Small Pixels
Video Coding Enhancement

Solution Overview

Document ID

WHITEPAPER - VIDEO CODING

Date

19/02/2026

Version

Confidential

Index

Small Pixels
Video Coding Enhancement

Solution Overview

ABSTRACT

This whitepaper describes Small Pixels' solutions for video coding enhancement. It provides a well-rounded picture of the service for new adopters, and a deeper understanding of the technologies behind Neural Networks for video coding enhancement.

The intended audience for this document includes people in technology roles, such as chief technology officers (CTOs), architects, developers, and operations team members. After reading this paper, you will understand the main concepts behind the Small Pixels' solutions for video coding enhancement, based on the concept of deep editing.

1. Introduction

This document gives a brief introduction to neural network-based video coding enhancement and a general idea behind the solutions provided by Small Pixels s.r.l.

This chapter gives an overview of the NN-Based video coding enhancement; in chapter 2 we present the Small Pixels solutions and products for video enhancement; conclusions are drawn in chapter 3. Relevant bibliographic information is provided at the end of the document.

1.1 Advanced Video Coding Enhancement via Deep Pre-Editing

The transmission of high-fidelity visual data over bandwidth-constrained networks represents a foundational challenge in modern digital communications. Historically, the telecommunications and streaming industries have relied on the iterative evolution of standard video codecs—ranging from the ubiquitous Advanced Video Coding (AVC/H.264) to High Efficiency Video Coding (HEVC/H.265), Versatile Video Coding (VVC/H.266), and the open-source AV1. 1 These algorithms operate as highly optimized rate-distortion machines. They structurally decompose video streams into coding tree units (CTUs) or macroblocks, utilizing intra-frame spatial prediction and inter-frame motion compensation to isolate residual signals. These residuals are then transformed into the frequency domain, typically via a Discrete Cosine Transform (DCT), and quantized to fit within a rigidly defined bit budget [DAS-2023].

These standards, besides their effective engineering, rely fundamentally on what quantization dictates when the allocated bit-rate is insufficient for the complexity of the scene, resulting in unavoidable truncation of the transform coefficients. This aggressive truncation strips the high-frequency energy from the video signal. To the human visual system, this loss of high-frequency data manifests as severe visual artifacts: blocking across macroblock boundaries, mosquito noise around high-contrast edges, color degradation, and generalized blurring [BERTINI-2022, TALEBI-2021].

Until now, the standard approach to mitigating these artifacts has been to deploy post-processing filters on the client side. Algorithms operating after the decompression stage attempt to smooth block boundaries and hallucinate missing textures. However, post-processing is fundamentally constrained by an information bottleneck; the algorithm is forced to reconstruct reality from a mathematically degraded source signal. The Small Pixels architecture approaches this problem from an entirely different vector, implementing a paradigm known as deep pre-editing [TALEBI-2021]. The core idea, originally introduced for JPEG still images by Hossein Talebi et al., is to intelligently modify the source signal before encoding such that the downstream codec operates in a more favorable regime at low bits, yielding fewer visible artifacts and/or lower bitrate for similar perceptual quality.

By deploying a learned neural network (NN) as a pre-encoding stage, the Small Pixels engine intelligently alters the incoming video frames before they are submitted to the standard codec. This

semantic and spatial adaptation actively reduces the native entropy of the video, transforming the content to become inherently more compressible. The following chapter provides a technical analysis of the deep pre-editing paradigm, presenting at high-level the formulations that enable end-to-end training, the architectural anatomy of the neural networks involved, the critical implementation of differentiable codec proxies, and providing an intuition of the whole process.

Small Pixels' method achieves superior video coding by intelligently modifying image content. This strategic editing is designed to meet three key objectives: *i)* preserving the majority of visual information; *ii)* optimizing the subsequent compression process to significantly reduce artifacts; and *iii)* ensuring the compressed, edited image remains visually appealing. By integrating these considerations, we created an innovative solution that maximizes the performance of the compression-decompression cycle within a specified bit budget.

