2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)

Joint rate-distortion optimization for simultaneous texture and deep feature compression of facial images

Yang Li
Institute of Digital Media
Peking University
Beijing, China
li.yang.@pku.edu.cn

Chuanmin Jia
Institute of Digital Media
Peking University
Beijing, China
cmjia@pku.edu.cn

Shiqi Wang
Department of Computer Science
City University of Hong Kong
Hongkong, China
shiqwang@cityu.edu.hk

Xinfeng Zhang
Ming-Hsieh Department of
Electrical Engineering, University of
Southern California, Los Angeles, USA
xinfengz@usc.edu

Shanshe Wang
Institute of Digital Media
Peking University
Beijing, China
sswang@pku.edu.cn

Siwei Ma
Institute of Digital Media
Peking University
Beijing, China
swma@pku.edu.cn

Wen Gao
Institute of Digital Media
Peking University
Beijing, China
wgao@pku.edu.cn

***Abstract*—The explosion of surveillance cameras in smart cites and the increasing demand of low latency visual analysis have pushed the horizon from the traditional image/video compression to feature compression. Due to the recent advances of face recognition, we investigate the simultaneous compression of facial images and deep features, which is demonstrated to be beneficial in terms of the whole system performance including visual quality and recognition accuracy. Herein, we propose the Texture-Feature-Quality-Index (TFQI) to measure the ultimate utility of the facial images based on automatic visual analysis and monitoring. Furthermore, based on TFQI, a bit allocation scheme is proposed to optimally allocate the given bits for images and features, such that the overall coding performance can be optimized. The proposed scheme is validated using the standard face verification benchmark, Labeled Faces in the Wild (LFW). Better rate-TFQI and rate-Accuracy performance compared to the traditional texture coding can be achieved, especially in the scenario of low bit-rate coding.**

***Index Terms*—Deep feature, Texture-Feature-Quality-Index, bit rate allocation, joint optimization**

I. Introduction

Recent years have witnessed the explosion of visual data and the related services. Surveillance cameras in smart cities are of great value to record the activities of the city in real-time. In the conventional compress-then-analyze (CTA) paradigm, the captured surveillance video data are compressed, and transmitted to the server side for analysis tasks.

With the recent advances of artificial intelligence, large-scale visual data can be automatically processed for high level analysis tasks. Limited by transmission bandwidth and storage capacity, low bit-rate coding method is often utilized in CTA paradigm. However, the low bit-rate texture coding may greatly affect the performance of analysis tasks [1]. One feasible solution is the analyze-then-compress (ATC) scheme, which extracts, compresses and transmits the feature extracted from original videos. Then good performance in the analysis tasks can be achieved at a low bit rate. However, the texture information is abandoned.

Feature coding algorithms play the most significant role in the ATC paradigm. In the literatures, many feature coding schemes have been proposed to further reduce the bit-rate required for feature transmission, for both traditional features (e.g., SIFT [2], SURF [3], [4]) and deep features. Baroffio *et al.* [5], [6] proposed a coding architecture designed for local features extracted from video sequences. Makar *et al.* [7], [8] proposed a temporally coherent key point detector in order to allow efficient inter-frame coding of canonical patches. Ding *et al.* proposed a HEVC-like deep feature coding method to make a trade-off between feature bitrate and analysis performance in [9]. In [10], a coding scheme tailored to both local and global binary features was proposed, which aims at exploiting both spatial and temporal redundancy by means of intra- and inter-frame coding.

As for surveillance video applications, human involved viewing and monitoring may also be required for further verification besides the automatic visual analysis. From such perspective, the framework of both feature and texture transmission was studied in [11], which demonstrates that it is feasible to transmit both hand-crafted features and video textures. Herein, we study the simultaneous compression of deep features and image textures, due to the promising applications of deep convolutional neural networks (CNNs) such as VGG [12], GoogleNet [13] and ResNet [14] in various visual analysis tasks. Moreover, the face recognition, which is the core application in video surveillance, is treated as the ultimate task in our study.

