Towards Analysis-friendly Face Representation with Scalable Feature and Texture Compression

Shurun Wang, Shiqi Wang, Member, IEEE, Wenhan Yang, Xinfeng Zhang, Member, IEEE, Shanshe Wang, Siwei Ma, Member, IEEE and Wen Gao, Fellow, IEEE

Abstract—it plays a fundamental role to compactly represent the visual information towards the optimization of the ultimate utility in myriad visual data centered applications. With numerous approaches proposed to efficiently compress the texture and visual features serving human visual perception and machine intelligence respectively, much less work has been dedicated to studying the interactions between them. Here we investigate the integration of feature and texture compression, and show that a universal and collaborative visual information representation can be achieved in a hierarchical way. In particular, we study the feature and texture compression in a scalable coding framework, where the base layer serves as the deep learning feature and enhancement layer targets to perfectly reconstruct the texture. Based on the strong generative capability of deep neural networks, the gap between the base feature layer and enhancement layer is further filled with the feature level texture reconstruction, aiming to further construct texture representation from feature. As such, the residuals between the original and reconstructed texture could be further conveyed in the enhancement layer. To improve the efficiency of the proposed framework, the base layer neural network is trained in a multi-task manner such that the learned features enjoy both high quality reconstruction and high accuracy analysis. We further demonstrate the framework and optimization strategies in face image compression, and promising coding performance has been achieved in terms of both rate-fidelity and rate-accuracy.

Index Terms—Feature compression, texture compression, scalable coding, multi-task learning.
representation of visual information into base and enhancement layers, where the base layer conveys compact features for analysis purpose, and enhancement layer accounts for the representation of visual signals based on the fine structure prediction from the base layer. The cross-modality reference from the visual signal and feature representation can greatly increase the degree of supported scalability and flexibility, such that various application requirements can be satisfied simultaneously. Such a framework also allows the direct access to the visual feature without introducing the data redundancy in server side processing. Due to these advantages, this paper provides a comprehensive study on this framework, and optimizes the proposed framework from multiple perspectives. The contributions of this paper are as follows,

- We comprehensively study the joint feature and texture compression based on a unified scalable coding framework in a hierarchical way. The framework enjoys the advantages of both feature and texture compression, and is implemented with the aim of effective facial data representation. Extensive experimental results show the proposed framework achieves better coding performance in terms of both rate-distortion and rate-accuracy.
- We optimize the base layer feature extraction based on a multi-task learning approach where both reconstruction capability and analysis accuracy are optimized, such that the representation capability has been enhanced with the comparable intelligent analysis performance. Furthermore, a compact feature representation model is proposed to efficiently compress the features for both intelligent analysis and texture description.
- We propose the Laplacian structure based mapping scheme from global feature to local texture, such that a reliable texture description that aligns the original input can be generated based on the base layer. The enhancement layer, which aims to reliably reconstruct the fine texture, is further compressed with an end-to-end coding strategy. As such, the base and enhancement layers seamlessly work together to achieve scalable representation of the visual data.

II. RELATED WORKS

There has been a tremendous development of visual signal compression techniques in recent years, and numerous standards have been developed including the still image coding standards from JPEG [1] to JPEG 2000 [2] and the continuous improvement for video compression from H.262/MPEG-2 [11], H.264/MPEG-4 AVC [3] till H.265/HEVC [4]. To further improve the coding performance, there are numerous algorithms developed for the future video compression standards including matrix weighted intra prediction [12], quadtree plus binary tree [13], extended coding unit partitioning [14], affine motion prediction [15], decoder-side motion vector refinement [16] and mode-dependent non-separable secondary transform [17]. Regarding the encoder optimization, various optimization algorithms have also been developed towards different targets, including the rate-distortion optimization for both signal [18] and feature quality [19]. For surveillance video data, which has been regarded as the biggest big data [20], the concept of golden-frame has been introduced for providing better reference quality [21], and based on this philosophy the background modeling based surveillance video compression techniques have been developed [22], [23]. In the meanwhile, along with the development of cloud computing, a cloud database has been introduced in image compression as an external reference to further remove the redundancy [24], [25]. In addition, efforts have been devoted to analyzing the bitstream without completely decoding, due to the abundant information implied in the bitstream [26], [27]. Due to the strong representation capability of deep learning, end-to-end compression framework based on deep neural networks has been developed. A vanguard work [28], which applied a recurrent neural network (RNN) to end-to-end learned image representation, has achieved a comparable performance compared with JPEG. Ballé et al. [29] proposed the generalized divisive normalization (GDN) based image coding with a density estimation model and achieved obvious compression performance promotion compared with JPEG 2000. Based on this method, the redundancy is further eliminated with a variational hyper-prior model [30], which surpassed BPG in terms of rate-distortion performance. The deep video compression (DVC) approach [31] based on an end-to-end model has also achieved better performance comparing with H.264/AVC. Though significant performance improvement has been observed, in such Compress-then-Analysis (CTA) paradigm, the main technical impediment to the video compression is that the increasing of coding efficiency is still far behind the growth rate in terms of the volume of data, such that in low bit rate coding scenarios the analysis performance could be severely degraded due to the quality degradation of textures.