1.2 The Rate-Distortion-Perception Bottleneck in Video Compression

To understand the necessity and efficacy of deep pre-editing [TALEBI-2021], we need to evaluate the limitations of traditional rate-distortion (R-D) theory. The design of standard video encoders is anchored in the minimization of an objective distortion metric, often computed in terms of pixel distortion, like the commonly used Mean Squared Error (MSE) or the Peak Signal-to-Noise Ratio (PSNR) metrics, subject to a strict constraint on the bit-rate. The encoder continuously searches for the optimal combination of block partitioning, motion vectors, and quantization parameters that yields the lowest MSE for a given number of bits.

However, recent advancements in information theory and deep learning for visual data compression have formalized the Rate-Distortion-Perception (RDP) tradeoff. This principle establishes a mathematical proof that strictly minimizing distortion (maximizing PSNR) is frequently fundamentally at odds with optimizing perceptual quality. The Mean Squared Error evaluates quality on a strict pixel-by-pixel basis, completely ignoring the complex, holistic manner in which the human visual system interprets semantic structures, textures, and temporal continuity [KHAN-2025, DING-2021]. Any deep learning training objective must manage the known tension between optimizing distortion metrics (PSNR/SSIM) and perceptual plausibility (as measured by perceptual metrics like MOS/VMAF/LPIPS). The perception–distortion tradeoff is formalized by [BLAU-2018].

When a standard codec like HEVC or VVC is starved of bits, it is designed to minimize the global MSE. Because stochastic, high-frequency textures—such as the chaotic movement of water, the fine grain of film, or the complex foliage of a forest—require massive amounts of bits to encode perfectly, the codec's quantization matrix zeroes out these coefficients. The codec effectively "spends" its limited bit budget trying to maintain the mathematical average of the pixel blocks, resulting in a decoded frame that is mathematically optimal according to PSNR, but perceptually disastrous, appearing flat, plastic, and riddled with rigid block artifacts.

Deep pre-editing operates by deliberately exploiting the RDP tradeoff. The Small Pixels pre-editing engine intentionally introduces a controlled, algorithmically determined imperceptible distortion into the pristine, uncompressed source video. This pre-distortion sacrifices absolute mathematical fidelity (the exact pixel-to-pixel equivalence with the original camera sensor data) in order to inject structural properties that the subsequent standard codec can process with maximum efficiency and, at the same time, reducing the actual distortion that is introduced by the codec itself when it is obligated by its own implementation to eliminate the details.

By intelligently operating on perceptually irrelevant regions, the pre-editor actively reduces the spatial variance of the frame. Consequently, the standard encoder operates in a vastly more forgiving mathematical regime. It can apply finer quantization steps to the remaining, semantically critical structures within the video. The ultimate result is a compressed-decompressed output that is vastly superior in visual quality and human perception. The Small pixels neural network is effectively acting as a sort of "plugin" for the encoder that substitutes its "dumb" raw lossy compression parameters of the standard implementation with a smarter NN solution.

2. Neural Network Based Video Coding Enhancement

NN-Based video coding enhancement is the task of improving the performance of video coding (i.e. video compression) quality of a given content using a neural network trained specifically for this task. Here, the input is a high-quality video, and the output is another high-quality video, perceptually similar for a human viewer but that can be compressed better by a video encoder, resulting in a final video coding that is both perceived as of higher quality by the user and that is more compressed. In the typical use case the output resolution and frame rate remains exactly the

same as the input: the goal of the neural network is to improve the performance of the following video encoder.

The input video is processed frame-by-frame, this basically reduces the latency of the whole system to 1 frame, allowing to use the solution both on live and VOD encoding [GALTERI-2019, VACCARO-2021]. Each frame is enhanced by a Neural Network and then passed to the video encoder that produces the final enhanced output video. This process is summarized in Figure 1.

