Semantically Structured Image Compression via Irregular Group-Based Decoupling

Ruoyu Feng[1*] Yixin Gao[1*] Xin Jin[2] Runsen Feng[1] Zhibo Chen [1,†]
[1]University of Science and Technology of China [2]Eastern Institute of Technology, Ningbo

Abstract

Image compression techniques typically focus on compressing rectangular images for human consumption, however, resulting in transmitting redundant content for downstream applications. To overcome this limitation, some previous works propose to semantically structure the bitstream, which can meet specific application requirements by selective transmission and reconstruction. Nevertheless, they divide the input image into multiple rectangular regions according to semantics and ignore avoiding information interaction among them, causing waste of bitrate and distorted reconstruction of region boundaries. In this paper, we propose to decouple an image into multiple groups with irregular shapes based on a ***customized group mask*** *and compress them independently. Our group mask describes the image at a finer granularity, enabling significant bitrate saving by reducing the transmission of redundant content. Moreover, to ensure the fidelity of selective reconstruction, this paper proposes the concept of* ***group-independent transform*** *that maintain the independence among distinct groups. And we instantiate it by the proposed* ***Group-Independent Swin-Block (GI Swin-Block)****. Experimental results demonstrate that our framework structures the bitstream with negligible cost, and exhibits superior performance on both visual quality and intelligent task supporting.*

1. Introduction

Image signal serves as a critical information carrier for various applications in modern society. Image compression techniques aim at converting images into compact representations (*i.e*. bitstreams) to save transmission and storage resources. Lossy image compression is one of the most practical techniques, as it allows for the restoration of important content while discarding a small amount of inessential information. In the past decades, traditional image compression standards [10,54,61,65,67] have been extensively studied and utilized. With the fast development of deep learning, neural image codecs [7, 8, 16, 27, 34, 40, 41, 50–52, 75] rapidly evolved and achieved promising results. Meanwhile, more and more media contents tend to be handled by machine vision algorithms, such as recognition [17,26,29,32,47], detection [43,44,55–57], and segmentation [5,9,11,12,28,46,48,69,73]. However, most compression methods are mainly developed for compressing regular rectangular images for human consumption, without considering the efficiency and functionality for downstream tasks or human-machine interaction scenarios.

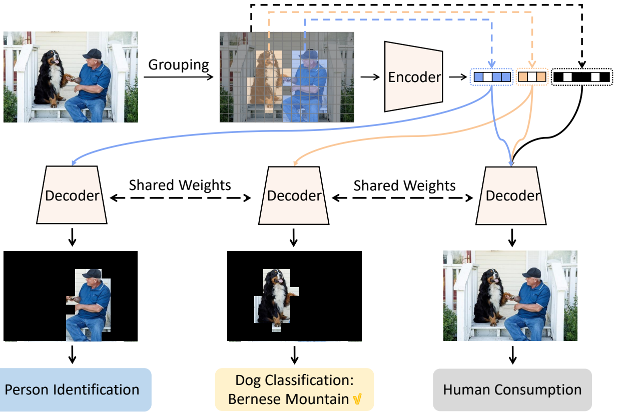


Figure 1. The input image is decoupled into groups according to distinct semantics. Then the semantically structured bitstream (SSB) is generated by compressing the image based on the partitioned groups. The SSB facilitates downstream applications by selective bitstream transmission and partial reconstruction, depending on the specific task requirements.

Recently, the field of image coding for machines (ICM) has emerged to develop a joint efficient and analytical framework for supporting intelligent analytics. End-to-end optimization of the trade-off between specific task loss and compression rate is a promising way [4, 13, 14, 18, 19, 38, 39, 49, 60], but it lacks generalization for massively diverse

* First two authors contributed equally.
† Corresponding author.

applications. To overcome this limitation, Feng *et al.* [21] propose to compress general and compact features learned by self-supervised learning under entropy constraints for supporting downstream tasks. Nevertheless, it requires retraining the task model with the proposed features as inputs, which makes the overall performance heavily depend on the effectiveness of feature extraction. The methods mentioned above are designed specifically for scenarios of compression for machine vision without considering situations in which human involvement is required. Semantically structured image compression (SSIC) [62] proposes to generate a semantically structured bitstream (SSB) by separately compressing rectangular regions of detected objects using a pre-prepared object detection toolbox. Although SSB is efficient in supporting intelligent tasks and human-machine interaction through partial transmission and reconstruction of the bitstream, its division approach based on rectangular regions may encounter problems with overlapped objects. SSIC addresses this issue by replacing the overlapping objects with a larger rectangular region, which can result in a waste of bitrate. In addition, SSIC generates the bitstream of each object by compressing the corresponding latent variables in the latent domain directly, without considering the interaction and dependence of features during the transform process. This can lead to blurry and distorted group boundaries in the partial reconstruction scenario, which in turn affects the reconstruction quality.

Going beyond the rectangular-based division, this paper proposes to decouple the image into multiple groups with irregular shapes based on a customized group mask. Then the SSB is generated by independently compressing these groups and can support various requirements via selective transmission and reconstruction. Notably, the generation of the group mask offers high flexibility in terms of shapes of groups, means of pre-analysis, and partition criteria, enabling customization to suit diverse application scenarios and requirements. Moreover, to avoid potential quality degradation in the partial reconstruction scenario, we propose the concept of the group-independent transform, which ensures the independence among groups in the latent representations and therefore the quality of the selective reconstruction will not be affected by the absence of other groups. More specifically, we instantiate it by carefully designing the Group-Independent Swin-Block (GI Swin-Block), which is an extension of Swin Transformer [47] tailored to our situation and requirements. GI Swin-Blocks make use of the hierarchical modeling capability of the Swin-Transformer [47, 75], achieving high coding efficiency under the premise of group independence. By combining the group mask based decoupling and the Group-Independent Swin-Block, our proposed method can efficiently support various downstream applications including human-machine interaction and machine vision tasks with only one bitstream generated.

