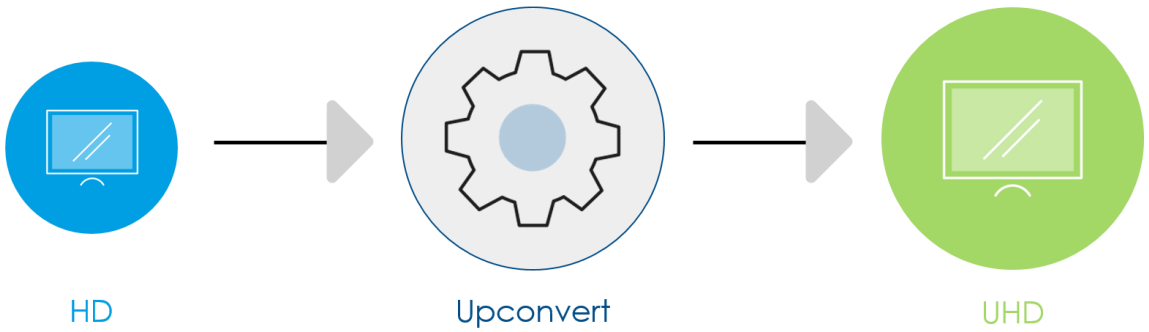
Increasing the value of existing content

Machine learning for HD to UHD up-conversion

MediaKind | Application Paper

As Ultra High Definition (UHD) TVs progressively fill households around the world (IHS reports 60% of TV sales in major markets are now UHD and predicts that by 2023, there will be 574 million 4K TV households[1]) it is no surprise, therefore, that UHD, together with High Dynamic Range (HDR), has become part of the mainstream media content ecosystem. However, as often happens with the introduction of new formats, there is a gap in availability of new format content until it becomes the default. This makes it difficult to launch full-time UHD channels, since both program content and advertising are not necessarily available in UHD.

A solution to this is to up-convert from the previous content format - in this case High Definition (HD) - in the same way when HD channels were newly introduced.



The problem with conventional up-conversion, though, is that it does not offer an improved resolution, so does not meet the expectations of the viewer at home, trying to watch on a UHD TV. The question, therefore, becomes: can we do better? If so, how?

Traditional approaches to up-conversion

UHD is a progressive scan format, with the native TV formats being 3840x2160, known as 2160p59.64 (usually abbreviated to 2160p60) or 2160p50. The corresponding HD formats, with the frame/field rates set by region, are either 1280x720 (720p60 or 720p50) or 1920x1080 (1080i30 or 1080i25).

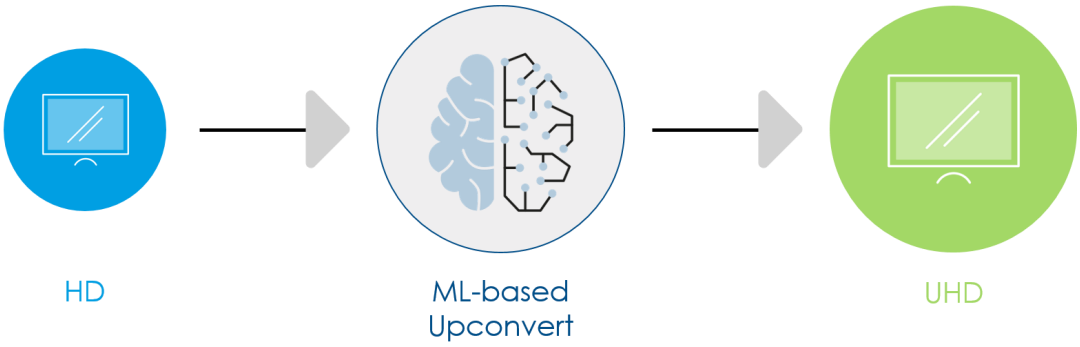
Conversion from HD to UHD for progressive images at the same frame rate is, in principle, fairly simple. It can be achieved using spatial processing only, typically in the form of a bi-cubic interpolation filter, which is a two-dimensional interpolation that is commonly used for image scaling for photographic images, graphics, etc, and uses a grid of 4x4 pixels as the source from which to interpolate intermediate locations within the center four pixels of that grid. The conversion from 1280x720 to 3840x2160 requires a 3x scaling factor in each dimension and is almost the ideal case for an upsampling filter as it repeats exactly for every source pixel grid location.