One of the major bottlenecks in applying the automatic computer vision algorithms in real applications lies in the large scale visual data, making the CTA paradigm impractical in many scenarios. By contrast, the convey of visual feature descriptors with which the computer vision tasks can be naturally supported has been widely studied during the past decade. The Analyze-then-Compress (ATC) paradigm [32] benefits the compact representation of the visual information as the features occupy much smaller space compared with textures, such that lower transmission bandwidth and higher analysis performance can be achieved simultaneously. There are several algorithms proposed for compactly representing the handcrafted features, such as hashing quantization [33], transform representation [34] and vector-based quantization [35]. Moreover, binary based representations have been proposed for fast matching and efficient transmission. In view of the necessity of compact feature representation, the ISO/IEC moving pictures experts group (MPEG) established the standard for a compact descriptor for visual search (CDVS), which standardizes the visual data descriptor extraction process and coding syntax at bit stream level for still images. Recently, MPEG also deployed the Compact Descriptors for Video Analysis (CDVA) standard targeting at video analysis and understanding, where both handcrafted and deep learning features have been involved. Compared with handcrafted features, deep learning features
Fig. 1: The framework of the proposed scalable feature signal compression.
the base layer extraction and compact representation, fine texture reconstruction and enhancement layer compression are delicately studied, in an effort to improve the representation capability at both feature and texture levels.

A. The Overall Architecture

We partition the representation of visual information into base and enhancement layers, as shown in Fig. 1. The base layer conveys compact deep learning features for analysis purposes, and enhancement layer accounts for representing the visual signals. The interactions between base and enhancement layers are naturally supported, such that the visual signal can be efficiently compressed based on the fine structure prediction from the base layer.

The proposed paradigm is able to shift compact feature representation as an integrated part of visual data compression, and offloads feature extraction to large-scale edge nodes. As such, the framework greatly improves the flexibility in visual signal representation to meet the requirements in different application scenarios. The proposed framework has several advantages. First, the base layer enables the front-end intelligence by extracting and compressing the deep learning features. As such, this could significantly economize the bandwidth when the ultimate receiver is machine vision instead of the HVS. The high fidelity deep learning features also guarantee the analysis performance as features could be extracted from the original visual signal. Second, the interactions between feature and texture are flexibly supported with the feature based texture reconstruction, which further removes the redundancy between the two different modalities. As such, the proposed scheme ensures promising compression performance in terms of rate-fidelity. Third, such a scalable representation framework allows the direct access to the visual feature without introducing the data redundancy and decoding complexity. In general, the visual analysis tasks may be performed more frequently than human involved monitoring, and at the decoder side the individual decoding of the base layer avoids the computational intensive video decoding and the subsequent feature extraction process. When further monitoring is required, the enhancement layer is then transmitted and reconstructed to provide such service without introducing additional redundancy.

The proposed framework is implemented with facial images that play important roles in many applications. More specifically, as shown in Fig. 1, given the input facial image $x$, the facial features $f_{raw}$ are extracted with deep neural network Facenet() trained by means of multi-task learning for simultaneous high-level visual analysis and low-level signal reconstruction. To compactly represent the base layer, an end-to-end deep learning feature compression scheme is proposed, where an encoder $Enc_{base}()$ and decoder $Dec_{base}()$ are learned. Subsequently, a feature-level signal reconstruction based on the compactly represented base layer $x_{base}$ is achieved with the deconvolution network $Rec_{base}()$. As such, the residual between original image $x$ and feature-level texture reconstruction $x_{enh}$ could be further conveyed in the enhancement layer $Enc_{enh}()$. At the decoder side, the base layer is reconstructed for the analysis purpose. Depending on the application circumstances, the enhancement layer is further reconstructed with $Dec_{enh}()$ such that the entire texture $x_{rec}$ is formed when monitoring/viewing is required.

B. Base Layer Construction

We adopt a typical deep neural network architecture Facenet with Inception Resnet V1 [50] for facial image feature extraction in the base layer. In particular, the deep learning based feature extraction embeds a face image into a hyperspace with 128 dimensions, which has achieved dramatic performance promotion towards face analysis such as verification and recognition. However, the feature of the pre-trained Facenet model is delicately designed for face analysis only such that it may not be appropriate for signal level reconstruction. In order to bridge the gap between analysis and signal level representation with deep learning features, we propose a multi-task learning based model, as illustrated in Fig. 2.