The main contributions of our approach are summarized as follows:

- We propose to decouple an image into multiple groups with irregular shapes for structuring the bitstream. Our group mask can decribe spatial division at a finer granularity manner than a typical rectangle, saving bitrate by reducing redundant content transmission.
- We propose the group-independent transform and instantiate it by carefully designing the Group-Independent Swin-Block (GI Swin-Block), which maintains powerful transformation capability and ensures the independence among groups in the latent representations.
- Experimental results demonstrate that our proposed model achieves state-of-the-art compression ability and superior downstream tasks performance, which is a codec with both high compression efficiency and functionality.

2. Related Works

2.1. Image Compression

Traditional Image Compression. Traditional image compression standards, such as JPEG [65], JPEG2000 [54], HEVC [61], and VVC Intra [10], have been extensively used in practice after several decades of development. These standards rely on transform coding [23], which decomposes the lossy image compression task into three parts: transform, quantization, and entropy coding. Each module of these standards is manually designed with multiple modes, and rate-distortion optimization is performed to determine the optimal mode. However, the completely hand-crafted structure of traditional codecs limits their flexibility and scalability to support various objectives, such as MS-SSIM and classification accuracy, as they cannot be optimized in an end-to-end manner.

Learned Image Compression. In recent years, learned image compression methods based on nonlinear transform coding [6] have achieved rapid progress. Early works in this area concentrated on enabling end-to-end training by developing differential quantization and rate estimation techniques [1, 7, 63]. Subsequently, a considerable amount of work design powerful neural network modules to enhance the transform [16, 75], quantization [25, 70], and entropy model [8, 27, 52, 53]. As a result, some neural image codecs [16, 24, 27, 52, 53] can achieve comparable or even superior performance to traditional coding standards like HEVC [61] and VVC [10]. Moreover, some works [2,51,58] introduce perceptual loss to effectively improve reconstruction quality. However, existing image compression methods focus on compressing the entire image without considering selective transmission and reconstruction of the bitstream for arbitrary reconstruction, leading to significant bandwidth waste when addressing downstream applications with different requirements.

2.2. Image Coding for Machines

Image coding for machines aims at compress source images to serve downstream tasks, such as recognition [17, 26, 29, 32, 47], detection [43, 44, 55–57], and segmentation [5, 9, 11, 12, 28, 46, 48, 69, 73]. Joint optimization of task loss and bitrate [3, 31, 38, 42, 66] in an end-to-end manner is a natural and promising way. Another direction is compress the features corresponding to downstream tasks [4,13,14,18,19,49,60]. However, these methods tend to be biased towards specific tasks and may lack generalizability to a wide range of intelligent applications. To address the limitation of generalization, a novel representation learning based method is proposed [21] to learn general and compact features by combining contrastive learning with entropy constraints. The learned features are used to replace the original images as the new source for compression and transmission, resulting in a significant improvement in coding efficiency across various intelligent tasks. Nevertheless, this approach necessitates re-training the task model using the proposed features as input, and its overall performance is heavily reliant on the effectiveness of feature extraction. In this paper, we propose a unified compression framework for machine vision support without task-oriented joint training or additional feature learning. Moreover, it can also conduct full or selective reconstruction for human perception according to requirements.

2.3. Object-oriented Image Compression

Object-oriented image compression is first proposed in Mpeg-4 Visual [20, 35, 59] by compressing visual object planes (VOP) with arbitrary shapes, targeting content-based interactivity and scalability for the image compression technique. However, due to the difficulty in annotating the complex and intricate shapes of the target and the lack of actual corresponding downstream applications, MPEG-4 Visual has not been widely used. Recently, with the fast development of deep learning [29, 64], the practical scenarios of image compression for downstream applications have increased rapidly. Sun *et al*. [33, 62] propose the semantically structured image compression (SSIC) of decoupling the image into objects and compressing the corresponding latent variables independently of the neural image codec to generate the semantically structured bitstream. Nevertheless, the rectangular-based division adopted by SSIC lacks generalization to irregular objects and flexibility for customization. Additionally, selecting the representation of objects directly in the latent space ignores the interaction among elements during the encoding transform process, leading to significant edge distortion of objects in partial reconstruction. This paper proposes to conduct semantically structured image compression by decoupling the image into irregular groups based on the group mask, which is flexible and customizable. Furthermore, we propose the group-independent transform conducted during both the encoding and decoding processes, which enables efficient partial or complete reconstruction of images under various requirements.

3. Method

3.1. Overview

The proposed method is an efficient and flexible version of semantically structured image compression [62]. It can satisfy multiple application requirements via transmitting and reconstructing partial spatial regions guided by a block-wise group mask. Our model takes the group mask as guidance to ensure that interactions of transform only occur within the same group, thereby achieving group independence during the redundancy removal process. The group mask, which is generated based on the pre-analysis, such as object detection, semantic segmentation, and saliency detection, provides high flexibility and customization for structuring the bitstream and is considered as side information. Then the entropy coding process is conducted on the variables of each group distinctly, resulting in a bitstream structured by semantics. The bitstream can be partially or fully transmitted according to the requirements of the decoder side, and then the receiver conducts entropy decoding on the bitstream and reorganizes the spatial-wise arrangement of latent variables according to the group mask and the group indexes.

The overall network structure is illustrated in Fig. 2, which incorporates the group mask and GI Swin-Block for guiding and instantiating the group-independent transform, respectively. Section 3.2 and Section 3.3 introduce the group mask and group independent, respectively. The detailed implementation of our network architecture is presented in Section 3.4.

3.2. Group Mask

Before compression, the group mask is generated according to the results of pre-analysis techniques such as object detection, instance segmentation, and saliency detection. Figure 3 provides an example of this process. More specifically, for the input image x with a height of H and a width of W, the spatial resolution of the group mask is the same as the input image, and it consists of $\frac{H}{B} \times \frac{W}{B}$ blocks, where B is the side length of each block. And the value of it indicates the index of the corresponding group.

