Snapshot High Dynamic Range Imaging via Sparse Representations and Feature Learning

Konstantina Fotiadou, Grigorios Tsagkatakis, and Panagiotis Tsakalides

***Abstract*—Bracketed High Dynamic Range (HDR) imaging architectures acquire a sequence of Low Dynamic Range (LDR) images in order to either produce a HDR image or an "optimally" exposed LDR image, achieving impressive results under static camera and scene conditions. However, in real world conditions, ghost-like artifacts and noise effects limit the quality of HDR reconstruction. We address these limitations by introducing a post-acquisition snapshot HDR enhancement scheme that generates a bracketed sequence from a small set of LDR images, and in the extreme case, directly from a single exposure. We achieve this goal via a sparse-based approach where transformations between differently exposed images are encoded through a dictionary learning process, while we learn appropriate features by employing a stacked sparse autoencoder (SSAE) based framework. Via experiments with real images, we demonstrate the improved performance of our method over the state-of-the-art, while our single-shot based HDR formulation provides a novel paradigm for the enhancement of LDR imaging and video sequences.**

***Index Terms*—High dynamic range imaging, deep learning, sparse stacked autoencoders, sparse representations.**

I. Introduction

THE rapid evolution of display technologies, moving up from HD to 4K and 8K, has created an enormous excitement in the imaging and multimedia communities. Nevertheless, despite the significant increase in spatio-temporal resolution and the introduction of 3D content, dynamic range enhancement has received relatively little attention, since only a handful approaches have been proposed in literature that combine spatial and temporal visual features, in order to synthesize high-quality HDR content [1]. However, in the past few years, High Dynamic Range (HDR) imaging technology is acknowledged as the next big thing in consumer imaging, with more and more image and video acquisition, content reproduction, and graphics applications supporting HDR features [2], including the adoption of HDR video streaming by YouTube.[1] Furthermore, recent studies showed that the HDR market is estimated to increase from 1.82 Billion USD in 2015 to 36.82 Billion USD by 2022.[2] Consequently, the demand for generating high quality HDR imaging systems has grown tremendously.

While consumer imaging systems typically acquire 8 or 12-bit raw images, HDR technology aims at increasing the bit depth, targeting either the reproduction in a high dynamic range display [1], [3], or the direct fusion into a single enhanced LDR frame [4]. In HDR imaging, one utilizes a bracketed sequence of multiple LDR images to generate the HDR version of the scene, which can be displayed into a conventional LDR device via a *tone-mapping* approach [5]–[9], or fused into an "optimally" exposed single LDR image using techniques such as *exposure fusion* [4], [10]. As a characteristic example we may consider the study presented in [4], where the authors design a novel gradient-based weighted least square algorithm that extracts and combines the necessary details from both the brightest and darkest regions of HDR scenes, in order to synthesize an edge-preserving multi-scale exposure fusion algorithm.

Additionally, under ideal settings where the camera and the scene are both static, the HDR process produces high-quality, artistic images, that capture a significantly larger content of the scene while supporting the application of higher level understanding methods. However, in real conditions, alignment, motion blurring, and low signal power present formidable challenges [11], [12]. As an illustrative example, consider the situation where imaging takes place under low-light conditions, where a single bright spot, *e.g.*, a lamp post, can completely mask the visual information captured by the camera. Even when no such bright shots appear in the image, the need for a short exposure time is mandated by the impact of motion blur associated with longer exposures [13]. Furthermore, for the case of video sequences, imaging dynamic scenes with varying degrees of illumination can lead to abrupt image brightness changes and visual artifacts [14]. Last, even though HDR displays will become readily available in the near future, the majority of content that is currently available, either professional or consumer, is stored in LDR formats. Hence, the development of an *LDR*-to-*HDR*

Manuscript received January 29, 2019; revised May 12, 2019 and June 27, 2019; accepted July 22, 2019. Date of publication August 5, 2019; date of current version February 21, 2020. This work was partially funded by the DEDALE project, contract 665044, within the H2020 Framework Program of the European Commission. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Han Hu. *(Corresponding author: Konstantina Fotiadou.)*

K. Fotiadou and P. Tsakalides are with the Department of Computer Science, University of Crete, Crete 74100, Greece, and also with the Institute of Computer Science - FORTH, Crete GR-70013, Greece (e-mail: kfot@ics.forth.gr; tsakalid@ics.forth.gr).

G. Tsagkatakis is with the Institute of Computer Science - FORTH, Crete GR-70013, Greece (e-mail: greg@ics.forth.gr).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2933333

[1] https://www.forbes.com/sites/johnarcher/2017/09/08/youtube-brings-hdr-to-mobile/

[2] http://www.marketsandmarkets.com/Market-Reports/high-dynamic-range-market-159950041.html

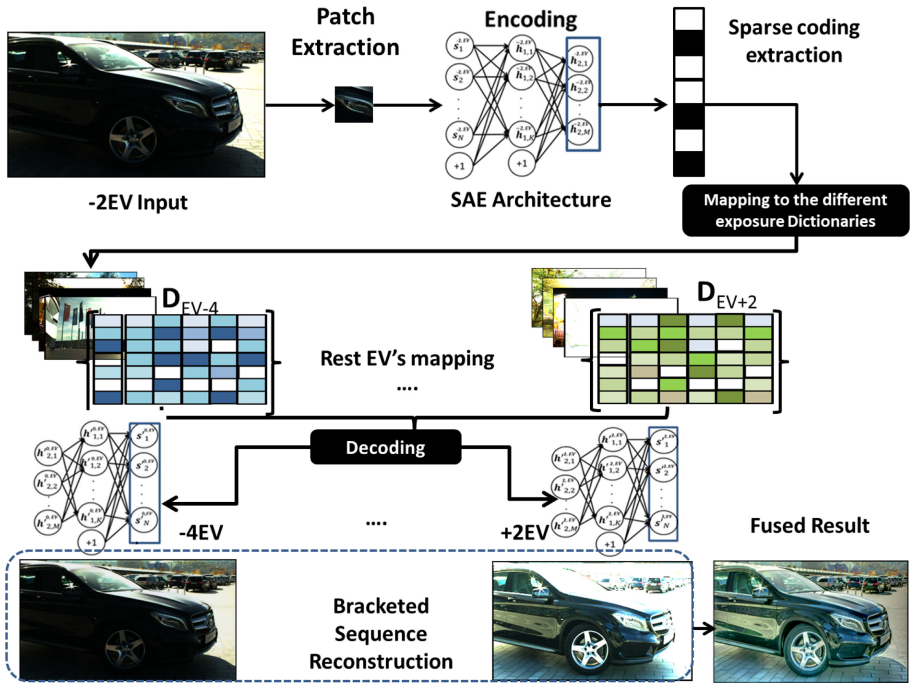


Fig. 1. Proposed Block Diagram: Our algorithm takes as input a snapshot image and reconstructs five different exposure versions of the scene. The resulting set of bracketed exposures is blended into a high-quality LDR image, via the exposure fusion technique that is explicitly explained in Section III.

transformation system for the conversion of the old LDR still images and video into HDR content, will attract significant interest within the computational photography and multimedia communities [15], [16].

A. Motivation and Contributions

To overcome limitations arising from traditional HDR photography and to address the enhancement of available LDR content, we propose a novel imaging model for post-acquisition enhancement of the dynamic range of static imagery and dynamic video. Compared to typical approaches, our technique considers as input a dramatically smaller number of LDR frames, even a single LDR exposure, thus allowing for Snapshot High Dynamic Range imaging and for the dynamic range enhancement of video. Formally, given an input image captured under a certain exposure condition, our task is to synthesize an extended sequence of bracketed LDR's which can either be combined to an "ideally" exposed LDR image or an HDR image. Fig. 1 presents the proposed system's block diagram, where a -2 EV image is utilized as input in order to reconstruct the bracketed sequence on the different EVs, from the darkest (-6 EV) to the brightest (4 EV).

At this point, it is crucial to emphasize that our algorithm is not able to directly synthesize the ideally exposed LDR image or the HDR image from a single input scene, since the results of this process cannot be numerically evaluated. Instead, we need first to estimate the bracketed sequence, and then proceed to the second step of our scheme that corresponds to the combination of the synthesized bracketed sequence. Additionally, operating from a single exposure opens the possibility of artificially enhancing the dynamic range of already acquired LDR imagery and video for reproduction in HDR displays. In addition to the objective of snapshot HDR imaging, we also explore the situation where a single exposure frame is acquired under low lighting conditions. In this task, we need to enhance a low power signal without increasing the noise by modeling the problem as estimating a long exposed image from its short exposure counterpart. In our experiments, we demonstrate that our system successfully recovers an extended bracketed sequence from extremely dark scenes, outperforming other state-of-the-art techniques. It is important to note that hallucinating information from saturated regions is outside the scope of this work.

Building on recent advances in optimization theory, we propose a scalable architecture based on the novel signal modeling and learning frameworks of Sparse Representations (SR) [17] and Joint Dictionary Learning. Sparsity has revolutionized multiple inverse imaging problems, including super-resolution, denoising, and deblurring, among others [18]–[23]. In this work, we employ *sparsity* as an efficient prior knowledge for producing differently exposed versions of an input scene. According to the SR framework, we argue that these different versions share similar sparse coding representations given an appropriately trained dictionary. To achieve this goal, we design a joint dictionary in a way that it captures the same statistical characteristics of the scene under different exposure values. Consequently, the features from the differently exposed images can be represented as a sparse linear combination of elementary dictionary atoms.

The objective of the proposed system is to take a small number of LDR images (even a single one!) and produce an extended sequence of LDR images of the same scene under different

exposure conditions. Exposure is quantified through the values which encode the relationship between exposure time $\mathbf{t}$ and relative aperture $\mathbf{N}$ (*f*-number), defined as: $EV = \log_2 \frac{\mathbf{N}^2}{\mathbf{t}}$. Large EVs lead to over-exposed saturated images, while small EVs produce under-exposed noisy ones. In our scheme, we consider the recovery of six differently exposed versions of the same scene, $\mathbf{S}_k$, where $k \in \{-6, -4, -2, 0, 2, 4\}$ EV.