To guarantee the low-latency interaction for analysis task while reserving the texture, we propose to be compressed,

978-1-5386-5321-0/18/$31.00 ©2018 IEEE

transmitted and decoded the texture and feature of the image independently. One critical problem is how to design the bit allocation strategy for features and textures to optimize the overall coding performance, which consists of both visual quality and the accuracy of the face recognition task. To address this issue, we first propose a metric Texture-Feature-Quality-Index (TFQI) to measure the overall performance. Based on TFQI, we subsequently propose a bit allocation scheme for rate-TFQI optimization, which distributes the target bits to texture and feature for the purpose of maximizing the TFQI. To validate the effectiveness of the proposed bit allocation scheme, experiments are conducted to compare the proposed method against the traditional CTA scheme. Extensive experimental results show that the proposed scheme can achieve better rate-TFQI performance and recognition accuracy compared with the conventional CTA scheme.

The rest of this paper is organized as follows. In Section II, the joint image and feature coding framework is described. Section III introduces the TFQI and the bit allocation scheme. Section IV gives the experimental results. Finally, Section V concludes this paper.

II. Joint Texture and Deep Feature Coding Framework

In the conventional CTA paradigm, low bit-rate coding often brings specific distortions to the reconstructed texture, degrading the feature quality which will result in bad analysis performance. As for ATC paradigm, the human involved viewing and monitoring are not supported because of lacking textures, which may limit its applications in reality. To address these issues, we propose to jointly compress the texture and feature extracted from the lossless image to optimize the overall performance.

The framework of proposed joint texture and feature coding is shown in Fig. 1. The acquired facial images or sequences are first detected by the face detection algorithms such as MTCNN [15], then the detected facial images are resized for deep learning feature extraction and further compression. Given the facial image, our proposed framework contains three modules including the joint bit allocation, texture image coding and deep feature coding. As for the texture image coding module, we follow the state-of-the-art video coding standard HEVC/H.265 [16], which is designed based on the hybrid prediction-transform coding framework. Regarding deep feature coding, the features are extracted from the lossless resized detected facial images. Specifically, the output of the last layer in the deep network is treated as the feature for compression and transmission. The extracted feature are further compressed with scalar quantization and entropy coding. The scalar quantization follows the quantization method in HEVC, where the quantization step Q_{step} is determined by the quantization parameter QP. Furthermore, consider the range of element of the feature vector in each dimension differ from different CNNs, deep features are normalized, then the quantization step is further manipulated based on the factor $s = 2^{10}$, which can be formulated as,

$$Q_{step} = \frac{2^{\frac{QP-4}{6}}}{s} = 2^{\frac{QP-4}{6} - 10}. \tag{1}$$

Then the quantization is performed by,

$$l = floor(\frac{c}{Q_{step}}), \tag{2}$$

where c and l denote normalized feature coefficient before and after quantization, respectively. The scale factor $s = 2^{10}$ is empirically chose according many experiments. The $floor$ means truncation, so that it can be expressed by fixed number of binary bits. Finally the quantized feature is Binarized and entropy encoded. The server side utilizes the decoded features for efficient face recognition task and the texture can also be utilized for human-involved monitoring.

To optimize the overall performance, a joint bit allocation module will be introduced in Section **??**. Finally, the bitstreams of features and textures are concatenated, enabling efficient recognition ability and good visual quality in bandwidth-limited scenarios.

III. TFQI-based Joint Bit Allocation

This section details the proposed metric TFQI and the joint bit allocation for image and feature. We first introduce the optimization goal of TFQI in Section III-A, then mathematical derivation of the TFQI based bit allocation scheme is then given in Section III-B.

