Saliency-guided deployment-adaptive compression for wildlife camera traps

Tianhong Xie
Harvey Mudd College
Claremont, CA, USA
tixie@hmc.edu

Justin Kay
Massachusetts Institute of Technology
Cambridge, MA, USA
jkay@mit.edu

Timm Haucke
Massachusetts Institute of Technology
Cambridge, MA, USA
haucke@mit.edu

Sara Beery
Massachusetts Institute of Technology
Cambridge, MA, USA
beery@mit.edu

Abstract

Automatically triggered cameras, also known as camera traps, are an indispensable tool for studying animal ecology. Retrieving data from remote camera traps is an ongoing challenge. Especially when deployed in remote areas, bandwidth of wireless transmission is extremely limited. We propose a method for efficient deep-learning based image compression to address this challenge. Our method utilizes a state-of-the-art autoencoder network to compress images into a compact latent representation. By simply transmitting these latents we demonstrate that in many cases we can reduce bandwidth compared to JPEG while achieving the same or better reconstruction quality. We also propose a method for deployment-specific fine-tuning of our autoencoder architecture to specialize models at the edge to specific environmental conditions. Specifically, we fine-tune the encoder with LoRA and saliency-guided compression. Our experiments demonstrate that this approach reduces bandwidth requirements and improves reconstruction error, particularly at locations with abundant imagery.

1 Introduction

Animals have an outsized impact on their environment, including on factors that are highly relevant to carbon sequestration rates, such as tree recruitment [19]. For studying the ecology of animals, automatically triggered cameras, also known as camera traps, are an indispensable tool. These devices enable researchers to monitor wildlife in their natural habitats with minimal human disturbance, providing invaluable data on animal behavior, population dynamics, and ecosystem health [26]. However, camera traps are often deployed in remote locations far from existing communication infrastructure, presenting significant challenges for data retrieval. Typically, data must be retrieved physically by exchanging SD cards, which can be time-consuming and impractical for large-scale deployments or long-term studies. Alternatively, data can be transmitted via satellite or radio links. This would be especially desirable for applications such as poaching prevention [14], where real-time information is needed for intervention.

Figure 1: Example camera trap deployment.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

A major limitation in utilizing satellite and radio links for transmitting camera trap data is the low available bandwidth and high cost of transmission. These limitations mean that highly efficient data compression techniques are needed in order for transmission to be a practical option for most camera trap users. From our conversations with users, we have found there is consensus that in order for transmission to be practical this bandwidth would need to be reduced to approximately 170 bytes per image. Unfortunately the compression ratios offered by traditional image codecs such as JPEG are typically several orders of magnitude away from this target—for example, a standard 720p JPEG image typically ranges from 100 KB to 800 KB depending on quality settings and image complexity.

In this work we investigate the potential for using efficient deep learning compression techniques to address this bottleneck. Our intuition is that there are several aspects of camera trap imagery that make learning-based methods particularly suitable: First, camera trap deployments are static, with cameras typically affixed to trees (see Fig. 1). This means that the backgrounds of transmitted imagery from a given site are very similar, and learning site-specific characteristics should be possible. Second, due to increased interest in computer-vision based processing of camera trap imagery, there are now commercially-available camera traps from companies like Conservation X Labs [17] that incorporate AI-ready edge-computing hardware, which could enable fine-tuning models on device as well as using existing computer vision models for detecting wildlife such as MegaDetector [5] to encourage compression techniques to focus on salient image regions.

We propose a novel approach to image compression for camera traps that (1) leverages the consistency of the monitored environment by fine-tuning site-specific encoders on-device, and (2) incorporates strong pre-trained wildlife detection models to guide the compression process.

We perform experiments to demonstrate that: (1) Our autoencoder-based approach is consistently able to achieve a better reconstruction than JPEG compression with a fixed number of bits per pixel, and (2) Location-specific fine-tuning consistently improves reconstruction performance. Overall our results demonstrate the feasibility of using autoencoders, LoRA fine-tuning, and saliency-guided compression to enable low-bandwidth connectivity for camera trap deployments.

2 Related Work

Existing works have shown the viability of using autoencoders [11] for improving lossy image compression [9, 28]. Traditional autoencoders consist of an *encoder* that compresses the input into a lower-dimensional latent space and a *decoder* that reconstructs the image from this compact representation. Recent work has extended classical transform coding [20] by replacing linear transforms with learned non-linear neural network-based transforms. Several works have improved entropy modeling by learning conditional priors. Theis et al. [24] use a fully factorized prior. Johnston et al. [13] add spatially-adaptive bitrates. Ballé et al. [4] build on these advancements and introduce a learned "hyperprior" to model spatial dependencies between different elements of the latent encoding. We use the method from Ballé et al as a starting point and build upon the implementation provided by [6].