However, these types of filters do not create new data, but merely interpolate smoothly, resulting in a better result than nearest-neighbor or bi-linear interpolation, but one that does not have the appearance of being a higher resolution.

© MediaKind
The content of this paper is based on the authors research and opinion based on their knowledge of the Media market and specific area of expertise.

Machine learning and image scaling

Machine Learning (ML) is a technique whereby a neural network learns patterns from a set of training data. For image processing, it is necessary to work in at least two dimensions, and the processing needs to consider the whole image. Images tend to be large, and it becomes infeasible to create neural networks that process this data as a complete set. As a result, a different structure is used for image processing, known as Convolutional Neural Networks (CNNs). CNNs are structured to extract features from the images by successively processing subsets from the source image and then processes the features rather than the raw pixels.

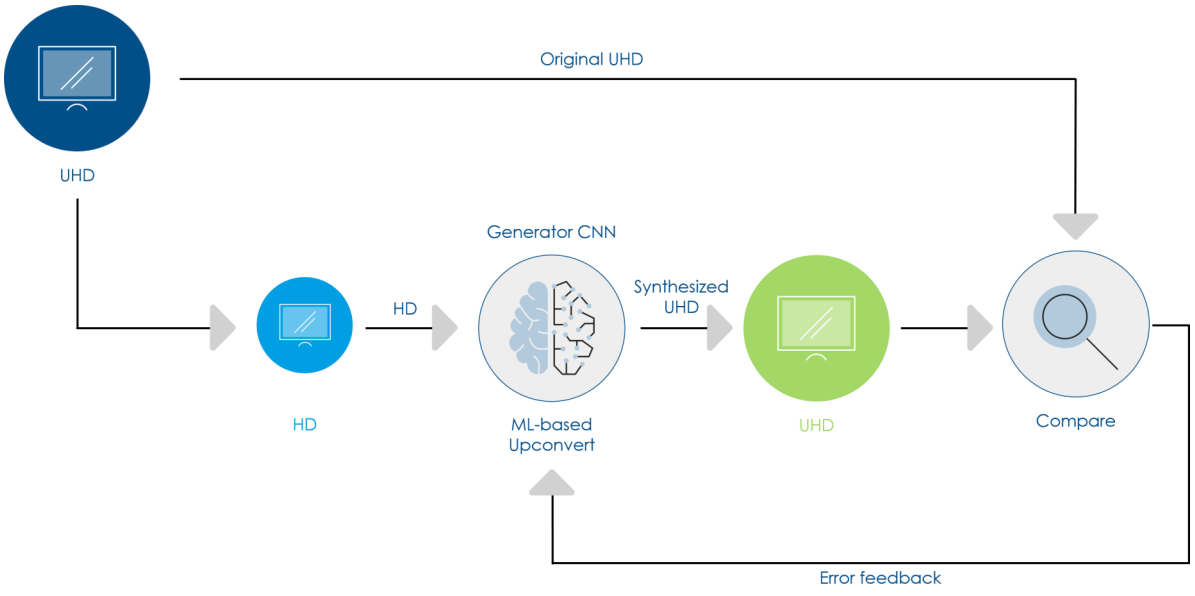


The interesting aspect of using CNNs is that, because of the inbuilt non-linearity in combination with feature-based processing, they can invent data that did not exist in the original image. In the extreme, it is possible for them to create content from scratch, for example creating artwork by learning a style and applying it to photographic images to obtain art stylized pictures[2].

However, in the case of up-conversion, the interest lies in the ability to create plausible new content that was not present in the original image, but which doesn't modify the nature of the image too much. The CNN used to create the UHD data from the HD source is known as the Generator CNN.

Training the neural network

In order for the Generator CNN to do its job, there must be a training process whereby a set of known data are input into the neural network – in this case patches of reference images – and a comparison made between the output and the correct image patch. Of course, in order to be able to do this, we need to know what "correct" means. Therefore, the starting point for training is a set of examples of high-resolution UHD representative images, which can be down-sampled to produce HD-representative images, then the results can be compared to the originals, as illustrated below.



© MediaKind
The content of this paper is based on the authors research and opinion based on their knowledge of the Media market and specific area of expertise.

