Region of Interest Scalable Image Compression Using Semantic Communications

Prabhath Samarathunga
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
prabhath.samarathunga@strath.ac.uk

Vishnu Gowrisetty
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
vishnu.gowrisetty@strath.ac.uk

Thanuj Fernando
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
thanuj.fernando.2023@uni.strath.ac.uk

Yasith Ganearachchi
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
yasith.ganearachchi@strath.ac.uk

Anil Fernando
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
anil.fernando@strath.ac.uk

***Abstract*—Growing consumer demand for media content over a wide range of devices has made scalable image compression vital in today's media landscape. Image compression is conventionally achieved by means of statistical signal processing, but since recently, deep learning techniques are seen to be widely as well. Capabilities of such systems also enable accurate identification of regions of interest in images, leading optimised performance in most applications. This paper proposes a region-of-interest scalable image compression system using semantic communications, where an autoencoder-based semantic encoder performs the base level compression, while a Semantic Mask Extracting Transformer (SeMExT) enables identification of regions of interest to create enhancement layers with different quality levels using a scalable JPEG encoder. When benchmarked against scalable JPEG across a variety of images, the proposed system demonstrates significantly improved compressive performance. The base layer achieved 61.4 times more compression on average, along with better rate-distortion performance at any given quality level.**

***Index Terms*—Deep Neural Networks, Image Compression, Region of Interest, Scalable Image Compression, Semantic Communications**

I. Introduction

Consumer demand for high-definition media content is growing at an staggering pace, with video content accounting for over 65% of internet traffic in 2022 with a growth of 24% over the previous year [1]. In this context, image and video compression continues to be relevant, with continuous evolution of coding standards and systems attempting to provide the optimum rate-distortion (RD) performance on an increasingly wide range of applications. This is further driven by increased diversity in end user devices, necessitating adaptive media transmission techniques such as scalable image coding. This is conventionally achieved by tuning quantization levels of compression parameters resulting in a selection of images which are optimized to different screen resolutions and communication bandwidths. Alternatively, image compression can be achieved using the concept of semantic communications, where the *semantic* represents the base level compressed image, augmented by encoded residuals at a range of enhancements.

Semantic communications, based on the initial concepts discussed in [2], attempts to communicate a message by transmitting its *semantic* as a compressed version of the original message in a setup where the sender and receiver can share a common context. Recent advancements in machine learning and artificial intelligence enables the practical implementation of semantic communication systems through multiple deep learning techniques, with autoencoders (AE) being of particular interest due to their ability to be trained unsupervised, in the absence of labelled data.

RD performance of semantic communication based scalable image compression systems can be improved significantly by identifying the region of interest (ROI) and applying residual coding just to it while retaining the background from the base layer. This paper proposes a scalable image compression system using principles of semantic communications to achieve base layer compression, while scalability in the ROI extracted through a semantic mask extracting transformer (SeMExT) is achieved using a scalable residual coding method, significantly boosting RD performance compared to conventional methods. The key contributions of this paper are:

- proposing a semantic communication based ROI scalable image compression system using an AE as the semantic coder
- demonstrating SeMExT capabilities in identifying ROI and creating ROI exclusion masks.

- demonstrating superior RD performance of the system over JPEG.

II. Related Work

Scalable image compression, and progressive image coding which it is built upon, is not a new concept and has been explored since the late 1990's, with a generic structure for scalable encoders using *T-layers* being proposed by [3]. While early work on scalable image compression is based on statistical image processing techniques, more recent work are based on deep neural networks, with some reporting achievement of bitrate savings between 37% and 80% over conventional systems for detection and segmentation of objects [4].

Similarly, ROI based image compression is a widely researched topic, with early investigations done using wavelet transforms [5] and distributed source coding with side information [6]. More recent work, such as [7], demonstrates better RD performance over conventional methods, but none of these systems propose scalable image compression based on the ROI.

Semantic communications, following the concepts proposed in [2] and gaining renewed attention with the advancement in deep neural networks [8], opens new possibilities in ROI identification. Hybrids of the conventional deep neural network architectures using semantic communication concepts provide opportunities of better identifying ROI and exploiting redundancies in images, which can be utilized in novel ROI scalable image compression systems.

III. Proposed System

The proposed ROI scalable image compression system consists of three main encoder-side components and two decoder-side components. The encoder-side is made up from an AE based semantic encoder for compressing the base layer, a SeMExT based ROI extractor for identifying and extracting the ROI, and a set of JPEG based residual encoders with different quality parameters for quantizing the residuals. The decoder-side consists of an AE based semantic decoder and a set of JPEG based residual decoders with a corresponding set of quality parameters.

