

3R-INN: How to be climate friendly while consuming/delivering videos?

Zoubida Ameur¹ , Claire-Hélène Demarty¹ , Daniel Ménard² , and Olivier Le Meur¹ 

¹ InterDigital, France, firstname.lastname@interdigital.com

² IETR, daniel.menard@insa-rennes.fr

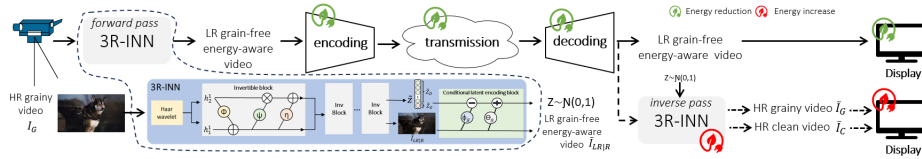


Fig. 1: 3R-INN: End-to-end **energy-aware** video distribution chain by **R**emoving grain, **R**escaling and **R**educing display energy.

Abstract. The consumption of a video requires a considerable amount of energy during the various stages of its life-cycle. With a billion hours of video consumed daily, this contributes significantly to the greenhouse gas (GHG) emission. Therefore, reducing the end-to-end carbon footprint of the video chain, while preserving the quality of experience at the user side, is of high importance. To contribute in an impactful manner, we propose 3R-INN, a single invertible network that does **three tasks at once**: given a high-resolution (HR) grainy image, it **R**escales it to a lower resolution, **R**emoves film grain and **R**educes its power consumption when displayed. Providing such a minimum viable quality content contributes to **reducing the energy consumption during encoding, transmission, decoding and display**. 3R-INN also offers the possibility to restore either the HR grainy original image or a grain-free version, thanks to its invertibility and the disentanglement of the high frequency, and without transmitting auxiliary data. Experiments show that, 3R-INN enables significant energy savings for encoding (78%), decoding (77%) and rendering (5% to 20%), while outperforming state-of-the-art film grain removal and synthesis, energy-aware and downscaling methods on different test-sets.

Keywords: Energy saving · Invertible network · Video distribution

1 Introduction

Over 75% of the world’s global GHG emissions comes from energy production, particularly from fossil fuels. The growing energy consumption of the media and entertainment industry, in particular streaming, strongly contributes to climate change, with more than 1.3% of GHG in 2020 [41]. Therefore, this industry has to move towards decarbonisation, energy efficiency and sustainability in

all its stages, *e.g.*, head-end (encoding), delivery (transmission) and end-user device (decoding and display). Taking apart the energy consumed while building the different necessary equipment, reduced energy consumption at the head-end translates into shorter encoding times and lower computing loads, while at the distribution level it translates into lower bit-rates. At the end-device level, significant gains can be achieved, as displays constitute the most power-hungry part of the whole chain [41]. In the specific case of emissive displays, *e.g.*, organic light-emitting diodes (OLEDs), the power consumption is pixel-wise and therefore directly dependent on the displayed content. Consequently, less energy-intensive images at display and shorter decoding times will also lead to lower energy consumption.

The encoding and decoding times are related to the content resolution and complexity. Downscaling the content before encoding and upscaling it after decoding while preserving the same quality of experience [7] is one straightforward solution to reduce the computational burden. Additionally, removing and modeling artistic noise, such as film grain, before encoding and synthesizing it after decoding, not only reduces encoding and decoding times, but also significantly reduces the bit-rate [32], while still preserving the artistic intent at the user side. Finally, as displays consume the largest proportion of the energy, providing energy-aware content, *i.e.*, that will consume less when displayed, is of significant importance, at least for OLED displays. Several studies addressed this issue by investigating how to reduce the content brightness [25, 26, 36].

Because climate change is a pressing issue, we believe that having a global vision on the overall energy consumption in the video chain and an holistic approach on how to reduce it is of the utmost importance. Therefore, in this paper, we propose an **end-to-end energy reduction of the video distribution chain**, while preserving a good quality of experience at the user side, by leveraging a deep learning invertible neural network (INN)-based model, called **3R-INN**. Prior to encoding a HR grainy image, our 3R-INN multi-task network **Rescales** it to a lower resolution, **Removes** film grain and **Reduces** its power consumption when displayed, by some reduction rate R . While saving energy along the video chain, 3R-INN also provides a visually-pleasant content intended to be displayed, following the new paradigm proposed in [37], which promotes to target a minimum viable video quality for transported videos. Within this same paradigm, the possibility to recover the original content is encouraged, with the counter part that it will consume more. This is feasible thanks to the invertibility of 3R-INN which allows to retrieve the original HR image from the clean energy-aware LR one by running inversely the framework. Furthermore, thanks to the modeling and disentanglement of the lost information in the forward pass, two versions, grainy and clean, of the original HR image can be restored, **without transmitting any auxiliary information**. With the idea that the energy consumed when applying an energy reduction processing should not exceed the amount of energy saved, we designed 3R-INN to be a **single network**, that replaces **three separate and potentially heavier processings**, showing that

the use of 3R-INN results in a positive energy balance, as soon as a content is displayed multiple times. In summary, our main contributions are four-folds:

- a first end-to-end solution for reducing the energy consumption of the video chain that exhibits a better energy balance compared to the sum of the corresponding three tasks in the state-of-the-art;
- a single network for the three tasks of rescaling, removing/synthesizing grain and reducing the energy at display, dedicated towards saving energy in the whole video chain;
- the provision of a visually pleasant, energy reduced version of the original image, and the capability to go back to the original HR grainy image with no transmission of additional metadata along the video chain;
- the best method so far for high-fidelity film grain synthesis, with no need of auxiliary data and the best method so far for downscaling and building energy-aware images.

In the following, we first review the state-of-the-art for rescaling, film grain removal/synthesis and energy-aware images (Section 2), before detailing our proposed solution (Section 3). In Section 4, we provide a compared analysis of the quantitative, qualitative and energy performances of the use of 3R-INN, against state-of-the-art solutions. In Section 5, we draw conclusions and perspectives.