In addition to the enhancement of dynamic range, we also investigate the case of enhancing low-light imagery. We address this problem as the mapping from a low-light image to the corresponding well-illuminated one by considering short and long exposures as prior knowledge with respect to the different illumination conditions. As opposed to the image denoising case, the proposed flexible scheme not only generates the complete bracketed sequence from low-light input images, but also from images acquired with larger exposure values, as well as from multiple differently exposed inputs. Finally, we should note that we are the first who combine the theoretical frameworks of Sparse Representations and Joint Dictionary Learning with Deep Feature learning formulations for HDR-based applications. In a nutshell, the key contributions of our work include:

i) a new framework for the synthesis of extended bracketed sequences from a small number of LDR frames, all the way to the extreme case of synthesis from a single snapshot;
ii) the automated extraction of appropriate features from raw pixel values through a deep learning approach, capable of modeling different exposure conditions, utilizing the efficient scheme of Stacked Sparse Autoencoders;
iii) the demonstration of proposed scheme for the enhancement of the dynamic range of LDR and low light frames;
iv) the experimental validation of our claims on challenging datasets, and a thorough comparison with several competitive state-of-the-art algorithms.

The remainder of the paper is structured as follows: Section II provides an overview of the related state-of-the-art concerning the low-light image enhancement and the HDR reconstruction techniques. Section III introduces the proposed multiple exposure sparsity based method. Section IV provides the experimental results concerning the stage of LDR synthesis and the step of the bracketed sequence blending. Additionally, this section highlights the performance of the proposed scheme for the problem of dynamic range enhancement of LDR video frames and compares with the state-of-the-art. Finally, extensions of our work are discussed in Section V.

II. Related Work

In the following section, we briefly overview state-of-the-art approaches for creating an enhanced version of a low-illumination scene and methods for producing HDR content from LDR imagery.

A. Low-Illumination Image Enhancement

Early approaches to the problem of low-light image enhancement exploit the family of histogram equalization techniques [24]. Although these techniques are straight-forward, they usually produce excessive contrast enhancement, leading to the introduction of severe noise, as well as saturation and leakage artifacts in the resulting image. Recently, the authors in [25] proposed a guided image contrast enhancement technique based on cloud images, in which the context-sensitive and context-free contrast is jointly improved via solving a multi-criteria optimization problem. Specifically, the context-sensitive contrast is improved via a unsharp masking on the filtered images, while the context-free contrast enhancement is achieved by a sigmoid transfer mapping. In order to automatically determine the contrast enhancement level, the optimization parameters are estimated by considering images with similar content. In [26], Huang *et al.* proposed an efficient method for contrast enhancement, combining transform-based gamma correction with traditional histogram equalization. Dong *et al.*, in [27], introduced an effective algorithm for the enhancement of low light videos by inverting the input dark frames and applying a *de-hazing* algorithm. Fu *et al.* proposed a Color Estimation Model (CEM), calculating the transform between night and day-time images [28].

Recently, in [8] the authors tested a window-based tone mapping approach that compresses the dynamic range, solving a global optimization problem. Apart from HDR compression, this method is also used for the enhancement of ordinary LDR images. Another night context enhancement technique, was demonstrated in [29], where the authors use the *Retinex Theory* as prior knowledge, for the enhancement of low-light images. Additionally, Raskar *et al.* proposed an image fusion approach, that enhances a dark image based on a mixed gradient field generated by the intensities of multiple pairs of night-time and their corresponding day-time images [30].

B. HDR Reconstruction

Traditional HDR approaches, utilize multiple LDR images, captured under different exposure settings in order to generate the HDR image [13], [31]. These approaches exhibit impressive performance when they are applied in static scenes, since they assume constant radiance at each pixel, for the different exposure conditions. However, in real world scenarios, scenes of interest typically contain moving objects, introducing motion blurring and leading to ghost-like effects, while the challenging problem of registration must also be addressed.

To overcome the registration and motion estimation problems of the traditional HDR approaches, the authors in [32], [33], construct the HDR image from a single shot, using spatially varying pixel sensitivities (SVE). The main idea is to place an optical mask, with cells of different exposures, adjacent to the image sensor, in order to control the amount of light that reaches each pixel. Motivated by the SVE acquisition architecture, Aguerrebere *et al.* [34] proposed a Gaussian Mixture Model framework for the reconstruction of the unknown (over-exposed or under-exposed) pixels and for the noise reduction in the dark regions of the scene. A recovery mechanism for the SVE frames was also recently proposed, based on the mathematical framework of low-rank Matrix Completion [35].

Sen *et al.* in [36], developed an energy minimization algorithm that considers a bracketed but not-aligned sequence of

LDR images of a dynamic scene and produces the corresponding aligned sequence, along with its HDR version. More recently, Zhao *et al.* [37] proposed a technique that generates HDR images using a modulo camera based on the assumption that a modulo camera is able to theoretically acquire unbounded radiance levels. The concept of versatile HDR Video production using multiple sensors was also recently proposed in [38]. More recent HDR-based context enhancement reconstruction were reported in [39]–[45]. Finally, the authors in [46] provide an interesting work on single image HDR reconstruction adhering to a convolutional neural network (CNN) learning-based architecture, while the authors in [47] provide a hybrid-loss learning approach for single image HDR, based on the novel scheme of generative adversarial networks and the pre-trained VGG19 architecture.

III. Multiple Exposure Sparsity Model

In this paper, we propose a novel scheme for synthesizing a bracketed exposure sequence from a small set of LDR images, or in the extreme case directly from a single image. Given a set of k differently exposed LDR images, $\mathcal{C} = \{1, \ldots, k\}$, our goal is to estimate the complement set of bracketed exposures, $\bar{\mathcal{C}}$. In order to achieve this, we employ the *Sparse Representations* framework [19], which states that features extracted from images captured under specific exposure conditions, can be represented as a sparse linear combination of elements from an over-complete dictionary that was constructed by training images captured under the same exposure conditions.

Formally, for each input LDR image, patches $\mathbf{s}_C \in \mathbb{R}^m$ are first extracted from the image and mapped to compact features through a non-linear operator $\mathcal{F}$. Subsequently, we seek to identify the sparse code vector $\mathbf{w} \in \mathbb{R}^n$, with respect to the corresponding dictionary matrix, $\mathbf{D}_C \in \mathbb{R}^{m \times n}$, with $m \ll n$, generated by features extracted from image patches with similar exposure conditions, according to: $\mathcal{F}(\mathbf{s}_C) = \mathbf{D}_C \mathbf{w}$.

Recovery of the sparse coding vector $\mathbf{w}$ is achieved by solving:

$$\min_{\mathbf{w}} ||\mathbf{w}||_0 \text{ subject to } ||\mathcal{F}(\mathbf{s}_C) - \mathbf{D}_C \mathbf{w}||_2^2 < \epsilon \qquad (1)$$

where ϵ stands for the acceptable approximation error which is related to the added noise, and solved by a greedy strategy such as the Orthogonal Matching Pursuit algorithm [48]. Alternatively, one can replace the non-zero counting ℓ_0 pseudo norm with the convex ℓ_1-norm: $||\mathbf{w}||_1 = \sum_i |\mathbf{w}_i|$, and solve the corresponding problem given by:

$$\min_{\mathbf{w}} ||\mathcal{F}(\mathbf{s}_C) - \mathbf{D}_C \mathbf{w}||_2^2 + \lambda ||\mathbf{w}||_1 \qquad (2)$$

where λ is a regularization parameter, a formulation known as LASSO [49], [50].

The joint training of the differently exposed dictionaries guarantees that the sparse coding vector is the same among all the representations. Consequently, given the optimal sparse code $\mathbf{w}^\star$, we recover the differently illuminated image patch by projecting it to the proper dictionary, encoding image features corresponding to the jth EV, according to $\mathbf{s}_j = \mathcal{F}^{-1}\{\mathbf{D}_j \mathbf{w}^\star\}$. $\mathcal{F}^{-1}$ is a non-linear inverse operator of $\mathcal{F}$ that transforms the feature vectors into their corresponding pixel values. In Section III-B we thoroughly discuss the feature learning procedure. The aggregation of all the differently illuminated image patches results in the exposed image $\mathbf{S}_j$.

The two main challenges associated with the proposed bracketed sequence estimation approach are related to the choice of an efficient sparsity measure for the sparse vector $\mathbf{w}$ and the proper construction of the dictionary matrices $\mathbf{D}_C$ in order to sparsify the input data. The following subsection describes an effective scheme for compact dictionary learning.

A. Joint Dictionary Construction

Consider a set $\mathbf{S}_k$, of differently exposed training images, namely, normally exposed, under-exposed, and over-exposed. We assume that these images are realized by the same statistical process under different exposure conditions, and as such, they have approximately the same sparse representation with respect to their corresponding dictionaries, $\mathbf{D}_k$. A straightforward strategy to create these dictionaries is to randomly sample multiple registered image patches extracted from all the differently exposed training images and use this random selection as the sparsifying dictionary [51]. However, such a strategy does not guarantee that the same sparse code can be utilized among all the representations. To overcome this limitation, we propose learning a compact dictionary from the differently exposed patches.

Given a large set of training features extracted by the operator $\mathcal{F}$ from multiple image patch pairs $\mathbf{S}_k, k \in \{-6, -4, -2, 0, 2, 4\}$, corresponding to the differently exposed image patches, our goal is to learn a joint feature dictionary $\mathbf{D}_f$, taking into account all the different exposure scenarios. Consequently, we formulate the joint dictionary learning problem as

$$\min_{\mathbf{D}_j, \mathbf{X}} ||\mathbf{P} - \mathbf{D}_f \mathbf{X}||_2^2 + \lambda ||\mathbf{X}||_1, \text{ s.t. } ||\mathbf{D}_f(:, i)||_2^2 \leq 1 \qquad (3)$$

where $\mathbf{D}_f \in \mathbb{R}^{n \times d}$, n denotes the concatenated number of features for the multiple exposure scenarios, d is the number of dictionary atoms, and $\mathbf{P} = \mathcal{F}\left[\mathbf{S}_f\right]$ corresponds to the set of features extracted from all the differently exposed training image patches. The aforementioned problem can be efficiently solved via the K-SVD dictionary learning algorithm [52], [53], alternating between two stages, namely, the sparse coding and the dictionary update.

