Underexposed Video Enhancement via Perception-driven Progressive Fusion

Qing Zhang, Yongwei Nie, Ling Zhang, Chunxia Xiao, Member, IEEE

Abstract—Underexposed video enhancement aims at revealing hidden details that are barely noticeable in LDR video frames with noise. Previous work typically relies on a single heuristic tone mapping curve to expand the dynamic range, which inevitably leads to uneven exposure and visual artifacts. In this paper, we present a novel approach for underexposed video enhancement using an efficient perception-driven progressive fusion. For an input underexposed video, we first remap each video frame using a series of tentative tone mapping curves to generate an multi-exposure image sequence that contains different exposed versions of the original video frame. Guided by some visual perception quality measures encoding the desirable exposed appearance, we locate all the best exposed regions from multi-exposure image sequences and then integrate them into a well-exposed video in a temporally consistent manner. Finally, we further perform an effective texture-preserving spatio-temporal filtering on this well-exposed video to obtain a high-quality noise-free result. Experimental results have shown that the enhanced video exhibits uniform exposure, brings out noticeable details, preserves temporal coherence, and avoids visual artifacts. Besides, we demonstrate applications of our approach to a set of problems including video dehazing, video denoising and HDR video reconstruction.

Index Terms—Underexposed video enhancement, visual perception, fusion, filtering

1 INTRODUCTION

With the rapid development of the digital photography technology, the acquisition of digital video is now an easy task. Recording episodes of our daily life in terms of video clips has become a popular lifestyle. However, it is even nontrivial for proficient photographers to capture high-quality videos in low light-level photographic environment. Thus, global or local underexposed video clips are inevitably created along with the daily photography of novice photographers. Due to the low visibility, these underexposed videos usually fail to present visually pleasing browsing. In the field of public safety, the widely used video surveillance systems also produce underexposed nighttime videos, which seriously weaken the system in traffic accidents analysis and crime forensics. To improve the visual appearance of the underexposed videos, video enhancement technique emerges. Currently, video enhancement is still an active research topic in the computer graphics and computer vision communities.

Underexposed video enhancement is a challenging problem. To date, there is no consensus standard for evaluating whether an enhanced video is visually pleasing. Based on human visual perception theory, we tailor three main objectives for our video enhancement technique: the original hidden details should be easy to identify in the enhanced video; the enhanced video should avoid introducing visual artifacts, such as flickering and uneven exposure, that originally do not exist in the source underexposed video; the enhanced video should be temporally consistent. We also demonstrate that our conceptually simple objectives for underexposed video enhancement can complement other relevant techniques, e.g., video dehazing.

Many approaches have been proposed for enhancing underexposed videos. These approaches can be simply classified into two categories: context-based approaches and context-free approaches. Context-based approaches [1], [2], [3] enhance low-quality nighttime videos by utilizing high-quality daytime videos. Since these methods require full day shooting with fixed cameras, they are generally limited to enhancing surveillance videos while failing to handle video clips without normally exposed exemplars.

In order to overcome the limitations of context-based approaches, context-free methods [4], [5] are developed. Bennett and McMillan [4] presented a Virtual Exposure Camera conceptual model for low dynamic range video enhancement, which comprises an adaptive spatio-temporal accumulation filter for reducing noise and a scale adaptive LDR tone mapping approach for making previously unwatchable details visually noticeable. Although this approach is capable of improving the visibility of underexposed videos, it may output videos with various types of disturbing visual artifacts, such as uneven exposure, degraded textures, and flickering artifacts, etc. Malm et al. [5] enhanced low light-level videos by first implementing a structure tensor based adaptive spatio-temporal filtering to reduce noise, and then applied the contrast limited histogram equalization [6] to widen the dynamic range. However, this method tends to degrade textures and produce ghosting artifacts.
despite its effectiveness in contrast enhancement. For underexposed videos, different regions usually require different exposures. Previous work [4], [5] rely on a single tone mapping curve to adapt the exposure to regions throughout the process. However, it is actually difficult for users to determine the optimal curve that best suits different input videos by tuning two or more parameters. Moreover, even with an optimal curve, it may also fail to ensure that all regions are well-exposed. Finally, single tone mapping curve may induce disturbing visual artifacts, such as distorted appearance. Recently, Yuan and Sun [7] have presented an automatic exposure correction method that is capable of estimating the desired exposure for different regions. The method demonstrates successful results in correcting exposure of static images. However, it may not be well suited to images with extremely uneven exposure, such as the input frame in the top row of Fig. 8.

