Distinguishing Ultra High Definition Images By Deep Neural Networks

Buqing Nie (516030910554), LingJia Meng (516030910553), Xinyu Wang (516030910557), Zixuan Chen (516030910545), and Hao Zhang (516030910559)

I. Introduction

ULTRA high definition (UHD) standard is obtaining increasing popularity among television industries, taking the place of traditional high-definition television (HDTV) [1]. While customers are enjoying bettwe visual performance of UHD videos, producers are competing to profit from the prospective UHD content production. Since UHD production requires advanced filming technologies and expensive equipments, malicious producers without qualification attempt to fabricate UHD videos by up-sampling videos with less resolution than UHD standards, such as 1080p and 720p.

We intend to distinguish ultra high definition contents from fake-UHD contents which are up-sampled from low-resolution sources by deep neural networks. Our design consists of a sample-and-vote strategy and classification models. The sample-and-vote strategy leverages the locality of differences between real-UHD and fake-UHD contents, in order to reduce the size of model input and increase the model efficiency. For classification models, we choose two classification models, and propose an enhanced convolutional neural network with Discrete Cosine Transform (DCT) as the first layer, named DCT-CNN. We comprehensively evaluate the performance of three models on two datasets generated from 200 real-4K images. Our experiments show that all three models achieve nearly perfect performance, and our proposed DCT-CNN model is better than traditional CNN model in both training efficiency and prediciton accuracy.

II. Background

A. Different Image Resolutions

Ultra high definition (UHD) includes 4K UHD and 8K UHD [1]. 4K and 8K refer to display resolution standrads with an aspect ratio of 16:9. Table I shows resolution standards and their corresponding pixels.

B. Related Works

Zhu et al.[2] proposed a dedicated algorithm to classify whether a video frame is naturally 4K or not base on the frequency analysis. They find that the widely used interpolation methods are unable to reproduce high frequency information. They compute the energy spectrum of DCT coefficients in vertical and horizontal directions and obtain normalized energy spectrum. Then they derive a frequency based metric (FBM) using the cumulative energy spectrum (CES). To avoid the interference of different level of image complexity, they propose adaptive thresholds with with the free energy modelling techniques for different images.

Resolutions	8K	4K	1440p	1080p	720p
Pixels	7680× 4320	3840× 2160	2560× 1440	1920× 1080	1280× 720

TABLE I
Resolutions and Corresponding Pixels

III. Design Overview

We distinguish ultra high definition (UHD) contents with our major focus on distinguishing real-4K images from false-4K ones. A real-4K image is an image generated (downsampled or compressed) from a source with resolution no less than 4K. A false-4K image, on the contrary, is an image generated from a source with resolution less than 4K, such as an 1080p- or 720p-resolution source. Real-4K images are different from false-4K ones mainly in local features. For example, figure 2 illustrates pixel differences between a real-4K image (figure 1, compressed when displayed), and its corresponding false-4K image. In figure 2, lighter regions indicate larger differences. This figure demonstrates that in the background area where colors only vary in a global scale, differences are much smaller than areas where colors vary in a local scale, such as hairs and clothes.

A 4K image, as a feature map, contains too much redundant information in our task to differentiate between real-4K and false-4K images. A 4K feature map has $3,840 \times 2,160 \times 3 = 24,883,200$ features, which consumes enormous computational resources. In order to decrease the size of our model, as well as to boost the model efficiency, we propose a sample-and-vote stratergy. A 4K image is first sampled as nine 224×224 representative sub-images. Then the classification model is applied to each sub-image, which tells us whether this sub-image belongs to a real-4K or a false-4K image. At last, a final decision is made by taking the majority vote from the nine sub-images. Section III-A provides more details of this strategy.

Distinguishing real-4K and false-4K images belongs to a classification problem. Therefore, a classification model is the core component of our design. To the best of our knowledge, no prior attempt has been made to classify real-4K and false-4K images by deep neural netowrks. Thus, we select two typical deep neural network classification models, including convolutional neural network (CNN) [3] and deep residual network (ResNet) [4], propose an enhanced CNN with Discrete Cosine Transform (DCT) as the first layer, named

Fig. 1. A real-4K image

Fig. 2. Pixel differences between a real-4K and its corresponding false-4K image (lighter regions indicate larger differences)

DCT-CNN, and comprehensively evaluate the performance of three models on this classificaiton task. More information of our selected models can be found in section III-B.

A. Sample-and-Vote Strategy

We apply a sample-and-vote strategy to every 4K image. For each 4K image, we obtain nine representative sub-images from the original 4K image. The nine images are specifically chosen to efficiently represent the overall information in a 4K image. As figure 3 shows, nine sub-images are cropped from nine distinct quarter points of a 4K image.

