rockwall	4.41	**4.85**	4	**5**	2.33	2.27
sunsetlake	**5.08**	4.82	5	5	2.33	2.34
train	**6.85**	3.62	**7**	3	2.3	2.1
wolfsketch	**8.25**	3.46	**9**	3	2.03	1.94

Table 3: Mean, median and standard deviation of Mturk scores for human and WebP reconstructions. Best result for each image is boldfaced.

Images

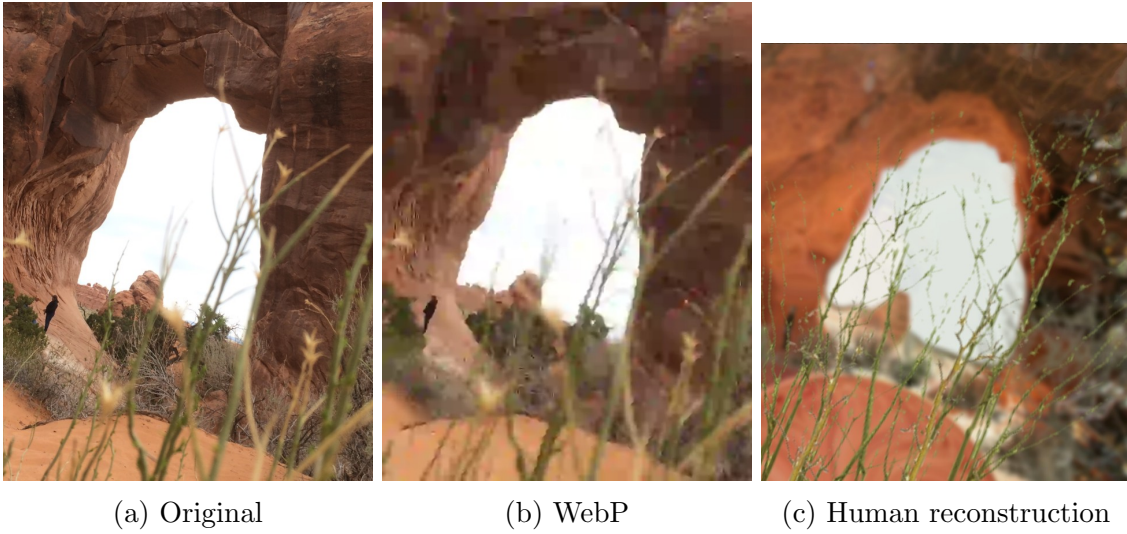This section contains all 13 original images along with their WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

Figure 8: Original arch image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

