# IDENTIFICATION OF HUMAN PAPILLOMAVIRUS FROM SUPER RESOLUTION MICROSCOPIC IMAGES GENERATED USING DEEP LEARNING ARCHITECTURES

Xiaohong W. Gao*[1], Xuesong Wen[1], Dong Li[1], Weiping Liu[2], Jichuan Xiong[2], Bin Xu[2], Juan Liu[2], Heng Zhang [2], Xuefeng Liu[2]

[1] Faculty of Science and Technology, Middlesex University, London, UK

[2] Nanjing University of Science and Technology, Nanjing, China

## Abstract

This chapter presents five deep learning architectures for identification of Human papillomavirus (HPV) through generation of super resolution (SR) images by 4 folds. Specifically, generative adversarial deep learning networks (GAN) and a texture-based vision transformer (TTSR) architecture are applied and evaluated. As such, the generated SR images are able to display the same way a high-resolution image offers in identification of HPV like structures. In comparison, TTSR appears to perform the best with PSNR and SSIM being 28.70 and 0.8778 respectively whereas 25.80/0.7910, 18.35/0.5059. 30.31/0.8013, and 28.07/0.6074 are observed for the methods of RCAN, Pix2Pix, CycleGAN, and ESRGAN respectively. With regard to sensitivity and specificity when detecting HPV clusters, TTSR also leads with 83.6% and 83.33% respectively. It appears the computational SR images are capable to differentiate distinguishing features of HPV like particles and to determine the effectiveness of anti-HPV agents, holding promise providing insights into the formation stage of a cancer from HPV in the near future.

Latest version of May 2022.

**Keywords:** Texture transformer, Generative adversarial network (GAN), super resolution, machine learning, Human papilloma virus (HPV) like particles (HPVLPs), cervical cancer.

# 1. Introduction

This chapter extends the work presented at ICMLA 2021 [1] and concerns with the determination of the presence of human papillomavirus (HPV) from generated super resolution (SR) images that are acquired using conventional light microscopes, leading to potentially contribution to the development of an effective anti-HPV drugs in the future. The state of the art deep learning techniques are applied and compared.

As one of the leading causes of virus-induced cancers, early detection of HPV plays a crucial role in providing timely, optimal and effective intervention before such a cancer develops. While conventional light microscopy constitutes one of inseparable tools applied for studying biological cell structures, its low resolution at ~100nm per pixel falls short of detecting HPV that typically has a size of ~50nm in diameter, giving rise to visualisation of HPV and subsequent evaluation of the efficacy of anti-HPV drugs at such sub-pixel level a challenging task if not overwhelmingly. This study aims to leverage this gap by endowing conventional microscopic images with a visible account of super resolution (SR) images through computationally up-sampling by four-fold (×4) by applying the burgeoning deep learning techniques.

# 2. Background

## 2.1 HPV Virus

The human papillomavirus, or HPV, refers to a group around 150 types of viruses, including high risk type of HPV that pertain to a number of epithelial cancer development and low risk types accounting for noncancerous papilloma such as HPV 6 and 11. As a small, non-enveloped, and double-stranded DNA virus, a papillomavirus has a diameter of 52–55 nm [2] and infects mucosal by inducing cellular proliferation.

A high risk HPV remains a leading cause of virus-induced cancers, mainly being discovered in cervical and head-and-neck cancers. Among those, HPV16 and HPV18 retain the two major types that account for 70% of cervical cancer cases [3,4,5] .

Since it usually takes several years to eventuate in cancer cells from high risk HPV infection to integrate into cells, it is a clinical urgency for early detection of these viruses not only to the respect in which more specific targeted treatments to improve patients' survival rates and hence reduce the mortality can be realised, but to the fact that HPV can be eradicated all together effectively before it progresses into associated cancers.

At present, the detection of HPV mainly relies on the molecular and cellular pathological evidence through labelling HPV oncogenes or oncoproteins using the approach of polymerase chain reaction (PCR), in situ hybridisation and immunohistochemical staining. This is because conventional light microscopes (e.g. Nikon C2plus Ti2 MS (Laser scan confocal, PMT)) can only depict sample structures at a maximum of ~70nm/pixel whereas an HPV sustains a size of ~50nm in diameter. As a result, direct identification of HPV at such a sub-pixel resolution visually confronts a great challenge to discern the details of infected individual cells or regions.

While a transmission electron microscope (TEM) [6] presents an option for HPV identification at high resolution, the sample preparation procedure tends to be time-consuming, sophisticated and technical, which requires dedicated personnel to operate. Consequently, TEM is usually not readily available at many research centres, which therefore necessitates a simple imaging technique that can offer high resolution to identify nano-size structures.

Furthermore, while there exist potential drugs, it is imperative to monitor cellular responses directly, such as apoptosis induction following drug treatment, and to evaluate its effectiveness, by which microscopy proffers quicker, easier and visually achievable than any traditional detection methods including PCR approach.

While the recent advent of high-resolution microscopy (HRM) holds a promise to be able to acquire information at nanoscale, it is not prevalent and again encounters the difficulties to set up the correct complex large array of parameters.

## 2.2 High Resolution Microscope (HRM)

In cell biology, fluorescence microscopy (MS), characterised with non-invasiveness, high sensitivity, and selectivity, composes an essential tool for imaging tagged biological structures. Due to the wave-like nature of light, the resolution of a conventional fluorescence microscope is limited laterally to about 200 nm and axially to about 600 nm per pixel, which is often referred to as the Abbe-diffraction limit [7] .

Consequentially, various methods have been developed to circumvent this limit of resolution, leading to the development of high-resolution microscopy (HRM, also termed as super resolution microscopy). HRM resorts to a series of techniques in light microscopy to allow images to be taken with a higher resolution than the one imposed by the diffraction limit [8], including optical/instrumental modifications and specific labelling of samples [9,10]. Current HRM methods rely on wide-field (WF), total internal reflection fluorescence (TIRF) or confocal microscope setups, which fundamentally differ in the way fluorescently labelled samples are excited and the manner those emitted photons are detected [11] . While HRM lends itself well to providing unprecedented access to the inner functions of cells and various biological processes, making up for the shortage of conventional MS, more accurate models often require exhaustive parameter search, sophisticated optical setups and high computational cost [12]. Significantly, conventional microscopy scanners are more prevalent, more economic and easier to set up.

Hence, in this study, computational approaches are employed to generate super resolution images. In particular, texture transformer based deep learning architecture will be evaluated and enhanced. In addition, the state of the art generative adversarial deep learning networks (GANs) are evaluated and compared. These systems are trained to transform low-resolution images (e.g. 20nm/px) to a high-resolution one (5nm/px) with four-fold (×4) increase of resolution.