Our proposed group mask can clearly divide the elements in the latent variable space after the one-pass down-

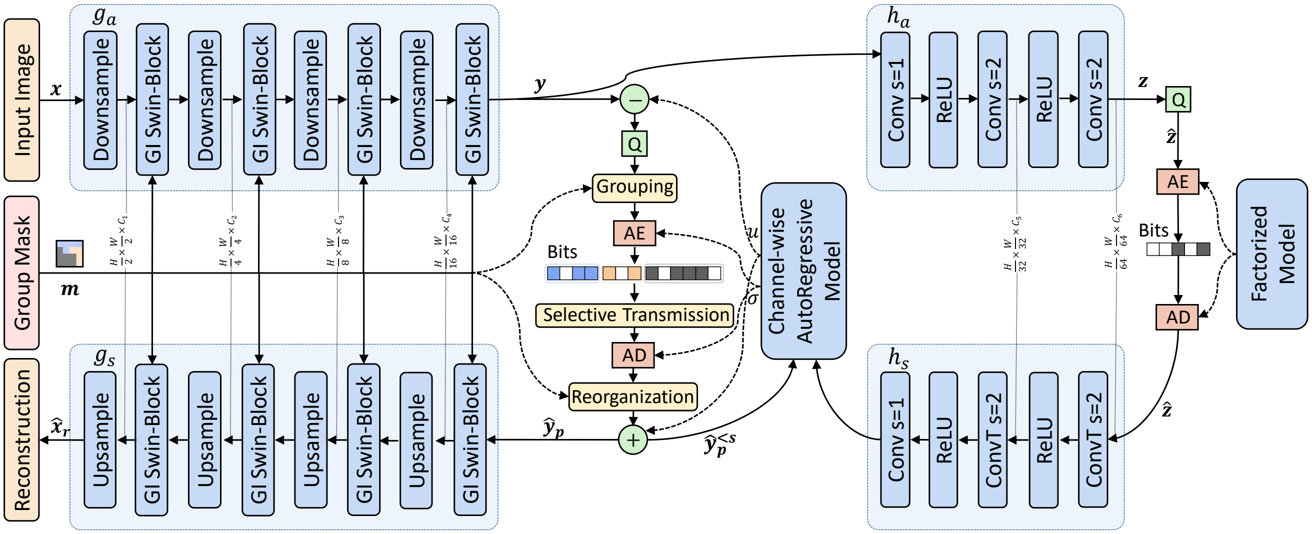


Figure 2. The network architecture of our proposed model with the channel-wise auto-regressive model (ChARM). ConvT denotes transposed convolution. AE and AD are respectively arithmetic encoding and arithmetic decoding. In Ours-Hyper model, we remove the ChARM component and instead output μ and σ directly from the hyperdecoder h_s.

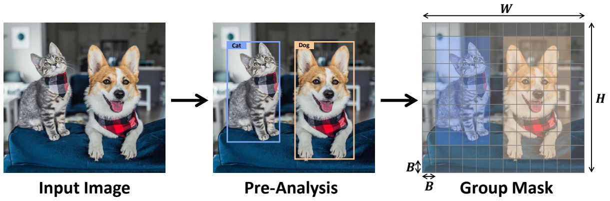


Figure 3. An example of group mask generation. The input image is pre-analyzed by object detection, then the group mask is generated based on the results of the pre-analysis.

sampling transform. Besides, downsampling the group mask by a factor of B before conducting compression can greatly reduce the overhead bitrate. The block size and the overhead bitrate have an inverse relationship. Furthermore, as shown in Figure 4, the generation of the group mask can be flexibly customized, in terms of the way of pre-analysis and the criteria of block allocation, *etc*.

3.3. Group-Independent transform

Lossy image compression based on transform coding can be divided into three modularized components: transform, quantization, and entropy coding. While quantization and entropy coding do not affect the generation and usage of the semantically structured bitstream (SSB), traditional transform applied to the entire image to remove spatial redundancy inevitably creates inter-group dependencies during compression. In situations of selective transmission and reconstruction, the incomplete representation that results from the reorganization of a subset of all groups can lead to inaccurate reconstruction due to the lack of inter-group dependencies. Therefore, we propose the concept of the group-independent transform for semantically structured bitstream generation. The basic idea is to constrain the transform to be conducted only within each group.

A straight and natural way is by using transformer [64] with the customized attention map corresponding to the group mask. However, image compression differs significantly from high-level understanding tasks, which the transformer excels at. The presence of numerous complex long-range dependencies can impede convergence and adversely affect the final performance. In order to achieve high coding efficiency while maintaining the group-independent property, and inspired by the strong hierarchical representation modeling capability of Swin-Transformer [47, 75], we extend the Swin-Block of Swin-Transformer to our proposed group-independent Swin-Block (GI Swin-Block). The GI Swin-Block serves as the core component of the transform. To be more specific, as shown in Figure 5a, by merging the window partition and the group partition into the group-independent window partition, allowing the self-attention to be computed within the partitioned local regions, enabling us to achieve high coding efficiency while maintaining the group-independent property. Additionally, the cross-window connections are introduced by merging the shifted window partition and group partition similarly, as depicted in Figure 5b.

3.4. Network architecture and pipeline

Figure 2 illustrate the proposed architecture. The main transform is performed in a staggered manner through

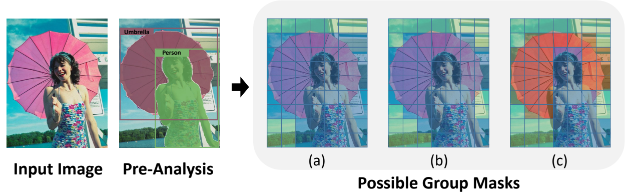


Figure 4. Illustration of flexibility to customize the group mask: (a). Taking overlapping objects as one group with the guidance of bounding boxes. (b). Taking overlapping objects as one group with the guidance of instance masks. (c) Taking overlapping objects as distinct groups with the guidance of instance masks.

the incorporation of upsample/downsample and group-independent Swin-Block (GI Swin-Block). Upsample contains a pixel shuffle operation and a convolution layer with 1×1 kernel size, whereas downsample contains a pixel unshuffle operation and a convolution layer with 1×1 kernel size. For entropy coding, we adopt the Mean & Scale (M&S) Hyperprior model [52] with the strong channel-wise auto-regressive model (ChARM) [53] to predict the probability distribution of latent variables.