The joint dictionary generated by (3) encodes multiple differently exposure features. During the testing phase, given a set of differently exposed frames $\mathbf{P}_\mathcal{C}$, we extract the sparse code vector $\mathbf{w}$ with respect to a truncated version of dictionary $\mathbf{D}_C \in \mathbb{R}^{m \times n}$, containing only the features corresponding to the particular exposures and we minimize

$$\min_{\mathbf{w}} ||\mathbf{P}_\mathcal{C} - \mathbf{D}_\mathcal{C} \mathbf{w}||_2^2 + \lambda ||\mathbf{w}||_1. \qquad (4)$$

The sparse coefficients $\mathbf{w} \in \mathbb{R}^n$ are directly projected to the complement dictionary, $\mathbf{D}_{\bar{\mathcal{C}}}$, in order to synthesize the features associated to each complementary exposure through

$$\hat{\mathbf{s}}_{\bar{\mathcal{C}}} = \mathcal{F}^{-1}(\mathbf{D}_{\bar{\mathcal{C}}} \mathbf{w}). \qquad (5)$$

To identify the appropriate exposure where the input mapping will take place, a simple histogram matching process is applied which has shown very good performance.

B. Feature Learning

Instead of relying on raw pixel values, we propose the application of feature learning on the input image patches in order to create descriptive representations that are able to encode the underlying characteristics of image patches and thus facilitate their sparse mapping to the appropriate dictionaries [54]–[56]. Unlike traditional strategies where hand-crafted feature extraction operators, such as SIFT, histograms, and gradients, are applied, our proposed approach considers a non-linear feature learning scheme based on *Sparse Autoencoders* (SAE) neural networks [57], [58]. The formulation considers both as the input and as the output, data $\mathbf{S}$ (image patches in our case), and encodes the information through a non-linear function $\sigma : \mathbb{R}^N \rightarrow \mathbb{R}^M$, such that each input vector $\mathbf{s} \in \mathbb{R}^N$ is mapped to a new feature space via M hidden units.

Formally, a single layer network consists of the input layer units $\mathbf{s} \in \mathbb{R}^N$, the hidden layer unit, $\mathbf{h} \in \mathbb{R}^M$, and the output units $\hat{\mathbf{s}} \in \mathbb{R}^N$, which for the case of SAE are set to be equal to the input units. The objective is to learn a set of weights $\mathbf{W} \in \mathbb{R}^{M \times N}$, along with an associated encoding bias $\mathbf{b} \in \mathbb{R}^M$, in order to generate compact and descriptive features, $\mathbf{h} = \sigma(\mathbf{W}\mathbf{s} + \mathbf{b}_1)$, able to accurately reconstruct the input patch $\mathbf{s}$. The function σ is usually selected to be the logistic sigmoid function, defined as: $\sigma(z) = \frac{1}{1+e^{-z}}$. Decoding of $\mathbf{h}$ is performed using the separate weight matrix $\mathbf{V} \in \mathbb{R}^{N \times M}$, that connects the hidden layer with the output units as $\hat{\mathbf{s}} = \sigma(\mathbf{V}\mathbf{h} + \mathbf{b}_2)$, where $\mathbf{b}_2$ stands for the decoding bias. Following standard approaches [59]–[61], we consider tied weights such that $\mathbf{W} = \mathbf{V}$.

The features are learned by minimizing the error of the loss function:

$$\mathcal{L}(\mathbf{S}, \hat{\mathbf{S}}) = \frac{1}{2} \sum_{j=1}^{N} ||\hat{\mathbf{s}}_j - \mathbf{s}_j||_2^2, \tag{6}$$

where $\mathbf{S}$ and $\hat{\mathbf{S}}$ correspond to the input and the output data, respectively. In order to restrict the average activation to a small desired value, the sparsity constraint is imposed by introducing a Kullback-Leibler divergence regularization term such that:

$$\mathcal{L}(\mathbf{S}, \hat{\mathbf{S}}) + \beta \sum_{j=1}^{M} \mathrm{KL}(\rho || \rho_j), \tag{7}$$

where β is a sparsity regularization parameter, M is the number of features, ρ is the average activation of $\mathbf{h}$, ρ_j is the average activation of the $\mathbf{h}_j$-th vector over the input N-data. Consequently, the network learns weights such that only a few hidden nodes are activated by a given input.

While single layer SAE architectures can learn a wide range of features, deep architectures are able to capture significantly more complex data structures. In this work, we consider deep learning architectures of *Stacked Sparse Autoencoders* (S-SAE) that are constructed by concatenating SAEs such that the outputs of each SAE hidden layer are directly introduced as the input of the successive SAE [62], following an efficient greedy training strategy [63]. This progressive feature encoding operation was shown to be able to encode highly abstract features which have been very successful in multiple classification problems, ranging from cancer detection to hyperspectral image analysis [64]–[66]. However, only a handful of approaches has been reported where deep feature learning architectures are employed in image enhancement [67], [68].

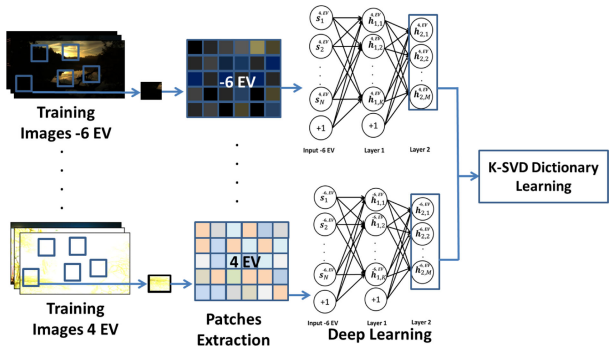


Fig. 2. The S-SAE feature and dictionary learning processes during training: The key objective of S-SAE is to find the most representative features encoded in the hidden layers, so that the output and the input layers are approximately equal. The activations of most deep hidden units are used as features that are introduced to K-SVD for dictionary learning.

In our implementation, we learn exposure-specific two-layer S-SAE networks from sampled patches of differently exposed images, and we train a joint dictionary, properly combining all the different exposure conditions. During the testing phase, we encode each input patch into a two-layer S-SAE feature vector, and represent it with respect to a dictionary matrix generated by features learned from correspondent training images. The extracted sparse code vector is directly mapped into the differently illuminated/exposed dictionaries in order to identify the features corresponding to the different illumination/exposure conditions. Finally, a decoding transformation is applied, in order to transform the feature vector into its pixel values.

C. Post Processing

To enforce compatibility between adjacent patches, we process the input image patches starting from the upper-left corner with a small overlapping factor in each direction. Subsequently, the reconstructed image appears with a slight blurring effect. In order to overcome this anomaly, we perform a back-projection technique, motivated by Yang *et al.* [18]. Specifically, we project the reconstructed scene $\mathbf{S}_0$ at the solution space $\mathbf{Y} = a\mathbf{H}\mathbf{S}$, where $\mathbf{Y}$ stands for the differently illuminated and blurred version of the input frame $\mathbf{S}$, $\mathbf{H}$ denotes the blurring operator which is a low-pass Gaussian filter in our case, and a is a small parameter that uniformly changes the illumination of the input frame. The parameter a was set manually after a cross validation process. The back-projection procedure is summarized as

$$\mathbf{S}^* = \arg\min_{\mathbf{S}} ||aH\mathbf{S} - \mathbf{Y}||_2 + c||\mathbf{S} - \mathbf{S}_0||_2, \tag{8}$$

(a) *Ground Truth -4 EV* (b) *Ground Truth 0 EV* (c) *Ground Truth 2 EV* (d) *Ground Truth 4 EV*

(e) *Rec -4 EV (0.81)* (f) *Rec. 0 EV (0.97)* (g) *Rec. 2 EV (0.84)* (h) *Rec. 4 EV (0.71)*

Fig. 3. Bistro Dataset. Top row: Ground truth images. Bottom row: Reconstructed images from a single -2 EV input image. The proposed reconstructions achieve high similarity with the ground truth bracketed frames, especially when reconstructing in a similar EV setting. The corresponding SSIM are shown in parentheses.

(a) *Ground Truth -2 EV* (b) *Ground Truth 0 EV* (c) *Proposed SAE, 0 EV, (0.97)* (d) *Proposed with Histogram Features, 0 EV, (0.87)*

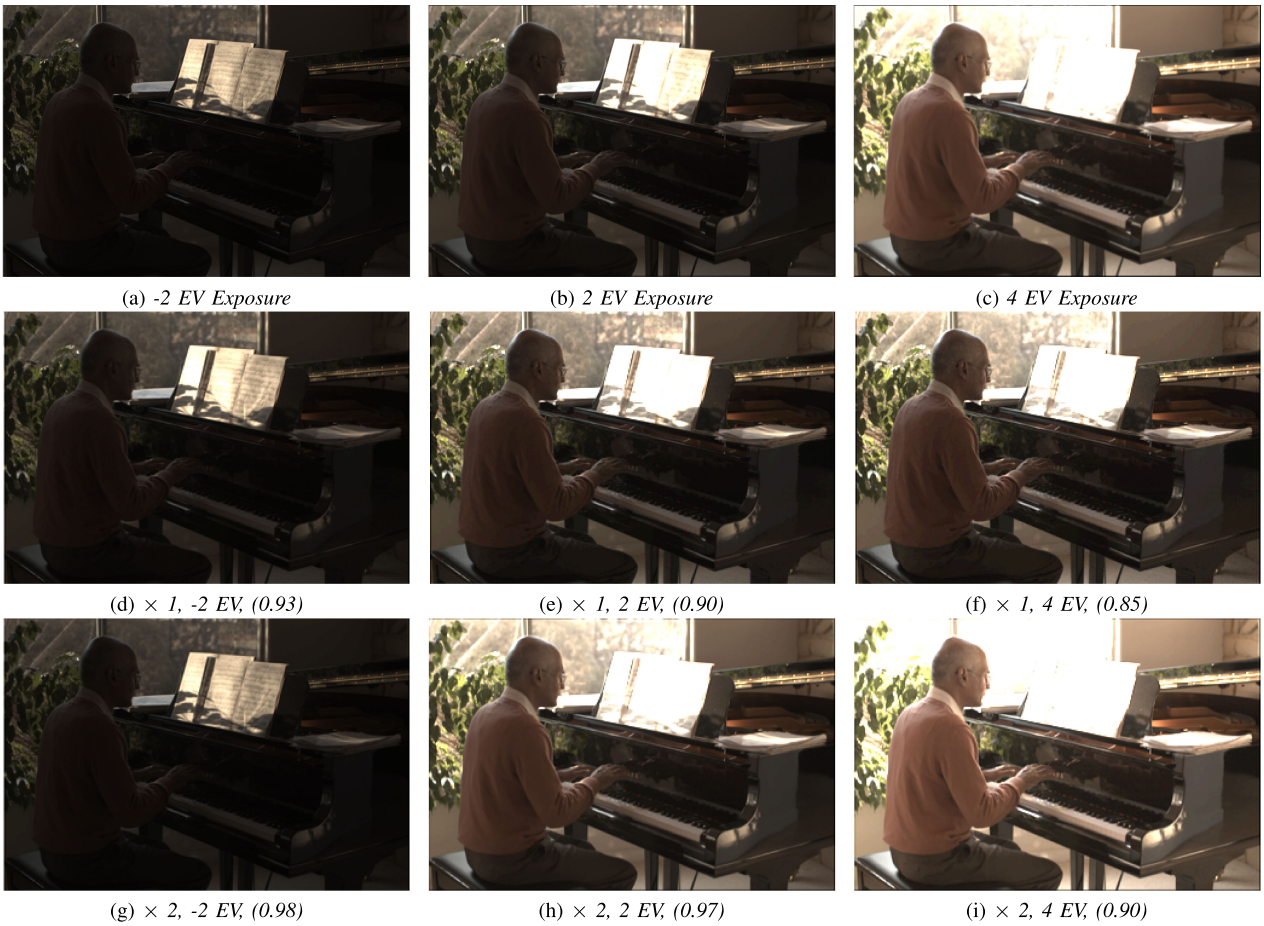
(e) *Ground Truth -2 EV* (f) *Ground Truth 2 EV* (g) *Proposed SAE, 2 EV, (0.84)* (h) *Proposed with Histogram Features, 2 EV, (0.67)*