Specifically, this study is different in training from many other super resolution networks where mathematical formulas, e.g. bicubic, are employed to generate low-resolution images from the available data. This work employs matching pairs of low- and high-resolution images that are obtained experimentally. In this way, real experimental setup in relation to concerned microscopes can be taken into account instead of focusing on a fixed algorithm. Consequently, the trained model is applied to upscale images obtained using conventional confocal microscopy (70nm to 120nm) to identify HPV or HPV like particles (HPVLP) and to evaluate the effectiveness of the developed drug for combating HPV virus. The ground truth is obtained using Transmission Electron Microscope (TEM) to identify HPVLP which can achieve at a resolution at 0.5nm/px. Fig. 1 epitomises the data sets collected in this study, from three modalities at varying scales, including conventional microscope (d & e), HRM (b&c) and TEM (a).

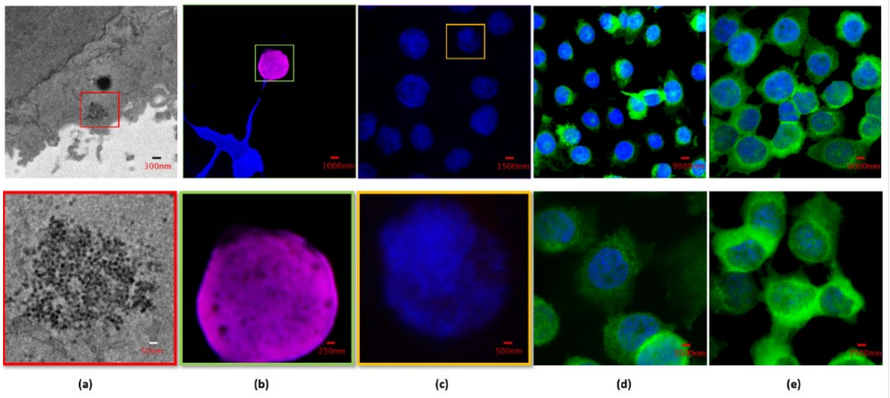


**Fig. 1.** Illustration of data sets applied in this study from three modalities, acquired at varying scales, including from (a): TEM, the ground truth; (b)(c): Data sets for training; (d)(e): low resolution datasets for testing. The bottom row of graphs of (a) to (c) are the selected regions on the top row of the same column that are acquired at a higher resolution.

## 2.3 Deep learning techniques for super resolution

In computer vision field, there are broadly two ways to contend with the fundamental low-level single image super-resolution (SISR) problem. One is from theoretical point of view and another is based on subjects' visual appearance evaluation. While SISR attempts to recover a high-resolution (HR) image from a single low resolution (LR) one, it appears that the application of deep learning neural networks (DNNs) can achieve state of the art results as pioneered by Dong et al. [13].

DNNs denote a class of computing machines that can learn a hierarchy of features by establishing high-level features from low-level ones based on biologically inspired human vision systems. One of these models is Generative Adversarial Network [14] (GAN), designating an approach to generative modelling using deep learning methods, such as convolutional neural networks (CNN) [15]. GAN performs an unsupervised machine learning and involves automatically discovering and learning the regularities or patterns from input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset.

Subsequently, a number of network architectures have been proposed to improve the super resolution (SR) performance mainly at improving Peak Signal-to-Noise Ratio (PSNR) values [16, 17, 18]. This, however, tends to be in disagreement with human observers' evaluation as pointed out in the network of SRGAN [19], one of the seminal works on the improvement of visual quality of generated SR images. Towards this end, several perceptual-driven methods are advanced, including incorporation of perceptual loss [20, 21] to optimise super-resolution models in a feature space instead of pixel space and segmentation of semantic images prior to recovering detailed textures [22]. Significantly, the application of GAN in SRGAN improves the overall visual quality of reconstruction over the PSNR-oriented methods considerably by encouraging the network to advocate solutions that look more like natural images. Specifically, an enhanced SRGAN (ESRGAN) [23] further improves the visual quality by reducing generated accompanying artefacts. ESRGAN introduces relativistic GAN [24], in which Residual-in-Residual Dense Block (RRDB) without batch normalization is utilised as basic network building blocks whereas SRGAN is built with residual blocks [25], offering consistently better visual quality with more realistic and natural textures. In generation of SR for fluorescence microscopic images, the application of a GAN-based system appears to demonstrate convincing results [12].

More recently, vision transformers (ViT) are emerging and starting to show potentials by performing computer vision tasks, such as image recognition [26, 27]. Built upon self-attention architectures and being a leading model in natural language processing (NLP), ViT appears to demonstrate excellent performance when trained on sufficient data, outperforming comparable state-of-the-art CNNs with four times fewer computational resources [28].

One of the advantages that that Transformers present is computational efficiency and scalability. It has become possible to train models of unprecedented sizes, with over 100 billion parameters [29].

In this work, the architecture of texture transformer (TTSR) [30] as well as GAN-based networks are evaluated for detection of HPV like particles (HPVLPs), or HPV viral factories. The four state of the art GAN models are ESRGAN [23], CycleGAN [31]. Pix2pix [32] and Pix2pixHD [33].

# 3. Methodology

## 3.1 Cell sample preparation for identifying HPVLP by confocal fluorescent microscopy and TEM scanning

The cervical cancer cell lines, CaSki containing HPV16 DNA sequences and C33a without HPV were used in this study as detailed in Appendix A. These laboratory samples are prepared in Middlesex University, UK.

The acquisition of EM data (Fig. 1(a)) for these samples then took place at Leicester University in the UK to provide ground truth. Before the scanning, these samples underwent a series of standard preparation processes (details are elaborated in Appendix B), then were viewed on a JEOL JEM-1400 transmission EM (TEM) (with an accelerating voltage of 120kV), a microscopic technique in which a beam of electrons is transmitted through a specimen to form an image (Fig. 1(a)). All these acquired images are saved as TIFF image format.

### 3.2 Microscopic Data Acquisition

As training datasets, both high resolution (HR) and low resolution (LR) images are obtained using Nikon A1plus Manual Microscope with settings of Laser scan Confocal and GaAsP. These fluorescence microscopic images (Fig. 1 (b) & (c)) were captured by scanning a microscopic slide carrying dual fluorescent labelling for HPV16 oncoprotein E6/E7 in green and nuclei in blue (CaSki Control) and HPV treated with drugs (C33a) (acting as normal images), which took place at Nanjing University of Science and Technology, China. Table 1 summaries the technical information about each microscope utilised in this study, including TEM, SRM and conventional microscope.