Now we explain why nine sub-images should be sufficient to classify a 4K image. We observe from masssive preliminary experiemnts that different pixels scatter across the whole image, while a large portion distribute in small regions with varying colors, such as edges. See figure 2, this figure displays pixel differences between a real-4K image and its corresponding false-4K image. Lighter regions indicate larger pixel differences. This figure shows that most pixel differences distribute at positions such as hairs and clothers, where colors vary locally. So we argue that we only need local information to differentiate between real-4K and false-4K images. Therefore, although a sub-image has much less pixels compared with a 4K image, it still retains adequate information for classification.

A sub-image is designed to have 224×224 pixels with 3 channels. A sub-image only has $224 \times 224 \times 3 = 150,528$ features, while a 4K image has $24,883,200$ features. So by sampling representative sub-images, we reduces the model input size to 0.6% the original size.

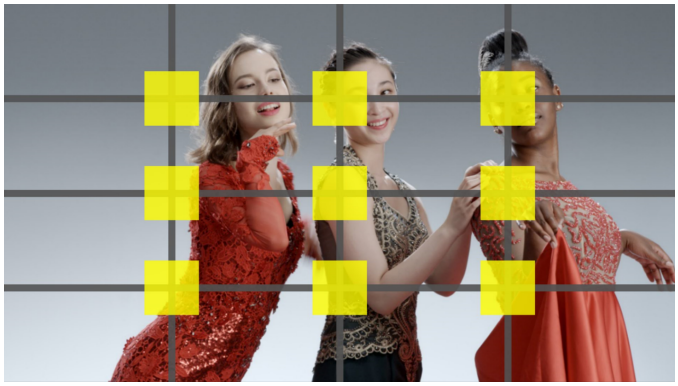


Fig. 3. Sub-image sampling from original 4K image. Nine sub-images (in yellow) are cropped from nine distinct quarter points of the original image.

Each sub-image is used as an isolated model input. The prediction of a single 4K image is drawn from the votes of all nine sampled sub-images. The majority of votes is taken as the final prediction.

B. Classification Models

We distinguish ultra high definition images with our major focus on distinguishing real-4K images from false-4K ones. This task is generally a classification problem. With adequate ground-truth data, we propose to solve this problem by utilizing deep neural networks as our classification model.

Deep neural networks have long been used as a solution to classification problems. Many types of neural networks have been developed to solve such problems. To the best of our knowledge, none of previous works has utilized deep neural networks to distinguish between real-4K and false-4K images. Therefore, we select 3 typical deep neural network netowrk classification models, and comprehensively evaluate the performance of these models on this classification task.

We intend to analyze the complexity of distinguishing real-4K and false-4K images by evaluating the performance of neural networks which are designed to handle classification problems of different complexities. We select two traditional models with increasing ability to handle complex classification problems. Specifically, we select (1) convolutional neural network (CNN), and (2) deep residual network (ResNet). We also propose an enhanced CNN with Discrete Cosine Transform (DCT) as the first layer, named DCT-CNN, as the third model in our experiments. A more detailed description of DCT-CNN is presented in section III-C

For ResNet we select the standard ResNet-18 architecture [5]. For CNN the architecture we use is listed as follows. (Batch normalization and ReLU layers are omitted. In 2d-convolutional and fully connected layers, “in” represents input channels, “out” represents output channels, and “kernel” represents kernel size.)

- Conv2d (in=3, out=64, kernel=5, padding=5)
- MaxPool2d

- Conv2d (in=64, out=128, kernel=5, padding=2)
- MaxPool2d
- Conv2d (in=128, out=16, kernel=5, padding=2)
- MaxPool2d
- FullyConnected (in=$16 \times 28 \times 28$, out=2)

C. DCT-CNN Model

We propose an enhanced CNN with DCT as the first layer, named DCT-CNN. The idea of adding a DCT layer as the first layer comes from our empirical analysis. When real-4K and false-4K images are transformed by DCT, their frequency-domain spectrum graphs show apparent differences, especially in the lower-left regions, i.e. high-frequency domain. (See figure 4.) Therefore we draw histograms of these high-frequency-domain spectrum, we observe that for a real-4K spectrum, high-frequency DCT values distribute evenly, while for a false-4K spectrum, high-frequency DCT values vary across values close to zero but mostly center at zero. Our observation in histograms explains for the apparent difference in frequency-domain spectrum between real-4K and false-4K images.

IV. Evaluation

A. Experimental Setup

1) Dataset: Our ground truth data are 200 real-4K images taken from 20 sequences of 4K videos. Our dataset is derived from these 200 real-4K images. We first generate false-4K images. False-4K images are up-sampled from source images with resolution less than 4K. To obtain images with resolution less than 4K, we down-sample each real-4K image to two low-resolution images, namely, 720 and 1080p images. Then each low-resolution image is up-sampled to 4K resolution by an up-sampling algorithm. We select Lanczos [6] as our up-sampling algorithm.