2 Related work

Rescaling The rescaling task helps saving resources, through the storage and transfer of downscaled versions of an original HR image/video. Recovering the original resolution while having pleasant LR content can be very challenging. For these purposes, to maximize the restoration performance while producing visually pleasant low-resolution (LR) content, several works learn jointly the two tasks, *i.e.*, downscaling and upscaling. In [22], an auto-encoder-based framework learns the optimal LR image that maximizes the reconstruction performance of the HR image. In [40], an unsupervised downscaling method with consideration on the upscaling process but no assumption on how the HR image is downscaled, allows to learn the essential information for upscaling in an optimal way. Following a different paradigm, the method called IRN [42] models the down- and up-scaling processes using an invertible bijective transformation. In a forward pass, IRN performs the downscaling process by producing visually pleasing LR images while capturing the distribution of the lost information using a latent variable that follows a specified distribution. Meanwhile, the upscaling process is made tractable such that the HR image is reconstructed by inversely passing a randomly drawn latent variable with the LR image through the network. However, the reconstruction is not image-adaptive due to the case-agnostic latent variable.

Film grain removal and synthesis To better preserve film grain while compressing video content efficiently, it is classically removed and modeled before encoding and restored after decoding [15, 32]. Hence, dedicated methods for film grain removal are proposed, based on either temporal filtering [10], spatio-temporal inter-color correlation filtering [19] or deep-learning encoder-decoder models [6]. On the other hand, several studies addressed the film grain synthesis

task. In [31], a Boolean in-homogeneous model [39] is used to model the grain, which corresponds to uniformly distributed disks. In AV1 codec [32], film grain is modeled by an autoregressive (AR) method as well as by an intensity-based function to adjust its strength. In VVC [34], a method based on frequency filtering is used. The grain pattern is first modeled thanks to a discrete cosine transform (DCT) applied to the grain blocks corresponding to smooth regions, and further scaled to the appropriate level, by using a step-wise scaling function. In [6], a conditional generative adversarial network (cGAN) that generates grain at different intensities is proposed. Yet, it does not perform any analysis on the original grain for a reliable synthesis. In [5], a deep-learning framework is proposed which consists of a style encoder for film grain style analysis, a mapping network for film grain style generation, and a synthesis network that generates and blends a specific grain style to a content in a content-adaptive manner.

Energy-aware images Many works target the reduction of the energy consumption of images while displayed on screens, especially for OLED displays. A first set of methods reduces the luminance through clipping or equalizing histograms [20, 21]. Other works directly scale the pixel luminance [26, 36, 38]. The most promising methods leverage deep learning models, trained with a combination of loss functions that minimize the energy consumption while maintaining an acceptable perceptual quality. In [43], a deep learning model trained with a variational loss for simultaneously enhancing the visual quality and reducing the power consumption is proposed. Authors in [38] describe an adaptive contrast enhancement (ACE) convolutional neural network, that performs contrast enhancement of luminance scaled images. In [33], an improved version of ACE, called Residual-ACE (R-ACE), is proposed to infer an attenuation map instead of a reduced image. In [26], authors revisit R-ACE to significantly reduce the complexity without compromising the performance. Different from the above methods, an invertible energy-aware network (InvEAN) [25] produces invertible energy-aware images and allows to recover the original images if required.

Invertible neural networks INNs learn the mapping $x = f(z)$, which is fully invertible as $z = f^{-1}(x)$, through a sequence of differentiable invertible mappings such as affine coupling layers [12] and invertible 1×1 convolutional layers [24]. INNs have direct applications in ambiguous inverse problems by learning information-lossless mappings [13, 27, 44]. The lost information in the forward process is captured by additional latent output variables. Thus, the inverse process is learned implicitly. A first application is the steganography, *i.e.*, concealing images into other images [9, 28]. In [44], an INN is used to produce invertible grayscale images, where the lost color information is encoded into a set of Gaussian distributed latent variables. The original color version can be recovered by using a new set of randomly sampled Gaussian distributed variables as input, together with the synthetic grayscale, through the reverse mapping. Similarly, an invertible denoising network (InvDN) transforms a noisy input into a LR clean image and a latent representation containing noise in [27]. To discard noise and restore the clean image, InvDN replaces the noisy latent representation with

another one sampled from a prior distribution during reversion. In [13], another INN further disentangles noise from the high frequency image information.

3 Proposed approach

With the target of reducing the overall energy consumption of the video chain, our 3R-INN framework, run in a forward pass at the encoder side, performs three invertible tasks simultaneously: 1) film grain removal, 2) downscaling and 3) display energy reduction, as illustrated in Figure 1. From a HR grainy image $I_G \in \mathbb{R}^{H \times W \times 3}$, 3R-INN outputs a visually pleasant grain-free LR energy-aware image $\tilde{I}_{LR|R} \in \mathbb{R}^{\frac{1}{2}H \times \frac{1}{2}W \times 3}$ with $R \in [0, 1]$ being the energy reduction rate, and 2 being the scaling factor corresponding to the best compromise between quality of the LR images, framework complexity and energy savings in the video chain. To ensure the process invertibility and the bijective mapping, the lost information mainly due to grain removal and downscaling is captured in a latent variable z distributed according to a standard Gaussian distribution $\mathcal{N}(0, 1)$. This can be formulated as: $[\tilde{I}_{LR|R}, z] = f_\theta(I_G)$ where θ is the set of trainable parameters of the 3R-INN network f . $\tilde{I}_{LR|R}$ is intended to be encoded, transmitted and displayed at the end-user device for an optimal energy consumption and quality of experience trade-off. The lost information z is further disentangled into two parts inside 3R-INN, by setting $\tilde{z}$ its internal representation as $\tilde{z} = [\tilde{z}_D, \tilde{z}_G]$ with $\tilde{z}_D$ and $\tilde{z}_G$ representing losses due to downscaling and grain removal, respectively.

In case the original content should be recovered, 3R-INN is run in an inverse pass at the decoder side (see Figure 1), as follows: $\tilde{I}_G = f_\theta^{-1}([\tilde{I}_{LR|R}, \tilde{z}])$. The original HR grainy content is then reconstructed with no need to transmit any auxiliary information in the video chain thanks to the modeling of the lost information. Moreover, thanks to the film grain and high frequency loss disentanglement, $\tilde{z} = [\tilde{z}_D, \tilde{z}_G]$, 3R-INN is also able to generate a clean HR version $\tilde{I}_C$ of the original content by setting $\tilde{z}_G = 0$. The overall architecture of the proposed framework is composed of three block types: one Haar Transformation block, several invertible blocks and a conditional latent encoding block, as illustrated in Figure 1.