In particular, this model is expected to be equipped with high analysis accuracy as well as strong capability in reconstructing the texture. As such, based upon the architecture of Facenet which is initialized with the pre-trained model[1], a multi-task based loss function is designed towards achieving the two tasks simultaneously.

$$\mathcal{L}_{bext} = -\sum_{i} y_i \log(y'_i) + \lambda_s \text{SATD}(x_s, x_{trans}),$$  \hspace{1cm} (1)$$

where $y_i$ and $y'_i$ denote the label and corresponding prediction of a softmax layer, respectively. $N$ is the number of training samples in each batch. The cross-entropy loss is responsible for ensuring the analysis accuracy. Moreover, $\text{SATD}(\cdot, \cdot)$ denotes the sum of absolute transformed differences based on the Hadamard transform, such that the difference between the original and reconstructed faces can be optimized. In particular, the target of fine texture reconstruction with the base layer is removing the redundancy between feature and texture, such that the performance of the enhancement layer can be significantly improved. As such, the SATD loss, which considers both the distortion and required coding bits of the residuals, is adopted here instead of the traditional $\ell_1$ or $\ell_2$ norm. To this end, a feature-texture transformation model is employed to transform the structure information from

\footnote{https://github.com/davidsandberg/facenet}
The architecture of the end-to-end feature compression framework is shown in Fig. 4. More specifically, the extracted deep learning feature \( f_{\text{raw}} \) is fed into a cascaded fully-connected layer for compact representation, where GDN and IGDN are utilized as the activation functions for the encoder \((\text{Enc}_{\text{fea}}(\cdot))\) and decoder \((\text{Dec}_{\text{fea}}(\cdot))\), respectively. Besides, the output of \( \text{Enc}_{\text{fea}}(\cdot) \), denoted as the latent code \( c_{\text{fea}} \), is further compressed with an arithmetic coding engine for entropy coding. As such, the whole process is formulated as follows,

\[
    c_{\text{fea}} = \text{Enc}_{\text{fea}}(f_{\text{raw}}), \quad f_{\text{rec}} = \text{Dec}_{\text{fea}}(c_{\text{fea}}). \quad (2)
\]

In general, the Euclidean distance has been widely adopted to measure the similarity of facial features in face analysis, such that the mean squared error (MSE) between the original feature \( f_{\text{raw}} \) and the reconstructed feature \( f_{\text{rec}} \) is involved in the loss function for training the end-to-end coding network. Moreover, since the decompressed deep learning feature is further applied in texture reconstruction, the feature coding process is further optimized with the guidance of texture quality by measuring the texture degradation originated from feature compression. As such, we introduce a new end-to-end compact feature representation scheme by combining the feature-structure transformation along with feature compression. More specifically, the feature-structure transformation module learned in the base layer generation is first utilized as the initialization of feature based texture generation, and fine-tuned together with the compression model. The fidelity of both the feature as well as the texture generated from the feature is preserved. Thus, the loss function of the end-to-end feature compression is expressed as follows,

\[
    \mathcal{L}_{b_{\text{com}}} = ||f_{\text{raw}} - f_{\text{rec}}||_2^2 + \lambda_1 ||c_{\text{fea}}||_1 + \lambda_2 \text{SATD}(x_s, x_{\text{trans}}), \quad (3)
\]

where \( f_{\text{raw}} \) and \( f_{\text{rec}} \) are the original and reconstructed features, and \( c_{\text{fea}} \) is the compact representation of deep learning features. \( x_{\text{trans}} \) is generated with feature-structure model and \( x_s \) is the downsampled version of original image \( x \) with the corresponding image size. The \( l_1 \) norm, denoted as \( || \cdot ||_1 \), is responsible for the approximation of the number of consumed bits in the compact representation. The \( l_2 \) norm measures the distortion of the base layer feature in feature reconstruction, denoted as \( || \cdot ||_2 \).

The compact representation \( c_{\text{fea}} \) is further clipped by the threshold \( r_{\text{clip}} \) in an element-wise manner to economize the representation expense by maintaining the range consistence under various compression ratios. Besides, random noise is applied to simulate the rounding of distortion, which could further strengthen the degree of robustness compared with the simple round operation.

Given the decoded features from the base layer, as illustrated in Fig. 5, the texture is further generated based on a generation model with a cascade of neural networks. The aim of the texture generation from the feature is to remove the redundancies between base and enhancement layer, such that the enhancement layer should be compression friendly by signaling the residuals between the input and generated texture. As such, again, the SATD has been adopted as

---

**Fig. 3:** The distribution of fine-structured Facenet features extracted from certain dimensions in various datasets. (a) VGG-Face2; (b) LFW.