Fig. 4. Bistro Dataset: Comparison between the SAE feature architecture and the hand-crafted histogram extraction feature scenario.

(a) *Ground Truth -4 EV* (b) *Ground Truth -2 EV* (c) *Ground Truth 0 EV* (d) *Ground Truth 4 EV*

(e) *Rec. -4 EV (0.84)* (f) *Rec. -2 EV (0.84)* (g) *Rec. 0 EV (0.83)* (h) *Rec. 4 EV (0.79)*

Fig. 5. Piano Dataset. Top row: Ground truth images. Bottom row: Reconstructed images from a single $+2$ EV input image. Our system reconstructs successfully the dark exposure value scenarios, preserving high similarity with the correspondent ground truth frames. The corresponding SSIM are shown in parentheses.

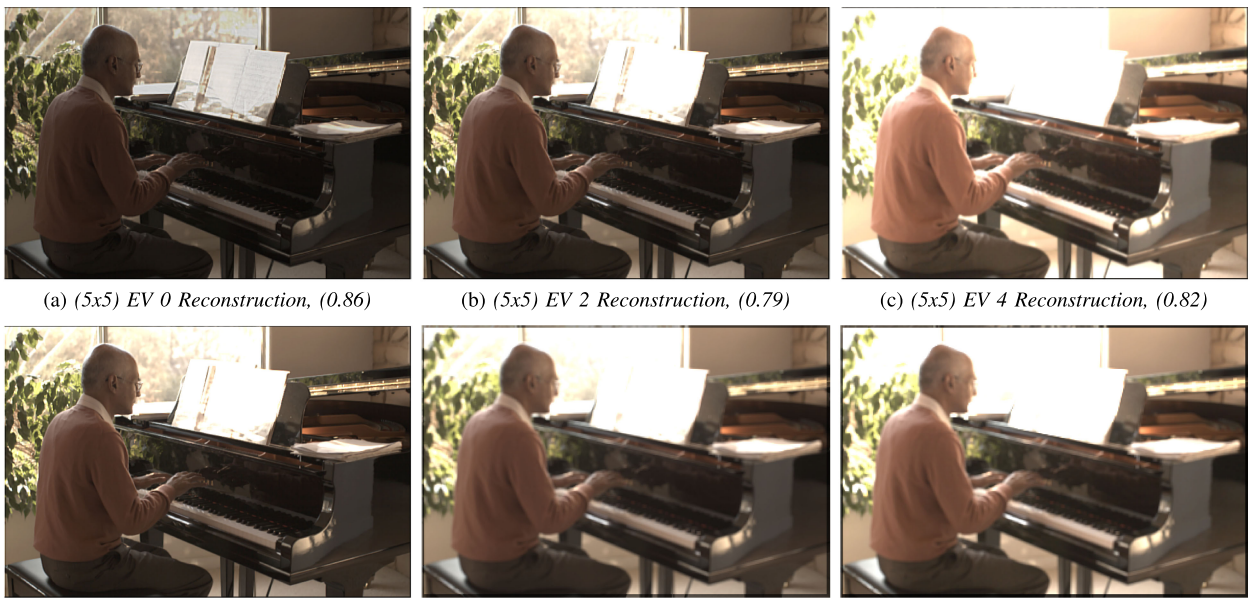


(a) *-2 EV Exposure* (b) *2 EV Exposure* (c) *4 EV Exposure*

(d) $\times$ *1, -2 EV, (0.93)* (e) $\times$ *1, 2 EV, (0.90)* (f) $\times$ *1, 4 EV, (0.85)*

(g) $\times$ *2, -2 EV, (0.98)* (h) $\times$ *2, 2 EV, (0.97)* (i) $\times$ *2, 4 EV, (0.90)*

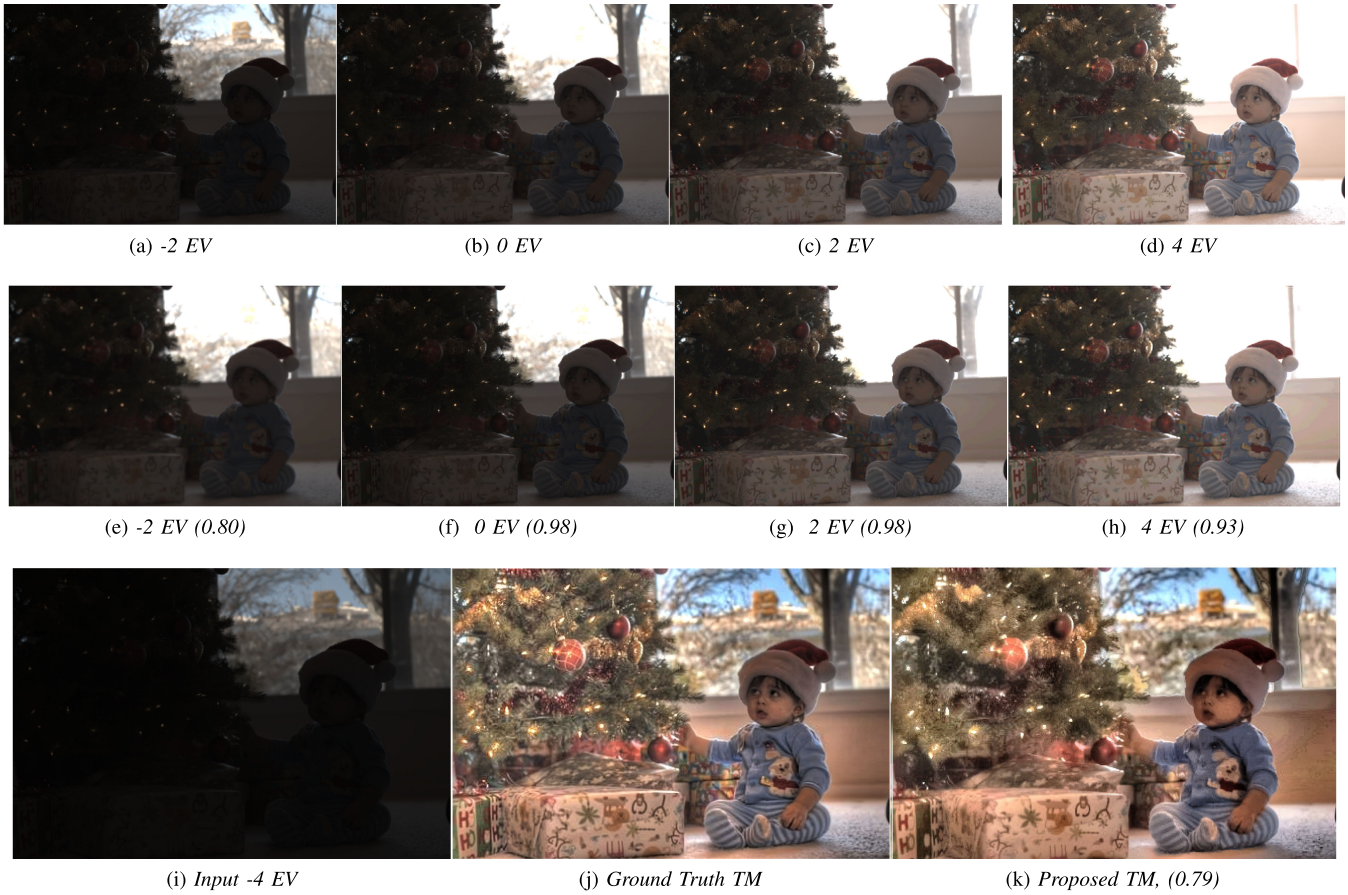
Fig. 6. Double vs. Single Input Scenario. First row: Ground-truth images acquired at -2, 2, and 4 EV, respectively. Second row: Reconstructed frames from a single exposure at -4 EV. Third row: Reconstruction from two exposures at -4 EV and 0 EV. We observe that both input cases approximate with high efficiency the ground truth bracketed sequence.



(a) *(5x5) EV 0 Reconstruction, (0.86)* (b) *(5x5) EV 2 Reconstruction, (0.79)* (c) *(5x5) EV 4 Reconstruction, (0.82)*

(d) *(3x3) EV 0 Reconstruction, (0.92)* (e) *(3x3) EV 2 Reconstruction, (0.85)* (f) *(3x3) EV 4 Reconstruction, (0.90)*

Fig. 7. Piano Dataset: Reconstruction performance using $3\times$ versus 5×5 patch sizes. We observe that both visually and in terms of the evaluation metric, the $3\times$ approach outperforms the scenario where we uses 5×5 overlapping grids.