Haar transform As removing film grain and downscaling an image significantly impacts high frequencies, it seems natural to first decompose the input HR image into low and high-frequency components. For that purpose, we chose the dyadic Haar wavelet transformation, similarly to [42, 44], because of its simplicity, efficiency and invertibility. Specifically, the Haar transform decomposes an input feature $f_{in} \in \mathbb{R}^{H \times W \times C}$ into one low-frequency $f_{low} \in \mathbb{R}^{\frac{1}{2}H \times \frac{1}{2}W \times C}$ and three high-frequency $f_{high} \in \mathbb{R}^{\frac{1}{2}H \times \frac{1}{2}W \times 3C}$ sub-bands. f_{low} , produced by an average pooling, represents the overall structure and coarse features of the image, while f_{high} contains finer details in the vertical, horizontal and diagonal directions, corresponding to film grain and edges. This splitting strategy allows to separate very early in the process the low frequency components from the information we aim to suppress. f_{low} and f_{high} serve as input to the following invertible blocks.

Invertible block As invertible blocks, we selected the coupling layer architecture proposed in [24]. A given input h^i is composed of two parts h_1^i and h_2^i , representing the three low-frequency and the nine high-frequency sub-bands of the color input channels RGB , respectively. These sub-bands are then processed by the i^{th} invertible block as follows:

$$\begin{aligned} h_1^{i+1} &= h_1^i + \phi(h_2^i) \\ h_2^{i+1} &= h_2^i \odot \exp(\psi(h_1^{i+1})) + \eta(h_1^{i+1}) \end{aligned} \quad (1)$$

where ϕ , ψ and η are dense blocks [18]. Given $[h_1^{i+1}, h_2^{i+1}]$, the inverse transformation can be easily computed by:

$$\begin{aligned} h_2^i &= (h_2^{i+1} - \eta(h_1^{i+1})) / \exp(\psi(h_1^{i+1})) \\ h_1^i &= h_1^{i+1} - \phi(h_2^i) \end{aligned} \quad (2)$$

Conditioned latent encoding block Invertible networks learn a bijective mapping between an input and an output distribution. In case of information loss, a latent variable $\tilde{z}$ is added to ensure the invertible property. This latent variable is assumed to follow a known distribution *i.e.*, a standard Gaussian distribution, to avoid transmitting additional information for the reconstruction, and to make the process case-agnostic. In our context, this would mean that the reconstruction of the HR grainy ($\tilde{I}_G$) or clean ($\tilde{I}_C$) images would not rely on the a priori knowledge of the LR image $\tilde{I}_{LR|R}$. To overcome this limitation and to enable an image-adaptive reconstruction during the inverse pass, the lost information $\tilde{z}$ is transformed into a Gaussian distributed latent variable z whose mean and variance are conditioned on $\tilde{I}_{LR|R}$. This is done through the use of a latent encoding block inspired from [44], whose structure is a one-side affine coupling layer that normalizes $\tilde{z}$ into a standard Gaussian distributed variable z as follows, with ϕ_g and θ_g being dense blocks:

$$z = (\tilde{z} - \phi_g(\tilde{I}_{LR|R})) / \exp(\theta_g(\tilde{I}_{LR|R})) \quad (3)$$

The reverse mapping can be formulated as:

$$\tilde{z} = z \odot \exp(\theta_g(\tilde{I}_{LR|R})) + \phi_g(\tilde{I}_{LR|R}) \quad (4)$$

Training objectives

3R-INN is first trained on the rescaling and film grain removal/synthesis tasks, before being fine-tuned on the energy reduction task.

Rescaling and film grain removal/synthesis tasks The **Forward Pass** optimization is driven by a fidelity loss $\mathcal{L}_{forw}$ to guarantee a visually pleasant clean LR image $\tilde{I}_{LR}$, and a regularization loss $\mathcal{L}_{reg}$ to guarantee that the latent variable z follows a standard Gaussian distribution. To guide f_θ to generate $\tilde{I}_{LR}$, a downsampled image I_{LR} of the HR clean image I_C is computed by a bicubic filter, and used as ground-truth to minimize $\mathcal{L}_{forw}$:

$$\mathcal{L}_{forw}(\tilde{I}_{LR}, I_{LR}) = \frac{1}{N} \sum_{i=1}^N \|\tilde{I}_{LR} - I_{LR}\|_2 \quad (5)$$

with N the batch size. Second, the log-likelihood of the probability density function $p(z)$ of the standard Gaussian distribution is maximized, with $D = \dim(z)$:

$$\mathcal{L}_{reg} = -\log(p(z)) = -\log\left(\frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}\|z\|^2\right)\right) \quad (6)$$

The **Inverse Pass** optimization consists of two fidelity losses $\mathcal{L}_{back_G}$ and $\mathcal{L}_{back_C}$, to restore $\tilde{I}_G$ and $\tilde{I}_C$, respectively. To this end, z is first decoded into $\tilde{z}$ by the latent encoding block conditioned by $\tilde{I}_{LR}$. Then the disentanglement of film grain (G) and fine details (D) is performed with $\tilde{z} = [\tilde{z}_D, \tilde{z}_G]$.

$\tilde{I}_G$ is reconstructed by considering all the information contained in $\tilde{z}$, *i.e.*, related to film grain and fine details:

$$\mathcal{L}_{back_G}(\tilde{I}_G, I_G) = \frac{1}{N} \sum_{i=1}^N \|f_{\theta}^{-1}(\tilde{I}_{LR}, z)_{||[\tilde{z}_D, \tilde{z}_G]} - I_G\|_1 \quad (7)$$

$\tilde{I}_C$ is restored by considering only the subset $\tilde{z}_D$ of $\tilde{z}$, *i.e.*, by using $\tilde{z} = [\tilde{z}_D, \tilde{z}_G = 0]$ as follows:

$$\mathcal{L}_{back_C}(\tilde{I}_C, I_C) = \frac{1}{N} \sum_{i=1}^N \|f_{\theta}^{-1}(\tilde{I}_{LR}, z)_{||[\tilde{z}_D, 0]} - I_C\|_1, \quad (8)$$

For both fidelity losses, the ℓ_1 norm is classically used as in [27, 42]. Finally, for the first two tasks, the following weighted sum is minimized:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{forw} + \lambda_2 \mathcal{L}_{reg} + \lambda_3 \mathcal{L}_{back_C} + \lambda_4 \mathcal{L}_{back_G} \quad (9)$$