Though a single tone mapping curve may fail to produce a globally well-exposed result, it typically more or less suits to a few underexposed regions. Thus, the key observation in this paper is that we can obtain a globally well-exposed video by adaptively keeping only locally best exposed regions derived from different tone mapping curves. Based on this observation, we construct multi-exposure image sequence containing different exposed versions of each video frame using a series of tone mapping curves. Guided by some visual perception quality measures encoding the desirable visual appearance, we adaptively locate locally best exposed regions from all multi-exposure image sequences and then seamlessly integrate them into a well-exposed video.

In this paper, we present a perception-driven context-free video enhancement method. The key idea in this work is to integrate visually best exposed regions derived from different tone mapping curves into a single well-exposed video. Our method is composed of three main stages. Firstly, each video frame is remapped by a series of tone mapping curves to construct a multi-exposure image sequence that contains different exposed versions of the frame. With some visual perception quality measures guiding us to adaptively locate locally best exposed regions from the multi-exposure image sequences, we then apply the proposed perception-driven progressive fusion to integrate all best exposed regions into a well-exposed video. Finally, we perform a texture-preserving spatio-temporal filtering on the well-exposed video to further remove the noise interference.

Our approach has four main advantages. The first advantage is that our approach can produce visually pleasing results while avoiding introducing visual artifacts. The second advantage is that our approach can adaptively process different input videos without tedious parameter setting. The third one is that our approach allows users to customize the desirable appearance by mixing different visual perception quality measures. The final one is that our approach can be easily extended to some practical applications, such as video dehazing, video denoising and HDR video reconstruction.

In summary, our work makes the following three main contributions:

- We develop a novel perception-driven progressive fusion for expanding dynamic range of underexposed videos, which can make hidden details perceptually noticeable while avoiding inducing visual artifacts.
- We introduce a texture-preserving spatio-temporal filter which can adaptively reduce the noise level while avoiding degrading textures and flickering artifacts.
- We extend our method to several practical applications, such as video dehazing, video denoising and HDR video reconstruction.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 gives the overview. Section 4 presents the technical details of the proposed underexposed video enhancement approach. Section 5 introduces several practical applications of our approach. Experimental results, comparisons and limitations are introduced in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

Here, we review the most related work in video enhancement, which can be roughly classified as histogram-based contrast enhancement, HDR video tone mapping, and underexposed video enhancement. Previous work in temporally consistent video filtering is also outlined.

Histogram-based Contrast Enhancement One of the simplest and most widely used techniques for contrast enhancement is histogram equalization. In general, histogram equalization increases the contrasts by finding a transformation function that evens out the original intensity histogram as much as possible. Here, we simply review current histogram equalization operators in two categories: global operators and local operators. The global operator [8] enhances the contrasts by stretching the original histogram to span the entire dynamic range. However, this method may cause significant loss of contrast for regions with high frequencies. To overcome this problem, some local operators [9], [6] are developed. Pizer et al. [9] redistributed the intensities using multiple adaptive histogram equalization techniques over distinct small regions. Though this method [9] can improve local contrast and bring out more details, it has a tendency to over amplify noise in relatively homogeneous regions. To limit the amplification, Zuiderveld et al. [6] presented the contrast limited adaptive histogram equalization by setting a limit on the derivative of the slope of the transformation function, but this method requires tentative parameter setting, and may produce ghosting artifacts.