(a) *-2 EV* (b) *0 EV* (c) *2 EV* (d) *4 EV*

(e) *-2 EV (0.80)* (f) *0 EV (0.98)* (g) *2 EV (0.98)* (h) *4 EV (0.93)*

(i) *Input -4 EV* (j) *Ground Truth TM* (k) *Proposed TM, (0.79)*

Fig. 8. Christmas Baby Dataset. First row: differently exposed ground-truth images. Second row: synthesized sequence from a snapshot image at -4 EV exposure. Third row: Tone-mapped HDR images using a ground-truth (middle) and a synthetic bracketed sequence (right).

which can be efficiently solved using any gradient descent technique.

D. Bracketed Sequence Blending

The final stage of our system involves the blending of the synthesized bracketed sequence into an enhanced LDR version of the imaged scene. We utilize Mertens *et al.* innovative exposure fusion [10] technique, producing directly an enhanced 8-bit result, bypassing the intermediate stage of HDR generation. Exposure fusion utilizes the bracketed input sequence and produces an enhanced LDR image by considering only the most significant parts of the bracketed sequence. Specifically, for each input LDR image $\mathbf{S}_k$, with k EV, exposure fusion estimates a scalar-valued weight map, $\mathbf{W}_k$, combining three measures: contrast, saturation, and well-exposedness. The fused result, $\mathbf{F}$, is obtained by merging the weight map with the input LDR sequence, composed of K images, as:

$$\mathbf{F} = \sum_{i=1}^{K} \mathbf{W}_k \cdot \mathbf{S}_k \tag{9}$$

Using Burt's and Adelson's [69] multi-resolution blending approach, the bracketed sequence images $\mathbf{S}_k$ are decomposed into a *Laplacian* pyramid, while their corresponding weight maps $\mathbf{W}_k$ into a *Gaussian* pyramid. Due to the proposed synthesis approach, the individual input images are aligned and ghost-free, so that the fused image $\mathbf{F}$ is obtained by appropriately collapsing the pyramids.

IV. Experimental Results and Applications

In this Section, we compare our bracketed results against the ground truth sequence. The recovered set of sequential exposures synthesizes the HDR version of the scene. We compare our HDR images against the HDR results produced from selected frames of the ground truth bracketed sequences, in terms of the *Peak Signal to Noise Ratio* (PSNR) (dB) metric defined as

$$\text{PSNR}(\mathbf{x}, \mathbf{y}) = 20 \log_{10} \frac{D}{\sqrt{MSE(\mathbf{x}, \mathbf{y})}}, \tag{10}$$

where $\text{MSE}(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2$, $\mathbf{x}$ and $\mathbf{y}$ denote the pixel values in the reference and the recovered images, while $D = \frac{\mathbf{S}_{max}}{\mathbf{S}_{min}}$ is the dynamic range of the input image, $\mathbf{S}$. Additionally, each enhanced (fused) 8-bit LDR result is compared to other state-of-the-art enhancement algorithms in terms of the

Structular Similarity Index Metric [70] (SSIM), a psychophysically modelled error metric defined as

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c1)\cdot(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)\cdot(\sigma_x^2 + \sigma_y^2 + c_2)}, \qquad (11)$$

where μ and σ stand for the mean value and the standard deviation, respectively.

A. LDR Sequence Synthesis

For the joint dictionary construction, we employ multiple registered patches from the differently exposed bracketed training sequence. The best performance is achieved by selecting a 3×3 patch size and training a two-layer S-SAE network for each exposure. The number of hidden units that we used in order to achieve a fair trade-off between performance and computational cost is 25 for the first layer and 16 for the second one. The resulting dictionaries are trained from 20.000 randomly sampled patches from the varying exposure training images, while the number of dictionary atoms was set to 1024 through a validation process.

To evaluate the proposed approach, a series of experiments was conducted, utilizing the cinematic dataset of Froehlich *et al.* [71], [72], and the sequences of C.S. Verma and Mon-Ju [73]. Fig. 3 illustrates the high-performance of our imaging system, applied on the challenging Bistro image in Froehlich's *et al.* dataset, against the accurate bracketed sequence. We observe that the synthesized sequence captures all the critical information that a single exposure is unable to do, in both excessively and poorly illuminated regions. For example, in sub-images (g) and (h), we may resolve the presence of the person in the scene.

Additionally, in order to further justify that the proposed SAE feature learning approach outperforms the hand-crafted feature extraction scenario, we perform the following experiment: we compare the performance of the proposed deep feature learning architecture versus the case where hand-crafted histogram features are extracted from the low dynamic range input image. Specifically, histogram features were extracted from the low dynamic range part. For this purpose, we have constructed coupled pair of dictionaries, corresponding to the different exposure conditions that are composed of both histogram features (i.e. for the mapping part) and pixel values (i.e. for the reconstruction part). Fig. 4 provides a characteristic comparison between the hand-crafted histogram features and the SAE deep feature learning architecture, for the Bistro dataset. Specifically, we illustrate the 0 and 2 EV reconstruction of the Bistro image, from a -2 EV input scene. We observe that both visually and in terms of the similarity index (SSIM), the proposed scheme using the SAE feature architecture outperforms the hand-crafted feature b scenario.

In addition to the aforementioned simulation results, Fig. 5 demonstrates the proposed system's ability to generate the full bracketed sequence from a bright image, captured at $+2$ EV. In this simulation, we highlight the effectiveness of the proposed system when it is applied on a well-illuminated scene. Consequently, the proposed machine learning technique is able to synthesize the bracketed sequence from both positive and negative exposure values, preserving all the important image features. In the specific example, we observe that all the significant visual information, such as the depicted man or the piano paper features, are successfully preserved as evidenced by looking at the ground truth bracketed image frames.

Fig. 6 provides a comparison between the reconstructions for the multiple versus the single input case. In the first scenario, our system considers as input a single shot captured at -4 EV and synthesizes the remaining frames of the bracketed sequence. In the second scenario, the proposed algorithm synthesizes the bracketed sequence from two input images captured at -4 and 0 EV. We observe that in neighbour exposure values our system achieves higher reconstruction quality in comparison with less similar exposure settings. Although the double input case provides better reconstruction quality compared to the single input case, the results recovered from the single input scenario, represent with higher accuracy the bracketed sequence.

Finally, Fig. 7 provides a characteristic example applied on the Piano dataset that examines the reconstruction quality in the scenario where we use larger grid size, i.e. of size 5×5, instead of 3×3, which is the default value for our study. According to the majority of state-of-the-art approaches that confront the problem of image enhancement, such as de-noising, super-resolution, and de-convolution among others, usually extract 3×3 overlapping patches from the input images. One of the main reasons for the selection of the specific patch size is the computational complexity. In the scenario where we choose small patch-size, the algorithm in both training and testing phases corresponds more efficiently in terms of computational cost, compared to the scenario where we select larger patch sizes, such as 5×5, or 7×7. Additionally, the best performance is achieved when we select the 3×3 grid, compared to the 5×5 and the 7×7. This is due to the fact that when we select larger patch sizes, the overlapping factor among adjacent pixels needs to be larger as well. Unfortunately, this phenomenon in the majority of cases leads into blurry reconstructions. However, in our case, when we utilize the 5×5 patch size, with 4 overlapping adjacent pixels, we still achieve high quality reconstructions, as it is demonstrated on Fig. 7. However, the best performance is achieved when we use 3×3 grid, with 1 overlapping factor between adjacent pixels.

B. HDR Generation and Tone-Mapping

The synthesized bracketed sequence can be directly used to generate a high quality HDR image. For this purpose, we utilize Sen's *et al.* [36] technique, for producing HDR images from a sequence of bracketed exposures. This method considers as input a sequence of coarsely aligned LDR scenes, and produces an HDR image, matching a reference image in the regions where it is sufficiently exposed. In order to compress the dynamic range and display the final HDR result, a tone-mapping procedure is followed, utilizing the commercial software package Photomatix [36].

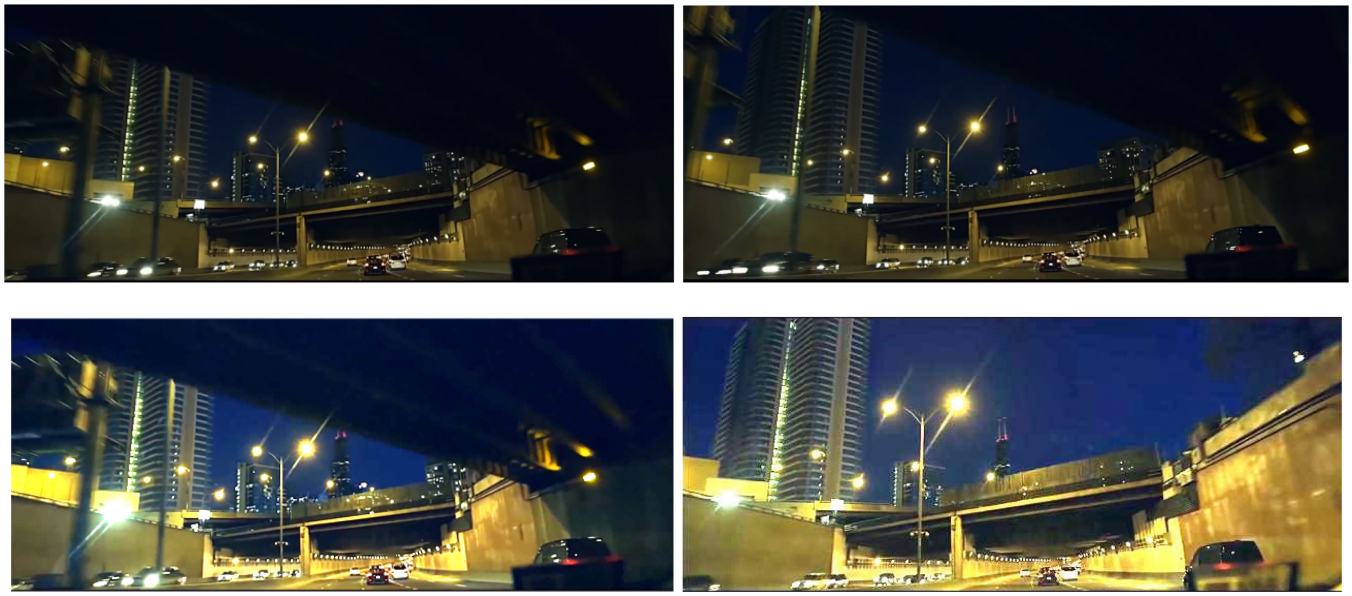


Fig. 9. Enhancement of Low Dynamic Range Video. Top row: Original video frames. Bottom row: Proposed system's fused reconstructed frames. We observe that under real life conditions, our scheme achieves a significant quality improvement when applied on already captured low light video, without amplifying the noise.