**Fig. 4:** The architecture of the end-to-end deep learning feature compression.

feature level to texture level at the base layer. In particular, it is composed of a series of deconvolutional layers. The output of transformation \( x_{\text{trans}} \) with size \( 32 \times 32 \) is regarded as reconstruction output and the target is approaching the downsampled version \( x_s \) from the original image \( x \). The base layer feature extraction model and the feature-structure transformation model are jointly trained.

**C. Base Layer Compression and Fine Texture Reconstruction**

The deep learning features in the base layer are compactly represented with an end-to-end coding framework. To this end, we first investigate the statistics of the features. The distributions of features from Labeled Face in Wild (LFW) [51] and VGG-Face2 [52] datasets are shown in Fig. 3. It is obvious that the distributions of the feature at every dimension are Gaussian-like and locate in similar range with zero expectations. As illustrated in [29], the features well match the characteristics of generalized divisive normalization (GDN) from the perspective of Gaussian densities. Inspired by the recent development of deep learning based image compression [53], we develop an end-to-end deep feature compression scheme that maps the input feature into a latent code for compact representation.
Figure 5: Texture reconstruction from the base layer with Laplacian pyramid structure. The kernel size $k$, filter number $f$ and stride size $s$ are shown in form of $k$-$f$-$s$. The green arrows indicate convolutional layers with kernel size 3, filter number 3 and stride size 1. The blue arrows indicate bilinear upsampling operations with factor 2.

At the receiver side, the elaborate reconstruction $x_{rec}$ in the proposed scalable compression framework is obtained by the combination of the texture construction from base and enhancement layers,

$$x_{rec} = x_{base} + x'_{enh_{rec}},$$

where $x'_{enh_{rec}}$ is the inv-normalized enhancement layer reconstructed from the latent code.

IV. EXPERIMENTAL RESULTS

To validate the efficiency of the proposed scheme, we compare the performance of the proposed scheme with the conventional compression schemes in terms of both rate-accuracy and rate-distortion performance. The implementation details of the proposed scheme are first provided. Subsequently, experimental results for feature compression and feature-level reconstruction are presented, to evaluate the effectiveness of compact feature representation in the base layer. The texture compact representation performance is further shown by comparing the performance with conventional visual data compression schemes.

A. Implementation Details

The proposed framework is implemented using TensorFlow as the deep learning toolbox, and the model parameters are initialized with the method in [54]. Besides, we set the learning rate as 0.0001 for all the neural network modules except for feature-level reconstruction, which exponentially decays from 0.0001 to 0.00001 every 5 epochs with a decay factor 0.9. The optimization method for all the modules is Adaptive Moment Estimation (Adam) [55]. The weighting factor of multi-task deep learning extraction $\lambda_s$ is set as 100. Regarding the deep learning feature compression, the compactness factor $\lambda_1$ is set from $1 \times 10^{-4}$ to $1 \times 10^{-7}$ to acquire various degrees of compact representation. The weighting factor for feature and texture reconstruction, $\lambda_2$, is set to 0.01. Moreover, the threshold for the latent code of deep learning feature compression is set to 20.0 and the random noise range is set from -0.5 to 0.5 correspondingly. Considering the enhancement layer compression, the bit rate constraint is set from 1 to $1 \times 10^{-2}$.

We implement the framework based on the face images as face plays critical roles in many visual analysis driven initiatives, especially in smart cities. Moreover, the current face recognition technologies are relatively mature to be deployed in machine vision centered applications, such that the proposed scheme can be recognized as the intersection between video coding for HVS and machine vision. The training dataset is VGG-Face2 [52], which contains over 3 million facial images. More specifically, VGG-Face2 includes over nine thousand subjects and every subject has over 360 images on average. Furthermore, a popular face verification dataset, Labeled Faces in the Wild (LFW) [51], is adopted as the test dataset. MTCNN [56] is utilized to crop and align the human face patch from original images. It is also worth mentioning that the texture reconstruction performance is evaluated at $256 \times 256$ scale and the FaceNet feature is extracted from images with $160 \times 160$, as illustrated in [58].
Fig. 6: Examples for feature-structure transformation in base layer. The first row is the downsampled images and the second row includes the feature-structure transformation outputs. The resolution of the output image from the network is $32 \times 32$, and they are upsampled for visualization. (a) VGG-Face2; (b) LFW.