**Table 1.** Detailed information on the microscopies employed in this study where ExW = excitation wavelength and EmW = emission wavelength.

| Modality | Scanning range (per pixel (px) | ExW<br>(nm) | EmW – Red<br>(nm) | EmW – Green<br>(nm) | EmW – Blue<br>(nm) | Pin-hole<br>(µm) |
|---|---|---|---|---|---|---|
| JEOL JEM-1400 Transmissive Electron Microscope | 0.5-6nm | | | | | |
| Nikon A1plus Manual Microscope (Laser scan confocal, GaAsP) | 5-30 nm | 405, 640nm | 700<br>(Cy5) | | 450<br>(Alexa Flour ) | 12.77 |
| Nikon C2plus Ti2 MS (Laser scan confocal, | 70-180 nm | 488, 405nm | | 510-590 | 450 | 20 |

| | | | | | | |
|---|---|---|---|---|---|---|
| PMT), (Plan Apo λ 100x Oil) | | | | (FITC) | (DAPI) | |

## 3.3 Visualisation of HPVLP from HR images

Different from TEM where HPV or HPV like particles (HPVLP) can be visualised directly as illustrated in Fig. 2 (middle row), conventional microscope exhibits HPVLP in a disguised way (Fig. 2 (b) & (h)) where those colourful pixels become unfathomable. Hence further processing takes place by undertaking Fast Fourier Transform (FFT) Gaussian filtering so that extraneous features can be filtered out by removing less frequent signals in the Fourier space as presented in Fig. 2 (c) and (i), which appear to be similar Fig. 2(f) for TEM images, the ground truth (GT) by depicting clusters of individual dark particles.

In the following sections, all the images displayed have undertaken the process of FFT coloured with *brgbcmyw* lookup table, representing black, red, green, blue, cyan, magenta yellow and white.

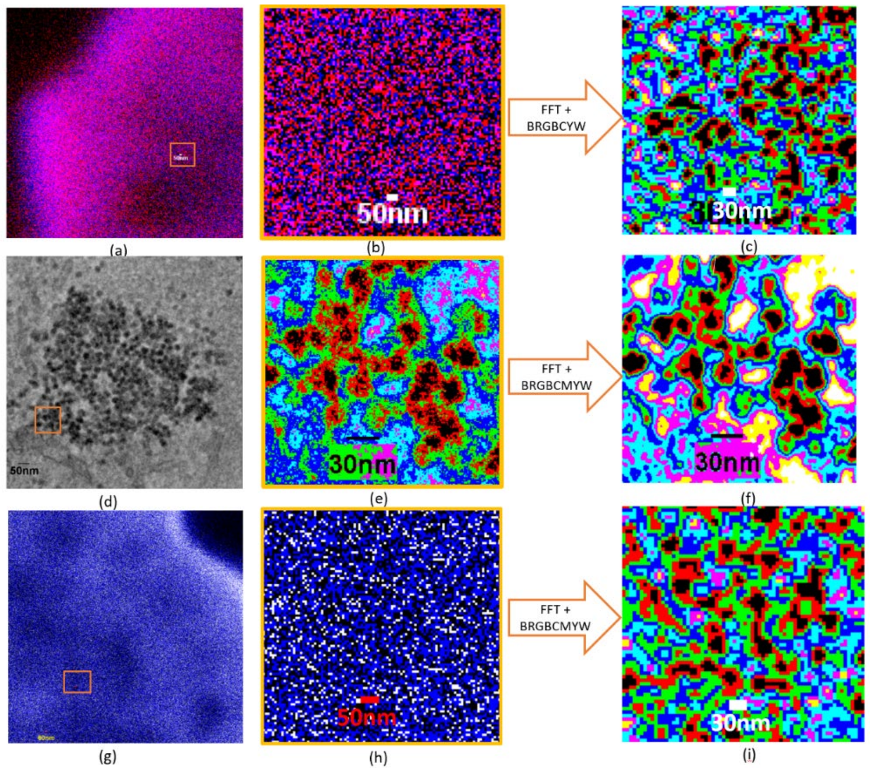


**Fig. 2.** Identification of HPVLP from HR images obtained with two different colour channels with reference to TEM data. (a)(g) HR images. (b)(h) Magnified regions in (a) and (g) respectively. (c)(i) Coloured and FFT-ed image of (b) and (h) respectively. (d) TEM data of ground truth as a reference. (e) Magnified region in (d). (f) Coloured and FFT-ed images of (e). *BRGBCMYW* = black, red, green, blue, cyan, magenta, yellow, white.

## 3.4 Implementation

The implementation is built upon Pytorch deep learning libraries [34, 35], through the application of Python language. The training and testing took place under Windows 10 system with one GPU Nvidia GeForce GTX1060 with 16 Gbyte memory. The training samples comprise 2431 images with 785 for validation and are of high resolution (x4). Test samples has 100 low resolution images containing 121 HPVLP clusters with 40 samples being normal. The input size is 256×256 pixels whereas the output size from

CycleGAN and Pix2pix remains the same. For Pix2pixHD and ESRGAN, the output size is 1024×1024 pixels, i.e., four time (x4) bigger. The training takes 50 epochs to complete for each network. For the application of TTSR, each patch is of 16×16 pixels.

Two common measures are employed to calculate the similarity between HR (ground truth (GT)) and SR, which are structural similarity (SSIM) (Eq. (1)) and peak signal-to-noise ratio (PSNR) (Eq. (2)).

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{x,y}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{1}$$

Where $\mu_x$ , $\mu_y$ are the averages of $x$, $y$, with $\sigma_x^2, \sigma_y^2$ being the variances of $x$ , $y$ respectively and $\sigma_{x,y}$ the covariance of $x$ and $y$. The variables of $c_1$ and $c_2$ are applied to stabilize the division when a small denominator occurs and are set to be $(0.01L)^2$ and $(0.03L)^2$ respectively, whereby L stands for the dynamic intensity range of an image, e.g. L=255 for a 8-bit image.

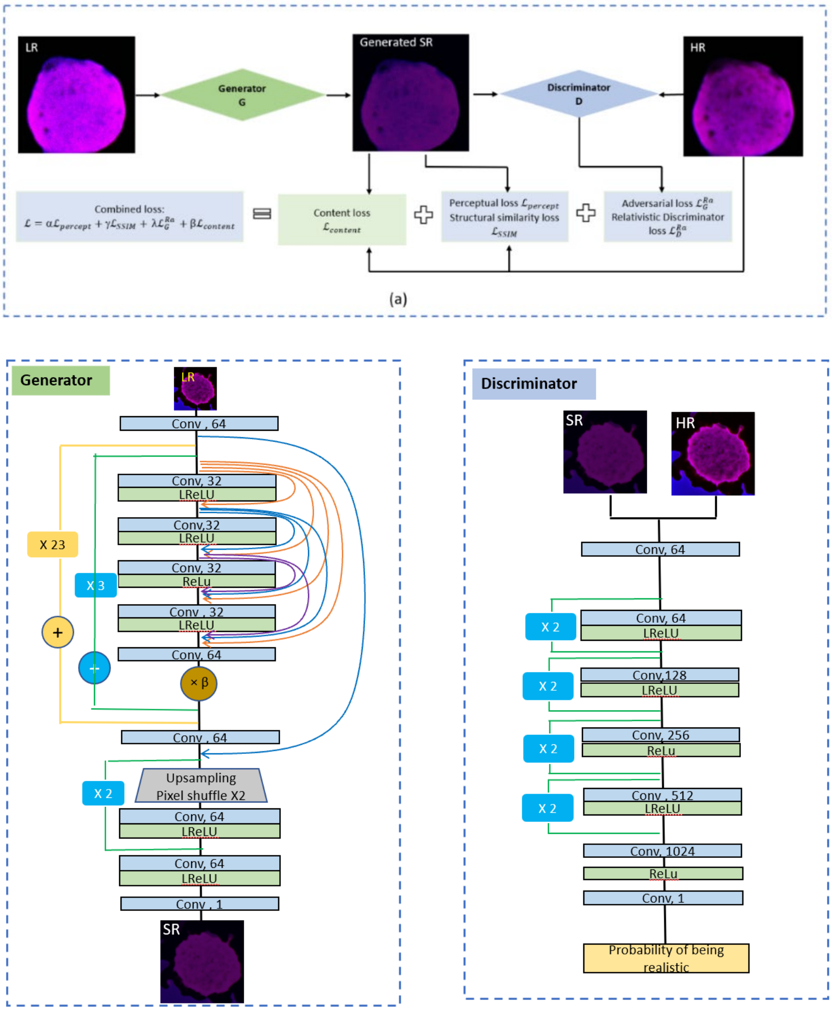
In Eq. (2), $MAX_I$ refers to the maximum possible value of an image ($I$) (e.g. 255 for 8-bit) and $MSE$ the mean squared error.