3 Methods

Autoencoder structure: We adopt the hyperprior model introduced by Ballé et al. [4], which has demonstrated state-of-the-art performance in image compression. The model architecture consists of four key components: (1) An **encoder** g_a transforms the input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a latent representation $\mathbf{y} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_y}$ using a series of convolutional layers; (2) A **hyperprior encoder** h_a processes $\mathbf{y}$ to generate a *hyperlatent* $\mathbf{z} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_z}$. This models the spatial dependencies in the latent space; (3) A **hyperprior decoder** h_s reconstructs parameters for the *entropy model* of $\mathbf{y}$ from $\mathbf{z}$; (4) Finally, a **decoder** g_s reconstructs the image $\hat{\mathbf{x}}$ from $\mathbf{y}$ using information from the hyperprior. The model is trained using a variational framework that balances compression rate and reconstruction fidelity. The overall loss function is defined as:

$$\mathcal{L} = R_y + R_z + \lambda D(\mathbf{x}, \hat{\mathbf{x}}), \tag{1}$$

where $R_y = -\log_2 p(\mathbf{y}|\mathbf{z})$ and $R_z = -\log_2 p(\mathbf{z})$ represent the "rates" (the bits required to represent each pixel) for the main latent representation and hyperlatent, respectively, $D(\mathbf{x}, \hat{\mathbf{x}})$ is the distortion measure (i.e. a reconstruction loss such as mean squared error or SSIM), and λ is a hyperparameter that balances rate and distortion.

A key component of this autoencoder framework is an ***entropy model*** that attempts to estimate the probability distribution of the image data being compressed, representing the likelihood of observing various components of a given image. The key intuition here is that probability distributions with lower entropy require fewer bits to represent [22]—this is especially relevant in our setting, as backgrounds are relatively static (i.e. they have low entropy with respect to the overall data distribution) and thus we can use fewer bits to represent them. In a typical autoencoder, the latent embedding $\mathbf{y} = g_a(\mathbf{x}; \theta)$ assumes all elements of the embedding are independent. This may be suboptimal due to spatial dependencies. Therefore, Ballé et al. introduce a learned *hyperlatent* $\mathbf{z} = h_s(\mathbf{y}; \theta)$, and then represent the probability distribution of $\mathbf{y}$ as $p(\mathbf{y}|\mathbf{z}) = \prod_{i=1}^{N} \mathcal{N}(y_i; \mu_i(\mathbf{z}), \sigma_i^2(\mathbf{z}))$. where N is the number of elements in $\mathbf{y}$, and $\mu_i(\mathbf{z})$, $\sigma_i^2(\mathbf{z})$ are the mean and variance predicted by the hyperprior decoder h_s. After entropy modeling, both $\mathbf{y}$ and $\mathbf{z}$ are entropy coded into a compressed bitstream using a method like arithmetic coding or range coding that has of a dynamic size based on the entropy of the image.

Location-specific fine-tuning: We fine-tune the autoencoder at each camera trap deployment site. In some sense, the goal is to overfit the model to each site such that we generate latent representations of each image with the lowest reconstruction error possible. We freeze the decoder and only fine-tune the encoder, allowing us to transmit the latent representations which can be later decoded by the fixed decoder from pre-training. We experiment with standard fine-tuning as well as fine-tuning with low-rank adaptation (LoRA [12]) to reduce GPU memory requirements (on average up to **70% decrease** in our experiments), which may be desirable for certain edge hardware.

Saliency-guided compression (SGC): The primary use case for transmitting data from camera traps is to enable users to verify the presence of wildlife. Therefore the reconstruction quality of background pixels is less important than foreground pixels, potentially enabling us to improve compression ratios further by using varying compression techniques between foreground and background.

To test this hypothesis, we utilize the pretrained wildlife detector MegaDetector [5] using the SpeciesNet codebase [10]. We utilize the bounding boxes predicted around animals by MegaDetector in two ways: (1) We modify the loss function from equation [1] to utilize different λ ratio values for pixels inside and outside predicted bounding boxes, downweighting background pixels; (2) We experiment with applying a simple Gaussian blur on background pixels, as blurring the background significantly reduces the entropy in the image while preserving most of the valuable image content. This allows for higher compression ratios due to the entropy model described above.