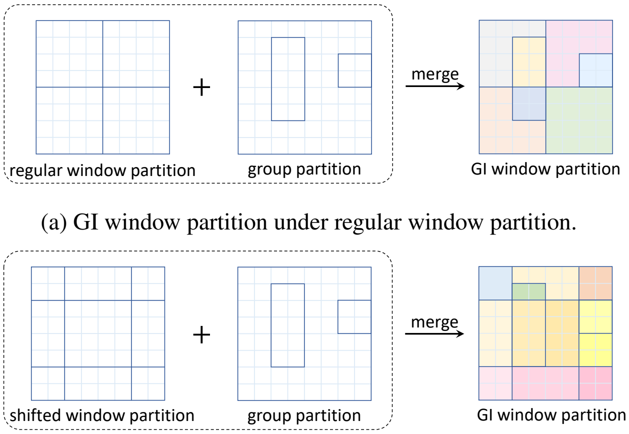
Specifically, given a source image $\boldsymbol{x}$, the encoder g_a converts it to latent representations $\boldsymbol{y}$ conditioned on the given group mask $\boldsymbol{m} = \sum_{i=1}^{N} \boldsymbol{m}_i$, where $i \in \{1, 2, 3, \dots, N\}$, indicating the index of the corresponding group in all N groups. In each $\boldsymbol{m}_i$, the values of positions corresponding to the i_{th} group are set as i, and all values of other positions are set as 0. Similarly, we can define $\boldsymbol{x} = \sum_{i=1}^{N} \boldsymbol{x}_i$ and $\boldsymbol{y} = \sum_{i=1}^{N} \boldsymbol{y}_i$ from the perspective of group, and transform only occurs within each group, which is given by

$$\boldsymbol{y}_i = g_a\left(\boldsymbol{x}_i | \boldsymbol{m}_i\right). \tag{1}$$

Thanks to the group-independent transform, we can easily implement the transform of all groups within one forward-pass, thus, the process can also be written as $\boldsymbol{y} = g_a\left(\boldsymbol{x} | \boldsymbol{m}\right)$.

Then the latents are quantized to discrete representations $\hat{\boldsymbol{y}}$ and selectively transmitted to the decoder side with predicted probability distribution through the entropy model. Concretely, the side information $\boldsymbol{z}$ is extracted by the hyper-encoder $\boldsymbol{z} = h_a\left(\boldsymbol{y}\right)$. The quantized hyper-latent $\hat{\boldsymbol{z}} = Q\left(\boldsymbol{z}\right)$ is modeled and entropy-coded with a learned factorized prior. In ChARM, latent $\hat{\boldsymbol{y}}$ is split into S (we choose S=10) slices along with channel dimension, and each slice $\hat{\boldsymbol{y}}_s$ is entropy-coded based on the previous slices $\hat{\boldsymbol{y}}_{<s}$. $\hat{\boldsymbol{y}}$ is modeled by a conditional Gaussian distribution convolving with a unit uniform noise.

After selectively entropy decoding the quantized latents, we have $\hat{\boldsymbol{y}}_p = \sum_{j \in \mathbf{P}} \hat{\boldsymbol{y}}_j$, where $\mathbf{P} \subseteq \{1, 2, \dots, N\}$ is the gather of required groups. The reconstruction image $\hat{\boldsymbol{x}}_p = \sum_{j \in \mathbf{P}} \hat{\boldsymbol{x}}_j$ will be calculated as follows



(a) GI window partition under regular window partition.

(b) GI window partition under shifted window partition.

Figure 5. Group-independent window partition of GI Swin-Block under regular and shifted window partition. Self-attention is conducted inside each local region.

$$\hat{\boldsymbol{x}}_j = g_s\left(\hat{\boldsymbol{y}}_j | \boldsymbol{m}_j\right). \tag{2}$$

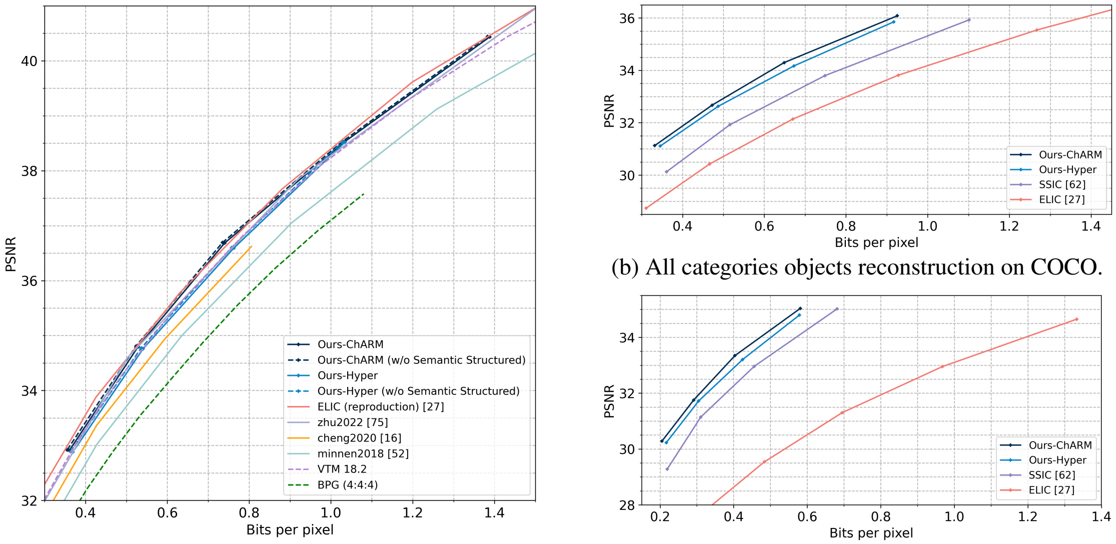
Similar to Equation (1), Equation (2) can also be written as $\hat{\boldsymbol{x}} = g_s\left(\hat{\boldsymbol{y}} | \boldsymbol{m}\right)$. Note that the rounding operation is non-differentiable, thus we apply uniform noise for learning the entropy model and use a rounded one as the input of the decoder g_s as in [53].