Energy-aware task After 3R-INN learns the film grain removal/synthesis and rescaling tasks, it is fine-tuned during the forward pass with additional power and fidelity losses, $\mathcal{L}_{pow}$ and $\mathcal{L}_{SSIM}$, to output an **energy-aware** grain-free LR image $\tilde{I}_{LR|R}$, *i.e.*, its power consumption is reduced by R compared to the power consumption of I_{LR} . Contrary to most works computing energy aware images, assuming a linear relationship between the power consumption P_Y of an image and its linearized luminance [33], we follow the model in [11] dedicated to RGBW OLED screens, and compute P_{RGBW} as the sum of the powers consumed by the four R, G, B, W leds. As in [25], the following power loss is then minimized:

$$\mathcal{L}_{pow} = \|\tilde{P}_{RGBW} - (1 - R) \times P_{RGBW}\|_1 \quad (10)$$

with $(1 - R) \times P_{RGBW}$ the desired target power and $\tilde{P}_{RGBW}$ the power of $\tilde{I}_{LR|R}$. To ensure a better visual quality of the energy-aware images, a structural similarity index measure (SSIM) loss is added and minimized as follows:

$$\mathcal{L}_{SSIM} = 1 - SSIM(\tilde{I}_{LR|R}, I_{LR}) \quad (11)$$

As the inverse pass objectives remain exactly the same, the total loss minimized in the fine-tuning stage is:

$$\mathcal{L}_{finetuned} = \mathcal{L}_{total} + \lambda_5 \mathcal{L}_{pow} + \lambda_6 \mathcal{L}_{SSIM} \quad (12)$$

Training details During training, we use the DIV2K training set [4] from the FilmGrainStyle740K dataset [5], which contains pairs of corresponding images with and without grain. To complement the DIV2K validation set, we evaluate 3R-INN on the BSDS300 test set [30] and Kodak24 dataset [14], which were augmented to add grainy versions of the images, by following the same process as in the FilmGrainStyle740K dataset³. Input images were randomly cropped into 144×144 and augmented by applying random horizontal and vertical flips. Other training parameters are: Adam optimizer [23, 35] with $\beta_1 = 0.9$, $\beta_2 = 0.999$; mini-batch size of 16; 500k (training of the first two tasks) + 5k (energy-aware fine-tuning) iterations; learning rate initialized as $2e-4$ and halved at [100k, 200k, 300k, 400k] mini-batch updates. Hyper-parameters are set to: $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6) = (40, 1, 1, 1, 1e10, 1e4)$ and eight successive invertible blocks are used. Scale and shift coefficients are learned through a five-layer densely connected convolutional block. Each convolutional filter is of size 3×3 , with padding 1, followed by a leaky ReLU activation layer with negative slope set to 0.2. The intermediate channel number of the convolutional blocks is fixed to 32. Dimensions of $\tilde{z}_D$ and $\tilde{z}_G$ were set to (8, 1), respectively.

4 Experiments

The goal of our paper is to reduce the overall energy consumption along the video distribution system using 3R-INN, which primarily supports the display of a grain-free energy-aware LR image, and then offers the possibility to recover the original version, *i.e.*, the grainy HR, as well as a clean HR version as a third option. Thus, we adopt the following evaluation use case: we first assess the energy savings achieved by using 3R-INN along the video chain, and evaluate its energy needs. Then, we assess its performances in terms of quality for the LR grain-free energy-aware images and the reconstructed HR grainy images against state-of-the-art methods. An evaluation of the reconstructed HR clean images and an ablation study are provided in the supplementary materials.

4.1 Energy consumption performance

Evaluation of energy savings To estimate the energy savings realized along the video distribution system (headend, delivery, decoding and display), we tested the full video transmission chain by applying 3R-INN on two JVET sequences *RaceHorses* (300 frames, 832×480 , 10s) and *BasketBall* (500 frames, HD) [8]. The LR clean energy-aware at $R = 20\%$ is encoded using VTM [3], in full intra mode. Although not reflecting real-world scenarios in which efficient hardware decoders are used, the choice of using a non-optimized software-based VVC decoder (VTM) will nevertheless enable to demonstrate that 3R-INN results in consistent energy savings. Fixed broadband transmission was assumed. We then decoded and displayed the sequences on an OLED screen, with all TV options disabled, including the ambient light setting. Similarly to state-of-the-art methods, no video specific optimization was conducted as 3R-INN works in a frame by frame manner. In particular, film grain is known to be temporally

³ The additional dataset is proposed at www.interdigital.com/data_sets/filmgrainstyle740k-dataset

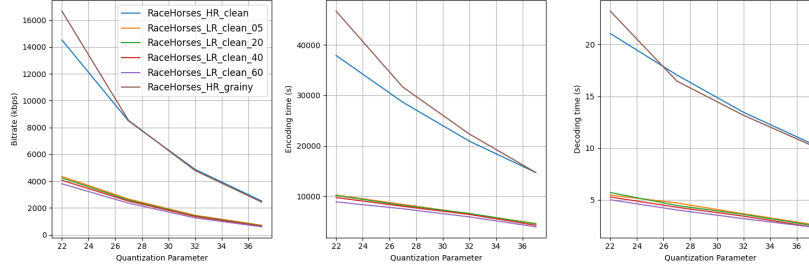


Fig. 2: Bit-rate, encoding and decoding times with and without using 3R-INN in terms of QP for sequence *RaceHorses*.

uncorrelated, hence its frame-based analysis. Resulting videos are provided in the supplementary material, illustrating 3R-INN capability to process temporal content.

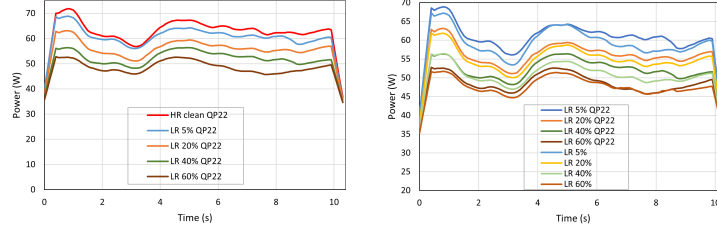


Fig. 3: Measured power consumption when displaying sequence *RaceHorses*. Left: Comparison between HR and LR versions at QP=22. Right: Comparison between LR versions before and after encoding/decoding.