TABLE I
ESTIMATION ERROR IN PSNR (DB) FOR THE HDR GENERATION FROM THE TWO-FRAME CASE AND THE FULL SYNTHESIZED SEQUENCE

Method	Ground Truth 2 frames	Bracketed Synthesized frames
Car	20.7	23.4
Bistro	24.3	31.8
Showgirl	28.3	30.7
Memorial	37.5	48.5
Baby	34.3	44.9
Piano man	34.6	34.4

Fig. 8 presents the ground truth and synthesized bracketed sequence of the *Christmas baby* scene. Utilizing the synthesized bracketed sequence, we created a high quality HDR image. The proposed tone-mapped result is compared against the tone-mapped result created by Sen's *et.al* algorithm. We observe that the tone-mapped images achieve highly similar results to the ground-truth case, maintaining detailed features while revealing the structure in the originally saturated regions (outside the window).

To justify the need for acquiring an extended LDR sequence, we compare the performance of HDR generation from either a set of two differently exposed images, or an augmented set containing additional differently exposed synthesized images. We consider two input images at -4 and 0 EV, along with the proposed synthesized sequence, and directly utilize them as input to Sen's *et al.* algorithm for the creation of the HDR image. The quantitative results are provided in Table I. These results suggest that synthesizing the extended bracketed sequence can have a dramatic effect on the quality of the produced HDR image, achieving a gain as high as 10 dBs.

C. Dynamic Range Enhancement of Video

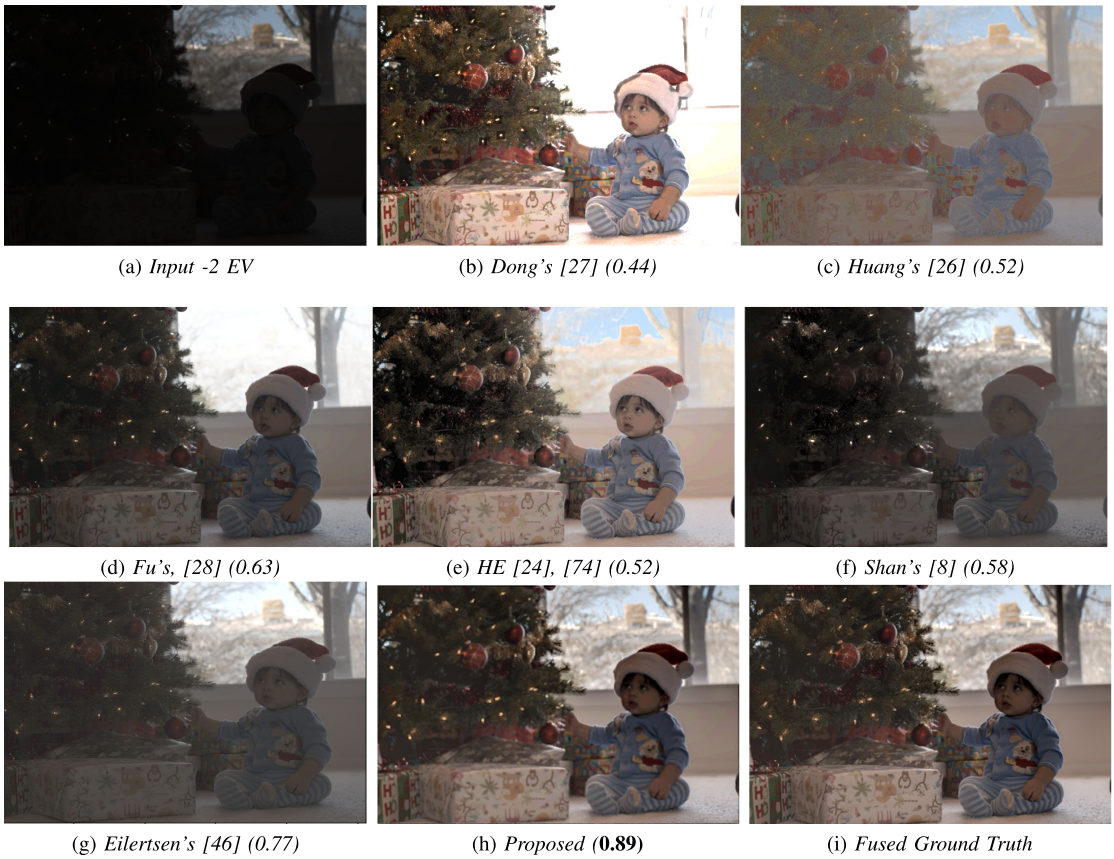
Concerning the case of video sequences enhancement, the dynamic nature of the scenes does not allow the acquisition of multiple exposures and thus the proposed method is a highly convenient option. For the case when multiple exposures are combined, we assume that the scene is static and thus no ghosting is present. Consequently, the proposed scheme is able to produce enhanced and high-quality dynamic range videos from conventional LDR video sequences. Considering an input LDR frame $\mathbf{I}_k$, we produce five differently exposed versions of the scene, $\mathbf{I}_{EV_i}, i = 1, ..., 5$, and blend them into an enhanced LDR result using an adaptation of the exposure fusion algorithm [10]. The stack of the varying exposure synthetic sequence is used to compute a scalar-valued weight map for each image, with respect to the $Well - Exposedness$ metric, defined as:

$$\mathbf{W}_{ij} = \exp\left(-\frac{(i-0.5)^2}{2\sigma^2}\right), \tag{12}$$

where $\sigma = 0.2$. The fused LDR video frame corresponds to the weighted blend of the bracketed sequence. Fig. 9 illustrates two characteristic frames obtained via the proposed scheme. We observe that the our system produces high quality moving scenes, without introducing motion artifacts. The reconstruction quality could be further improved by considering the temporal consistency between adjacent video frames.

D. Comparison With State-of-the-Art

In the last Section, we compare the proposed scheme's performance against several enhancement techniques, including the



(a) *Input -2 EV* (b) *Dong's [27] (0.44)* (c) *Huang's [26] (0.52)*

(d) *Fu's, [28] (0.63)* (e) *HE [24], [74] (0.52)* (f) *Shan's [8] (0.58)*

(g) *Eilertsen's [46] (0.77)* (h) *Proposed* (**0.89**) (i) *Fused Ground Truth*

Fig. 10. Christmas Baby Sequence: Dong's and Fu's techniques saturate the region outside the window, while Huang's and HE produce extremely uniformly bright results. Shan's approach preserves the details outside the window, without producing a sufficiently enhanced result, while our scheme properly enhances the dark regions and preserves details.

traditional Histogram Equalization (HE) [24], [74], serving as the baseline method, Dong's *et al.* [27] enhancement algorithm of low-light video frames, Fu's *et al.* Color Estimation Model (CEM) [28], Shan's *et al.* globally optimized linear windowed tone mapping approach [8], and Huang's *et al.* AGCWD contrast enhancement technique [26]. In order to achieve a fair comparison with the competing methods, we optimized their parameters to achieve the best performance. Concerning Shan's *et al.* algorithm [8], we modified their default parameter values for the *ordinary image enhancement* application as follows $\beta_1 = 0.5, \beta_2 = 0.4, \beta_3 = 0.08$.

Fig. 10 evaluates the performance of the various methods on the Christmas baby test scene, extracted from Sen's *et al.* [36] dataset. Although Huang's [26], Shan's [8] and Eilertsen's [46] techniques accurately estimate the background window details, they also lead to color leakage effects. Our method produces an accurate approximation of the ground truth, revealing significant details in about all regions of the scene. Additionally, Fig. 11, illustrates the fused sequence results applied on the Car scene taken from Froehlich's *et al.* dataset [71], [72]. We observe that state-of-the-art approaches introduce color and zipping artifacts, occasional lead to saturation effects, and fail in the recovery of background information. On the other hand, our system reconstructs a high quality image, preserving the necessary information details concerning both the foreground car and the background. For a complete quantitative set of results, SSIM and PSNR values are comprehensively provided in Table II for all methods, clearly demonstrating the improved performance of the proposed approach.

E. Sensitivity Analysis

In this Section, we provide the sensitivity analysis of the proposed multiple exposure sparsity model. Specifically, Fig. 12 reports the PSNR values for the reconstruction of Bistro $+2$EV scene from a -2EV low-dynamic range image, as a function of the number of training example for the joint dictionary learning process. Results indicate that the performance of the proposed learning scheme monotonically increases as a function of the number of training examples, as expected. However, after a certain number of training examples, the PSNR value reaches a stable plateau, achieving its highest PSNR value of 37.7 dB. Consequently, for the specific application, using a larger number than 10^5 training examples in the joint dictionary learning process offers marginal improvement to the reconstruction quality.

(a) *Input -2 EV* (b) *Dong's [27] (0.68)* (c) *Huang's [26] (0.61)*

(d) *Fu's [28] (0.48)* (e) *Shan's [8] (0.62)* (f) *HE [24], [74] (0.70)*

(g) *Eilertsen's [46] (0.72)* (h) *Proposed* (**0.83**) (i) *Fused Ground Truth*

Fig. 11. Car Sequence: Comparison with other state-of-the-art approaches. Unlike the other methods, our system reconstructs successfully both the car's wheel and the region below the car, preserving the background details and providing a smooth, well-enhanced visual result.