As mentioned in section III-A, our design leverages a sample-and-vote strategy, where nine sub-images are sampled from each 4K image. For each real-4K and false-4K images, nine sub-images are cropped from nine distinct quarter points. Figure 3 illustrates the sub-image sampling process.

With 200 real-4K images, we generated 200 false-4K images up-sampled from 1080p source images. From real-4K images we sampled $1,000$ sub-images as positive data samples and from false-4K images we sampled $1,000$ sub-images as negative data samples. These $2,000$ data samples constitute the False-1080p-to-4K dataset, denoted as D_{1080p}. Similarly, we also generated 200 false-4K images up-sampled from 720p source images, and obtained a False-720p-to-4K dataset comprised of 1000 positive data samples and 1000 negative data samples, denoted as D_{720p}.

Since classification models are trained on 224×224 sub-images but the final decision is made on a 4K image, we show model performance on both image sizes, where the prediction of a 4K image is made by taking majority votes from all of its nine sub-images. We denote the prediction results on sub-images by $D(224 \times 224)$ and 4K images by $D(3840 \times 2160)$. In summary, we use the following four notations to represent different prediction performance:

Models	Batch size	Learning rate	Training iterations
CNN	64 / 16	0.0001	100
DCT-CNN		0.0001	100
ResNet18		0.001	50

TABLE II
TRAINING HYPERPARAMTERS

Dataset	index	Batch size 64	Batch size 16
$D_{1080p}(224 \times 224)$	F1-score	0.9921	0.9994
	accuracy	0.9922	0.9994
	recall	0.9883	0.9988
	precision	0.9960	1.0
$D_{720p}(224 \times 224)$	F1-score	0.9921	0.9994
	accuracy	0.9922	0.9994
	recall	0.9883	0.9988
	precision	0.9960	1.0
$D_{1080p}(3840 \times 2160)$	F1-score	0.9975	1.0
	accuracy	0.9975	1.0
	recall	1.0	1.0
	precision	0.9950	1.0
$D_{720p}(3840 \times 2160)$	F1-score	0.9975	1.0
	accuracy	0.9975	1.0
	recall	1.0	1.0
	precision	0.9950	1.0

TABLE III
CONVOLUTIONAL NEURAL NETWORK CLASSIFICATION RESULTS

- $D_{1080p}(224 \times 224)$: performance of predicting sub-images on the False-1080p-to-4K dataset
- $D_{720p}(224 \times 224)$: performance of predicting sub-images on the False-720p-to-4K dataset
- $D_{1080p}(3840 \times 2160)$: performance of predicting 4K images on the False-1080p-to-4K dataset
- $D_{720p}(3840 \times 2160)$: performance of predicting 4K images on the False-720p-to-4K dataset

2) Hyperparameters: Table II shows the hyperparamters that we selected for our experiments. For each model we selected two different batch sizes in order to analyze the influence of batch size on the classification performance.