$$PSNR = 20\log_{10} MAX_I - 10\log_{10} MSE \tag{2}$$

In addition, sensitivity and specificity are employed to calculate the accuracy of detection of HPV clusters from both HR and SR images.

### 3.5 GAN Architectures for Super Resolution Images

Conventionally, GAN based networks are favoured for generation of super resolution images. In this work, four GAN-oriented architectures (ESRGAN, CycleGAN. Pix2pix and Pix2pixHD) are evaluated. As illustrated in Fig. 3, the architecture of ESRGAN comprises two sub-networks, a generator and a discriminator where the generator contains twenty-three Residual-in-Residual Dense Blocks, to transform low-resolution images (e.g. 20nm/px) to a high-resolution one (5nm/px) with four-fold (×4) increase of resolution.


LR
Generator G
Generated SR
Discriminator D
HR
Combined loss: $\mathcal{L} = \alpha\mathcal{L}_{percept} + \gamma\mathcal{L}_{SSIM} + \lambda\mathcal{L}_G^{Ra} + \beta\mathcal{L}_{content}$
Content loss $\mathcal{L}_{content}$
Perceptual loss $\mathcal{L}_{percept}$ Structural similarity loss $\mathcal{L}_{SSIM}$
Adversarial loss $\mathcal{L}_G^{Ra}$ Relativistic Discriminator loss $\mathcal{L}_D^{Ra}$
(a)
Generator
LR
Conv , 64
X 23
X 3
Conv, 32 LReLU
Conv,32 LReLU
Conv, 32 ReLu
Conv , 32 LReLU
Conv, 64
+
+
× β
Conv , 64
X 2
Upsampling Pixel shuffle X2
Conv, 64 LReLU
Conv, 64 LReLU
Conv , 1
SR
Discriminator
SR
HR
Conv, 64
X 2
Conv, 64 LReLU
X 2
Conv,128 LReLU
X 2
Conv, 256 ReLu
X 2
Conv , 512 LReLU
Conv, 1024
ReLu
Conv, 1
Probability of being realistic

(b) (c)

**Fig. 3.** The architecture of GAN applied in this study. (a) Overall diagram. (b) Generator. (c) Discriminator.

Hence, the total loss ($\mathcal{L}_{GAN}$ ) for the generator in Eq. (3) integrate perception loss ($\mathcal{L}_{percept}$), structural similarity (SSIM) loss, content loss ( $\mathcal{L}_{content}$) and adversarial loss ($\mathcal{L}_G^{Ra}$) .

$$\mathcal{L}_{GAN} = \alpha\mathcal{L}_{percept} + \gamma\mathcal{L}_{SSIM} + \lambda\mathcal{L}_G^{Ra} + \beta\,\mathcal{L}_{content} \tag{3}$$

In Eq. (3), SSIM (Eq.(1)) index is defined in Eq. (4) to measure the similarity between two images in relation to spatial structure where $x_f, x_r$ indicate fake (SR) and real (HR) images respectively.

$$\mathcal{L}_{SSIM} = 1 - SSIM(x_f, x_r) \tag{4}$$

The α, γ, λ and β in Eq. (3) are the coefficients to balance different loss terms and are set to be 1, 0.1, 0.005 and 0.1 respectively for perceptual, SSIM, adversarial and content losses in this study.

In addition, Discriminator (D) loss is formulated in Eq. (5),

$$\mathcal{L}_D^{Ra} = -\mathbb{E}_{x_r}\left[\log\left(D_{Ra}\left(x_r, x_f\right)\right)\right] - \mathbb{E}_{x_f}\left[\log\left(1 - D_{Ra}\left(x_r, x_f\right)\right)\right] \tag{5}$$

where

$$D_{Ra}\left(x_r, x_f\right) = \sigma(C(x_r) - \mathbb{E}_{x_f}[C(x_f)] \tag{6}$$

In Eq. (6), σ refers to a sigmoid function whereas $C(x)$ represents the non-transformed discriminator output. $\mathbb{E}_{x_f}[.]$ takes the average of all fake data in a mini-batch (8 in this study).

The adversarial loss for the generator presents a symmetrical form as expressed in Eq. (7).

$$\mathcal{L}_G^{Ra} = -\mathbb{E}_{x_r}\left[\log\left(1 - D_{Ra}\left(x_r, x_f\right)\right)\right] - \mathbb{E}_{x_f}\left[\log\left(D_{Ra}\left(x_f, x_r\right)\right)\right] \tag{7}$$

where $x_f = G(x_i)$, indicates the generator output of fake image $x_f$, and $x_i$ stands for the input of an LR image.

The content loss that is calculated in Eq. (8), evaluates the 1-norm distance between recovered image $G(x_i)$ from the generator (Fig. 3(b)) and the ground-truth $x_r$ using mean pixel-wise absolute error.

$$\mathcal{L}_{content} = \mathbb{E}_{x_i} \| G(x_i) - x_r \|_1 \tag{8}$$

The perceptual loss function [21] runs by summing up all the squared errors between all the pixels in a feature layer and taking the mean as given in Eq. (9).

$$\mathcal{L}_{percept} = \mathbb{E}_{x_i} \left\| \ell_{feature}^{\phi,j}(G(x_i), x_r) \right\|_1 \tag{9}$$

where

$$\ell_{feature}^{\phi,j}(G(x_i), x_r) = \frac{1}{C_j H_j W_j} \left\| \phi_j(G(x_i)) - \phi_j(x_r) \right\|_2^2 \tag{10}$$

In Eq. (9), ϕ represents a pre-trained model built upon VGG19 [36] whereas $\phi_i(x)$ in Eq. (10) refers to the activations of the $j_{th}$ layer of the network $\phi$ with a feature map of the shape $C_j \times H_j \times W_j$.

