MHV23
May 7-10, Denver, CO, USA

Cloud based AI-driven Video Super-Resolution

Julien Le Tanou

Nelson Francisco

MediaKind

Using Deep-Learning for Video Super-Resolution

UHD Growth

The global 4K TV market is expected to grow at a compound annual growth rate **(CAGR) of 21.3% in 2022.** The 4K TV market will continue to grow to $396.34 billion in 2026 *
An increasing number of other devices support UHD resolutions.

Limited Availability

The majority of Linear TV content shows still come through at a 720p or 1080i resolution. Most true 4K content comes from streaming providers like Netflix, Disney+ and Amazon Video, and even that is limited.

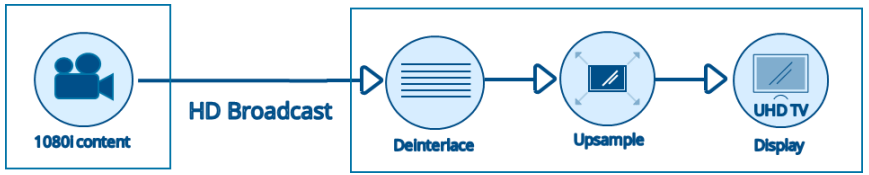
Demand

Content providers need technological advancements to react to market **demand** quickly, with high quality offerings from the source.

**Based on Absolute Research - "Global 4k Ultra-High Definition (UHD) Technologies Market", Sept. 2022*

Video Super-Resolution

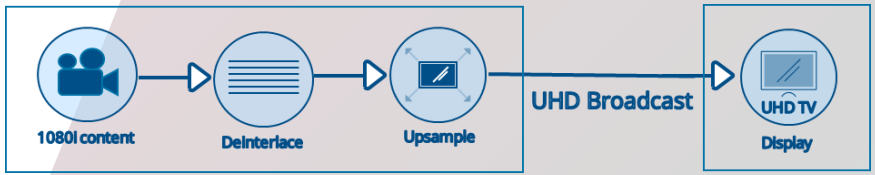
Viewing device side Upscaling



- No content provider added value, so limited scope for monetization;
- Upscaling depends on viewing devices capabilities;
- Inefficient as upscaling must be performed on every device.

- Lower resolutions mean lower bandwidths.

Content provider side Upscaling



- Scope for monetization since content provider is delivering a higher quality video service;
- Image rendering does not depend on the devices capabilities;
- Upscale is only performed once. Can leverage on high performance hardware (or cloud resources).

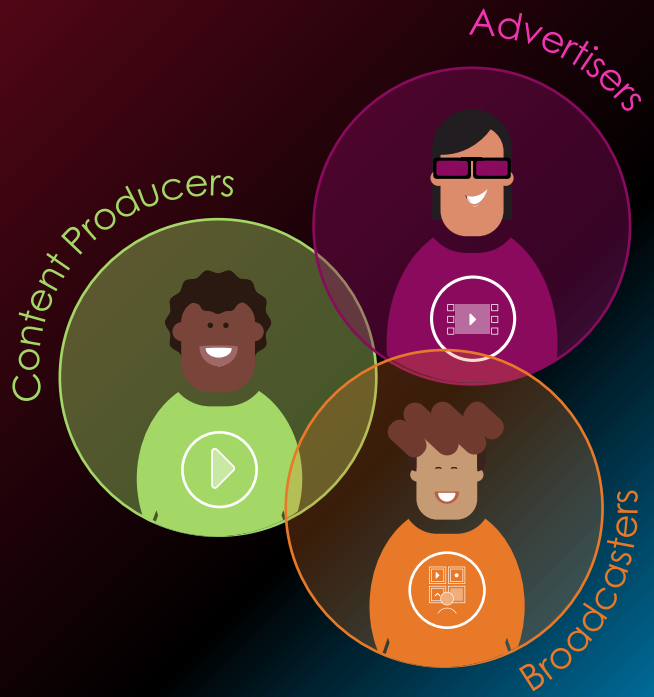
- Typically requires higher bandwidths as video is delivered at higher resolution.

Using Deep-Learning for Video Super-Resolution

- Initial approaches to Super-Resolution mainly relied in simple interpolation techniques (Bicubic, Lanczos);
- Some progress was achieved with more sophisticated methods (statistical, prediction-based, patch-based, or edge-based methods);
- Developments on deep-learning techniques truly revolutionized this research field.