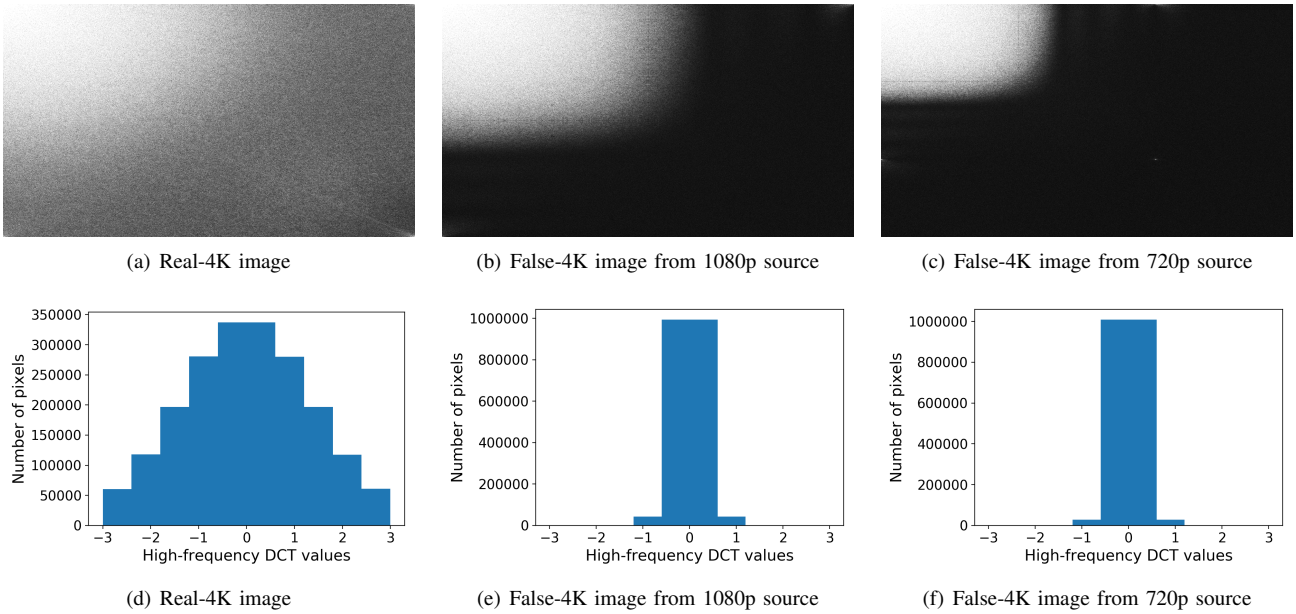
B. Results

Table III, IV, and V present the performance of CNN, DCT-CNN, and ResNet-18 respectively. We evaluate the model performance from four indexes, including F1-score, accuracy, recall rate, and prediction precision in the testing dataset. We also evaluate the influence of batch size on the model performance by conducting experiments for different settings of batch size, incldluing 64 and 16.

Our results show that all of three models achieve nearly perfect performance in all settings. Also, by enhancing CNN with DCT, the DCT-CNN model achieves better performance than CNN in all settings. Therefore, we conclude that our design perfectly distinguishes real-4K and false-4K images, and the proposed DCT-CNN model strengthens the performance of CNN in prediction accuracy. We also observe that training DCT-CNN needs much fewer iterations to converge than CNN, so the DCT-CNN model is better than CNN in training efficiency.

V. Conclusion

We propose to distinguish ultra high definition (UHD) contents with our major focus on distinguishing real-4K and



(a) Real-4K image (b) False-4K image from 1080p source (c) False-4K image from 720p source

(d) Real-4K image (e) False-4K image from 1080p source (f) False-4K image from 720p source

Fig. 4. Frequency-domain spectrum and corresponding high-frequency DCT value distribution for a real-4K image (figure 1), its false-4K image up-sampled from 1080p source, and another false-4K image up-sampled from 720p source

Dataset	Index	Batch size 64	Batch size 16
$D_{1080p}(224 \times 224)$	F1-score	1.0	1.0
	accuracy	1.0	1.0
	recall	1.0	1.0
	precision	1.0	1.0
$D_{720p}(224 \times 224)$	F1-score	1.0	1.0
	accuracy	1.0	1.0
	recall	1.0	1.0
	precision	1.0	1.0
$D_{1080p}(3840 \times 2160)$	F1-score	1.0	1.0
	accuracy	1.0	1.0
	recall	1.0	1.0
	precision	1.0	1.0
$D_{720p}(3840 \times 2160)$	F1-score	1.0	1.0
	accuracy	1.0	1.0
	recall	1.0	1.0
	precision	1.0	1.0

TABLE IV
DCT-CNN CLASSIFICATION RESULTS

Dataset	Index	Batch size 64	Batch size 16
$D_{1080p}(224 \times 224)$	F1-score	1.0	0.9997
	accuracy	1.0	0.9997
	recall	1.0	0.9994
	precision	1.0	1.0
$D_{720p}(224 \times 224)$	F1-score	1.0	1.0
	accuracy	1.0	1.0
	recall	1.0	1.0
	precision	1.0	1.0
$D_{1080p}(3840 \times 2160)$	F1-score	1.0	1.0
	accuracy	1.0	1.0
	recall	1.0	1.0
	precision	1.0	1.0
$D_{720p}(3840 \times 2160)$	F1-score	1.0	1.0
	accuracy	1.0	1.0
	recall	1.0	1.0
	precision	1.0	1.0

TABLE V
DEEP RESIDUAL NETWORK CLASSIFICATION RESULTS

false-4K images. Our design is comprised of a sample-and-vote strategy, which reduces the training cost and increase the prediction efficiency, and a classification model. For the classification model we select convolutional neural network (CNN), and deep residual network (ResNet), and propose an enhanced CNN with Discrete Cosine Transform (DCT) as the first layer, based on our observation about the frequency-domain spectrum differences between real-4K and false-4K images. We comprehensively evaluate the performance of our design on two datasets generated from 200 real-4K images. Our experiments show that all three models achieve nearly perfect performance, and our proposed DCT-CNN model is better than traditional CNN model in both training efficiency and prediciton accuracy.

REFERENCES

[1] Wikipedia contributors, "Ultra-high-definition television — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Ultra-high-definition_television&oldid=932656068, 2019, [Online; accessed 28-December-2019].

[2] W. Zhu, G. Zhai, J. Liu, J. Wang, and X. Yang, "Distinguish true or false 4k resolution using frequency domain analysis and free-energy modelling," in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. IEEE, 2016, pp. 197–202.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[5] ——, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] C. Lanczos, "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators," 1950.