4. Experiments

4.1. Experiment Setup

Image Codec. We train our high bitrate models (>= 0.6bpp) on the Flickr 2K dataset, which has the same setting as [74] and train our low bitrate models (< 0.6bpp) on the COCO 2017 training set [45]. Ours-Hyper are trained with 3.2M iterations, while Ours-ChARM models are trained by initializing their weights from higher bitrate hyperprior models for another 1.2M iterations. Each batch contains 8 random 256×256 crops from the training dataset. The learning rate is set as $5e-5$ and is decayed by a factor of 10 at $2.8M$ iterations. Training loss $L = R + \beta D$ is the weighted combination of the rate-distortion trade-off, balanced by the Lagrange multiplier β. We adopt the mean squared error (MSE) in the RGB color space as the distortion metric D. The bitrate range of our image codecs is achieved by applying different β that $\beta \in \{512, 1024, 2048, 4096\}$. During the training stage, we randomly generate the group masks for the training of the group-independent transform, the detailed operations are shown in the supplementary material.

Evaluation Datasets and Protocol. For the entire image reconstruction quality, we use the widely-used Kodak dataset [37] to evaluate the coding efficiency of our models. We evaluate the selective reconstruction quality on objects



(a) Entire image reconstruction on Kodak.

(b) All categories objects reconstruction on COCO.

(c) Human category objects reconstruction on COCO.

Figure 6. Rate distortion comparison of compression efficiency on both entire and partial reconstruction scenarios. The PSNR of partial reconstructions is only calculated on the pre-detected bounding boxes.

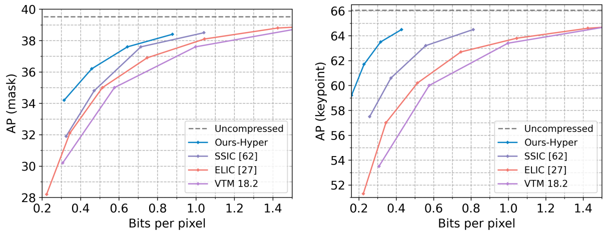


Figure 7. Performance comparison on instance segmentation (left) and pose estimation (right) on COCO.

of all categories and human objects using 40 and 20 images, which are randomly selected from the COCO 2017 validation set. The PSNR of the region of interest serves as the metric for measuring the objective quality. Additionally, the bits per pixel (bpp) are calculated by dividing the total bits of the transmitted bitstream by the number of pixels in the region of interest. To verify the effectiveness of our proposed method in supporting downstream tasks, we conduct experiments on instance segmentation and pose estimation using the COCO 2017 validation set. [45]. This dataset is widely used for dense prediction tasks and comprises 118K training, 5K validation, and 20K test-dev images. We evaluate the performances of instance segmentation and human keypoint detection using Mask R-CNN (X101-FPN backbone) and Keypoint R-CNN (X101-FPN backbone), respectively, implemented in the Detectron2 toolbox [68].

4.2. Quantitative Results

For comparisons, we choose recent neural image codecs [16,27,52,75], powerful classical image codecs [10, 61], and the functional codec that supports semantically structured bitstream [62]. Among them, ELIC [27] is reproduced and performs closely as their report. It is critical to emphasize that the encoder does not possess prior knowledge regarding the regions of interest required by the decoder. And we utilize object detection as the pre-analysis by default, which is one of the most fundamental visual analysis tools. The semantically structured bitstreams are generated based on the bounding boxes detected by the detector of Mask R-CNN (X101-FPN backbone), and are selectively decoded and reconstructed into images for evaluation according to the requirements of relative semantics. In all experiments of quantitative results, the block size B of group masks is set to 32 to balance the trade-off between performance and overhead bitrate, and we take the union of overlapping groups as one group. More experiments compared with other methods of image coding for machines are presented in the supplementary.

Reconstruction Quality. For entire image reconstruction, Figure. 6a demonstrates that our proposed model achieves state-of-the-art rate-distortion performance, surpassing VTM 18.2 in terms of PSNR at all bitrates. Notably, structuring the bitstream according to semantics introduces negligible effects. Therefore, our model has a strong potential to achieve both high coding efficiency and functionality simultaneously. For partial image reconstruction in scenarios of specific interest, as shown in Figure. 6b and Figure. 6c, our model has achieved a significant improvement compared to other codecs. Specifically, the semantically structured bitstream enabled our model and SSIC [62] to avoid transmitting and decoding bitstreams corresponding to the entire image, which is required of ELIC [27]. Additionally, our model's flexible block-level group mask and group-independent transform spatially decouple latents from the semantic level, resulting in sharper and more re-