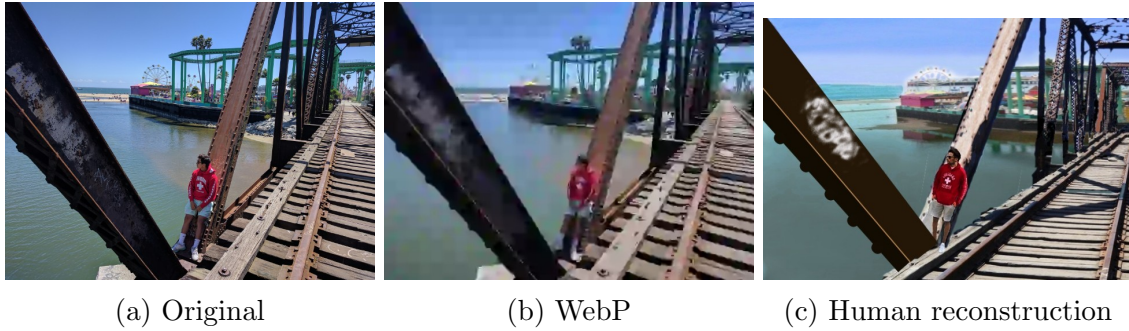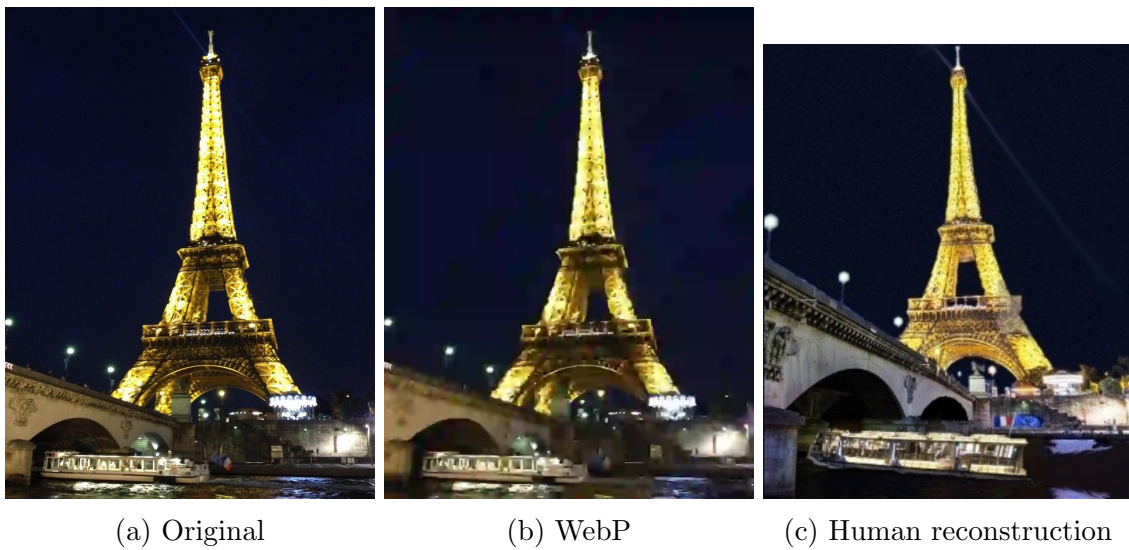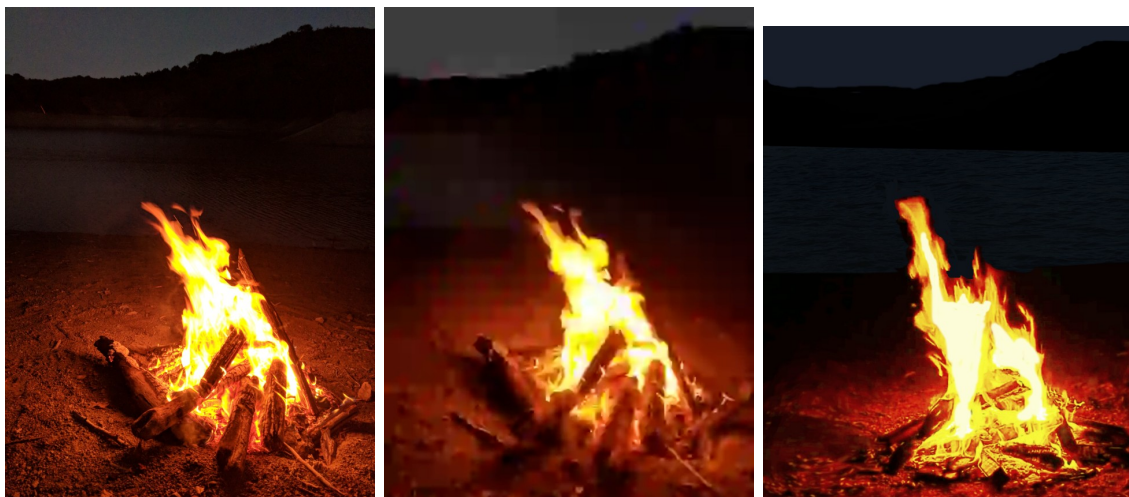Figure 9: Original balloon image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

Figure 10: Original beachbridge image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

Figure 11: Original eiffeltower image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