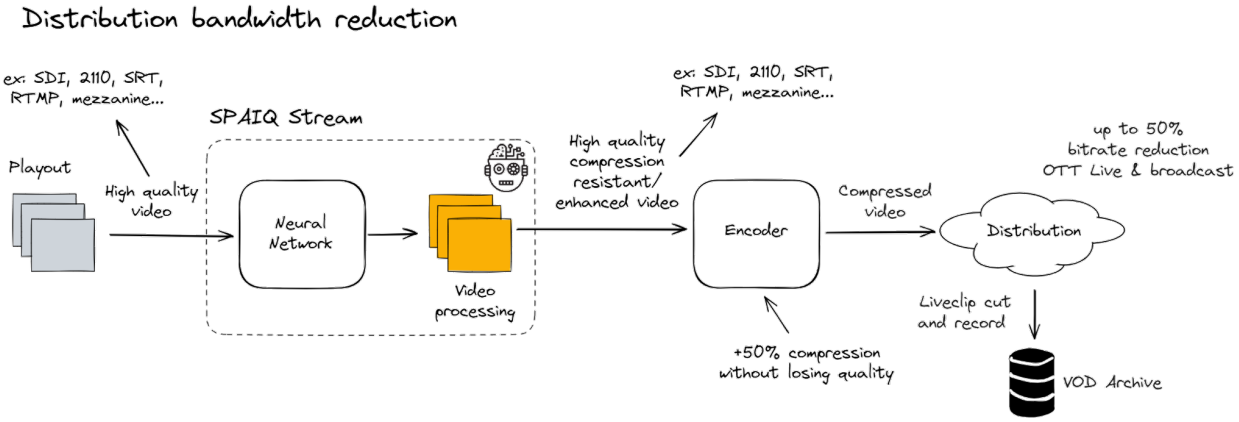

Figure 1: NN-Based Video coding enhancement process

The video coding enhancement system runs before a standard encoder. This keeps the standards bitstream and all clients unchanged. It also allows deployment in existing transcoding/packaging pipelines. Small Pixels' solution is designed to maintain that "no client changes" is a core value. After an enhancement network has been trained and deployed in production, it does not need to undergo further changes (i.e., subsequent training).

2.1 Formulation of the Optimization Problem

The engineering objective of the Small Pixels deep pre-editing pipeline is to construct a universal, feed-forward deep neural network capable of performing content adaptation automatically on an arbitrary incoming video stream. This process must occur in real-time on the server side prior to the application of standard encoding.

To formally define this operation, let $z_t \in \mathbb{R}^{H \times W \times C}$ where z_t denote a high-quality, uncompressed video frame at time index *t*, possessing height *H*, width *W*, and *C* color channels (typically RGB or YUV). Let the operation of standard video coding, encompassing both compression to a bitstream and subsequent decompression back to pixel space, be denoted by the function $C_B(\cdot)$, where *B* represents the strict bit budget constraint. The conventional transmission pipeline yields a decoded frame $\widehat{x_t} = C_B(z_t)$, which carries native compression artifacts.

Deep pre-editing introduces a parametric neural editing network, designated as $T(\theta, z_t)$, where θ encompasses the learned weights and biases of the network. The network ingests the original frame and outputs a pre-edited frame $x_t = T(\theta, z_t)$. The mathematical objective is to optimize the parameters θ such that the final compressed-decompressed frame $C_B(x_t)$ strictly satisfies three concurrent constraints:

1. **Semantic Proximity**: The pre-edited frame must maintain a high degree of semantic and visual similarity to the original frame z_t, preventing the network from hallucinating entirely new content.
2. **Bit-Rate Compressibility**: The pre-edited frame x_t must demand fewer bits to encode than the original z_t, or alternatively, yield a higher perceptual quality at the identical bit budget *B*.
3. **Visual Artifact Suppression**: The final output $C_B(x_t)$ must exhibit superior perceptual quality, specifically targeting the elimination of blockiness, ringing, color bleeding, etc.