A. TFQI formulation

In this work, the TFQI metric is proposed to measure the overall performance. Due to the lack of existing metric which can measure the quality of both texture and feature, we firstly choose a simple and intuitive linear weighted model here. The distortions from these two aspects are weighted-averaged according to their normalized weight coefficients. The weighted combination reflects the overall distortion of texture and feature to a certain extent. In particular, with the consideration of just-noticeable distortion model [17], slight distortion is hard to noticed by human eyes or recognition algorithms, similar with metric PSNR, we then utilized the $-log()$ function to smooth the curve in the less distorted part so that it is more practical. Besides that, this operation makes TFQI reflects better overall coding performance by larger values. As such, the TFQI can be formulated as follows,

$$TFQI = -log(w_1 D_1 + w_2 D_2), \quad \sum_{i=1}^{2} w_i = 1, \tag{3}$$

where w_1 and w_2 are weight coefficients, D_1 and D_2 denote distortion in texture and feature, respectively. Specifically, D_1 denotes the MSE between reconstructed and original images, and D_2 denotes the error rate of face recognition. The proposed TFQI has the ability to measure both of the visual quality and the performance of face recognition task. The balance weights w_1 and w_2 can be manually adjusted based

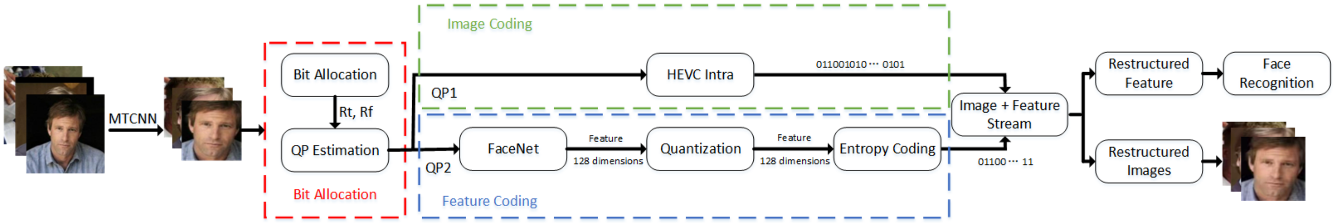


Fig. 1. The joint coding framework of facial images and feature.

on different trade-off demands for visual quality and face recognition, such that this metric can be applied to different practical applications.

B. TFQI-based joint bit allocation scheme

The objective of the proposed bit allocation scheme is to partition the target bits for textures and features to maximize the overall performance in terms of TFQI.

Since both the image quality MSE and feature quality accuracy can be expressed as functions of their corresponding bit-rates, and different bit levels can be achieved by adjusting the QP values for texture and feature coding. We are able to formulate the proposed joint optimization for feature coding and texture coding by using the TFQI definition in Eqn.(3) as follows,

$$\max_{QP_1,QP_2} TFQI(QP_1,QP_2), \quad s.t. \quad R_t+R_f \leq R_c, \tag{4}$$

where QP_1 and QP_2 denote quantization parameters for texture coding and feature coding. R_t and R_f are the bits allocated for texture and feature, and R_c represents the given bits. To solve the above optimization problem, a Lagrange multiplier λ is introduced to convert the constrained problem into unconstraint problem:

$$\begin{aligned}\max_{QP_1,QP_2} J &= TFQI(QP_1,QP_2)+\lambda * R \\ &= -log(w_1 D_t + w_2 D_f)+\lambda(R_t+R_f).\end{aligned} \tag{5}$$

Here, J denotes the joint rate-TFQI cost.

By calculating the partial derivative of J and setting them to zero, we can obtain two following equations:

$$\begin{cases} \dfrac{\partial J}{\partial R_t} = -\dfrac{1}{(w_1D_t+w_2D_f)ln10}\dfrac{\partial w_1 D_t}{\partial R_t}+\lambda=0 \\ \dfrac{\partial J}{\partial R_f} = -\dfrac{1}{(w_1D_t+w_2D_f)ln10}\dfrac{\partial w_2 D_f}{\partial R_f}+\lambda=0 \end{cases} \tag{6}$$

By solving the linear equations in Eqn.(6), we have

$$\lambda = -\frac{\partial w_1 D_t}{\partial R_t} = -\frac{\partial w_2 D_f}{\partial R_f}, \tag{7}$$

with the bits constraint in Eqn.(4):

$$R_t(QP_1)+R_f(QP_2)=R_c. \tag{8}$$

When w_1 and w_2 are set for different applications, the optimal bits allocation R_t and R_f can be calculated based on Eqn.(7) and Eqn.(8). Therefore, the optimal values of QP_1 and QP_2 can be determined.