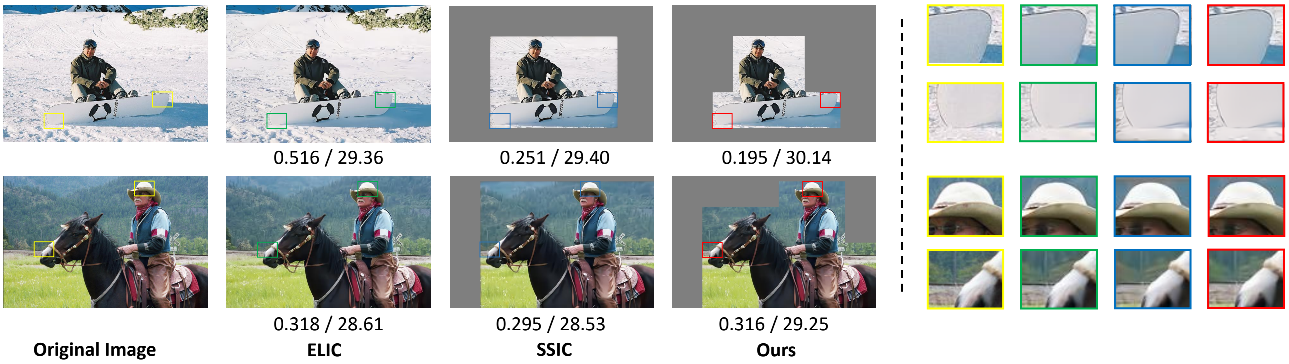


Figure 8. Visualization of reconstructed images for human consumption from the COCO 2017 dataset, in the scenario where foreground objects are the regions of interest. The numbers below the images correspond to bits per pixel and PSNR.

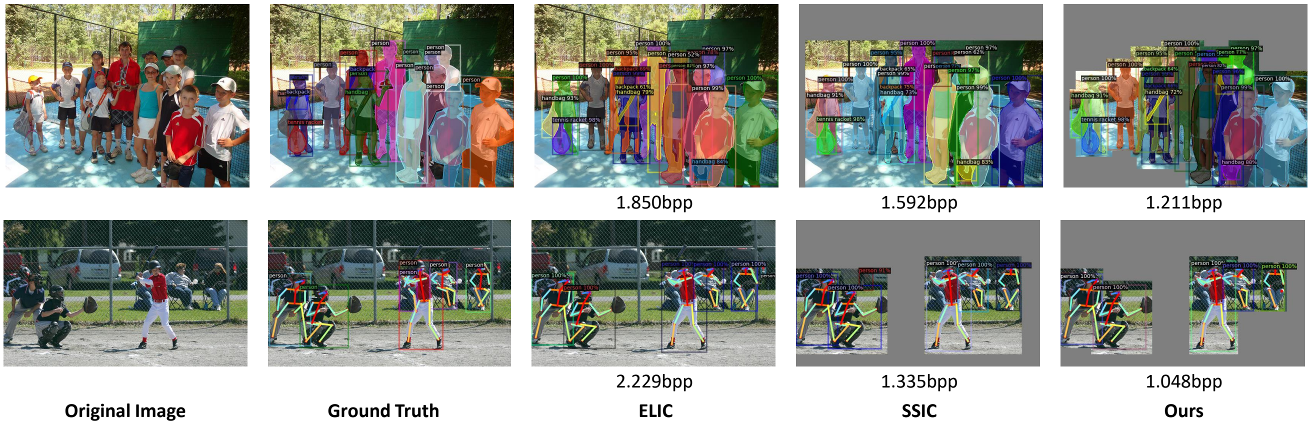


Figure 9. Visualization of reconstructed images for downstream tasks from the COCO 2017 dataset. The first line corresponds to instance segmentation, and the second one corresponds to pose estimation.

alistic boundaries compared to SSIC [62]. Moreover, in the case of overlapping regions, our model can significantly save bitrate by grouping them into an irregular group instead of replacing them with a larger bounding box, which would introduce more distortion.

Downstream Task Supporting. Figure. 7 demonstrates our excellent performance on instance segmentation and pose estimation. On instance segmentation, our model achieves significantly improved performance at low bitrates (<0.6bpp) compared with other methods, which can be attributed to the faithful reconstruction of RoI boundaries in our approach. It is worth noting that the superior performance of our method on pose estimation is due to the ability of our model to preserve fine details and boundaries, which are critical for accurately localizing human keypoints. In addition, human objects are a minority in the image and are often sparsely distributed, and compared with SSIC, our method can significantly reduce bitrate by avoiding transmitting the latents corresponding to the rectangular area containing all overlapped objects.

4.3. Qualitative Results

Reconstruction for Human Consumption. When the reconstructed images are mainly intended for human consumption with specific semantic interests, as shown in Figure 8, bitstreams generated by codecs designed for compressing regular rectangular images are required to be fully transmitted and serve full reconstruction, regardless of the requirements and the content of images. Although SSIC [62] partially addresses this issue through generating SSB through bounding box based division, it lacks flexibility and may introduce additional irrelevant disturbance when merging overlapping objects into larger bounding boxes. Moreover, SSIC generates the bitstream by directly dividing the corresponding positions in the latent variable space after the transform based on ConvNets, which can lead to lost dependencies during selective reconstruction, resulting in blurred and distorted region boundaries. With group mask based partitioning and group-independent transform, our method decouples images to generate the SSB more efficiently while ensuring that no distortion or blurring is introduced after selective reconstruction, leading to both significant bitrate saving and a better visual experience.

Reconstruction for Intelligent Tasks. When using reconstructed images for downstream intelligent tasks, selectively transmitting and decoding the semantically structured bitstream based on the prior knowledge of relative seman-

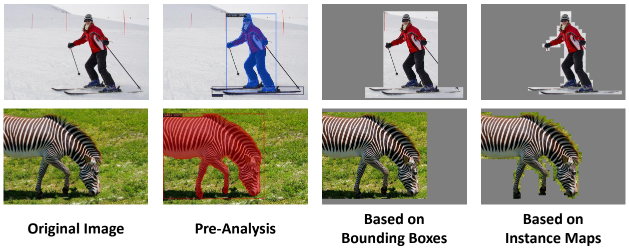


Figure 10. Selective reconstructions from the semantically structured bitstream. The bitstream can be generated based on different pre-analysis methods.