B. Base Layer Construction

We first evaluate the performance of feature extraction in terms of face verification accuracy and feature-level reconstruction. The original facial images are downsampled into $160 \times 160$ with Bicubic downsampling because of the input size demand of Facenet model. The multi-task learning based Facenet model achieves a comparable analysis performance in terms of face verification, which is 99.10%. It is worth mentioning that the accuracy of the original Facenet feature is 99.6% [38], such that ignorable performance degradation in terms of accuracy has been observed. Moreover, regarding the feature-structure transformation on the basis of deep learning features, the results are shown in Fig. 6 for both train and test datasets.

The deep learning features acquired by the proposed multi-task learning feature extractor could bear critical visual structure information such that the textures could be promisingly constructed by means of feature-structure transformation. The proposed feature extractor reveals strong information preservation characteristics for the facial information, including skin color, facial posture, light distribution, facial organs and facial expression. Moreover, facial texture with high discriminating capability could also be recovered from the multi-task learned deep learning feature, such as mustache, sunglasses and hair. However, the facial texture with low discriminating capability and background details will be discarded, and this phenomenon is consistent with the task of facial analysis.

C. Base Layer Compression and Fine Texture Reconstruction

Experiments have been conducted to verify the effectiveness of the proposed end-to-end deep learning feature compression model by comparing it with a series of feature compression algorithms. First, a scalar quantization algorithm (SQ) used in [10] is adopted. On the top of this strategy, a deep learning based feature enhancement model (SQ-E), proposed in [42] is also involved. We verify the effectiveness of end-to-end compression model in terms of rate-accuracy performance, which is shown in Table 1. More specifically, the proposed algorithm is denoted as PRO and the accuracy of the original deep learning feature, which is extracted via multi-task learning, is 99.10%. Except for SQ and SQ-E, we also compared the performance with some other feature compression algorithms, including PQ [57], OPQ [58], DBH [59] and DHN [41]. It is worth mentioning that the original Facenet features without multi-task learning are adopted here as the input of the compared feature compression methods. As such, although the performance of the proposed deep learning model trained with multi-task learning is marginally lower than that of comparison models, the proposed method could still deliver better rate-accuracy performance. This provides useful evidence regarding the effectiveness of the proposed base layer compression method.

In order to demonstrate the feature-level texture reconstruction capability, we reconstruct the facial images based on the compressed features, and the comparison results are shown in Fig. 7. As shown in Fig. 7 the facial texture could be recovered with the main structural information. As such, the detail texture information, such as wrinkle, hair and background could be further conveyed by enhancement layer. The Peak Signal to Noise Ratio (PSNR) of feature-level reconstruction in VGG-Face2 and LFW is 22.26 dB and 22.18 dB respectively at 0.0052 bpp in average, which demonstrates the the base layer texture reconstruction could further remove the redundancy between the base and enhancement layers.

D. Enhancement Layer Compression

The final reconstruction of the facial image could be obtained by combining the reconstruction signals from both base and enhancement layers. As shown in Fig. 8, the accuracy of feature compression would reach the saturation point when the bitrate of the base layer has exceeded 0.0055 bpp. As such, the base layer compression for deep learning features is fixed at the saturated point to guarantee the analysis performance.
TABLE I: Base layer compression performance comparison in terms of rate-accuracy.

| BPP | PQ | Acc(%) | BPP | OPQ | Acc(%) | BPP | DBH | Acc(%) | BPP | DHN | Acc(%) | BPP | SQ | Acc(%) | BPP | SQ-E | Acc(%) | BPP | PRO | Acc(%) |
|-----|----|--------|-----|-----|--------|-----|-----|--------|-----|-----|--------|-----|-----|--------|-----|------|--------|-----|-----|--------|
| 0.0020 | 98.48 | 0.0020 | 98.40 | 0.0020 | 97.48 | 0.0020 | 98.25 | 0.0024 | 50.00 | 0.0024 | 50.41 | 0.0016 | 96.28 | 0.0027 | 98.74 | 0.0039 | 98.87 |
| 0.0040 | 99.13 | 0.0040 | 99.17 | 0.0040 | 98.70 | 0.0040 | 98.00 | 0.0033 | 87.60 | 0.0033 | 87.82 | 0.0050 | 98.15 | 0.0050 | 98.15 | 0.0050 | 98.15 |
| 0.0080 | 99.13 | 0.0080 | 99.17 | 0.0080 | 98.43 | 0.0080 | 99.08 | 0.0050 | 99.11 | 0.0082 | 99.23 | 0.0082 | 99.26 | 0.0052 | 99.00 | 0.0052 | 99.00 |

Fig. 8: Performance of face recognition based on different feature compression levels, together with the accuracy of the original feature.

and economize the base layer representation expenses simultaneously.