The formulation of the optimization loss function used to train the network θ across a comprehensive dataset of uncompressed videos is designed to cater to these aspects. Among its main components are:

- $C_q(\cdot)$ that represents a differentiable proxy of the target video codec operating at a specific quantization parameter or quality factor *q*. A differentiable proxy is absolutely mandatory, as standard codecs cannot pass backpropagation gradients.
- $dist(\cdot, \cdot)$ that serves as the proximity measure between original and compressed frame. Crucially, simple L_1 and L_2 pixel distances are insufficient, as they heavily penalize the quality of the images leading to blurred versions. Instead, the Small Pixels implementation utilizes a combination of deep perceptual loss based on feature extraction advanced differentiable perceptual metrics like Learned Perceptual Image Patch Similarity (LPIPS) [ZHANG-2018] and the Video Multimethod Assessment Fusion (VMAF) LI-2025].

This carefully designed loss function continuously drives the output away from blocky, blurry states and toward aesthetically pleasing, high-quality reconstructions.

2.2 Overcoming the Non-Differentiable Codec Barrier

The most formidable obstacle in training an end-to-end deep pre-editing network is the mandatory presence of the standard video codec, $C_q(\cdot)$, within the optimization loss function. Deep learning relies entirely on backpropagation, which computes gradients using the chain rule of calculus. Standard video codecs, however, are fundamentally non-differentiable systems due to their core reliance on discrete quantization and entropy coding.

Quantization is the mathematical process of dividing continuous transform coefficients by a quantization matrix and subsequently rounding the result to the nearest integer. The rounding operation $\lfloor x \rceil$ operates as a step function. The mathematical derivative of a step function is exactly zero everywhere, except at the instantaneous step boundaries, where it is undefined. Because the gradient is zero, backpropagation halts completely at the quantization stage. If gradients cannot flow backward from the quality assessment metric through the codec to the pre-editing network, the CNN cannot update its weights; it cannot learn how its spatial warps and smoothing operations affect the final compressed output.

Resolving this mathematical blockade requires the construction of fully differentiable proxies of the video codec, to be used exclusively during the offline training phase. This specialized component is one of the most important elements of the training procedure developed by Small Pixels, along with differentiable proxies for artifacts and other elements typical of a video codec like block-based motion-compensated prediction [CHADHA-2021].

2.3 Advanced Differentiable Proxy Codecs

For intra-frame coding architectures, the differentiable proxy must replicate the exact functionality of standard transform coding while maintaining a continuous flow of non-zero gradients. The color conversion phases (RGB to YUV) and the transform phases (Forward and Inverse DCT) are inherently linear operations executed via matrix multiplication, and thus possess well-defined, easily computable derivatives.

The critical bottleneck remains the quantization phase. The Small Pixels training architecture leverages more recent state-of-the-art advancements in differentiable codec approximations to ensure perfect alignment with the true codec across the entire spectrum of compression ratios. As detailed in recent research on differentiable JPEG models [REICH-2024], standard approximations often fail to model critical discretizations and bounding limits, leading to poor gradient flow at extreme compression levels.

To achieve a robust differentiable pipeline, the proxy incorporates several advanced mechanisms:

- Differentiable Clipping: The proxy implements differentiable clipping applied to both the pixel-space codec image and the quantization tables, preventing gradient explosion at the extremes of the color gamut.
- Differentiable Flooring: The architecture introduces differentiable flooring for both the quantization table scale and the quantization matrix values themselves
- Straight-Through Estimators (STE): Rather than relying solely on polynomial approximation, the most robust models utilize a Straight-Through Estimator. In the forward pass, the STE performs the exact, discrete rounding operation required by the standard codec, ensuring the forward loss calculation is perfectly accurate. In the backward pass, the STE bypasses the zero-derivative of the rounding function, passing the gradient straight through as if the operation had been the identity function.