Figure 2 reports the average encoding/decoding times and bit-rates, for different quantization parameters (QP), for the original HR clean and grainy *RaceHorses* sequences, and for the resulting LR versions with different $R \in \{5\%, 20\%, 40\%, 60\%\}$. Up to $QP = 27$, encoding and decoding the HR grainy video is more time and bit-rate demanding than for the HR clean version. For higher QPs, encoding time is still higher, however, bit-rate and decoding time are similar, because grain was removed during the encoding process. This confirms that compressing a grainy video while preserving film grain requires encoding at low QPs (which is far from the real-world scenario), leading to high and impractical bit-rates. On the contrary, encoding LR grain-free versions, whatever the value of R , shows substantially lower times and bit-rates, and consequently reduces the energy at the head-end, transmission and decoding stages.

Figure 3 presents actual measures of energy consumptions on an OLED LG-42C2 screen, for $R \in \{5\%, 20\%, 40\%, 60\%\}$, for the sequence *RaceHorses*. On the left plot, we compare the consumption of the encoded/decoded LR and HR clean sequences at $QP = 22$. This proves that displaying an energy-aware video at different reduction rates significantly reduces the display power consumption. The average gains of power are 6.8%, 21.5%, 33.3% and 44.2%. The right

plot compares the consumption of the LR sequences for different R , before and after encoding/decoding (QP = 22). For each R , we observe a non-significant impact of the compression on the display consumption. This demonstrates that energy-aware images are to some extent robust to compression in terms of power values. Similar results are obtained for the sequence *BasketBall* (shown in the supplemental material).

Table 1 illustrates the end-to-end energy savings at each level of the video chain, for the sequence *RaceHorses*, at QP22 and $R = 20\%$, according to the energy model in [17, 29]. Note that a range of power consumption values are considered for both encoding and decoding where the boundary values represent a very optimized vs. a non optimized power consumption encoder, as well as hardware and software decoders, respectively (the detailed computation is provided in the supplemental material). 3R-INN allows 74% and 78% of total end-to-end savings with respectively highly power optimized and non-optimized encoders/decoders. This corresponds to savings of 78% for head-end, 19% for delivery and *ca.* 77% for decoding. From this, we draw several observations: **Head-end:** As expected, the encoding energy consumption E_c significantly depends on the incoming resolution. The HR sequences are the most energy-demanding, particularly when they contain film grain, making encoding at a lower resolution a wise choice. **Delivery:** The transmission energy consumption E_t does not strongly depend on bitrate, but more on the power consumed by the infrastructure. Thus, transmitting HR instead of LR content results in a relatively small energy gain. **Decoding:** The gain in energy consumption for the decoding operation is significant, even if, in absolute value, it remains quite low. **Display:** Displaying the energy-aware clean LR video at $R = 20\%$ results in 11% of energy consumption reduction compared to the original grainy HR video. With the removal of the static consumption of the screen, the achieved energy reduction is higher and nearly reaches the target rate of 20%. In absolute value, the energy consumption of display E_D is significant compared to that of the other components of the video chain, except for the encoder. However, since the content is encoded once and displayed several times, the display energy gains are further multiplied by the number of displays. Assuming that the sequence *RaceHorses* is viewed by 10% of Netflix subscribers [2] for one hour, using 3R-INN would save 156 GWh of energy, equivalent to the monthly consumption of 176 American citizens [1].

Energy cost of using 3R-INN 3R-INN is a **single network** that replaces **three** separate NN architectures for the tasks of grain removal /synthesis, rescaling and building energy-aware images. In that sense, Table 2 reports a comparison of its complexity and energy performance against those of the sum of the

Table 1: End-to-end energy savings along the video chain for a fixed broadband access. Case study of the sequence *RaceHorses* (300 frames, 832×480 , 30fps, 10s).

	Headend	Delivery	Decoding	Display	Total
	Encoding time (s), E_c	Bitrate (kbps) (QP=22), E_t	Decoding time (s), E_d	Power (W), E_D	Power (w)
Original (Grainy HR)	46682, [2.59Wh, 1167kWh]	16668, 0.0055Wh	23.2, [0.005Wh, 0.15Wh]	60.8, 0.168Wh	[2.7685Wh, 1167.0003kWh]
Original (Clean HR)	37901, [2.10Wh, 947kWh]	14516, 0.0053Wh	21.0, [0.004Wh, 0.14Wh]	59.7, 0.165Wh	[2.2743Wh, 947.0003kWh]
Ours (Energy-aware clean LR) $R = 20\%$	10150, [0.56Wh, 253kWh]	4237, 0.0045Wh	5.7, [0.001Wh, 0.039Wh]	54.4, 0.151Wh	[0.7165Wh, 253.0001kWh]
Reduction in % Grainy HR vs Ours	[78%, 78%]	19%	[80%, 74%]	11%	[74%, 78%]
Reduction in energy Grainy HR vs Ours	[2.03Wh, 914kWh]	0.0010Wh	[0.004Wh, 0.11Wh]	0.017Wh	[2.052Wh, 914.0002kWh]

best-performing networks for the three tasks, on sequence *RaceHorses*, in terms of number of parameters, number of Multiply-ACcumulate operations (MACs), power consumption and *number of equivalent displays*. Power consumption is approximated using the model in [16], which we assume is valid for all the networks used in this study. The number of equivalent displays represents the number of displays/users that are needed to counterbalance the needed energy to run the network(s). Computation details are provided in the supplementary materials.

While producing a clean, 20% energy-reduced LR image from an HR grainy one, the cost of running 3R-INN in a forward pass is compared with the successive use of three NN architectures: StyleFG analyzer module to model film grain, IRN to remove it and downscale the image, and InvEAN to reduce its energy consumption. From the first part of Table 2, 3R-INN counts significantly fewer parameters/operations and needs less power than the combination of the three networks. Moreover, solely comparing with the savings at the display side (≈ 6 Watts, see Table 1), the power gain is tangible as soon as the content is displayed 110 times for 3R-INN vs. 161 times for the combination.

In case recovering an HR grainy image is required, 3R-INN is simply run in an inverse pass, thus, its complexity and power savings remain the same. In contrast, InvEAN and IRN are first run in an inverse pass to recover a clean HR version; then, film grain is synthesized using StyleFG synthesizer module. This time, 3R-INN use represents 58% less power than the combination (660 vs. 1573W), and requires fewer equivalent displays (110 vs. 262) to offset its power requirements.