Underexposed Video Enhancement As described earlier, current techniques for underexposed video enhancement can be classified as context-based approaches and context-free approaches. Context-based approaches [1],
[2], [3] enhance nighttime surveillance videos by exploiting the context in daytime images. Although these methods are very practical in improving visibility of nighttime videos, they are generally limited to enhancing full day surveillance videos captured by fixed cameras while failing to process video clips without available reference images or videos.

Context-free approaches [4], [5] are developed to enhance underexposed video with no auxiliary contexts through preconditioned point-wise intensity mapping. Bennett and McMillan [4] proposed a Virtual Exposure Camera conceptual model for low dynamic range video enhancement, which also developed an adaptive spatio-temporal accumulation filter for reducing noise and a scale adaptive LDR tone mapping approach for expanding dynamic range. Although this approach is capable of restoring visibility of challenging underexposed videos, it may fail to ensure the temporally consistency, and may induce disturbing visual artifacts. Malm et al. [5] enhanced low light-level videos by first implementing a structure tensor based adaptive spatio-temporal filtering to reduce noise, and then applied the contrast limited histogram equalization [6] to widen the dynamic range. However, this method tends to producing results with blurred textures and ghosting artifacts.

**HDR Video Tone Mapping** Tone mapping is another common dynamic range tuning technique, and have attracted broad interests recently. Researchers have developed lots of tone mapping operators for displaying the high dynamic range (HDR) images or videos to dynamic range limited devices. A variety of tone mapping operators [10], [11], [12], [13], [14] have been presented for manipulating dynamic range of static HDR images. However, only a handful of the presented operators are suitable for processing video sequences because extending tone mapping operators from static images to video sequences pose new challenges.

Based on operator presented in [12], Kang et al. [15] devised a temporal tone mapping technique using statistics from adjacent frames to produce tone mapped HDR videos that vary smoothly in time. Van et al. [16] presented a dynamical cone model to render HDR videos, which can compress the dynamic range while reducing the noise. Lee et al. [17] extended gradient domain tone mapping operator presented in [11] to HDR video by incorporating motion field into Poisson equation. Recently, Gabriel et al. [18] have given evaluation of tone mapping operators for HDR video.

Though underexposed LDR video enhancement has much in common with tone mapping of HDR video, HDR video tone mapping operators are actually not applicable in enhancing underexposed LDR videos that usually appear in the form of low-quality videos with poor visibility. In addition, HDR video tone mapping mainly focuses on compressing the dynamic range, which departs from our goal of expanding dynamic range.

**Temporally Consistent Video Filtering** Video filtering is an important topic in computer graphics and computer vision. Compared with image filtering [19], it is a more challenging topic since extending image-based filtering algorithms to video sequences is nontrivial, and is especially difficult for low-quality underexposed videos we are targeting. Lee et al. [20] constructed a spatio-temporal anisotropic diffusion framework to remove noise by a weighted average on the spatio-temporal pixels. Bhat et al. [21] presented a gradient-domain optimization framework for image and video filtering. However, this method is generally expensive to compute since it needs to solve a linear system on each video frame. In [22], the non-local means method was introduced to remove noise by averaging spatio-temporal pixels based on patch similarities. Though the method is capable of producing impressive noise removing results, it requires a tedious parameter setting to adapt the smoothness to different video sequences. Instead, our texture-preserving spatio-temporal filtering can automatically adapt the smoothness according to estimated noise level.

### 3 Overview

Let $I_t$ ($t = 0, 1, ..., T$) be the frames of an input underexposed video sequence. For frame $I_t$, the video enhancement problem can be formulated as finding a suitable mapping function $f_t$ that maps the underexposed frame $I_t$ to a normally exposed frame $I'_t$:

$$I'_t = f_t(I_t)$$

The great challenge lies in exploring a suitable mapping function that can restore all underexposed regions without introducing uneven exposure, temporally inconsistency, and other visual artifacts.