4 Experiments

We perform experiments using the Snapshot Kgalagadi camera trap dataset[1]. The dataset contains 3,611 sequences of camera trap images, totaling 10,222 images from 20 different camera sites. For each experiment, we fine-tune one model per site and report performance over the union of all sites.

We used the a pre-trained scale hyperprior model introduced by Ballé et al. [4]. We finetune the encoder using both standard fine-tuning and low-rank adaptation. We report results for eight different values of λ, similar to the original work, each corresponding to a specific trade-off between compression ratio and reconstruction quality. The reconstruction quality between the reconstructed image and the original image is measured by structural similarity index (SSIM). We compare our approach to JPEG compression at different quality levels as well as to the pretrained autoencoders provided by Ballé et al.. We also test our approach with and without background blurring. When using blurring, we report SSIM only on foreground pixels in all experiments for fair comparison.

Results: As shown in Figure 2, by fine-tuning the autoencoder using saliency-guided compression **(SGC)**, the fine-tuned models significantly outperforms Ballé et al's baseline model and JPEG in both SSIM and compression ratio. The LoRA fine-tuned model performs slightly worse, which is expected since LoRA typically trades some performance for improved efficiency [7].

SGC greatly helped our case. Without SGC, meaning a 1:1 ratio between λ_{animal} and $\lambda_{\text{background}}$, the improvement in compression ratio after fine-tuning is less than 2%. By changing λ_{animal} and $\lambda_{\text{background}}$ ratio to 1:0.001, we inform the model that background is less important and can be compressed more. An example of this can be shown in Figure 3. Visually, in the $\lambda_{\text{background}} = 0.001$ model, the background is more compressed, and the animal has a slightly sharper contour. This is both backed up by numeric numbers in a higher SSIM and compression ratio. As shown in Figure 2,

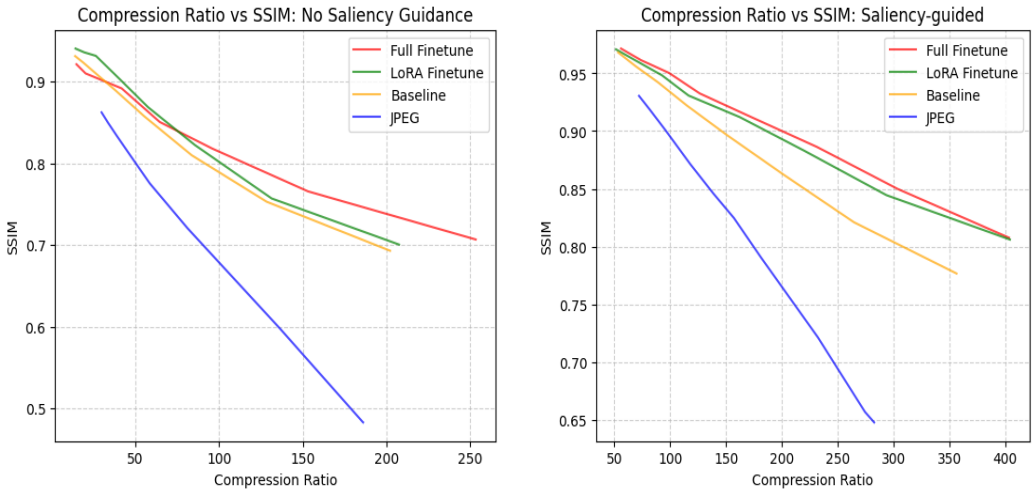
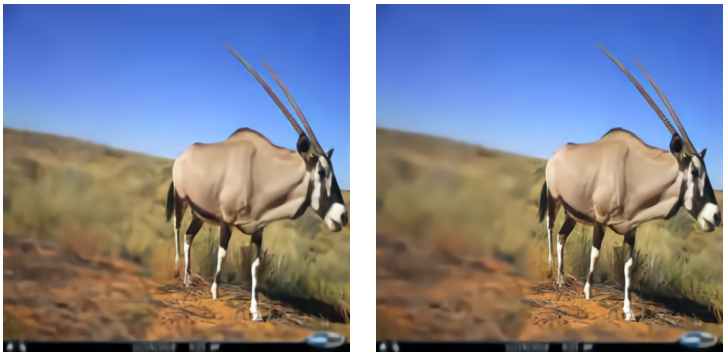


Figure 2: Results of SSIM vs. compression ratio for different configurations

peak compression ratio with saliency is 403.44. Without saliency, it is only 253.3 and has a higher reconstruction error.