4.2 Quantitative and qualitative evaluation of LR images

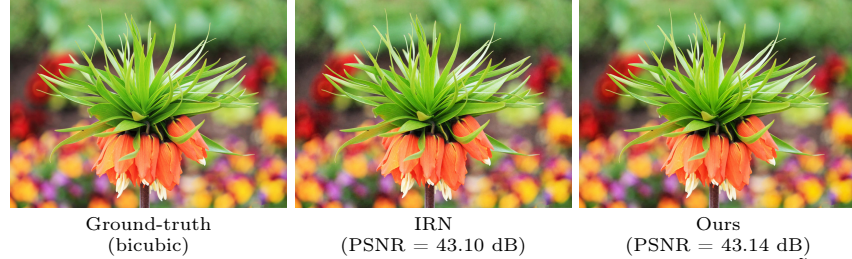
Film grain removal and downscaling The quantitative and qualitative evaluations of the LR clean image $\tilde{I}_{LR|R=0}$, *i.e.*, corresponding to an energy reduction rate $R = 0$, are given in Table 3 and Figure 4, respectively. The reference image is the bicubic rescaling of the HR clean image. Although quite similar to the experimental protocol used for IRN in [42], we here assess the ability of the network both to rescale and to remove film grain. Thus, for a fair comparison, we re-train IRN on our dataset to perform both tasks, given grainy HR images as input. The latter is designed and optimized for clean HR image reconstruction only, although it can reconstruct grainy HR images if the original high-frequencies z are provided. However, in the video chain context, this would lead to the transmission of heavy metadata, whereas 3R-INN operates without any transmission of metadata. Results show that the proposed method outperforms IRN in terms

Table 2: Comparison of complexity and power performance of 3R-INN against the use of three independent neural networks for sequence *RaceHorses*. A corresponding power equivalent in terms of number of displays is also given.

Output	Network(s)	Parameters	GMACs	Power(W)	Display equivalent #
Clean LR at R=20%	StyleFG analyzer + IRN + InvEAN	23.3M	334	967	161
	3R-INN (forward)	1.7M	230	660	110
Grainy HR	StyleFG synthesizer + IRN + InvEAN	36.3M	616	1573	262
	3R-INN (inverse)	1.7M	230	660	110

Table 3: Comparison between generated LR clean images $\tilde{I}_{LR|R=0}$ and a bicubic rescaling of the HR clean image as ground-truth.

Method	DIV2K		BSDS300		Kodak24	
	PSNR $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$
IRN [42]	39.06	0.942	38.95	0.953	38.75	0.947
Ours	39.63	0.951	39.79	0.964	39.71	0.957

**Fig. 4:** Comparison between a bicubic downscaling, IRN and the clean LR $\tilde{I}_{LR|R=0}$. of PSNR and SSIM. They also outline its good generalization, as even better performances are observed on BSDS300 and Kodak24 datasets.

Energy-aware images For $R > 0$, we evaluate the visual quality of the LR clean energy-aware image $\tilde{I}_{LR|R}$ against state-of-the-art energy-aware methods, *i.e.*, a global linear scaling of the luminance (LS), R-ACE [33], DeepPVR [26] and InvEAN [25]. To solely evaluate the energy-aware task, and for a fair comparison, existing methods were evaluated while taking as input the output of our method after the fine tuning step with $R = 0$. All evaluation metrics in the following were calculated with this image as reference. Table 4 reports PSNR-Y and SSIM metrics at four reduction rates, on three test sets. Two conclusions can be drawn. First, when the power consumption model P_Y is used for a fair comparison with state-of-the-art methods, the proposed method outperforms LS and R-ACE methods, while being similar to DeepPVR and slightly below InvEAN. When the power consumption model P_{RGBW} is used, the quality scores of 3R-INN are significantly better, and especially for the PSNR-Y. This can be explained by the fact that our model does not learn to reduce the image luminance, contrary to state-of-art methods. The latter in turn were not trained to optimize P_{RGBW} ; this may explain their lower performances.

This trend is confirmed by Figure 6 which plots SSIM scores as function of the actual reduction rate, computed with P_{RGBW} . PSNR plots are provided in the supplemental material. Figure 5 shows a qualitative comparison of energy-

Table 4: PSNR-Y and SSIM quality scores for the energy-aware task for four reduction rates R . 3R-INN results are presented for two power consumption models, *i.e.* P_Y (for comparison with state-of-the-art methods) and P_{RGBW} , corresponding to RGB and RGBW OLED screens, respectively. InvEAN model is not available at $R = 5\%$ in [25].

Method	DIV2K				BSDS				Kodak24			
	R=5%	R=20%	R=40%	R=60%	R=5%	R=20%	R=40%	R=60%	R=5%	R=20%	R=40%	R=60%
LS	39.34/0.999	27.01/0.991	20.33/0.958	16.06/0.877	39.64/0.999	27.31/0.990	20.67/0.955	16.35/0.867	39.38/0.999	27.05/0.991	20.41/0.957	16.09/0.875
R-ACE [33]	41.53/0.995	26.59/0.967	20.05/0.901	15.92/0.788	40.55/0.997	26.90/0.978	20.24/0.915	16.12/0.806	40.70/0.997	26.74/0.983	20.08/0.930	15.98/0.830
DeepPVR [26]	39.37/0.996	27.12/0.983	21.04/0.952	15.81/0.890	39.63/0.997	27.53/0.989	21.13/0.950	16.36/0.894	39.27/0.997	27.17/0.989	20.61/0.955	16.00/0.892
InvEAN [25]	-	27.75/0.994	21.17/0.973	17.07/0.932	-	28.25/0.993	21.74/0.973	17.72/0.931	-	27.92/0.993	21.42/0.973	17.37/0.932
Ours (P_Y)	39.55/0.987	27.32/0.980	20.62/0.949	16.43/0.883	40.06/0.994	27.65/0.986	20.94/0.955	16.77/0.883	40.02/0.992	27.43/0.985	20.70/0.954	16.51/0.886
Ours (P_{RGBW})	47.68/0.998	38.02/0.993	29.15/0.974	23.66/0.945	48.33/0.999	38.36/0.995	30.47/0.983	24.96/0.961	47.47/0.998	37.39/0.994	29.63/0.982	24.18/0.958