DL Super-Resolution methods mainly fall into the following categories:

- PSNR-based models;
- Generative Adversarial Networks;
- Flow methods;
- Transformer models;
- Diffusion models.



Our solution - Atlas AI upscale

- Uses a GAN architecture as it provides a great balance between picture quality and cost;
- Uses multiple frames/fields with feature alignment, to achieve consistent and stable detail enhancement;
- Allows simultaneous deinterlacing and upscaling, to provide support to some of the most common legacy formats;
- Cloud vendor agnostic;
- Offered as a transcoding service:
  - Update source content;
  - Run transcoder in GPU accelerated VM instance;
  - Leverages in the GPU video encoding to reencode the content;
  - Retrieve upscaled content from secure bucket.

Ultimate Performance

Significantly outperforms traditional upscaling and deinterlacing methods

Monetize

Complement UHD offerings with high-quality content generated from existing libraries and leverage existing production pipelines all the way up to a final up-conversion stage

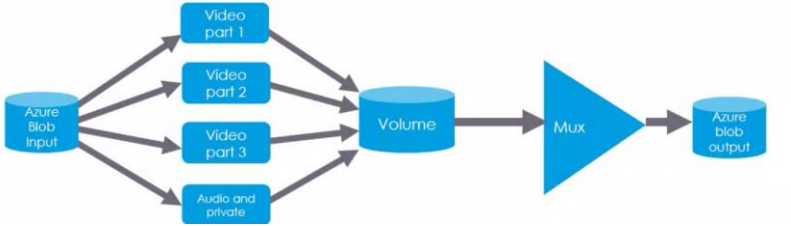
Scalable costing

A cloud-based AI-driven upscaling solution that retains low and scalable operational costs

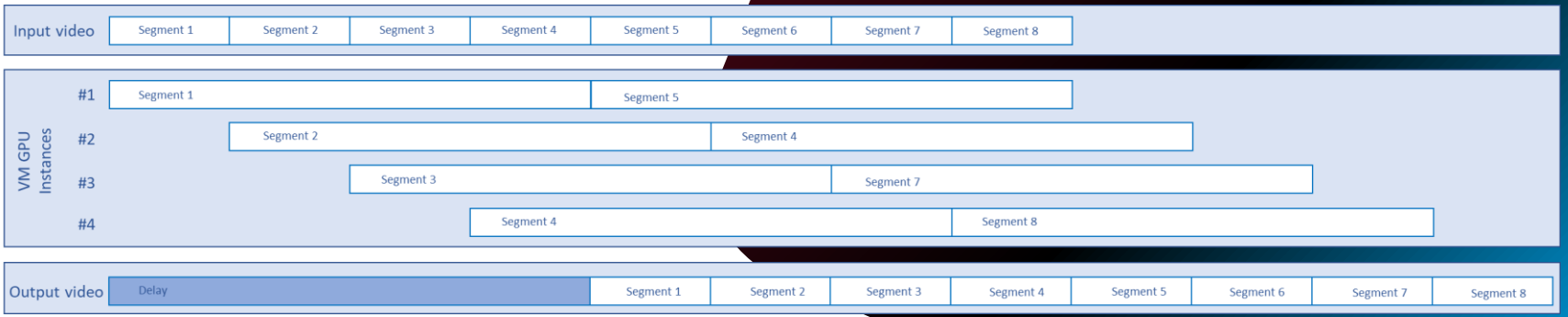
Real world applications




- Real-time operation can be achieved by parallelizing the inference to run on multiple concurrent GPUs:



- This approach allows to keep the same cost per hour (N VMs running for T/N time), with the introduction of an additional delay into the stream:



Why GANs?

- The popularity of GANs has been decreasing, with flow and diffusion methods being now considered the state-of-the art.
- However, the improved performance often comes at the cost of increased computational complexity, that can be prohibitive for practical implementations.
- To illustrate the trade-offs, we selected a state-of-the-art method from each category for a head-to-head comparison.