The original image is semantically encoded using an AE creating a latent space representation (LSR) containing the *semantic* of the image, which can be decoded to reconstruct the base layer using a pre-trained decoder. The original image is fed to the SeMExT in parallel, which creates a mask enabling extraction of ROI which can be an object of interest or an area of interest depending on image contents. The residual of the ROI is then obtained and is quantized using JPEG residual encoding with a quality parameter of 75% to represent the first enhancement level. The second and third enhancement levels are obtained using a similar process with quality parameters of 90% and 100%. The overall architecture of the proposed system is shown in Fig. 1.

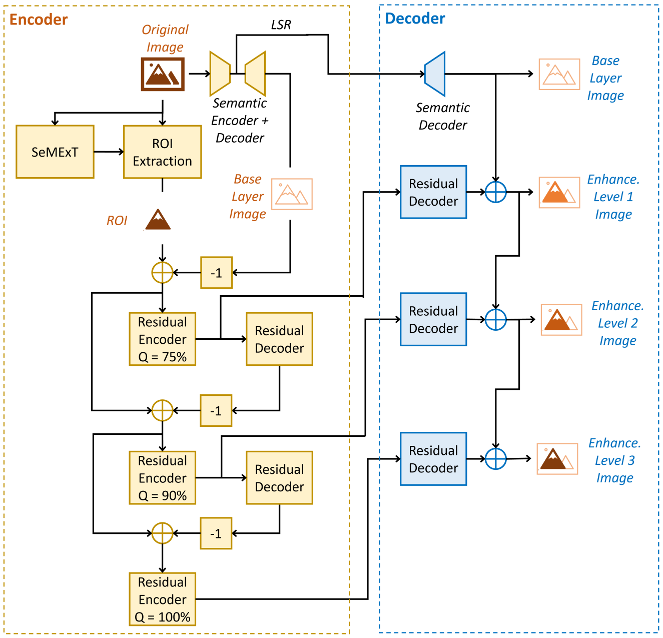

Fig. 1. Architecture of Proposed System

A. AE based semantic encoding

The semantic encoder of the proposed system is composed of the trained encoder from an convolutional AE using a rectified linear activation function for the input and a sigmoid activation function for the output, as shown in Fig. 2. The semantic decoder is the trained decoder network from the same AE, and both are trained using a series of images, of spatial resolution 192×256 in RGB colour space. The AE training is carried out through 100 epochs each with a batch size of 10, learning rate of 0.001, and to optimize the binary cross-entropy loss.

B. SeMExT based ROI extraction

SeMExT is based on the concept of vision transformers [9] and was created to extract the *semantic* aspects of images and preserve them in the form of a semantic segmentation map, and enables dynamic identification and extraction of the ROI in the proposed system. The design is trained using COCO data set [10] with weights associated with the Swin Transformer [11] to extract image specific features with a combination of supervised and self-supervised methods of learning. During the supervised learning phase the model is optimised to predict segmentation masks based on input images, and during the self-supervised learning phase the model is trained to gain meaningful representations by successfully completing pretext tasks, which includes predicting image rotations or colour changes.

The SeMExT encoder consists of 3 self-attention layers and 2 convolutional layers, with the parameter count of increasing with the complexity of the image. The SeMExT decoder consists of a dynamic instance interactive head (DIIHead), which interacts encoded characteristics from the transformer

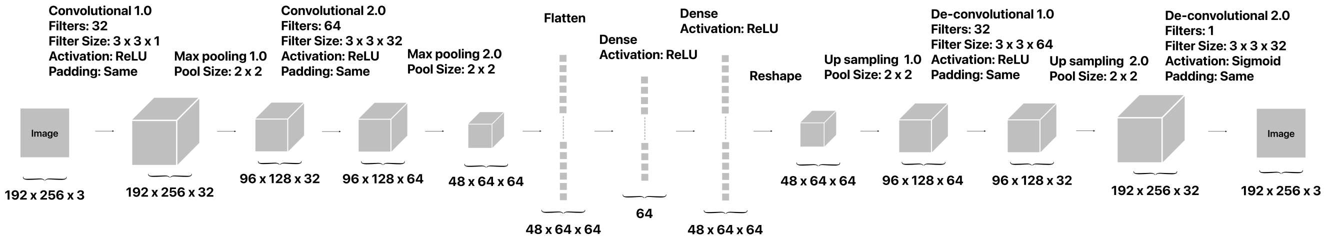

Fig. 2. Architecture of the Autoencoder used as Semantic Encoder

with learnt instance queries to provide instance-aware semantic features making it possible to discriminate between various instances of the same class, and a dynamic mask head (DMHead) which forecasts a segmentation mask for each instance using the instance-aware features from DIIHead. To determine whether a pixel belongs to a specific instance, the convolutional layer matches out the probabilities enabling creation of the desired instance segmentation map for the ROI.