In Fig. 4, the network of Pix2pixHD [33] is depicted, where the training takes place to translate original images (A) to its corresponding processed maps (B) that is undergone FFT Gaussian filtering and coloured using a colour lookup table, *brgbcmyw*.

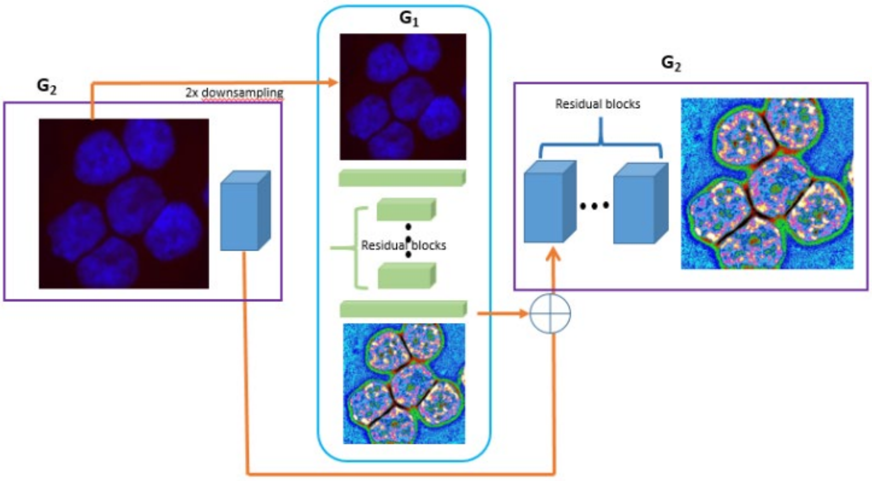


**Fig. 4.** The architecture of Pix2pixHD.

By decomposing the generator into two sub-networks, $G_1$ and $G_2$ as global generator and local enhancer networks respectively, the Pix2pixHD network first trains a residual network $G_1$ on lower resolution images. Then, another residual network $G_2$ is appended to $G_1$ so that the two networks are trained jointly on high resolution images. Specifically, the input to the residual blocks in $G_2$ (right most graph) is the element-wise sum of the feature map from $G_2$ (left graph) and the last feature map from $G_1$. In this work, the input image has a resolution of 256×256 whereas the output of the coloured map has 1024×1024 pixels, a four-folder increase. The overall loss integrates both GAN (Eq.(3)) loss and feature matching loss as formulated in Eq.(11) [33].

$$\min_G((\max_{D_1,D_2,D_3} \sum_{k=1,2,3} \mathcal{L}_{GAN}(G, D_k)) + \lambda \sum_{K=1,2,3} \mathcal{L}_{FM}(G, D_k)) \qquad (11)$$

Where λ is a factor that controls the importance of the above two parts, $\mathcal{L}_{FM}$ refers to the feature matching loss and $D_k$ $(k = 1,2,3)$ the feature extractor, extracting features from each of the three blocks in Fig. 4. During the test, the increasing size is set to be 4, so that Pix2pixHD will up-sample, fine tune and output a high resolution image four time larger than the input one.

## 4. Texture transformer network

In Fig. 5, the architecture of texture transformer (TT) [30] is presented and applied in this study [2]. TT comprises four components: the learnable texture extractor (LTE), the relevance embedding module (RE), the hard-attention module for feature transfer (H) and the soft-attention module for feature synthesis (S). In the Fig., low-resolution (LR) and LR↑ represent the input and its 4 times higher (x4) bicubic-up-sampled counterpart respectively whereas Ref images are x4 high resolution (HR) ground truth (GT) images acquired experimentally. In addition, bicubic down-sampling (4x) and up-sampling (x4) on Ref images take place to obtain Ref↓↑ which is domain-consistent with LR↑. These images are the input to the texture transformer that produces a synthesized feature map, for the prediction of HR by generating super resolution (SR) images.

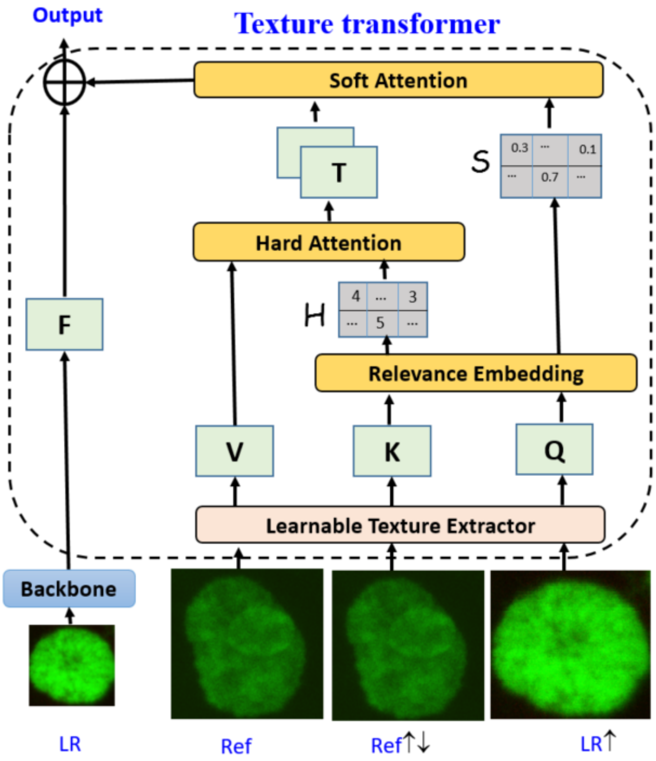


**Fig. 5.** The texture transformer [4] architecture where LR=low resolution, Ref=GT (x4), LR↑ = up-sampling by x4 from LR employing bicubic formula, and ↑↓ refers to up-sampling and down-sampling by 4 folds.

Specifically, the learnable texture extractor (LTE) extracts texture features of Q (query), K (key) and V (value), the three basic elements of the attention mechanism inside a transformer. Subsequently, the relevance embedding component estimates the similarity between Q and K as calculated in Eq. (12), where $q_i$ $(i \in [1, H_{LR} \times W_{LR}])$ and $k_j (j \in [1, H_{Ref} \times W_{Ref}])$, representing the patches unfolded from Q and K respectively, where $H_{LR}$ and $W_{LR}$ refer to the height and width of a LR image.

$$r_{i,j} = \langle \frac{q_i}{\|q_i\|}, \frac{k_i}{\|k_i\|} \rangle \tag{12}$$

Based on Eq. (12), hard attention is calculated to transfer features to a hard-attention map $H$ from the most relevant position in $V$ for each query $q_i$ as formulated in Eq. (13), where $h_i$ ($i \in [1, H_{LR} \times W_{LR}]$).

$$h_i = arg\max_j r_{i,j} \tag{13}$$

On the other hand, the component of soft attention is to synthesize features from the transferred HR texture features $T$ and L$R$ features $F$ that is obtained from a Dense neural network (DNN) backbone (e.g. ResNet50) from $LR$ images, as computed in Eq. (14), forming a soft-attention map S of confidence for $T$.

$$s_i = \max_j r_{i,j} \tag{14}$$

As a result, the output of the texture transformer is calculated in Eq. (15).

$$F_{out} = F + Conv(Concat(F,T)) \odot S \tag{15}$$

In Eq. (5), $\odot$ indicates element-wise multiplication between feature maps to ensure that HR texture features from $Ref$ are transferred into LR, enhancing the process of texture generation.