To solve Eqn.(7) and Eqn.(8), the $D_t - R_t$ and $D_f - R_f$ models are established. In our experiments, We first compress the texture image of the dataset with different bit-rates, then the exponential model is applied to fit the $D_t - R_t$. The $D_f - R_f$ model can be obtained in a similar manner. Specifically, they are represented as follows:

$$\begin{aligned} D_t &= a_1 * e^{b_1 R_t} \\ D_f &= a_2 * e^{b_2 R_f} + c_2 * e^{d_2 R_f} \end{aligned} \tag{9}$$

where a_i, b_i, c_i and d_i are parameters of the exponential model. In particular, these parameters are closely related to the dataset, they should be retrained in different dataset but the general tendency of the curve will not change. Fig. 2 plots the relationship between the fitted relationship and actual one on the LFW dataset.

IV. Experimental Results

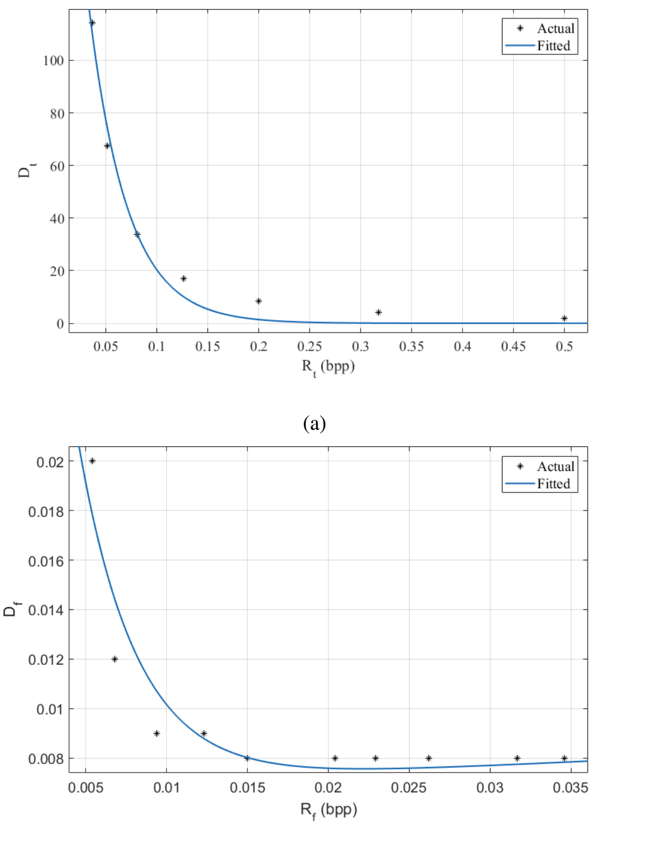
In this section, experiments are conducted to demonstrate the effectiveness of the proposed joint bit allocation scheme based on rate-TFQI optimization. Section IV-A describes the experimental settings. Section IV-B and Section IV-C evaluate our approach from the perspectives of rate-TFQI performance and rate-accuracy, respectively.

A. Experiment settings

In the experiments, the LFW dataset [18] is utilized. These pictures are cropped and resized to 160 × 160 by MTCNN for deep feature extraction. Fig. 3 shows examples of these facial images in the LFW dataset, including the original and processed images.

The cropped facial images are compressed by the HEVC reference software HM-16.0 with the All-Intra (AI) configuration. Seven QP values are selected to cover from low bit rate to high bit rate coding , which are 22, 27, 32, 37, 42, 47, and 51.