C. Scalable image compression

The scalability of the system is achieved by masking and extracting the ROI from the original image using SeMExT, and then obtaining its residual compared to the decoded base layer image. This effectively results in the residual of the ROI, which is then quantized with a quality parameter of 75% to obtain the first enhancement layer. Next the residual between the input and decoded output of the first residual encoder is encoded with a quality parameter of 90% using the second residual encoder to obtain the second enhancement layer. The same process is repeated using the input and decoded output of the second residual encoder and the third residual encoder with a quality parameter of 100% to obtain the third enhancement layer.

The decoder side will reverse this process, with the decoded base layer and decoded first enhancement layer added to create the first enhanced image with ROI quality improved to 75%. The first enhanced image and decoded second enhancement layer added to create the second enhanced image with ROI quality improved to 90%. In the final stage, the second enhanced image and decoded third enhancement layer added to create the third enhanced image with ROI quality improved to 100%. The quality of outside the ROI will remain the same as the base layer.

D. Performance benchmarking

To benchmark the performance of the proposed ROI scalable image compression system, the same test images are scalably compressed using the proposed system and scalable JPEG codec. The RD performance of each is evaluated based on the peak signal to noise ratio (PSNR) and structural similarity index (SSIM) against the compression efficiency in bits/pixel (bpp).

IV. Results and Discussion

A sample set of RGB covering a wide range of complexities and spatial features with spatial resolution 192×256 are scalably encoded using the proposed system and scalable JPEG for comparison. In both cases a base layer image is created, and is then augmented by three enhancement layers encoded with scalable JPEG codec with quality levels of 75%, 90% and 100% respectively, using the same ROI masks generated using SeMExT for both systems.

Fig. 3 shows the average RD performance of the proposed system and JPEG codec over the set of test images, with three examples shown in Fig. 4a, where it is clearly evident that for a given PSNR or SSIM value the proposed system provides superior compression performance, especially in the base layer. The base layer itself from the proposed system requires nearly 61.4 times less bpp on average to encode the images while maintaining better RD performance. In terms of scalability, the proposed codec is able to outperform JPEG in each quality level with better RD performance, despite using the same residual coding.

Fig. 4b demonstrates the ROI extracted by SeMExT for each example image, and demonstrates its capability and flexibility in identifying ROI from a wide variety of images. The visual quality of the output images generated by the proposed system (Fig. 4c and Fig. 4d) and JPEG (Fig. 4e and Fig. 4f) shows that a noticeably lesser amount of artefacts are generated by the proposed system compared to JPEG codec.

The results clearly demonstrate that the proposed system has superior compressive performance compared to scalable JPEG, with an average compression gain of 61.4 in the base

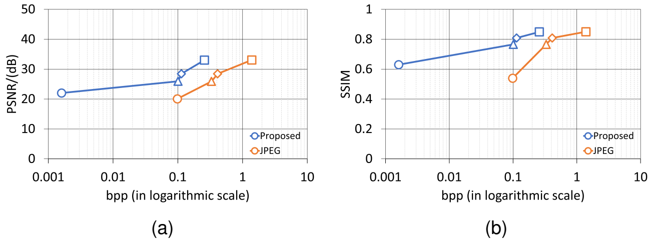

Fig. 3. Average RD Performance comparison of proposed framework and JPEG codec when comparing full images (a) PSNR vs. bpp (b) SSIM vs. bpp Symbology: ◯ - base layer, △ - ROI enhancement level 1 (75%), ◇ - ROI enhancement level 2 (90%), □ - ROI enhancement level 3 (100%).