By modeling these specific bounds and discretizations in a fully differentiable setting, the proxy achieves an incredibly strong approximation of the standard codec over the entire quality range. This ensures the pre-editing network learns the exact nuances of the downstream codec.

3. Conclusion and Future Outlook

The development of standard video compression has reached an undeniable point of diminishing returns: the incremental rate-distortion gains offered by the iterative adjustments of standard codecs are wholly insufficient to overcome the hard physics of network bandwidth limitations.

Deep pre-editing represents a foundational, necessary evolution in the digital video distribution pipeline. By actively, imperceptibly and intelligently modifying the source content through deep neural networks, i.e. exploiting the semantic capabilities of neural networks to reduce

uncompressible stochastic noise and to align complex geometry perfectly with block-based transforms, the paradigm fundamentally reduces the raw data entropy presented to the encoder.

Training end-to-end utilizing a combination of perceptual quality metrics and highly advanced, accurate differentiable codec proxies, these pre-editing networks learn to exploit the Rate-Distortion-Perception tradeoff. They proactively shield the video content from the harshest, most visually destructive effects of low-bitrate quantization. Crucially, because this sophisticated enhancement occurs entirely prior to the standard encoding phase, it requires zero modifications to global decoding hardware infrastructure, ensuring immediate backward compatibility across billions of existing edge devices.

REFERENCES

[BERTINI-2022] BERTINI, Marco, GALTERI, Leonaro, SEIDENARI, Lorenzo, URICCHIO, Tiberio, & DEL BIMBO, Alberto. Fast and effective AI approaches for video quality improvement. In: Proceedings of the 1st Mile-High Video Conference. 2022

[BLAU-2018] BLAU, Yochai; MICHAELI, Tomer. The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018

[CHADHA-2021] CHADHA, Aaron; ANDREOPOULOS, Yiannis. Deep perceptual preprocessing for video coding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. p. 14852-14861, 2021

[DAS-2023] DAS, Tanni; CHOI, Kiho; CHOI, Jaeyoung. High quality video frames from VVC: A deep neural network approach. *IEEE Access*, 2023, 11: 54254-54264.

[DING-2021] DING, D., MA, Z., CHEN, D., CHEN, Q., LIU, Z., & ZHU, F. Advances in video compression system using deep neural network: A review and case studies. Proceedings of the IEEE, 109(9), 1494-1520, 2021.

[GALTERI-2019] GALTERI, L., SEIDENARI, L., BERTINI, M., DEL BIMBO, A. (2019). Towards Real-Time Image Enhancement GANs. In: Proc. of Computer Analysis of Images and Patterns (CAIP), 2019

[KHAN-2025] KHAN,Muhammad Umar Karim, CHADHA,Aaron, ANAM,Mohammad Ashraful, ANDREOPULOS,Yiannis. Perceptual Video Compression with Neural Wrapping Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025

[LI-2025] LI,J., ZHOU,C., CHEN,Y. and LU,G., Differentiable VMAF: A trainable metric for optimizing video compression codec, In Proc. of IEEE International Symposium on Circuits and Systems (ISCAS),, 2025

[REICH-2024] REICH, C., DEBNATH, B., PATEL, D., & CHAKRADHAR, S. Differentiable jpeg: The devil is in the details. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024

[TALEBI-2021] TALEBI, Hossein, KELLY, D., LUO, X., DORADE, I. G., YANG, F., MILANFAR, P., & ELAD, M. Better compression with deep pre-editing. IEEE Transactions on Image Processing, 2021

[VACCARO-2021] VACCARO,Federico, BERTINI,Marco, URICCHIO,Tiberio, DEL BIMBO,Alberto. Fast Video Visual Quality and Resolution Improvement using SR-UNet. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2021

[ZHANG-2018] ZHANG, R.; ISOLA, P.; EFROS, A.A.; SHECHTMAN, E.; WANG, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018