TABLE II
QUANTITATIVE PERFORMANCE EVALUATION OF THE COMPETING METHODS IN TERMS OF SSIM AND PSNR METRICS

Methods	Dongs's [27]		Huang's [26]		CEM [28]		Shan's [8]		HE [24], [74]		Eilertsen's [46]		**Proposed**	
Image	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Car	11.9	0.68	15.1	0.61	11.5	0.48	17.2	0.62	15.9	0.70	18.3	0.72	**20.4**	**0.83**
Bistro	16.1	0.61	15.7	0.44	12.0	0.28	16.8	0.46	18.1	0.60	18.1	0.52	**20.9**	**0.73**
Showgirl	14.9	0.48	16.5	0.64	14.8	0.58	14.9	0.55	18.2	0.79	16.3	0.72	**18.7**	**0.80**
Memorial	17.8	0.62	16.5	0.64	20.5	0.79	15.6	0.66	21.8	0.72	23.4	0.79	**24.6**	**0.86**
Baby	12.3	0.44	17.8	0.52	17.6	0.63	17.9	0.58	17.2	0.52	21.3	0.77	**23.8**	**0.89**
Piano	9.6	0.83	18.3	0.66	16.7	0.79	18.2	0.72	21.0	0.83	19.8	0.77	**21.0**	**0.83**
Average	13.76	0.61	16.65	0.58	15.51	0.59	16.76	0.59	18.7	0.69	19.5	0.71	**21.56**	**0.82**

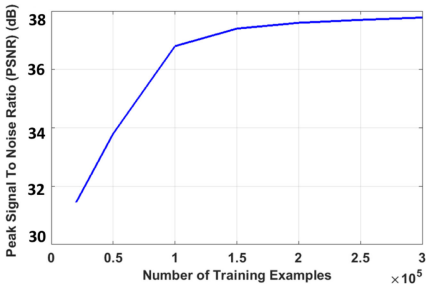


Fig. 12. (2 EV Reconstruction of the Bistro Dataset): In this simulation, we illustrate the PSNR of the proposed algorithm as a function of the number of training examples. We observe that after approximately a fixed number of training examples, the proposed joint sparse-based dictionary learning procedure reaches a stable plateau.

We also empirically validate the convergence of the proposed joint dictionary learning algorithm under different sparsity (i.e. λ parameter) values. Specifically, we illustrate the convergence behavior of both the EV -2 and EV$+2$ dictionaries, i.e., $\mathbf{D}_{EV-2}$, $\mathbf{D}_{EV2}$ in Fig. 13, which depicts the normalized reconstruction errors for the two dictionaries as a function of the number of iterations. Experimental results demonstrate that both dictionaries converge after a small number of iterations. In Fig. 14 we examine the impact of sparsity parameter λ, on the reconstruction quality of the $+2$EV Bistro Scene. We observe that as the value of the parameter sparsity λ increases, the reconstruction quality drops. In all tested cases, the highest PSNR value is achieved for $\lambda = 0.2$.

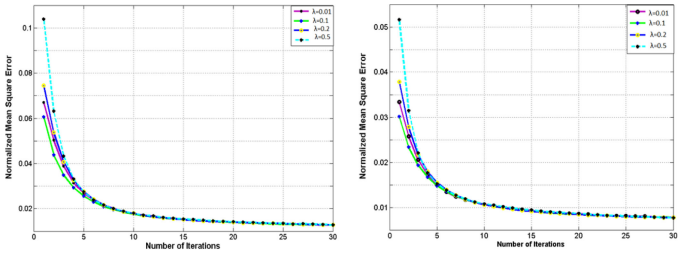


Fig. 13. (From Left to Right): -2EV and $+2$EV Dictionary Convergence Behavior: In this experiment, we illustrate the convergence of -2 and $+2$ EV as a function of the number of iterations. We observe that after approximately 5 iterations our scheme achieves a stable reconstruction error. Additionally, this behavior is observed under variant λ parameter's values, indicating the robustness of our system.

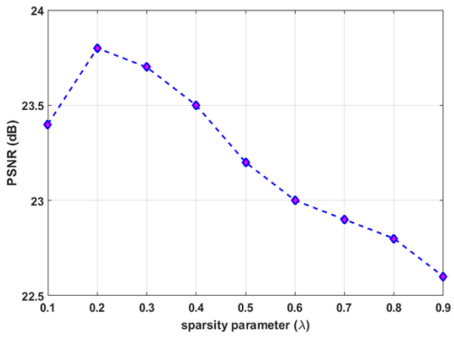


Fig. 14. In this experiment, we demonstrate the impact of the sparsity regularization parameter on the reconstruction performance. We observe, that for $\lambda = 0.2$, our system achieves its optimal performance.

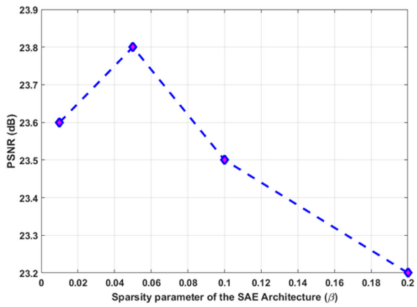


Fig. 15. In this experiment, we demonstrate the impact of the sparsity regularization parameter β of the Sparse Autoencoders Architecture, on the reconstruction performance. We observe, for $\beta = 0.05$, our system achieves its optimal performance.

Moreover, the impact of the sparsity values β, for the Sparse Autoencoders scheme, is examined in Fig. 15, where the number of hidden nodes for the two-layer architecture was 25 and 16, for the first and second layers respectively. For the joint dictionary learning process, the parameter λ was set to its highly achieved value of 0.2, while the joint dictionaries are composed of 1024 atoms. We observe, that after a specific number of β parameter's values, i.e. 0.05, the reconstruction quality marginally drops. However, in terms of numerical comparison the reconstruction

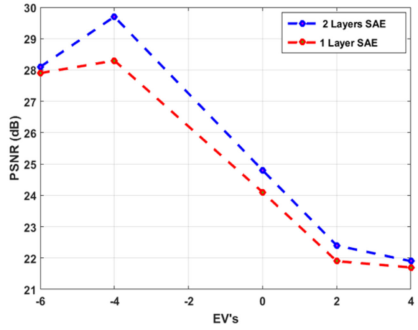


Fig. 16. In this experiment, we demonstrate the reconstruction performance of the Bistro Bracketed Sequence, from a -2EV input image, under two different SAE architectures: (i) two-hidden layers, composed of 25 and 16 hidden nodes, and (ii) a single-layer composed of 25 hidden nodes. We observe, that the two-layer architecture outperforms the single one in all reconstructed EV's.

quality is deteriorated by only 0.6 db, still validating a high quality recovery of the grouth truth $+2$EV Bistro scene.

Finally, in Fig. 16 we illustrate the reconstruction performance of the Bistro Bracketed Sequence from a -2EV input scene, in the scenario where we use 2 hidden layers composed of 25, and 16 hidden nodes, versus the case where we use a single-layer architecture composed of 25 units. Experimental results demonstrate that using 2-hidden layers outperforms the case where we use a single-layer, justifying the demand of deeper architectures. Additionally, we should note that the same sensitivity analysis is also implemented for all experimented datasets, including the video-sequences.

Additionally, to further validate that the proposed scheme overcomes standard limitations of imaging systems, such as ghosting and noise effects, we compare ourselves versus the state-of-the-art BM3D [75] denoising algorithm. Specifically, Fig. 17 illustrates the application of the state-of-the-art BM3D [75] denoising algorithm on a -4 EV input image. In order to modify the exposure of the BM3D denoised -4EV scene, we used Photoshop's built-in re-adjustments. Specifically, Photoshop software is able to modify heuristically the exposure value of any image by applying modifications on the Camera Response Function (CRF) of the input scene.

The response function of a scene is usually not provided by the imaging instrument, and additionally is far beyond the standard Gamma curve. Specifically, the response function of an image that is captured by any imaging system corresponds to the measured intensity values which are related to scene radiance. Traditionally, the CRF can be determined by establishing a mapping of intensity values between images taken with different exposures [76]. This mapping is called intensity mapping function. In our scenario, we observe that the results from the denoised scene are extremely blurry compared to the proposed high quality results, while the proposed approach, capitalizing on the sparse representations framework, implicitly performs denoising.

Finally, to illustrate the benefits over simple exposure adjustments, we extracted the Camera Response Function (CRF) from

(a) *Proposed, -4EV, Input* (b) *Proposed, 2EV, 18.9dB* (c) *Proposed, 4EV, 20dB*

(d) *BM3D [75], -4EV* (e) *BM3D [75], 2EV, 12dB* (f) *BM3D [75], 4EV, 9.9dB*

Fig. 17. First row: Input image and proposed reconstructions. Second row: Denoised -4 EV result by the BM3D [75] algorithm, along with the exposure adjustments using Photoshop, and PSNR metric refers to the different exposure scenarios, e.g. $(-4, 2)$ EV, denotes two input images at -4 and 2 EV.

(a) *CRF, -2EV, 20.5dB* (b) *CRF, 0EV, 13.5dB* (c) *CRF, 2EV, 5.4dB* (d) *CRF, 4EV, 2.1dB*

Fig. 18. Exposure Adjustments of the Camera Response Function (CRF) starting from -4 EV.

a bracketed training sequence, and applied the method in [1] to re-expose the input image on different exposures. Fig. 18 demonstrates and evaluates in terms of the PSNR, the impact of the CRF applied on an input -4 EV. Compared to the results from our deep learning formulation, we may notice that re-exposed images using Camera Response Function (CRF) adjustments are inferior in terms of quality and visual perception, justifying the need for learning-based approach.

V. Conclusions

In this work we proposed an innovative scheme for dynamic range enhancement of static imagery and dynamic video, from a limited sequence of low dynamic range images, or in the extreme case, directly from a single shot. The synthesized bracketed sequence is employed for the production of a high quality HDR image, or it can be directly fused into an enhanced LDR image. The proposed method produces high quality imagery, revealing valuable details, without intensifying noise effects in under-exposed areas, or saturating the overexposed regions. Furthermore, instead of processing raw pixel values, we introduced an intelligent feature learning scheme, based on the Stacked Sparse Autoencoders framework. Quantitative and qualitative observations indicate that the proposed approach outperforms state-of-the-art techniques, while offering novel capabilities for video sequence enhancement.

References

[1] Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Human visual system-based saliency detection for high dynamic range content," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 549–562, Apr. 2016.

[2] E. François *et al.*, "High dynamic range and wide color gamut video coding in HEVC: Status and potential future enhancements," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 63–75, Jan. 2016.

[3] Y.-T. Lin, C.-M. Wang, W.-S. Chen, F.-P. Lin, and W. Lin, "A novel data hiding algorithm for high dynamic range images," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 196–211, Jan. 2017.

[4] F. Kou *et al.*, "Intelligent detail enhancement for exposure fusion," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 484–495, Feb. 2018.

[5] I. R. Khan, S. Rahardja, M. M. Khan, M. M. Movania, and F. Abed, "A tone-mapping technique based on histogram using a sensitivity model of the human visual system," *IEEE Trans. Ind. Electron.*, vol. 65, no. 4, pp. 3469–3479, Apr. 2018.

[6] Z. Wei, C. Wen, and Z. Li, "Local inverse tone mapping for scalable high dynamic range image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 550–555, Feb. 2018.