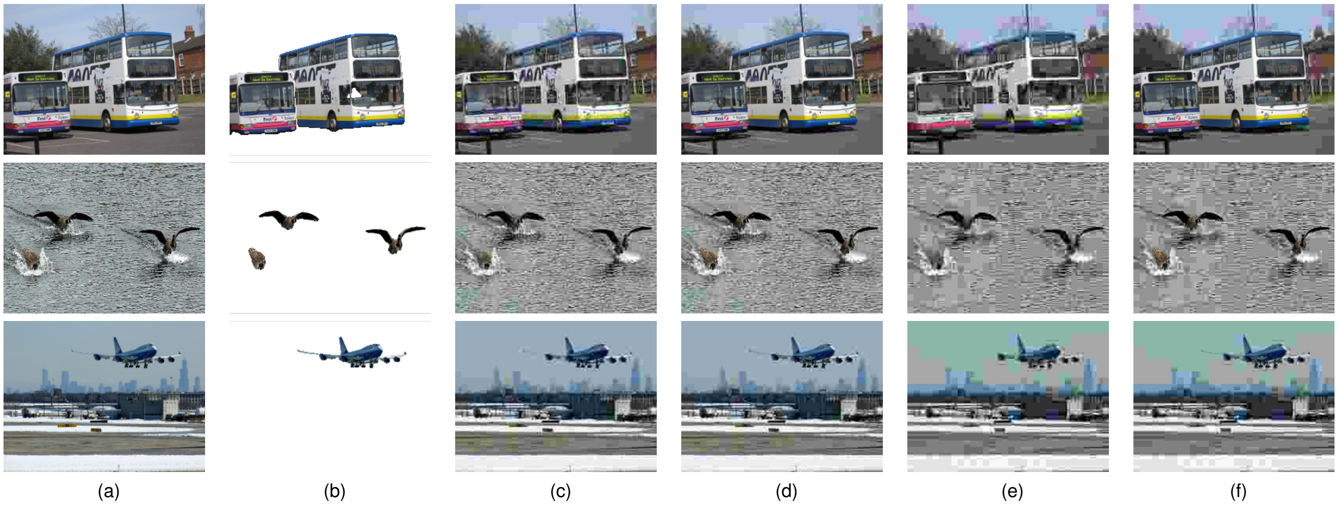

Fig. 4. Example image outputs from proposed system and JPEG. (a) Original image, (b) ROI extracted from SeMExT, (c) Base layer from proposed system, (d) ROI 100% enhanced from proposed system, (e) Base layer from JPEG, (f) ROI 100% enhanced from JPEG.

layer, 3.3 in enhancement level 1, 3.6 in enhancement level 2 and 5.4 in enhancement level 3, despite the same quantization of residuals being used for both instances. Therefore, the proposed system can be employed for image communication tasks for most human perception applications as well as for any machine-to-machine (M2M) application in situations constrained by bandwidth and storage capacity.

The main challenge in the proposed ROI scalable image compression system is its dependency on JPEG codec to perform the scalable image compression. Further investigations need to be carried out using alternatives to JPEG for residual encoding such as HEIF, as well as on distributed coding based solutions as well as neural network based solutions. Future work on the proposed system also aims to expand the system to support higher quality and more complex images, and to extend the capabilities to ROI based video coding, further evolving it towards a human and machine perception based scalable media compression framework.

V. Conclusion

This work introduces a ROI scalable image compression system based on semantic communications, where an AE based semantic encoder performs base level image compression supported by SeMExT to dynamically identify and extract the ROI and a JPEG based quantizer for scalably coding the residuals. When benchmarked against scalable JPEG codec, the proposed system exhibits significantly improved RD performance for all residual coding quality levels tested, with base level compression being 61.4 times more compared to JPEG. This makes the proposed system an ideal candidate for most human vision and machine vision applications where ROI carries more significance than the background.

Further enhancements for the proposed system can be explored by developing more efficient residual coding mechanisms, including neural network based methods, as well as by extending the results towards ROI scalable video compression. These future improvements, along with further optimization, will enable the propose system to outperform even state-of-the-art conventional scalable coding systems when ROI based scalability is required.

References

[1] Sandvine, *Phenomena: The Global Internet Phenomena Report*. Plano, TX: Sandvine, 2023. [Online]. Available: https://www.sandvine.com/phenomena

[2] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949.

[3] J.-R. Ohm, “Advances in scalable video coding,” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 42–56, 2005.

[4] H. Choi and I. V. Bajić, “Scalable image coding for humans and machines,” *IEEE Transactions on Image Processing*, vol. 31, pp. 2739–2754, 2022.

[5] A. Bruckmann and A. Uhl, “Selective medical image compression techniques for telemedical and archiving applications,” *Computers in Biology and Medicine*, vol. 30, no. 3, pp. 153–169, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482500000044

[6] G. Ding, F. Yang, Q. Dai, and W. Xu, “Distributed source coding theorem based region of interest image compression method,” *Electronic Letters*, vol. 41, no. 22, 2005.

[7] Y. Zhu, “Application-oriented region of interest based image compression using bit-allocation optimization,” *Journal of Electronic Imaging*, vol. 24, no. 1, p. 013014, 2015. [Online]. Available: https://doi.org/10.1117/1.JEI.24.1.013014

[8] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, “Semantic Communications: Principles and Challenges,” *arXiv e-prints*, p. arXiv:2201.01389, Dec. 2021.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.

[10] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll’a r, and C. L. Zitnick, “Microsoft COCO: common objects in context,” *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 9992–10 002. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00986