(a) $\lambda_{\text{background}} = 1$ model (b) $\lambda_{\text{background}} = 0.001$ model

Figure 3: Baseline vs. finetuned model

Hardware validation: We validated our approach end-to-end on real hardware. A Jetson Orin AGX performed on-device training (full fine-tuning and LoRA) and compressed images to embeddings, which were sent via Ethernet to a Raspberry Pi. The Pi transmitted embeddings over an SX1262 long-range radio HAT [2] to a laptop with a matching HAT, achieving reliable reception at $\sim$400 m and successful image reconstruction. Unlike satellite links, long-range radios are low-cost and fee-free. Future work includes stress-testing hardware and quantizing models for lightweight deployment.

5 Conclusion

We have presented a novel method for efficient, site-specific fine-tuning of autoencoder-based image compression for wildlife camera traps. By leveraging LoRA for parameter-efficient updates and applying saliency-guided compression, we enable bandwidth-efficient model adaptation at the edge. Our experiments demonstrate the potential for this approach to offer significant compression gains and improved reconstruction quality compared to conventional codecs like JPEG. We hope to build upon this work toward a scalable solution for remote environmental monitoring with constrained satellite bandwidth. By enabling better camera traps pipelines, we can monitor and safeguard wildlife that keeps ecosystems climate-resilient.

References

[1] Snapshot kgalagadi dataset. https://lila.science/datasets/snapshot-kgalagadi. Accessed: 2025-08-19.

[2] Sx1262 868m lora hat. https://www.waveshare.com/wiki/SX1262_868M_LoRa_HAT, n.d. Accessed: 2025-07-02.

[3] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc V Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. *Advances in neural information processing systems*, 30, 2017.

[4] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1611.01704.

[5] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019.

[6] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

[7] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Conference*, 2024. arXiv preprint arXiv:2405.09673.

[8] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Deep convolutional autoencoder-based lossy image compression. In *2018 Picture Coding Symposium (PCS)*, pages 253–257. IEEE, 2018.

[9] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3146–3154, 2019.

[10] Tomer Gadot, Ștefan Istrate, Hyungwon Kim, Dan Morris, Sara Beery, Tanya Birch, and Jorge Ahumada. To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. *IET Computer Vision*, 2024.

[11] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[13] Nick Johnston, Damien Vincent, Jean-Baptiste Alayrac, Johannes Balle, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4394–4402, 2017. URL https://arxiv.org/abs/1703.10114.

[14] Jacob Kamminga, Eyuel Ayele, Nirvana Meratnia, and Paul Havinga. Poaching detection technologies—a survey. *Sensors*, 18(5):1474, 2018.

[15] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. On-device training under 256kb memory. *Advances in Neural Information Processing Systems*, 35: 22941–22954, 2022.

[17] Lisa Palmer. Hacking conservation, 2019.

[18] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2922–2930, 2017.

[19] William J Ripple and Robert L Beschta. Trophic cascades in yellowstone: the first 15 years after wolf reintroduction. *Biological Conservation*, 145(1):205–213, 2012.

[20] Jorma Rissanen and Glen Langdon. Universal modeling and coding. *IEEE Transactions on Information Theory*, 27(1):12–23, 1981.

[21] Shibani Santurkar, David Budden, and Nir Shavit. Generative compression. *arXiv preprint arXiv:1703.01467*, 2017. URL https://arxiv.org/abs/1703.01467.

[22] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[23] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Sigurd Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3738–3746, 2016.

[24] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *International Conference on Learning Representations (ICLR)*, 2017. URL `https://arxiv.org/abs/1703.00395`.

[25] George Toderici, Sean M. O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5306–5314, 2017. doi: 10.1109/CVPR.2017.577.

[26] Franck Trolliet, Cédric Vermeulen, Marie-Claude Huynen, and Alain Hambuckers. Use of camera traps for wildlife studies: a review. *Biotechnologie, Agronomie, Société et Environnement*, 18(3), 2014.

[27] Martin J. Wainwright and Eero P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. *Advances in Neural Information Processing Systems (NeurIPS)*, 12:855–861, 2000.

[28] Lei Zhou, Chunlei Cai, Yue Gao, Sanbao Su, and Junmin Wu. Variational autoencoder for low bit-rate image compression. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2617–2620, 2018.