The features are extracted from the 160 × 160 facial images using the FaceNet [13], each face is compactly represented by a 128 dimensional float-point vector which will be compressed by the proposed feature coding approach. To illustrate the performance of proposed framework, we compare the performance between our scheme and the conventional CTA scheme using the TFQI metric. When calculating the accuracy, 6000 image pairs are generated, half of them are of the same



(a)

(b)

Fig. 2. Relationship between the rate and distortion. (a) The variations of D_t under different coding bits. $D_t = 293.4e^{-26.7R_t}$, $R-squared = 0.9826$. (b) The variations of D_f under different coding bits (The coding bits are used to represent the compressed feature, which are extracted from lossless image) $D_f = 0.048e^{-276.8R_f} + 0.0068e^{3.882R_f}$, $R-squared = 0.8393$.

Fig. 3. Examples of facial images in the LFW dataset. First row: the original images from LFW dataset. Second row: the detected facial images (160×160 pixels) by MTCNN.

people while half are of different people, the predicted results calculate by the pre-trained model are compared with the groundtruth to measure the accuracy.

B. Evaluation on the rate-TFQI performance

Firstly, we compare the proposed method with the conventional CTA method in terms of the rate-TFQI performance. As shown in Fig. 4, the rate-TFQI performance is evaluated under different values of w_1, for both the proposed and conventional CTA scheme. We can see that the proposed scheme achieves better performance in terms of TFQI than the conventional method, especially at lower bit-rate conditions. Moreover, the lower the given bit-rate, the better performance our scheme compared to the conventional one.

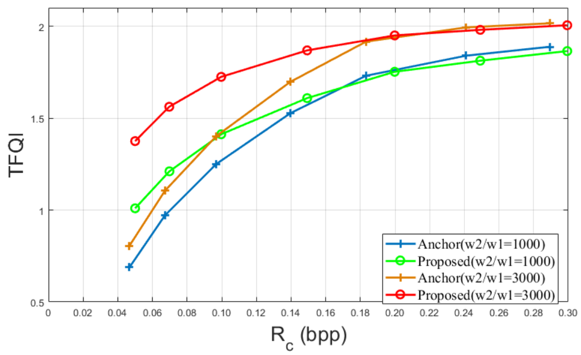


Fig. 4. Performance comparisons in terms of rate-TFQI.

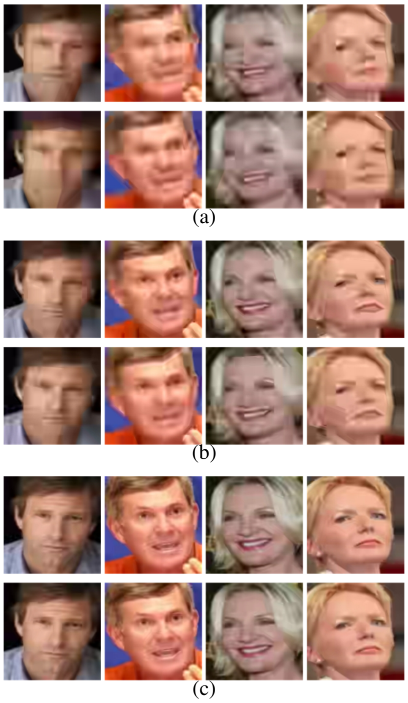
Fig. 5 shows visual quality of example facial images, encoded by HM-16.0 with the proposed and conventional schemes at the same bit-rate. Though the coding bits of texture is lower in the proposed scheme due to the consumption of the feature representation, one can barely observe the difference between them, especially for relatively high bit-rate. In our case, there is no obvious difference between such two frameworks in the sense of visual quality which indicates that the proposed framework can achieve comparable visual result with texture only framework. With the help of joint optimization for texture and feature coding, we significantly promote the performance of CV tasks while maintaining similar visual quality, which is the ultimate objective proposed framework. The proposed scheme achieves a good trade-off between visual quality and the analysis accuracy.

C. Evaluation on the rate-accuracy performance

A great advantage of our scheme is that it can greatly improve the performance of analysis tasks especially at low bit-rate while reserving the texture information. In this paper, this is reflected in the accuracy rate of face recognition tasks on the LFW dataset. Fig. 6 shows the rate-accuracy curves of the proposed and conventional scheme. We can observe that the proposed scheme can greatly improve the accuracy at the relatively low bit-rate due to the severe distortion introduced in the texture encoding process. With the increase of the coding bits, the accuracy of conventional scheme also increases while the accuracy of the proposed scheme maintains at a saturated level. Finally, if the coding bits is high enough, the accuracy of these two schemes converge to the same point. As such, our scheme can effectively improve the accuracy rate in a wide range of bit rate.