First, we compare the rate-distortion performance of the proposed framework with the conventional paradigm in terms of PSNR and MS-SSIM. For a fair comparison, we utilize an end-to-end deep learning image compression scheme [30], as the enhancement layer compression is also trained in an end-to-end manner with the SATD as the loss function. The coding expenses in terms of bpp are obtained based on the sum of base and enhancement layers in the proposed scheme. Moreover, the PSNR and MS-SSIM are computed as the average value over the whole test dataset LFW. The experimental results are shown in Fig. 9. An obvious compression performance improvement compared with the traditional scheme with deep learning based image compression has been observed, and 4.57% bit rate savings on average are obtained on whole LFW. The results also prove the effectiveness of the proposed scalable compression framework, along with the efficiency towards large-scale facial analysis.

In addition to the improvement of the compact representation performance against deep learning based image compression, we also conduct experiments to demonstrate the adaptation and efficiency compared with traditional image compression standards, including JPEG, JPEG 2000 and HEVC intra. Following the same pipeline, we adopt these standards in the enhancement layer as well for a fair compression. The experimental results are shown in Fig. 10, Fig. 11 and Fig. 12 where the proposed framework could achieve 10.25%, 10.14% and 8.42% compression performance improvement. These results also provide useful evidence regarding the adaptation capability and the robustness of the proposed scalable framework. Furthermore, better compression performance improvement is revealed at low bit rate scenarios, due to feature-level texture construction with multi-task based feature learning.

Moreover, examples of reconstructed images with different methods are shown in Fig. 13 for visual quality comparisons. It is obvious that the proposed scalable framework could achieve promising reconstruction results by delivering a better quality of reconstructed facial image compared with traditional schemes. In addition, it is worth mentioning that the proposed framework can achieve obvious compression performance improvement compared with the deep learning based image compression model which is also specifically trained in facial image compression dataset.

Fig. 9: Compression performance comparisons between the proposed framework and traditional scheme with same deep learning image compression model [30] in terms of PSNR and MS-SSIM. The compared deep learning based image compression model is trained with VGG-Face2 dataset. (a) Rate distortion performance in terms of PSNR; (b) Rate distortion performance in terms of MS-SSIM.

Fig. 10: Compression performance comparisons between the proposed framework and traditional scheme in terms of PSNR and MS-SSIM. The enhancement layer is compressed with JPEG and the compared method is JPEG as well. (a) Rate distortion performance in terms of PSNR; (b) Rate distortion performance in terms of MS-SSIM.
applications. As pedestrian and vehicle, to bring a significant boost to these, will extend the proposed philosophy to other scenarios such as visual analysis and signal representation, and in the future we paradigm. The research results will bridge the gaps between enhancement layers which account for analysis and viewing, the generation, compression and optimization of the base and scale edge nodes. The novelty of the proposed scheme presentation, enabling the offloading of feature extraction to large-compression as an integrated part of compact visual data repre-

Fig. 11: Compression performance comparisons between the proposed framework and traditional scheme in terms of PSNR and MS-SSIM. The enhancement layer is compressed with JPEG 2000 and the compared method is JPEG 2000 as well. (a) Rate distortion performance in terms of PSNR; (b) Rate distortion performance in terms of MS-SSIM.

![Graph](image1)

Fig. 12: Compression performance comparisons between the proposed framework and traditional scheme in terms of PSNR and MS-SSIM. The enhancement layer is compressed with HEVC and the compared method is HEVC as well. (a) Rate distortion performance in terms of PSNR; (b) Rate distortion performance in terms of MS-SSIM.

V. CONCLUSIONS

We propose a hierarchical coding scheme that shifts feature compression as an integrated part of compact visual data representation, enabling the offloading of feature extraction to large-scale edge nodes. The novelty of the proposed framework lies in the generation, compression and optimization of the base and enhancement layers which account for analysis and viewing, respectively. The superior performance of the proposed scheme is demonstrated using both rate-accuracy and rate-distortion, and significant bit rate savings have been achieved compared to the advanced image compression algorithms in the traditional paradigm. The research results will bridge the gaps between visual analysis and signal representation, and in the future we will extend the proposed philosophy to other scenarios such as pedestrian and vehicle, to bring a significant boost to these applications.

REFERENCES

[1] G. K. Wallace, “The JPEG still picture compression standard,” IEEE Transactions on Consumer Electronics, vol. 38, no. 1, pp. xviii–xxiv, 1992.

[2] M. Rabbani, “JPEG2000: Image compression fundamentals, standards and practice,” Journal of Electronic Imaging, vol. 11, no. 2, p. 286, 2002.

[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 560–576, 2003.

[4] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649–1668, 2012.

[5] A. Mohan, K. Gauen, Y.-H. Lu, W. W. Li, and X. Chen, “Internet of video things in 2030: A world with many cameras,” in IEEE International Symposium on Circuits and Systems. IEEE, 2017, pp. 1–4.