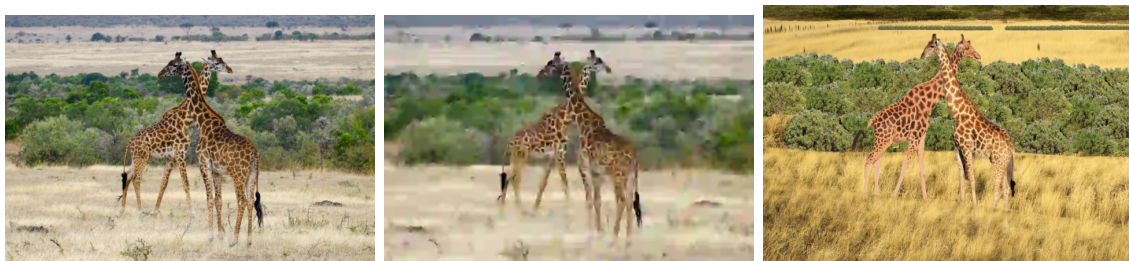Figure 12: Original face image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

Figure 13: Original fire image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

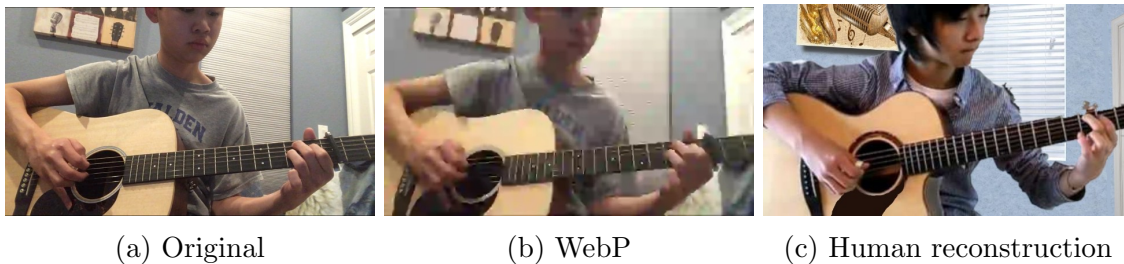Figure 14: Original giraffe image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

Figure 15: Original guitarman image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

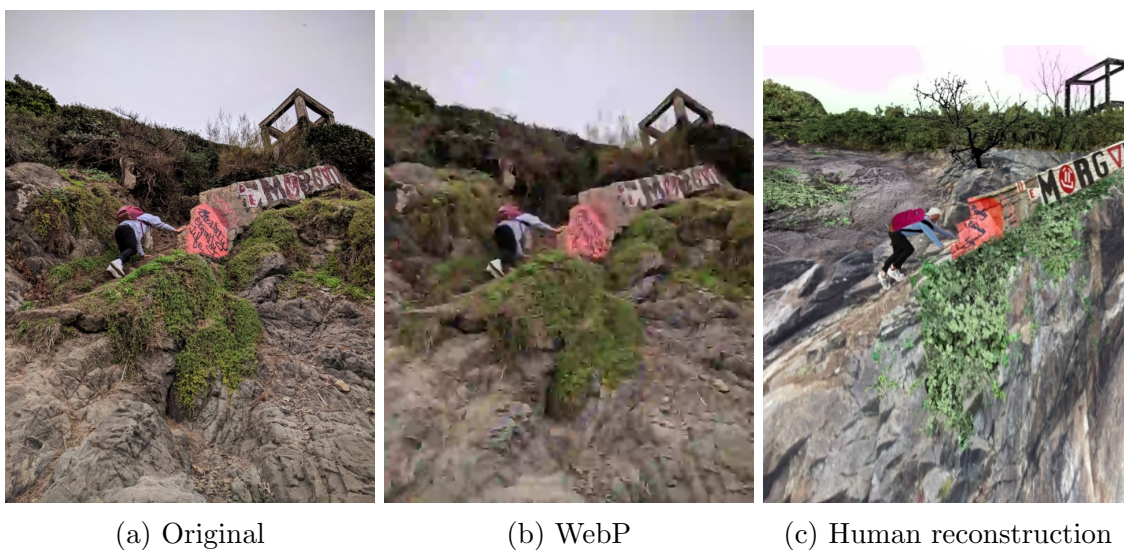Figure 16: Original intersection image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

Figure 17: Original rockwall image along with the WebP and human reconstructions.