V. Conclusion

In this paper, a joint optimization framework is proposed for textures and features compression. To guide the joint optimization process, we first propose a novel TFQI metric for measuring the overall performance of visual quality and recognition


(a)

(b)

(c)

Fig. 5. Comparisons of the visual quality. First row: reconstructed images by the conventional scheme. Second row: the reconstructed images by the proposed scheme. (a) 0.05 bpp; (b) 0.07 bpp; (c) 0.15 bpp. The number of coding bits here is obtained by averaging on the whole LFW dataset, and these four images are Aaron_Eckhart, Mark_Brown, Dagmar_Dunlevy and Gabi_Zimmer.

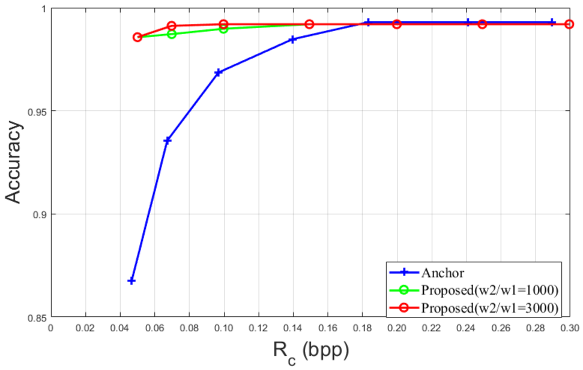


Fig. 6. Performance comparisons in terms of rate-accuracy.

accuracy. Based on the TFQI, we formulate the joint coding scheme as a constrained bit allocation problem. Extensive experiments show that the proposed scheme achieves better rate-TFQI performance and rate-accuracy performance at a relatively low bit-rate than conventional CTA approach. The future work will focus on the framework extension for video coding and other analysis tasks such as retrieval and tracking.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (61571017), National Postdoctoral Program for Innovative Talents (BX201600006), Top-Notch Young Talents Program of China, High-performance Computing Platform of Peking University, which are gratefully acknowledged.

References

[1] W. Gao, Y. Tian, T. Huang, S. Ma, and X. Zhang, "The ieee 1857 standard: Empowering smart video surveillance systems," *IEEE Intelligent Systems*, vol. 29, no. 5, pp. 30–39, 2014.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer vision–ECCV 2006*, pp. 404–417, 2006.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[5] L. Baroffio, M. Cesana, A. Redondi, S. Tubaro, and M. Tagliasacchi, "Coding video sequences of visual features," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 1895–1899.

[6] L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, "Coding visual features extracted from video sequences," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2262–2276, 2014.

[7] M. Makar, S. S. Tsai, V. Chandrasekhar, D. Chen, and B. Girod, "Interframe coding of canonical patches for mobile augmented reality," in *Multimedia (ISM), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 50–57.

[8] ——, "Interframe coding of canonical patches for low bit-rate mobile augmented reality," *International Journal of Semantic Computing*, vol. 7, no. 01, pp. 5–24, 2013.

[9] 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.

[10] L. Baroffio, A. Canclini, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, "Coding local and global binary visual features extracted from video sequences," *IEEE transactions on image processing*, vol. 24, no. 11, pp. 3546–3560, 2015.

[11] X. Zhang, S. Ma, S. Wang, X. Zhang, H. Sun, and W. Gao, "A joint compression scheme of video feature descriptors and visual content," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 633–647, 2017.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] 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.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] 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.

[16] G. J. Sullivan, J. 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.

[17] X. Yang, W. Ling, Z. Lu, E. P. Ong, and S. Yao, "Just noticeable distortion model and its applications in video coding," *Signal Processing: Image Communication*, vol. 20, no. 7, pp. 662–680, 2005.

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.