Method type	Name	Reference	Repository
PSNR-based. Uses ESRGAN but without adversarial loss.	RRDB	[Wang et al., 2018] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. **ESRGAN: Enhanced super-resolution generative adversarial networks.** ECCV, 2018.	https://github.com/cszn/KAIR
GAN	ESRGAN	[Wang et al., 2018] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. **ESRGAN: Enhanced super-resolution generative adversarial networks.** ECCV, 2018.	https://github.com/cszn/KAIR
Flow-based	SRFLOW	[Lugmayr et al., 2020] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. **SRFLOW: Learning the super-resolution space with normalizing flow.** In ECCV, pages 715–732. Springer, 2020.	https://github.com/andreas128/SRFlow
Transformer	SWIN2R	[Conde at al., 2022] Marcos Conde, Ui-Jin Choi, Maxime Burchi and Radu Timofte. **Swin2SR: SwinV2 Transformer for Compressed Image Super-Resolution and Restoration.** arXiv preprint arxiv.2209.11345 arXiv, 2022.	https://github.com/mv-lab/swin2sr
Diffusion	SRDIFF	[Li et al., 2021] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li and Yueting Chen. **SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models.** arXiv preprint arXiv:2104.14951, 2021.	https://github.com/LeiaLi/SRDiff

Why GANs?

- Images on the DIV2K validation dataset are around 2040x1536 (varying slightly and showing both portrait and landscape orientations).
- Used 4x scaling parameter: 510x384 -> 2040x1536. Although practical implementations will be more on the 2x or 3x ballpark (HD to 4K conversion), this seems to be a standard comparison on most papers, as it highlights the potential/issues of each method.
- Tests run in a Nvidia RTX3080Ti with 12Gb of VRAM.

Method	PSNR ↑	SSIM ↑	LR-PSNR ↑	LPIPS[1] ↓	Network Parameters ↓	Memory Usage [Mb] ↓	Throughput ↑
Bicubic	26.693	0.766	38.697	0.421	-	-	-
RRDB	29.475	0.844	53.737	0.262	16M	6213	2.1 fps
ESRGAN	26.636	0.764	42.613	**0.119**	16M + 34M discriminator	6213	2.1 fps
SRFLOW	27.109	0.756	52.066	0.124	40M	9895	0.8 fps
SWIN2R	**29.621**	**0.848**	**54.605**	0.256	12M	7165	1 fps
SRDIFF	27.125	0.785	49.617	0.132	13M	9101	3 it/s. 100 iterations per image 31s image
Our model	26.775	0.7687	50.844	0.127	**1.6M + 34M discriminator**	**5689**	**6 fps**

[1] Richard Zhang, Phillip Isola, Alexei Efros, Eli Shechtman and Oliver Wang. **The unreasonable effectiveness of deep features as a perceptual metric**. CVPR, 2018.

- SWIN2R achieved the best overall PSNR and SSIM but is quite complex and has relatively low LPIPS (images can be smooth).
- ESRGAN has the best LPIPS (sharper detail) and a relatively low complexity.
- SRDIFF has a good balance on the metrics with a reasonable footprint, but the need for multiple iterations make it very complex.
- SRFLOW also shows a good balance but a relatively high complexity.