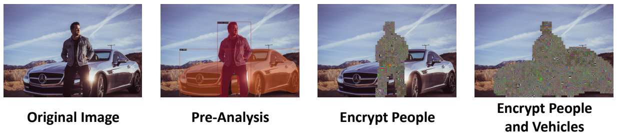
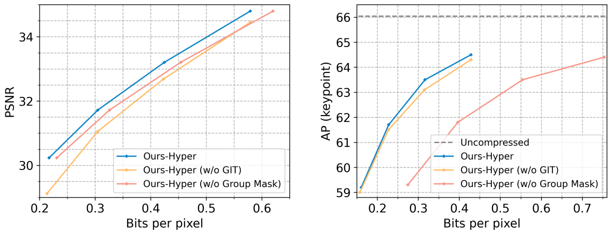


Figure 11. Illustration of semantically-aware encryption based on our proposed method.

tics can significantly save bitrate. As illustrated in Figure 9, conventional codecs designed for compressing rectangular images require the entire bitstream to be transmitted and decoded. Downstream models then perform intelligent analytics on the fully reconstructed image, resulting in a waste of bitrate. Similar to image reconstruction for human consumption, SSIC can perform selective transmission to save bitrate. However, its bounding box based partitioning may not be optimal for bitrate savings, and the resulting blurred and distorted region boundaries and irrelevant content can further impede downstream intelligent analytics. (misidentifying the car as a person shown in Figure 9). Our method can more efficiently generate the SSB and support various downstream tasks with specific requirements with both accuracy and coding efficiency.

Customizablity and Flexibility. It's flexible to customize the semantically structured bitstream based on different semantical partition criteria. For instance, as shown in Figure 10, the generation and usage of semantically structured bitstreams can be based on either object detection or instance segmentation, depending on the specific requirements. It should be noted that the methods of pre-analysis are not limited to object detection and instance segmentation as shown in Figure 10, but can include other techniques such as saliency detection [22, 30, 72], semantic segmentation [5, 48, 69, 71, 73], panoptic segmentation [15, 36], *etc*., and even human annotation.

Semantically-Aware Encryption. As presented in Figure 11, our proposed method can be applied to bitstream encryption, enabling selective and even layered encryption with semantic priors based on the user's security level. In some cases, the selectively encrypted SSB can allow for secure transmission and storage of sensitive information while minimizing the impact on visual quality and downstream analytics. Please refer to the supplementary material for details of implementation.



(a) RoI reconstruction. (b) Intelligent task supporting.

Figure 12. Ablation study.

5. Ablation Study

In order to show the effectiveness of our proposed group mask and group-independent transform (GIT) for both RoI-aware selective reconstruction and downstream intelligent analytics, we conducted the ablation study at different bitrates. Specifically, we use people as the region of interest for RoI reconstruction, and the results for taking all foreground objects as RoI are presented in the supplementary material. Additionally, we conducted an ablation study on pose estimation of COCO 2017 for intelligent task support, and the results on instance segmentation are also included in the supplementary material.

Group Mask. Figure 12 shows that group mask based partitioning saves significant bitrate compared with naive bounding box based partitioning. Meanwhile, both the visual quality of RoI reconstruction and downstream task performances would not be impeded.

Group-Independent transform. As shown in Figure 12a, group-independent transform is crucial for the visual quality of RoI reconstruction. And the downstream task performance would be impeded by the blurred and distorted region boundaries without group-independent transform, which is shown in Figure 12b.

6. Conclusion

In this work, we propose to generate semantically structured bitstream with strong functionality based on group masks, which are highly flexible, customizable, while maintaining lightweight overhead. Moreover, we first propose the concept of group-independent transform and instantiate it by designing the Group-Independent Swin-Block (GI-Swin Block) to ensure independence among distinct groups, resulting in pleasing reconstructions of RoI regions. Particularly, our proposed method outperforms VTM-18.2 and achieves comparable coding efficiency with the SOTA neural image codecs, while offering strong functionality. The experimental results have demonstrated the effectiveness of our proposed method across several applications, including human consumption of regions of interest, downstream task support, and semantically-aware encryption.

Acknowledgement. This work was supported in part by NSFC under Grant U1908209, 62021001. This work was also supported in part by ZJNSFC under Grant LQ23F010008.

References

[1] 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. *NeurIPS*, 30, 2017. 2

[2] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *ICCV*, pages 221–231, 2019. 2

[3] Mohammad Akbari, Jie Liang, and Jingning Han. Dsslic: Deep semantic segmentation-based layered image compression. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2042–2046. IEEE, 2019. 3

[4] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *ECCV*, pages 584–599. Springer, 2014. 1, 3

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12):2481–2495, 2017. 1, 3, 8

[6] Johannes Ballé, Philip A Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2020. 2

[7] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *ICLR*, 2017. 1, 2

[8] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *ICLR*, 2018. 1, 2

[9] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, pages 9157–9166, 2019. 1, 3

[10] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *TCSVT*, 2021. 1, 2, 6

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017. 1, 3

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 1, 3

[13] Zhuo Chen, Kui Fan, Shiqi Wang, Lingyu Duan, Weisi Lin, and Alex Chichung Kot. Toward intelligent sensing: Intermediate deep feature compression. *TIP*, 29:2230–2243, 2019. 1, 3

[14] Zhuo Chen, Kui Fan, Shiqi Wang, Ling-Yu Duan, Weisi Lin, and Alex Kot. Lossy intermediate deep learning feature compression and evaluation. In *ACM MM*, pages 2414–2422, 2019. 1, 3

[15] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, pages 12475–12485, 2020. 8

[16] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *CVPR*, pages 7939–7948, 2020. 1, 2, 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. 1, 3

[18] Ling-Yu Duan, Vijay Chandrasekhar, Jie Chen, Jie Lin, Zhe Wang, Tiejun Huang, Bernd Girod, and Wen Gao. Overview of the mpeg-cdvs standard. *TIP*, 25(1):179–194, 2015. 1, 3

[19] Ling-Yu Duan, Yihang Lou, Yan Bai, Tiejun Huang, Wen Gao, Vijay Chandrasekhar, Jie Lin, Shiqi Wang, and Alex Chichung Kot. Compact descriptors for video analysis: The emerging mpeg standard. *IEEE MultiMedia*, 26(2):44–54, 2018. 1, 3