[6] L.-Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, “Overview of the MPEG-CDVS standard,” IEEE Transactions on Image Processing, vol. 25, no. 1, pp. 179–194, 2015.

[7] L.-Y. Duan, Y. Lou, Y. Bai, T. Huang, W. Gao, V. Chandrasekhar, J. Lin, S. Wang, and A. C. Kot, “Compact descriptors for video analysis: The emerging MPEG standard.” IEEE MultiMedia, vol. 26, no. 2, pp. 44–54, 2018.

[8] L.-Y. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, “Video coding for memories: A paradigm of collaborative compression and intelligent transform prediction for next generation video coding,” arXiv preprint arXiv:2001.03569, 2020.

[9] Y. Li, C. Jia, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, “Joint rate-distortion optimization for simultaneous texture and deep feature compression of facial images,” in 2018 IEEE Fourth International Conference on Multimedia Big Data. IEEE, 2018, pp. 1–5.

[10] S. Wang, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, “Scalable facial image compression with deep feature reconstruction,” in 2019 IEEE International Conference on Image Processing. IEEE, 2019, pp. 2691–2695.

[11] I. Recommendation, “Generic coding of moving pictures and associated audio information: Video,” 1995.

[12] J. Pfaff, B. Stallenberger, M. Schafer, P. Merkle, P. Helle, T. Hinz, H. Schwarz, D. Marpe, and T. Wiegand, “CE3: Affine linear weighted intra prediction (CE3-4.1 CE3-4.2),” in document JVET-N0217, in Proc. of 14th JVET meeting, 2019.

[13] J. An, Y. Chen, K. Zhang, H. Huang, Y. Huang, and S. Lei, “Block partitioning structure for next generation video coding,” MPEC doc. m37524 and ITU-T SG16 Doc. COM16-C666, 2015.

[14] M. Wang, J. Li, L. Zhang, K. Zhang, H. Liu, S. Wang, S. Kwong, and S. Ma, “Extended coding unit partitioning for future video coding,” IEEE Transactions on Image Processing, 2019.

[15] S. Lin, H. Chen, H. Zhang, S. Maxin, H. Yang, and J. Zhou, “Affine transform prediction for next generation video coding,” MPEC doc. m37525 and ITU-T SG16 Doc. COM16-C1016, 2015.

[16] X. Chen, J. An, and J. Zheng, “EE3: Decoder-side motion vector refinement based on bilateral template matching,” in Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-E0052, 5th Meeting, 2017.

[17] X. Zhao, J. Chen, and M. Karczewicz, “Mode-dependent non-separable secondary transform,” ITU-T SG16/Q6 Doc. COM16-C1044, 2019.

[18] B. Li, H. Li, L. Li, and J. Zhang, “A domain rate control algorithm for high efficiency video coding,” IEEE Transactions on Image Processing, vol. 23, no. 9, pp. 3841–3854, 2014.

[19] J. Chao, R. Huiel, E. Steinbach, and D. Schroeder, “A novel rate control framework for SIFT/SURF feature preservation in H.264/AVC video compression,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 6, pp. 958–972, 2015.

[20] T. Huang, “Surveillance video: The biggest big data,” Computing Now, vol. 7, no. 2, pp. 82–91, 2014.

[21] M. Paul, W. Lin, C.-T. Lau, and B.-S. Lee, “Explore and model better i-frames for video coding,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 9, pp. 1242–1254, 2011.

[22] X. Zhang, T. Huang, Y. Tian, and W. Gao, “Background-modeling-based adaptive prediction for surveillance video coding,” IEEE Transactions on Image Processing, vol. 23, no. 2, pp. 769–784, 2013.

[23] F. Chen, H. Li, L. Li, D. Liu, and F. Wu, “Block-composed background reference for high efficiency video coding,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 12, pp. 2639–2651, 2016.

[24] H. Yue, X. Sun, J. Yang, and F. Wu, “Cloud-based image coding for mobile device toward thousands to one compression,” IEEE Transactions on Multimedia, vol. 15, no. 4, pp. 845–857, 2013.

[25] Z. Shi, X. Sun, and F. Wu, “Photo album compression for cloud storage using local features,” IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 4, no. 1, pp. 17–28, 2014.
Fig. 13: Examples of reconstructed faces with different models. The proposed scheme corresponds to the proposed image compression framework with the employed coding scheme as the codec for the enhancement layer, denoted as “Proposed (codec)”. By contrast, the traditional compression scheme for comparison is denoted with “codec”.

[26] L. Zhao, Z. He, W. Cao, and D. Zhao, “Real-time moving object segmentation and classification from hevc compressed surveillance video,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1346–1357, 2016.