The difference between the Original UHD image and the Synthesized UHD image is calculated by the Compare function, which is then fed back as an error signal to the Generator CNN. Over repeated training processes, the Generator CNN learns how to better create an image that is increasingly similar to an Original UHD image.

CNNs are formed from a series of different layers that perform operations. For example, Convolution Layers take grids of data and apply 3D learnable non-linear filtering (the color components are the third dimension). There will be multiple such filters applied for each set of the same source data subset. Different types of layers are applied sequentially – not least to prevent the size of the network from becoming excessive. There are pooling layers whose purpose is to reduce the size of the data by summarizing or eliminating. There can also be other types of layers for other purposes.

There are academic examples of CNNs for up-scaling, known as Super-Resolution CNNs, which formed an initial basis from which a more suitable CNN architecture was created and more suited to the up-conversion task for TV content, where management of levels is also important.

The training process is highly dependent on the data set used for training, and the neural network will try to fit the characteristics that it has seen during training onto the current image. This is intriguingly illustrated in Google's AI Blog[3], where a neural network presented with a random noise pattern as a starting point will introduce shapes that are like the ones it saw during training. It is therefore important that a diverse and representative set of content is used for training. In the process of research at MediaKind, we used a publicly available image set for training, using patches from about 800 different images.

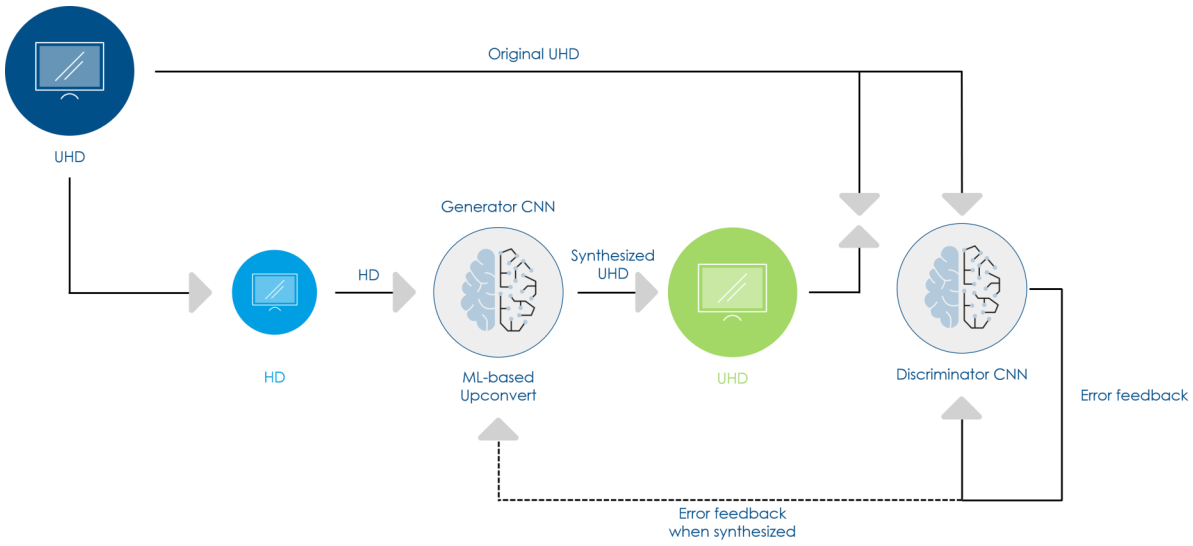
The compare function

The way that the comparison is performed affects the way that the Generator CNN learns to process the HD source data. A simple method would be to perform a simple arithmetic difference – sum of absolute differences – between the Original UHD image and the Synthesized UHD Image. The problem with this relates to training data set imbalance. In this case, the specific imbalance is that more areas of real pictures have relatively little fine detail, so the data set is biased towards regenerating a result that looks like that – which turns out to be remarkably similar to the use of a bicubic interpolation filter.

This doesn't really achieve the objective of creating plausible fine detail.