The network is trained based on the overall loss ($\mathcal{L}_{TT}$) of three loss functions [30], reconstruction loss ($\mathcal{L}_{rec}$), perceptual loss ($\mathcal{L}_{percept}$ ) (Eq.(9)), and adversarial loss ($\mathcal{L}_{GAN}$ ) (Eq.(3)), which are computed by employing $L_1$ loss, penalization of gradient norm [37], and perceptual similarity between predicted SR and HR as well as predicted textures of SR and T respectively.

$$\mathcal{L}_{TT} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{percept} + \lambda_3\mathcal{L}_{GAN} \tag{16}$$

Where

$$\mathcal{L}_{rec} = \frac{1}{CHW}||I^{x_r} - I^{x_f}||_1 \tag{17}$$

And $(C, H, W)$ is the size of $x_r$.

## 5. Results

This work is probably one of the first to identify human papilloma (HP) viral particles in CaSki cell line using microscopic images. Table 2 lists the comparison with four GAN-based networks for classification of detected HPV clusters between high resolution (x4) images of ground truth (GT) and generated SR images. The drug treated samples are referred as normal, which have little trace of PHVLP clusters.

**Table 2.** Sensitivity and specificity for four approaches in percentage (%).

| | **Sensitivity (HPV)** | **Specific-ity (HPV)** | **Sensitivity (Treated)** | **Specificity (Treated)** |
|---|---|---|---|---|
| CycleGAN | 74.28 | 85.98 | 75.40 | 79.60 |
| Pix2pix | 65.74 | 80.26 | 85.00 | 74.48 |
| ESRGAN | 70.14 | 81.63 | 77.5 | 77.01 |
| ESRGAN (with SSIM in loss) | 74.46 | 81.63 | 77.5 | 79.77 |
| Pix2pixHD | **86.66** | **84.84** | **82.14** | **85.22** |

In Table 3, the average PSNR and SSIM between HR and SR images together with sensitivity and specificity for detecting HPV clusters from SR images are provided for several deep learning architectures. Similar to TTSR, RCAN [19] is another explainable attention-base network based on very deep residual channel attention network (RCAN).

**Table 3.** Average PSNR and SSIM for different SR methods together with sensitivity and specificity when counting HPV clusters.

| | PSNR | SSIM | Sensitivity | Specificity |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| RCAN | 25.80 | 0.7910 | 79.80 | 83.33 |
| Pix2pix | 18.35 | 0.5059 | 65.74 | 80.26 |
| CycleGAN | 30.31 | 0.8013 | 74.28 | 85.98 |
| ESRGAN | 28.07 | 0.6074 | 74.46 | 81.63 |
| TTSR | **28.70** | **0.8778** | **83.6** | **83.33** |

To quantify the quality of generated high resolution images, spatial frequency spectrum analysis is commonly employed, which unveils the frequency extrapolation nature of the developed GAN system and the closeness of similarity between generated fake image and GT real image in appearance. Fig. 6 exemplifies such an example demonstrating a real (6(a)) and fake (6(b)) images and their respected frequency spectrum (in log scale) at Fig. 6(c) and 6(d). The cross-section of radially averaged power spectrum showing in Fig. 6(e) indicates an overall good agreement with largely closeness of spatial frequency spectrum.

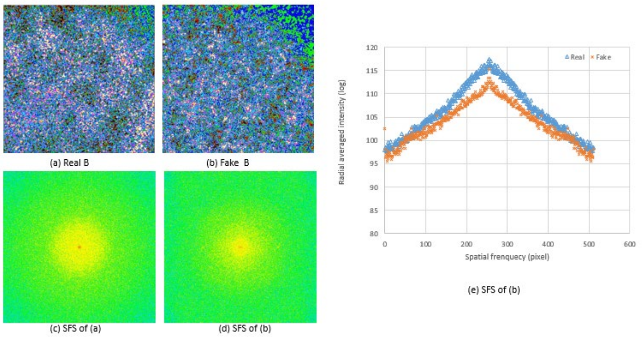


**Fig. 6.** Comparison of spatial frequency spectrum between generated fake image and GT for Pix2pixHD. (a) Real B; (b) Fake B; (c) (d) Spatial frequency spectrum (SFS) of (a) and (b) respectively; (e) Plot of cross-section of spatially averaged power spectrum of both real (in blue) and fake images (in orange) (in log scale).

In Fig. 7, visual comparison between ESRGAN and Pix2pixHD is presented, where top row is for ESRGAN and bottom row for Pix2pix2D.

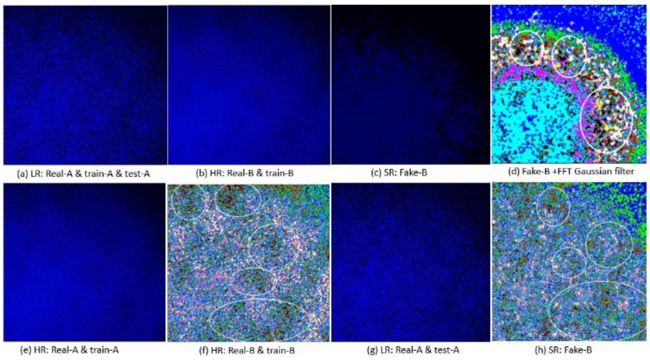

(a) LR: Real-A & train-A & test-A
(b) HR: Real-B & train-B
(c) SR: Fake-B
(d) Fake-B +FFT Gaussian filter
(e) HR: Real-A & train-A
(f) HR: Real-B & train-B
(g) LR: Real-A & test-A
(h) SR: Fake-B

**Fig. 7.** Comparison the final results (last row) for ESRGAN (top row) and Pix2pixHD (bottom row) approaches. The circles are identified HPV clusters.

Fig. 7 demonstrates that by directly translating through the training from raw images to their corresponding filtered maps (e.g. Pix2pixHD) can lead to more accurate results with regard to detect HPV clusters, which is also evidenced in Table 2. In Fig. 7(h), 4 out of 5 HPV clusters (white circle) are detected in comparison with 7(f) (GT) whereas in 7(d) only 3 clusters are located.

Understandably, both CycleGAN and Pix2Pix networks are developed for image translation and only applicable at the same resolution between input and output. Their advantage is that they do not require matching paired training data. Both methods perform similar in terms of detection of HPV clusters as given in Table 2, as also illustrated in Fig. 8.