[27] D. Edmundson and G. Schaefer, “An overview and evaluation of JPEG compressed domain retrieval techniques,” in *Proceedings ELMAR-2012*. IEEE, 2012, pp. 75–78.

[28] G. Toderici, S. M. O’Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, “Variable rate image compression with recurrent neural networks,” in *ICLR*, 2016.

[29] J. Ballé, V. Laparra, and E. Simoncelli, “End-to-end optimized image compression,” in *International Conference on Learning Representations*, 2017.

[30] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” in *International Conference on Learning Representations*, 2018.

[31] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, “DVC: An end-to-end deep video compression framework,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 006–11 015.

[32] A. Redondi, L. Baroffio, M. Cesana, and M. Tagliasacchi, “Compress-then-analyze vs. analyze-then-compress: Two paradigms for image analysis in visual sensor networks,” in *IEEE International Workshop on Multimedia Signal Processing*. IEEE, 2013, pp. 278–282.

[33] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, “Supervised hashing with kernels,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2074–2081.

[34] V. Chandrasekhar, G. Takacs, D. Chen, S. S. Tsai, J. Singh, and B. Girod, “Transform coding of image feature descriptors,” in *Visual Communications and Image Processing 2009*, vol. 7257. International Society for Optics and Photonics, 2009, p. 725710.

[35] H. Jegou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[37] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.

[38] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[39] L. Ding, Y. Tian, H. Fan, Y. Wang, and T. Huang, “Rate-performance-loss optimization for inter-frame deep feature coding from videos,” *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5743–5757, 2017.

[40] L. Ding, Y. Tian, H. Fan, C. Chen, and T. Huang, “Joint coding of local
and global deep features in videos for visual search,” IEEE Transactions on Image Processing, vol. 29, pp. 3734–3749, 2020.

[41] H. Zhu, M. Long, J. Wang, and Y. Cao, “Deep hashing network for efficient similarity retrieval,” in Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[42] S. Wang, W. Yang, and S. Wang, “End-to-end facial deep learning feature compression with teacher-student enhancement,” arXiv preprint [arXiv:2002.03627] 2020.

[43] Z. Chen, K. Fan, S. Wang, L. Duan, W. Lin, and A. C. Kot, “Toward intelligent sensing: Intermediate deep feature compression,” IEEE Transactions on Image Processing, vol. 29, pp. 2230–2243, 2020.

[44] Y. Hu, S. Yang, W. Yang, L.-Y. Duan, and J. Liu, “Towards coding for human and machine vision: A scalable image coding approach,” arXiv preprint [arXiv:2001.02915] 2020.

[45] S. Xia, K. Liang, W. Yang, L.-Y. Duan, and J. Liu, “An emerging coding paradigm VCM: A scalable coding approach beyond feature and signal,” arXiv preprint [arXiv:2001.03904] 2020.

[46] Y. Lou, L. Duan, S. Wang, Z. Chen, and W. Gao, “Front-end smart visual sensing and back-end intelligent analysis: A unified infrastructure for economizing the visual system of city brain,” IEEE Journal on Selected Areas in Communications, vol. PP, no. 99, pp. 1–1, 2019.

[47] L. Duan, Y. Lou, S. Wang, W. Gao, and Y. Rui, “AI-oriented large-scale video management for smart city: Technologies, Standards, and Beyond,” IEEE MultiMedia, vol. 26, no. 2, pp. 8–20, April 2019.

[48] G. Wen, Y. Tian, T. Huang, S. Ma, and X. Zhang, “The IEEE 1857 standard: Empowering smart video surveillance systems,” Intelligent Systems IEEE, vol. 29, no. 5, pp. 30–39, 2014.

[49] Y. Lou, L. Duan, Y. Luo, Z. Chen, T. Liu, S. Wang, and W. Gao, “Towards efficient front-end visual sensing for digital retina: A model-centric paradigm,” IEEE Transactions on Multimedia, 2020.

[50] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, Inception-Resnet and the impact of residual connections on learning,” in Thirty-first AAAI conference on artificial intelligence, 2017.

[51] B. Gary, R. Manu, B. Tamara, and L. Erik, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[52] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in International Conference on Automatic Face and Gesture Recognition, 2018.

[53] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, “Image and video compression with neural networks: A review,” IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[54] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[55] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint [arXiv:1412.6980] 2014.

[56] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[57] H. Jgou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117–128, Jan 2011.

[58] T. Ge, K. He, Q. Ke, and J. Sun, “Optimized product quantization,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 4, pp. 744–755, April 2014.

[59] W. Liu, J. Wang, R. Ji, Y. Jiang, and S. Chang, “Supervised hashing with kernels,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 2074–2081.