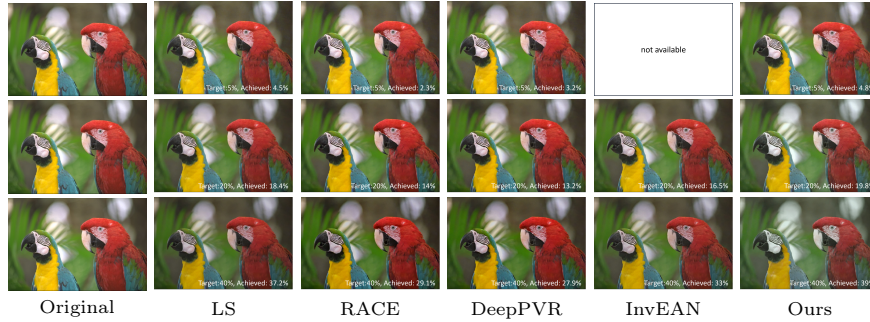


Fig. 5: Comparison of generated energy-aware images with the state-of-the-art, for $R \in \{5\%, 20\%, 40\%\}$ from first to third lines. Achieved rates computed by the power model in [11] are provided. InvEAN model is not available at $R = 5\%$ in [25].

aware images. 3R-INN and LS respect the reduction rate targets better than other methods. Our method also exhibits a different behavior for high values of R , once again keeping the luminance but modifying the colors. The subjective comparison is however difficult since the achieved energy reduction rates vary from one method to another. Although not fully dedicated to the energy-reduction task, 3R-INN performs well compared to existing methods and similarly to InvEAN, it offers the possibility to recover the original image without any side-information.

4.3 Quantitative and qualitative evaluation of HR grainy images

The reversibility property of 3R-INN is an important feature. To evaluate this property, we evaluated the HR grainy reconstruction with state-of-the-art film grain synthesis methods: VVC (Versatile Video Coding) implementation [34], Deep-FG [6] and Style-FG [6]. Table 5 summarizes the quantitative results for $R = 0$, in terms of fidelity of the synthesized grain using learned perceptual image patch similarity (LPIPS), JSD-NSS and the KL divergence (KLD) [45], these last two being computed between the histograms of ground-truth and HR grainy images. All methods perform analysis and synthesis except Deep-FG for which we generated 5 versions of grain, one per available intensity level, and kept only the best performing image for each metric in the comparison.

Results show that the proposed method outperforms quantitatively VVC [34] and Deep-FG [6]. It also performs better than Style-FG [6] for LPIPS and KLD

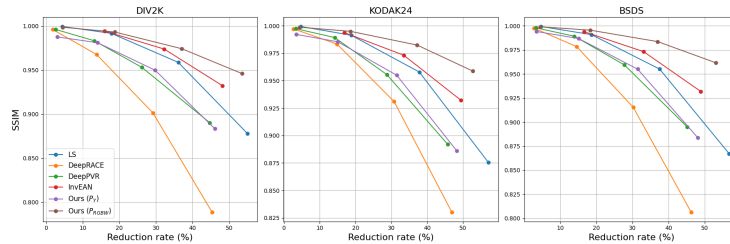


Fig. 6: SSIM scores as function of the target power reduction, for the different energy-aware methods.

Table 5: Comparison between reconstructed HR grainy images and ground-truth for different methods on DIV2K validation set.

	Analysis	Auxiliary data	JSD-NSS ↓	LPIPS ↓	KLD ↓
VVC [34]	✓	set of params	0.0148	0.2981	0.0327
Deep-FG [6]	x	x	0.0134	0.3722	0.0260
Style-FG [5]	✓	style vector	0.0024	0.1592	0.0232
Ours	✓	none	0.0088	0.0445	0.0177

metrics which are representative of the quality of generated grain. The lower JSD-NSS value for Style-FG [6] could be explained by the fact that it is a GAN-based network which models the data distribution at the expense of the output quality. The qualitative comparison in Figure 7 confirms these observations (additional results in the supplemental material). Another advantage of 3R-INN is that no auxiliary data is required for grain synthesis, unlike VVC and Style-FG, which transmit a set of parameters and a style vector respectively. Similar results are obtained for $R > 0$ and are presented as supplemental material.

**Fig. 7:** Qualitative evaluation of HR synthesized grainy images for different methods, with LPIPS values between parenthesis.

5 Conclusion

This paper presents 3R-INN, the first network that enables to reduce the overall energy consumption in the video transmission chain. Given an HR grainy image, 3R-INN delivers a minimum viable quality, low-resolution, grain-free and energy-aware image, thus reducing the energy required for encoding, transmission, decoding and display. With multiple views of the same content, 3R-INN achieves a positive energy balance, far more efficient than current state-of-the-art systems. Furthermore it does not need to transmit auxiliary information to reconstruct the original grainy content, since all the lost information including details, film grain and brightness was encoded and disentangled in a standard Gaussian distribution, through a latent encoding block conditioned on the LR image. Experimental results demonstrate that 3R-INN outperforms the existing methods by a large margin for film grain synthesis, and achieves state-of-the-art performance in the rescaling and energy-aware tasks. For the latter, a fine-tuning for each value of energy reduction rate target R was conducted. Conditioning the network on R to avoid fine-tuning different networks for each value of R , will therefore be investigated in the future, as an extension of current work. Some subjective test will also be conducted to assess the acceptability by end users of the provided LR energy-aware images.