(a) Original   (b) WebP   (c) Human reconstruction

Figure 18: Original sunsetlake image along with the WebP and human reconstructions.

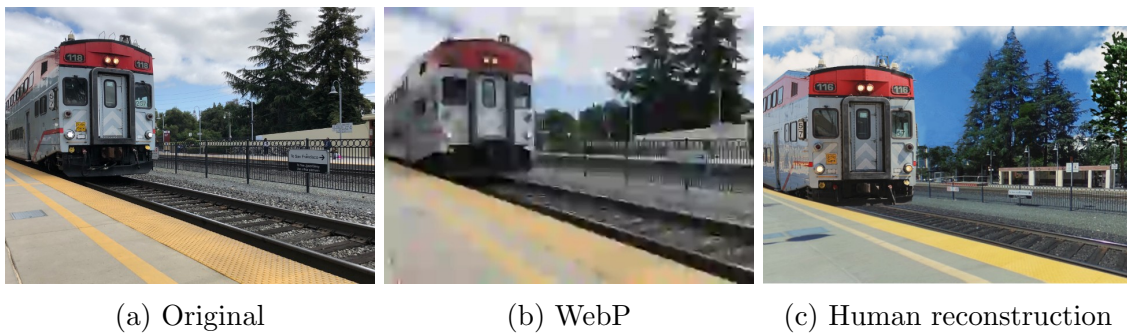(a) Original   (b) WebP   (c) Human reconstruction

Figure 19: Original train image along with the WebP and human reconstructions.

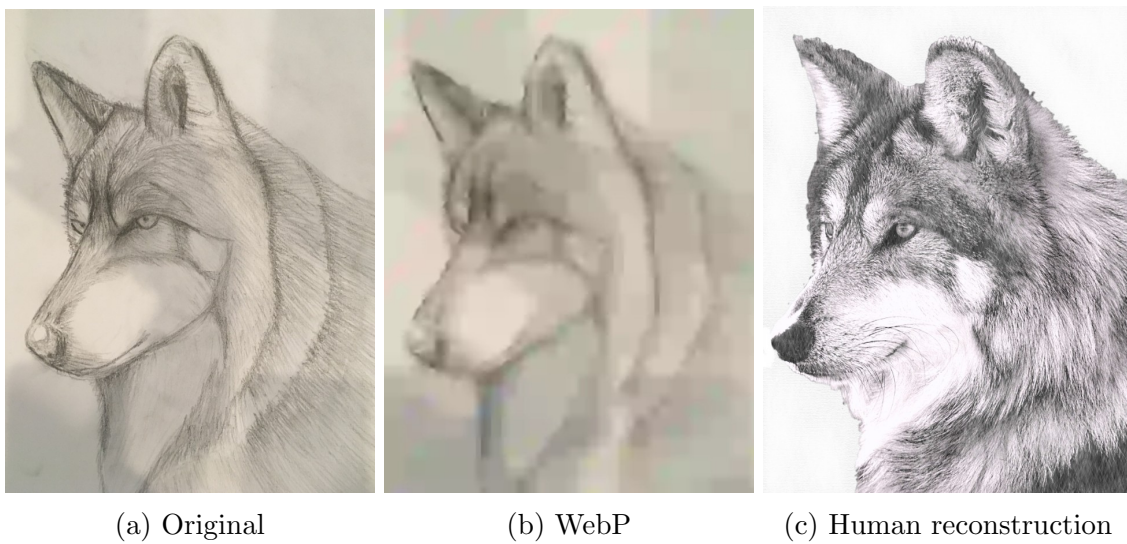(a) Original   (b) WebP   (c) Human reconstruction

Figure 20: Original wolfsketch image along with the WebP and human reconstructions.