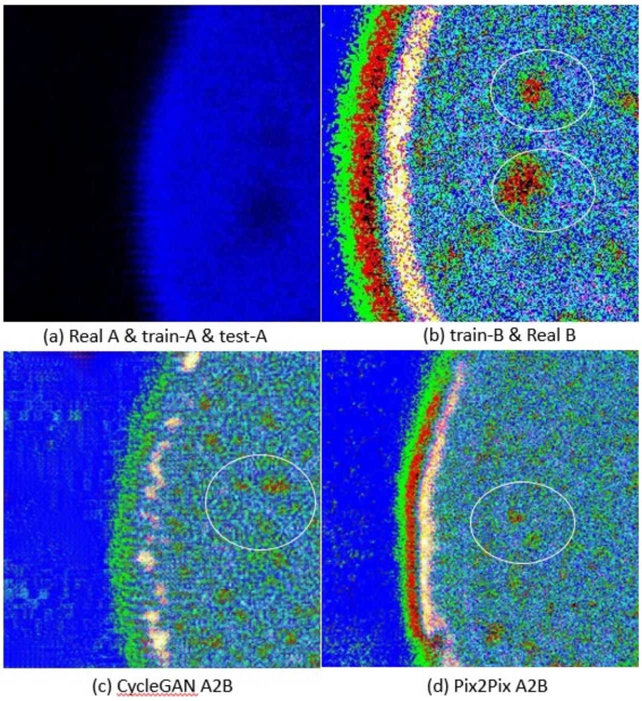


**Fig. 8.** Comparison results between CycleGAN (8(c)) and Pix2Pix (6(d)). White circle refers to HPV clusters. (a) & (b): training datasets.

In Fig. 9, comparison is made to show the detection between LR, SR and HR images where SR is generated using TTSR approach. It shows that the left top squared regions from both LR (d) and SR (f) fail to show HPV clusters whereas bottom right regions from all 3 resolution images detect a HPV cluster.

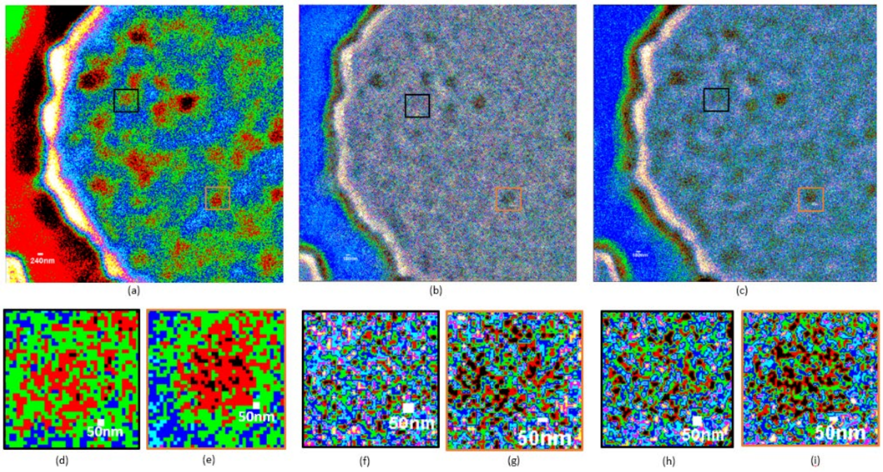


**Fig. 9.** The generated SR image using TT depicts HPVs in comparison with GT. (a): LR; (b) generated SR; (c) HR; (d) (f) (h): regions showing little trace of HPV; (e)(g)(i): regions exhibit HPV.

## 6. Discussion and Conclusion

This work constitute one of the first to identify Human papillomavirus (HPV) like structures (LP) from conventional light microscopic images through the application of state-of-the-art deep learning techniques, making a step further to allow fluorescence microscopy living up to their expectations. While a conventional light microscope has established to be an essential tool in studying cell structures, super resolution appears to be the desideratum for discerning these structures at nanoscale. High risk HPVs, such as HPV16 and HPV18, have been confirmed to be associated with some cancers' pathogenesis, especially cervical cancers. Hence early detection of HPV can assist to identify pre-cancerous lesions that subsequently can be treated before the onset of cancer develops. While detecting HPV from biological specimens is valuable in diagnosing HPV associated cancers and monitoring therapeutic effects, direct visualization of HPV or HPVLP in cells or tissue samples cannot be achieved through traditional molecular and proteomic methods. Hence microscopic imaging has been resorted to for directly observing HPV or virus particles, which calls for the increase of microscope's resolution (hence reducing pixel sizes) that is usually above 100nm/px [39] whereas a typical HPV has a size of ~50nm in diameter. While TEM provides a solution to capture those nanostructures, it is not readily available in additional to lengthy and complex sample preparation procedures.

This study investigates four state of the art generative adversarial deep learning networks (GAN) as well as vision transformer (ViT) to differentiate HPV clusters for microscopic images. The GAN-oriented architectures are CycleGAN, Pix2pix, ESRGAN and Pix2pixHD. Between GAN-based models, Pix2pixHD performs the best with sensitivity and specificity being 86% and 84% for detecting HPV clusters and 82% and 85% for detecting normal (i.e. drug treated) cells. For the other three networks, the averaged sensitivity and specificity are 78% 76% and 76% for CycleGAN, Pix2pix, ESRGAN respectively. When compared with Texture-based Transformer (TT), TT appears to perform the best with 83.6% an 83.33% sensitivity and specificity respectively. In addition, TTSR has the 2nd highest PSNR (28.70) and the highest SSIM (0.8778).

In the future, further to collecting more samples for both training and evaluation, the study of  the effectiveness of anti-HPV drugs will be ascertained through the employment of ViT networks to produce SR images. ViT-based approach not only is explainable by building upon human attention mechanism but also requires less data resources.

While it presents advantageous to use TEM and HR microscopy (MS) at different research centres, it limits the number of acquired datasets due to sample preparation and travelling. On the other hand, this practice has led the trained system being robust by taking in the information from different scanners, especially when the test LR datasets come from different cohort of MS scanners.

Furthermore, the authors will take the findings of effectiveness of generated super resolution (x4) forward and will consider to further to increase other scales, e.g. x8, on the premise of availability of data pairs of both LR and

HR images in the future, raising the prospect of unravelling the insights of formation from HPV to cancer while maintaining the prosperity of conventional light fluorescence microscopy.

## Acknowledgment

This work is financially supported by The Royal Society in the UK (Ref: IEC\NSFC\181557). Their support is gratefully acknowledged. The authors would also like to thank Dr. Natalie Allcock from University of Leicester, UK for providing the service of acquisition of TEM data.

## References:

[1] Gao XW, Wen X, Li D, Liu W, Xiong J, Xu B, Liu J, Zhang H, Liu X, Evaluation of GAN architectures for visualisation of HPV viruses from microscopic images, ICMLA 2021, Virtual, Dec 13-16, 2021.

[2] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, Human Papillomaviruses. Lyon (FR): International Agency for Research on Cancer; 2007. https://www.ncbi.nlm.nih.gov/books/NBK321770/. Retrieved in August 2020.

[3] Braaten KP, Laufer MR, Human Papillomavirus (HPV), HPV-Related Disease, and the HPV Vaccine. *Rev Obstet Gynecol*. 2008;1(1):2-10.