References

1. Energy consumption household. <https://www.energybot.com/blog/average-energy-consumption.html>.
2. Netflix subscribers. <https://www.usnews.com/news/business/articles/2024-01-23/netflixs-subscriber-growth-surges-as-streaming-service-unwraps-best-ever-holiday-season-results>.
3. Vtm-19.0. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tags/VTM-19.0
4. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
5. Ameer, Z., Demarty, C.H., Le Meur, O., Ménard, D., François, E.: Style-based film grain analysis and synthesis. In: Proceedings of the 14th Conference on ACM Multimedia Systems. pp. 229–238 (2023)
6. Ameer, Z., Hamidouche, W., François, E., Radosavljević, M., Menard, D., Demarty, C.H.: Deep-based film grain removal and synthesis. IEEE Transactions on Image Processing (2023)
7. Bonninaeu, C., Hamidouche, W., Travers, J.F., Déforges, O.: Versatile video coding and super-resolution for efficient delivery of 8k video with 4k backward-compatibility. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2048–2052 (2020). <https://doi.org/10.1109/ICASSP40776.2020.9054716>
8. Boyce, J., Suehring, K., Li, X., Seregin, V.: Jvet-j1010: Jvet common test conditions and software reference configurations. In: 10th Meeting of the Joint Video Experts Team. pp. JVET–J1010 (2018)
9. Chen, Z., Liu, T., Huang, J.J., Zhao, W., Bi, X., Wang, M.: Invertible mosaic image hiding network for very large capacity image steganography. arXiv preprint arXiv:2309.08987 (2023)
10. Dai, J., Au, O.C., Pang, C., Yang, W., Zou, F.: Film grain noise removal and synthesis in video coding. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 890–893. IEEE (2010)
11. Demarty, C.H., Blondé, L., Le Meur, O.: Display power modeling for energy consumption control. In: 2023 IEEE International Conference on Image Processing (ICIP). IEEE (2023)
12. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. arXiv preprint arXiv:1605.08803 (2016)
13. Du, W., Chen, H., Zhang, Y., Yang, H.: Hierarchical disentangled representation for invertible image denoising and beyond. arXiv preprint arXiv:2301.13358 (2023)
14. Franzen, R.: Kodak lossless true color image suite. source: <http://r0k.us/graphics/kodak> 4(2), 9 (1999)
15. Gomila, C.: Sei message for film grain encoding. JVT document, May 2003 (2003)
16. Herglotz, C., Brand, F., Regensky, A., Rievel, F., Kaup, A.: Processing energy modeling for neural network based image compression. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 2390–2394. IEEE (2023)
17. Herglotz, C., Kränzler, M., Schober, R., Kaup, A.: Sweet streams are made of this: The system engineer’s view on energy efficiency in video communications [feature]. IEEE Circuits and Systems Magazine 23(1), 57–77 (2023)
18. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)

19. Hwang, I., Jeong, J., Choi, J., Choe, Y.: Enhanced film grain noise removal for high fidelity video coding. In: 2013 International Conference on Information Science and Cloud Computing Companion. pp. 668–674. IEEE (2013)
20. Kang, S.J.: Image-quality-based power control technique for organic light emitting diode displays. *Journal of Display Technology* **11**(1), 104–109 (2015)
21. Kang, S.J., Kim, Y.H.: Image integrity-based gray-level error control for low power liquid crystal displays. *IEEE Transactions on Consumer Electronics* **55**(4), 2401–2406 (2009). <https://doi.org/10.1109/TCE.2009.5373816>
22. Kim, H., Choi, M., Lim, B., Lee, K.M.: Task-aware image downscaling. In: Proceedings of the European conference on computer vision (ECCV). pp. 399–414 (2018)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems* **31** (2018)
25. Le Meur, O., Demarty, C.H.: Invertible energy-aware images. *IEEE Signal Processing Letters* (2023)
26. Le Meur, O., Demarty, C.H., Blondé, L.: Deep-learning-based energy aware images. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 590–594. IEEE (2023)
27. Liu, Y., Qin, Z., Anwar, S., Ji, P., Kim, D., Caldwell, S., Gedeon, T.: Invertible denoising network: A light solution for real noise removal. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13365–13374 (2021)
28. Lu, S.P., Wang, R., Zhong, T., Rosin, P.L.: Large-capacity image steganography based on invertible neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10816–10825 (2021)
29. Malmudin, J.: The power consumption of mobile and fixed network data services-the case of streaming video and downloading large files. In: *Electronics Goes Green*. vol. 2020 (2020)
30. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 2, pp. 416–423. IEEE (2001)
31. Newson, A., Delon, J., Galerne, B.: A stochastic film grain model for resolution-independent rendering. In: *Computer Graphics Forum*. vol. 36, pp. 684–699. Wiley Online Library (2017)
32. Norkin, A., Birkbeck, N.: Film grain synthesis for av1 video codec. In: 2018 Data Compression Conference. pp. 3–12. IEEE (2018)
33. Nugroho, K.A., Ruan, S.J.: R-ace network for oled image power saving. In: 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech). pp. 284–285. IEEE (2022)
34. Radosavljevic, M., François, E., Reinhard, E., Hamidouche, W., Amestoy, T.: Implementation of film-grain technology within vvc. In: *Applications of Digital Image Processing XLIV*. vol. 11842, pp. 85–95. SPIE (2021)
35. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237 (2019)
36. Reinhard, E., Demarty, C.H., Blondé, L.: Pixel value adjustment to reduce the energy requirements of display devices. *SMPTE Motion Imaging Journal* **132**(7), 10–19 (2023)

37. Robinson, D.: Greening of streaming: The less accord: Low energy sustainable streaming. In: Proceedings of the 2nd Mile-High Video Conference (MHV'23). p. 115 (2023)
38. Shin, Y.G., Park, S., Yoo, M.J., Ko, S.J.: Unsupervised deep power saving and contrast enhancement for oled displays. arXiv preprint arXiv:1905.05916 (2019)
39. Stoyan, D., Kendall, W.S., Chiu, S.N., Mecke, J.: Stochastic geometry and its applications. John Wiley & Sons (2013)
40. Sun, W., Chen, Z.: Learned image downscaling for upscaling using content adaptive resampler. IEEE Transactions on Image Processing **29**, 4027–4040 (2020)
41. Trust, T.C.: Carbon impact of video streaming. <https://www.carbontrust.com/en-eu/node/1537> (2021)
42. Xiao, M., Zheng, S., Liu, C., Wang, Y., He, D., Ke, G., Bian, J., Lin, Z., Liu, T.Y.: Invertible image rescaling. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 126–144. Springer (2020)
43. Yin, J.L., Chen, B.H., Peng, Y.T., Tsai, C.C.: Deep battery saver: End-to-end learning for power constrained contrast enhancement. IEEE Transactions on Multimedia **23**, 1049–1059 (2020)
44. Zhao, R., Liu, T., Xiao, J., Lun, D.P., Lam, K.M.: Invertible image decolorization. IEEE Transactions on Image Processing **30**, 6081–6095 (2021)
45. Zhu, F., Chen, G., Hao, J., Heng, P.A.: Blind image denoising via dependent dirichlet process tree. IEEE transactions on pattern analysis and machine intelligence **39**(8), 1518–1531 (2016)