Generative Adversarial Neural Networks

Generative Adversarial Neural Networks (GANs) are a relatively recent concept[4], where a second neural network, known as the Discriminator CNN, is used and is itself trained during the training process of the Generator CNN. The principle is that the Discriminator learns to detect the difference between features that are characteristic of Original UHD images and Synthesized UHD images. During the training process, the Discriminator sees either an Original UHD image or a Synthesized UHD image, with the detection correctness fed back to the discriminator and, if the image was a Synthesized one, also fed back to the Generator.



© MediaKind
The content of this paper is based on the authors research and opinion based on their knowledge of the Media market and specific area of expertise.

As the training proceeds, each CNN is attempting to beat the other: the Generator learns how to better create images that have characteristics that appear like Original full-resolution images, while the Discriminator becomes progressively better at detecting what the Generator produces.

The result is the synthesis of details that have features characteristic of original UHD images.

Hybrid GAN approach

With a GAN approach, there is no real constraint to the ability of the Generator to create new detail everywhere. The problem with this is that the Generator can create images that, while containing plausible features, can diverge from the original image in more general ways. A better answer is to use a combination of the mathematical difference and the Discriminator's correctness as the feedback to the Generator CNN. This retains the detail regeneration, but also prevents excessive divergence. This construct produces results that are subjectively better than conventional up-conversion techniques.

What about interlace?

While frame-based processing is interesting, the reality is that most HD content is interlaced rather than progressive. Conversion from 1080i60 to 2160p60 is necessarily more complex than from 720p60. Starting from 1080i, there are three basic approaches to up-conversion:

- Process only from the corresponding field
- De-interlace and process from the frame
- Process from multiple fields directly

Once again, we need training data and this must come from 2160p video sequences, such that a set of fields can be created and downsampled, each field coming from one frame in the original 2160p sequence, to ensure the fields are not temporally co-located.

Somewhat surprisingly, the results from field-based up-conversion tended to be better than using de-interlaced frame conversion, despite using sophisticated motion-compensated de-interlacing, with the results from the frame-based conversion being dominated by the artifacts from the de-interlacing process. However, it is clear that some potentially useful data from the opposite fields of the interlace are not contributing to the result, and therefore the field-based approach is missing some data that could produce a better result.

Hybrid GAN with multiple fields

A solution to this is to use multiple fields' data as the source data directly into a modified Generator CNN, letting the GAN learn how best to perform the deinterlacing function. This approach was adopted and re-trained with a new set of video-based data, where adjacent fields were also provided.

© MediaKind
The content of this paper is based on the authors research and opinion based on their knowledge of the Media market and specific area of expertise.

This led to some very impressive results, with both high apparent spatial resolution and good temporal stability. These are, of course, best viewed as a video sequence, however an example of one frame from a test sequence shows the comparison:

Bi-cubic up-sampling from one field

Deinterfaced fields (w3f) followed by GAN frame-based up-sampling

3-field Hybrid GAN up-sampling

Original UHD source

Conclusions

Up-conversion using a hybrid GAN approach with multiple fields has been shown to be effective across a range of different content. This offers a realistic means by which content that has more of the appearance of UHD can be created from both progressive and interlaced HD source. This, in turn, can enable an improved experience for the viewer at home when watching a UHD channel, even when some of that content does not exist natively as UHD.

References:

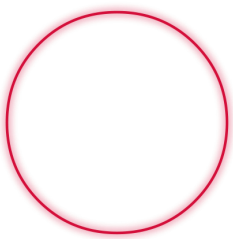
1 Advanced Television, "IHS: 4K display shipments top 60% in key markets," 14 10 2019. [Online]. Available: https://advanced-television.com/2019/10/14/ihs-4k-display-shipments-top-60-in-key-markets/

2 G. Surma, "Style Transfer - Styling Images with Convolutional Neural Networks," 13 01 2019. [Online]. Available: https://towardsdatascience.com/style-transfer-styling-images-with-convolutional-neural-networks-7d215b58f461

3 A. Mordvintsev, C. Olah and M. Tyka, "Inceptionism: Going Deeper into Neural Networks," 2015. [Online]. Available: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

4 I. e. a. Goodfellow, "Generative Adversarial Nets," *Neural Information Processing Systems Proceedings,* vol. 27, 2014.

© MediaKind
The content of this paper is based on the authors research and opinion based on their knowledge of the Media market and specific area of expertise.



Acquire
amazing.

Deliver
dynamic.

Experience
extraordinary.

mediakind.com