[4] Chelimo C, Wouldes TA, Cameron LD, Elwood JM, Risk factors for and prevention of human papillomaviruses (HPV), genital warts and cervical cancer, J. Infect, 66:207–217, 2013.

[5] Song B, Ding C, Chen W, Sun H, Zhang M, Chen W, Incidence and mortality of cervical cancer in China, 2013. *Chin J Cancer Res*., 29 (6):471-476, 2017.

[6] *Crewe AV, Isaacson M, Johnson D,* A Simple Scanning Electron Microscope*, Rev. Sci. Instrum., 40 (2): 241–246, 1969.*

[7] *Born M, Wolf E, Principles of Optics. 7th edition,* Cambridge University Press, Cambridge, UK, 1997.

[8] Hell SW, Microscopy and its focal switch. Nat Methods 6:24–32, 2009.

[9] Rust MJ, Bates M, Zhuang X, Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM), Nat Methods 3:793–795, 2006.

[10] Hell SW, Far-field optical nanoscopy, Science, 316:1153–1158, 2007.

[11] Schermelleh, L, Ferrand, A, Huser T, *et al.*, Super-resolution microscopy demystified, Nat Cell Biol ., 21, 72–84, 2019.

[12] Wang H, Rivenson Y, Jin Y, Wei Z, Gao R, Bentolila L, Kural C, Ozcan A, Deep learning enables cross-modality super-resolution in fluorescence microscopy, Nature Methods, 16:103-110, 2019.

[13] Dong C, Loy CC, He K, Tang X, Learning a deep convolutional network for image super-resolution, In: ECCV 2014, 2014.

[14] *Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y, (2014).* Generative Adversarial Networks*, Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672–2680, 2014.*

[15] *LeCun Y, Bengio Y, Hinton G, Deep Learning, Nature, 521: 436-444, 2015.*

[16] Kim J, Lee JK, Lee KM. Deeply-recursive convolutional network for image super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[17] Haris M, Shakhnarovich G, Ukita N, Deep networks for super resolution. In: CVPR, 2018.

[18] Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y, Image super-resolution using very deep residual channel attention networks. In: ECCV18, 2018.

[19] Ledig C, Theis L, Husz ́ar F, et al. Photo-realistic single image super-resolution using a generative adversarial network, arXiv:1609.04802, 2016.

[20] Johnson J, Alahi A, Li FF, Perceptual losses for real-time style transfer and super-resolution. In: ECCV'16, 2016.

[21] Bruna J, Sprechmann P, LeCun Y, Super-resolution with deep convolutional sufficient statistics, In: ICLR'15, 2015.

[22] Wang X, Yu K, Dong C, Loy CC: Recovering realistic texture in image super-resolution by deep spatial feature transform, In: CVPR'18, 2018.

[23] Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Loy CC, ESR-GAN: Enhanced super-resolution generative adversarial networks. In: Proc. ECCV Workshops, 2018.

[24] Jolicoeur-Martineau A, The relativistic discriminator: a key element missing from standard GAN, arXiv preprint arXiv:1807.00734, 2018.

[25] He K, Zhang X, Ren S, Sun J, Deep residual learning for image recognition. In: CVPR'16, 2016.

[26] Dosovitskiy A., et al., An image is worth 16x16 words: transformers for image recognition at scale, ICLR 2021.

[27] Gao XW, Wen X, Li D, Liu W, Xiong J, Xu B, Liu J, Zhang H, Liu X, Detection of human papillomavirus (HPV) from super resolution microscopic images applying a texture transformer network, SPIE Medical Imaging 2022, 20-24 Feb. 2022, San Diego, USA.

[28] Vaswani A, Shazeer N, Parmar N, et al., Attention is all you need. In NIPS, 2017.

[29] Brown TB, Mann B, Ryder N, Subbiah M, et al., Language models are few-shot learners. arXiv, 2020.

[30] Yang F, Yang H, Fu J, et al., Learning texture transformer network for image super-resolution, CVPR 2020.

[31] Zhu JY, Park T, Isola P, Efros AA, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

[32] Isola P, Zhu JY, Zhou, T, Efros AA, Image-to-Image Translation with Conditional Adversarial Networks, ICCV 2017, 2017.

[33] Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B, High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, In CVPR 2018.

[34] Pix2pixHD, https://github.com/NVIDIA/pix2pixHD. Accessed in January, 2021.

[35] ESRGAN-Pytorch, https://github.com/wonbeomjang/ESRGAN-pytorch. Retrieved in August 2020

[36] Simonyan K, Zisserman A, Very deep convolutional networks for large-scale image recognition, In ICLR'15, 2015.

[37] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. In NeurIPS, pages 5767–5777, 2017.

[38] Zhang Y, Li K, Wang L, et al, Image super-resolution using very deep residual channel attention networks, ECCV2018, 2018

[39] Netherton C, Moffat K, Brooks E, Wileman T, A guide to viral inclusions, membrane rearrangements, factories, and viroplasm produced during virus replication. Advances in Virus Research, 70:101-182, 2007.

## Appendix A. *Sample preparation for confocal fluorescent microscopy*

The cervical cancer cell lines, CaSki containing HPV16 sequences and C33a without HPV were used in this study. Both cell lines were grown on cover slips in a six-well culture plate containing RPMI culture media with 10% Fetal calf serum and 1% penicillin/streptomysin. They were kept in a humidified incubator with 95% air and 5% Carbon dioxide for two days until the cells reached 70-80% confluence. The cells were washed by PBS three times with 1 minute each time before they were fixed by 4% paraformaldehyde for 10 mins. They were then exposed to 0.1% Triton-100 for 5 mins following PBS washes. 50% house serum was then added for 8 mins, then it was removed. Next, 200ul 1 in 100 dilution of anti-mouse HPV E6/E7 antibody (Abcam, UK) was added in and the cells were left at the room temperature for 2 hours before they were washed again and 100ul biotinyted 2$^{nd}$ antibody (ABC Universal Kit, Vector lab, UK) was added. After 30 mins, the cells were washed again then tertiary antibody was added and left for 20 minutes. Finally, 100 ul TSA/FITC amplification reagent exposed to the cells for 6 minutes in the darkness before the cells were washed. DAPI containing mounting media was added on the labelled slides and cells attached on the cover slips were sealed inside for microscopic viewing.

## Appendix B. *Sample preparation for TEM scanning*

Cells were grown and prepared in a similar way as described above in Section S1 until the procedure reached at fixation step. Instead of fixing cells by paraformaldehyde, 2.5% glutaraldehyde in PBS buffer at pH 7 was used and the cells were fixed for 3 hours at RT before further steps taking place.

In addition, before the scanning, these samples undertake a series of standard preparation processes, including (1) flat embedding into EM capsules and polymerise for 6 hours at 16$^{o}$C; (2) sectioned to 70nm thick using a Leica

UC7 ultramicrotome, (3) collected onto copper mesh grids and (4) stained in lead citrate for 5 minutes.