[20] Touradj Ebrahimi and Caspar Horne. Mpeg-4 natural video coding–an overview. *Signal Processing: Image Communication*, 15(4-5):365–385, 2000. 3

[21] Ruoyu Feng, Xin Jin, Zongyu Guo, Runsen Feng, Yixin Gao, Tianyu He, Zhizheng Zhang, Simeng Sun, and Zhibo Chen. Image coding for machines with omnipotent feature learning. In *ECCV*, pages 510–528. Springer, 2022. 2, 3

[22] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *TPAMI*, 34(10):1915–1926, 2011. 8

[23] Vivek K Goyal. Theoretical foundations of transform coding. *IEEE Signal Processing Magazine*, 18(5):9–21, 2001. 2

[24] Zongyu Guo, Zhizheng Zhang, Runsen Feng, and Zhibo Chen. Causal contextual prediction for learned image compression. *TCSVT*, 32(4):2329–2341, 2021. 2

[25] Zongyu Guo, Zhizheng Zhang, Runsen Feng, and Zhibo Chen. Soft then hard: Rethinking the quantization in neural image compression. In *ICML*, pages 3920–3929. PMLR, 2021. 2

[26] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *NeurIPS*, 34:15908–15919, 2021. 1, 3

[27] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *CVPR*, pages 5718–5727, 2022. 1, 2, 6

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1, 3

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 3

[30] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8. Ieee, 2007. 8

[31] Yueyu Hu, Shuai Yang, Wenhan Yang, Ling-Yu Duan, and Jiaying Liu. Towards coding for human and machine vision: A scalable image coding approach. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 3

[32] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *ECCV*, 2022. 1, 3

[33] Xin Jin, Ruoyu Feng, Simeng Sun, Runsen Feng, Tianyu He, and Zhibo Chen. Semantically video coding: Instill static-dynamic clues into structured bitstream for ai tasks. *arXiv preprint arXiv:2201.10162*, 2022. 3

[34] Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *CVPR*, pages 4385–4393, 2018. 1

[35] Aggelos K Katsaggelos, Lisimachos P Kondi, Fabian W Meier, Jörn Ostermann, and Guido M Schuster. Mpeg-4 and rate-distortion-based shape-coding techniques. *Proceedings of the IEEE*, 86(6):1126–1154, 1998. 3

[36] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 8

[37] Eastman Kodak. Kodak lossless true color image suite (photocd pcd0992). http://r0k.us/graphics/kodak, 1993. 5

[38] Nam Le, Honglei Zhang, Francesco Cricri, Ramin Ghaznavi-Youvalari, and Esa Rahtu. Image coding for machines: An end-to-end learned approach. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1590–1594. IEEE, 2021. 1, 3

[39] Nam Le, Honglei Zhang, Francesco Cricri, Ramin Ghaznavi-Youvalari, Hamed Rezazadegan Tavakoli, and Esa Rahtu. Learned image coding for machines: A content-adaptive approach. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1

[40] Mu Li, Wangmeng Zuo, Shuhang Gu, Jane You, and David Zhang. Learning content-weighted deep image compression. *TPAMI*, 2020. 1

[41] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *CVPR*, pages 3214–3223, 2018. 1

[42] Xin Li, Jun Shi, and Zhibo Chen. Task-driven semantic coding via reinforcement learning. *TIP*, 2021. 3

[43] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ECCV*, 2022. 1, 3

[44] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 1, 3

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5, 6

[46] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768, 2018. 1, 3

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1, 2, 3, 4

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1, 3, 8

[49] Siwei Ma, Xiang Zhang, Shiqi Wang, Xinfeng Zhang, Chuanmin Jia, and Shanshe Wang. Joint feature and texture coding: Toward smart video representation via front-end intelligence. *TCSVT*, 29(10):3095–3105, 2018. 1, 3

[50] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *CVPR*, pages 4394–4402, 2018. 1

[51] Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *arXiv preprint arXiv:2006.09965*, 2020. 1, 2

[52] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *NeurIPS*, 2018. 1, 2, 5, 6

[53] David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343. IEEE, 2020. 2, 5

[54] Majid Rabbani and Rajan Joshi. An overview of the jpeg 2000 still image compression standard. *Signal processing: Image communication*, 17(1):3–48, 2002. 1, 2

[55] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1, 3

[56] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017. 1, 3

[57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28:91–99, 2015. 1, 3

[58] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *ICML*, pages 2922–2930. PMLR, 2017. 2

[59] Thomas Sikora. The mpeg-4 video standard verification model. *TCSVT*, 7(1):19–31, 1997. 3

[60] Saurabh Singh, Sami Abu-El-Haija, Nick Johnston, Johannes Ballé, Abhinav Shrivastava, and George Toderici. End-to-end learning of compressible features. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3349–3353. IEEE, 2020. 1, 3

[61] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *TCSVT*, 22(12):1649–1668, 2012. 1, 2, 6

[62] Simeng Sun, Tianyu He, and Zhibo Chen. Semantic structured image coding framework for multiple intelligent applications. *TCSVT*, 2020. 2, 3, 6, 7

[63] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017. 2

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3, 4

[65] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992. 1, 2

[66] Shurun Wang, Shiqi Wang, Wenhan Yang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Towards analysis-friendly face representation with scalable feature and texture compression. *TMM*, 2021. 3

[67] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *TCSVT*, 13(7):560–576, 2003. 1

[68] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 6

[69] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021. 1, 3, 8

[70] Yibo Yang, Robert Bamler, and Stephan Mandt. Improving inference for neural image compression. *NeurIPS*, 33:573–584, 2020. 2

[71] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 8

[72] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, pages 1265–1274, 2015. 8

[73] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 1, 3, 8

[74] Lei Zhou, Zhenhong Sun, Xiangji Wu, and Junmin Wu. End-to-end optimized image compression with attention mechanism. In *CVPR workshops*, page 0, 2019. 5

[75] Yinhao Zhu, Yang Yang, and Taco Cohen. Transformer-based transform coding. In *ICLR*, 2022. 1, 2, 4, 6