[7] A. Rana, G. Valenzise, and F. Dufaux, "Learning-based tone mapping operator for efficient image matching," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 256–268, Jan. 2019.

[8] Q. Shan, J. Jia, and M. S. Brown, "Globally optimized linear windowed tone mapping," *IEEE Trans. Visualization Comput. Graph.*, vol. 16, no. 4, pp. 663–675, Jul./Aug. 2010.

[9] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 640–648, 2005.

[10] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion," in *Proc. 15th Pacific Conf. Comput. Graph. Appl.*, 2007, pp. 382–390.

[11] Z. Su, K. Zeng, L. Liu, B. Li, and X. Luo, "Corruptive artifacts suppression for example-based color transfer," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 988–999, Jun. 2014.

[12] H. Xu, G. Zhai, X. Wu, and X. Yang, "Generalized equalization model for image enhancement," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 68–82, Jan. 2014.

[13] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proc. 24th Annu. Conf. Comput. Graph. Interactive Techn.*, 2008, pp. 369–378.

[14] F.-L. Zhang *et al.*, "Detecting and removing visual distractors for video aesthetic enhancement," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 1987–1999, Aug. 2018.

[15] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content," 2018, *arXiv:1803.04815*.

[16] F. Banterle *et al.*, "High dynamic range imaging and low dynamic range expansion for generating HDR content," *Comput. Graph. Forum*, vol. 28, pp. 2343–2367, 2009.

[17] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, 1st ed. New York, NY, USA: Springer, 2010.

[18] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[19] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[20] H. Zhang, J. Yang, Y. Zhang, and T. S. Huang, "Sparse representation based blind image deblurring," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2011, pp. 1–6.

[21] D.-A. Huang, L.-W. Kang, Y.-C. F. Wang, and C.-W. Lin, "Self-learning based image decomposition with applications to single image denoising," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 83–93, Jan. 2014.

[22] L.-W. Kang, C.-C. Hsu, B. Zhuang, C.-W. Lin, and C.-H. Yeh, "Learning-based joint super-resolution and deblocking for a highly compressed image," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 921–934, Jul. 2015.

[23] W. Yang *et al.*, "Consistent coding scheme for single-image super-resolution via independent dictionaries," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 313–325, Mar. 2016.

[24] M. Kaur, J. Kaur, and J. Kaur, "Survey of contrast enhancement techniques based on histogram equalization," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 7, 2011.

[25] S. Wang *et al.*, "Guided image contrast enhancement based on retrieved images in cloud," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 219–232, Feb. 2016.

[26] S.-C. Huang, F.-C. Cheng, and Y.-S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1032–1041, Mar. 2013.

[27] X. Dong *et al.*, "Fast efficient algorithm for enhancement of low lighting video," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2011, pp. 1–6.

[28] H. Fu, H. Ma, and S. Wu, "Night removal by color estimation and sparse representation," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 3656–3659.

[29] A. Yamasaki, H. Takauji, S. Kaneko, T. Kanade, and H. Ohki, "Denighting: Enhancement of nighttime images for a surveillance camera," in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[30] R. Raskar, A. Ilie, and J. Yu, "Image fusion for context enhancement and video surrealism," in *Proc. 3rd Int. Symp. Non-Photorealistic Animation Rendering*, 2005, pp. 85–152.

[31] E. Reinhard *et al.*, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. San Mateo, CA, USA: Morgan Kaufmann, 2010.

[32] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2000, vol. 1, pp. 472–479.

[33] S. K. Nayar and V. Branzoi, "Adaptive dynamic range imaging: Optical control of pixel exposures over space and time," in *Proc. IEEE 9th Int. Conf. Comput. Vis.*, 2003, pp. 1168–1175.

[34] C. Aguerrebere, A. Almansa, Y. Gousseau, J. Delon, and P. Muse, "Single shot high dynamic range imaging using piecewise linear estimators," in *Proc. IEEE Int. Conf. Comput. Photography*, 2014, pp. 1–10.

[35] G. Tsagkatakis and P. Tsakalides, "Efficient high dynamic range imaging via matrix completion," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2012, pp. 1–6.

[36] P. Sen *et al.*, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Trans. Graph.*, vol. 31, no. 6, 2012, Art. no. 203.

[37] H. Zhao, B. Shi, C. Fernandez-Cull, S.-K. Yeung, and R. Raskar, "Unbounded high dynamic range photography using a modulo camera," in *Proc. IEEE Int. Conf. Comput. Photography*, Apr. 2015, pp. 1–10.

[38] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system," *ACM Trans. Graph.*, vol. 30, 2011, Art. no. 41.

[39] K. Gu *et al.*, "Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 432–443, Mar. 2016.

[40] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb. 2013.

[41] Y. Song *et al.*, "Naturalness index for a tone-mapped high dynamic range image," *Appl. Opt.*, vol. 55, no. 35, pp. 10084–10091, 2016.

[42] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, Jun. 2017.

[43] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "Blique-tmi: Blind quality evaluator for tone-mapped images based on local and global feature analyses," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 323–335, Feb. 2019.

[44] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.

[45] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[46] G. Eilertsen, J. Kronander, G. Denes, R. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. Graph.*, vol. 36, no. 6, 2017, Art. no. 178.

[47] K. Moriwaki, R. Yoshihashi, R. Kawakami, S. You, and T. Naemura, "Hybrid loss for learning single-image-based HDR reconstruction," 2018, *arXiv:1812.07134*.

[48] J. Tropp *et al.*, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[49] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., Ser. B (Methodological)*, vol. 58, pp. 267–288, 1996.

[50] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *J. Roy. Stat. Soc., Ser. B (Stat. Method.)*, vol. 73, no. 3, pp. 273–282, 2011.

[51] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. no. 46.

[52] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[53] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD and its non-negative variant for dictionary design," in *Proc. Opt. Photon. Int. Soc. Opt. Photon.*, 2005, pp. 591411–591411.

[54] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[55] G. Papandreou, I. Kokkinos, and P.-A. Savalle, "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 390–399.

[56] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[57] A. Ng, "Sparse autoencoder," *CS294A Lecture Notes*, vol. 72, pp. 1–19, 2011.

[58] J. Ngiam *et al.*, "On optimization methods for deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 265–272.

[59] E. Hosseini-Asl, J. M. Zurada, and O. Nasraoui, "Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2486–2498, Dec. 2016.

[60] A. Makhzani and B. Frey, "K-sparse autoencoders," 2013.
[61] A. Lemme, R. Felix Reinhart, and J. J. Steil, "Online learning and generalization of parts-based image representations by non-negative sparse autoencoders," *Neural Netw.*, vol. 33, pp. 194–203, 2012.
[62] A. Ng, J. Ngiam, C. Yu Foo, M. Yifan, and S. Caroline, "Stanford deep learning tutorial," 2013. [Online]. Available: http://deeplearning.stanford.edu/wiki/index.php/UFLDL_Tutorial
[63] Y. Bengio *et al.*, "Greedy layer-wise training of deep networks," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, vol. 19, pp. 153–160, 2007.
[64] H.-C. Shin, Matthew R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1930–1943, Aug. 2013.
[65] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013.
[66] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.
[67] Z. Cui, H. Chang, S. Shan, B. Zhong, and X. Chen, "Deep network cascade for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 49–64.
[68] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.
[69] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.
[70] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[71] J. Froehlich *et al.*, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," in *Proc. IS&T/SPIE Electron. Imag. Int. Soc. Opt. Photon.*, 2014, Art. no. 90230X.
[72] J. Froehlich *et al.*, "HdmHDR-2014 project," 2014. [Online]. Available: https://hdr-2014.hdm-stuttgart.de/
[73] C. S. Verma and Mon-Ju, "LDR dataset," [Online]. Available: http://pages.cs.wisc.edu/csverma/CS766_09/HDRI/hdr.html
[74] Y.-T. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE Trans. Consum. Electron.*, vol. 43, no. 1, pp. 1–8, Feb. 1997.
[75] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "BM3D image denoising with shape-adaptive principal component analysis," in *Proc. Signal Process. Adaptive Sparse Struct. Representations*, 2009.
[76] M. D. Grossberg and S. K. Nayar, "Modeling the space of camera response functions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1272–1282, Oct. 2004.

Konstantina Fotiadou is currently working toward the Ph.D. degree in computer science from the Computer Science Department, University of Crete (UoC), Crete, Greece, under the supervision of Prof. P. Tsakalides. She received the M.Sc. degree in computer science from the Computer Science Department, University of Crete, and the B.Sc. degree in applied mathematics from the Department of Applied Mathematics, UoC. Since 2012, she has been with the Signal Processing Laboratory, FORTH-ICS, as a Research Assistant. Her main research interests include image processing and computational photography applications with an emphasis on hyperspectral image enhancement techniques.

Grigorios Tsagkatakis received the Diploma and M.S. degrees in electronics and computer engineering from the Technical University of Crete, Crete, Greece, in 2005 and 2007, respectively, and the Ph.D. degree in imaging science from the Rochester Institute of Technology, New York, NY, USA, in 2011. He is currently working as a Research Associate with the Signal Processing Laboratory, Institute of Computer Science (ICS) of the Foundation for Research and Technology Hellas. His research is focused on topics related to signal processing and machine learning with applications in remote sensing and astrophysics and has been funded by numerous national and European projects. He has coauthored more than 50 peer-reviewed conference papers, journals, and book chapters; has been awarded one U.S. patent; and three best paper awards. He has served on the organizing committee of a number of workshops including Cyber-Physical Systems for Smart Water Networks and the workshop on Computational Intelligence in Remote Sensing and Astrophysics.

Panagiotis Tsakalides received the Diploma degree in electrical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1990, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1995. He is currently a Professor in computer science and the Vice Rector for finance at the University of Crete, Heraklion, Greece, and the Head of the Signal Processing Laboratory, FORTH-ICS. He has extensive experience of transferring research and interacting with the industry. During the last 10 years, he has been the Project Coordinator in seven European Commission and 12 national research and innovation projects totaling more than 5 M C in actual funding for FORTH-ICS and the University of Crete. His research interests include the field of statistical signal processing with an emphasis on non-Gaussian estimation and detection theory, sparse representations, and applications in sensor networks, audio, imaging, and multimedia systems. He has coauthored more than 180 technical publications in these areas.