Result comparison – example

Original 2040x816

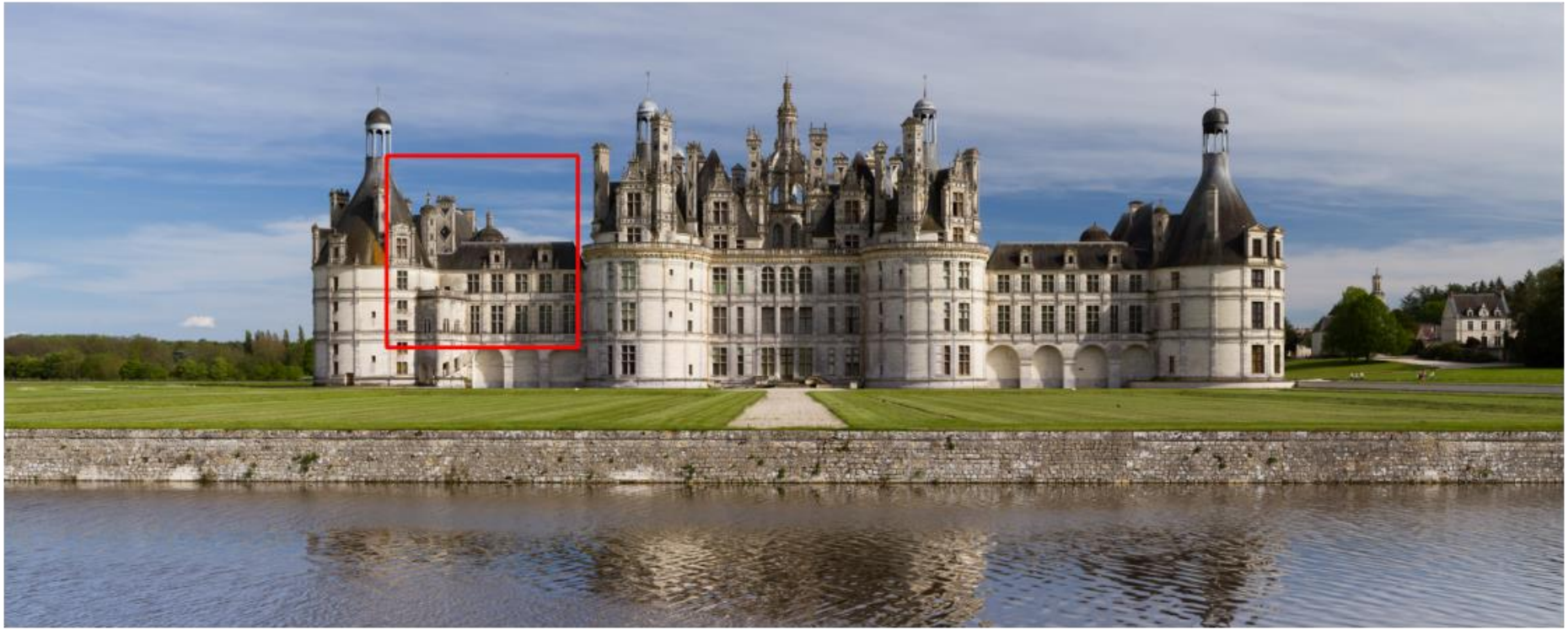


Image #830 from the DIV2K dataset – 250x250 pixels detail

Result comparison – detail

Original

Bicubic 23.60dB

RRDB 24.89dB

ESRGAN 20.52dB

SWIN2R 24.97dB

SRFLOW 22.30dB

SRDIFF 21.85dB

Our method 22.21dB

Result comparison – Real world applications

- For reference, lets assume we wanted to upscale a 60fps 510x384 -> 2040x1536 video, processing each frame independently, in a single GPU.

Instance name	Fps	Time to process 1h of content	Total VM Cost/h
Our method	6	10h	$30
ESRGAN	2.1	28.5h	$85
SWIN2R	1	60h	$180
SRFLOW	0.8	75h	$225
SRDIFF	0.03	2000h	$6000

- A VM with a Nvidia A10 GPU (such as NV36ads_A10_v5) has a cost of around $3/h and would offer comparable performance to the RTX3080Ti.

For a real-world 1080i to 4K conversion, and assuming proper feature alignment is used for video upscaling, fps of comparable methods would drop further, increasing costs even more.

Super-Resolution Video Upscale

Frame from Hybrid_Log_Gamma_High_Dynamic_Range_Demonstration_Pack_BT2020_v6_supplement_04 - Timelapse

Super-Resolution Video Upscale

Detail A

Conventional upscale:
YADIF deinterlacer + Bicubic Upscaling

Atlas Upscale:
Joint AI deinterlace and upscale

Super-Resolution Video Upscale

Detail B

Conventional upscale:
YADIF deinterlacer + Bicubic Upscaling

Atlas Upscale:
Joint AI deinterlace and upscale

Our cloud-based solution

- Currently being trialed by several European customers.
- Trained over a wide range of content for stable and reliable operation.
- Primary target on HD to 4K conversion, but other formats supported.
- SaaS deployment – no initial investment.
- Targeted a cost around $250 per hour of content processed (including data egress and ingress).
- Currently only offered as VOD but may be offered for live applications in the future.

Our Atlas AI Upscale won a NAB 2023 Product of the Year awards under the CAPITALIZE - AI/Machine Learning category


LIVE
without limits
MediaKind