Fig. 1 illustrates the main motivation of our proposed approach. Given an underexposed video, we first perform a per-frame multi-exposure by using a series of tone mapping curves (For simplicity, we only show three tone mapping curves). In this fashion, we construct a multi-exposure image sequence for each original underexposed video frame. We then adaptively locate locally best exposed regions from the multi-exposure image sequence based on some predefined visual perception quality measures. Finally, a temporally consistent multiscale fusion technique is applied to progressively integrate all these best exposed regions into a high-quality well-exposed frame. In this case, among the three virtual exposed versions of the input frame, the umbrella in the top, the basketball in the middle, and the floor in the bottom are identified as best exposed regions (shown in non-transparent), as shown in Fig. 1.

Our video enhancement approach is composed of three main stages. In the first stage, a per-frame multi-exposure is performed to construct a multi-exposure image sequence for each video frame. In the second stage, we implement a perception-driven progressive fusion to seamlessly integrate all best exposed regions
in multi-exposure image sequences into globally well-exposed video. Lastly, we further perform a texture-preserving spatio-temporal filtering on this well-exposed video to obtain a high-quality noise-free video. Fig. 2 shows an overview.

Fig. 2. Overview of the proposed underexposed video enhancement approach.

4 UNDEREXPOSED VIDEO ENHANCEMENT

In this section, we give the technique details of the proposed video enhancement approach. We first introduce a per-frame multi-exposure used for constructing multi-exposure image sequence. Then, we defined some visual perception quality measures and analyze how to locate the locally best exposed regions. Thereafter, we formulate a perception-driven progressive fusion framework. Finally, we present the texture-preserving spatio-temporal filtering.

4.1 Per-frame Multi-exposure

Considering the Weber’s law of the human visual perception to just-noticeable differences, we design a LDR tone mapping operator to adaptively adjust the perceived intensity. Specifically, the mapping function is defined as:

\[ f(x, \alpha) = x + \frac{x}{\sqrt{x^2 + \alpha^2}}(x_{\text{max}} - x) \]  

where \( x \in [0, 1] \) and \( x_{\text{max}} = 1.0 \) denote the input normalized intensity and the upper bound of the intensity range, respectively. \( \alpha \) is the factor that controls the intensity climb. Fig. 3 illustrates the family of the mapping functions for different \( \alpha \). As shown in Fig. 3, a small \( \alpha \) corresponds to a large intensity climb, and there is typically not much increase in intensity for a large \( \alpha \).

Similar to [10], [23], we manipulate the tone of the input frame by mapping the intensity channel only and keeping the color channels unchanged. We first compute an intensity image \( I_t = (20I_r + 40I_g + I_b)/61 \) and chrominance ratios \( (\rho_r, \rho_g, \rho_b) = (I_r, I_g, I_b)/I_t \), where \( I_r, I_g \) and \( I_b \) refer to the RGB channels. We then apply the mapping function on \( I_t \) to get mapped intensity \( I_t' \). Finally, we multiply \( I_t' \) by the chrominance ratios \( (\rho_r, \rho_g, \rho_b) \) to obtain the output RGB channels.

Fig. 3. Family of mapping functions for \( \alpha \) varying from 0.1 to 1.0.
Fig. 4. Locating best exposed regions. To locate best exposed regions, we first compute visual friendliness map for each input image. Since higher visual friendliness value indicates that the pixel is more appealing to human eyes, we can then identify best exposed regions based on the visual friendliness map. In this case, we visualize each normalized visual friendliness map according to the rightmost color bar. Images courtesy of Jacques Joffre.

\[ \alpha_{\text{max}} \] are typically set to 0.1 and 1.0, and we use \( M = 6 \) as the default value.

4.2 Best Exposed Regions Locating

Intuitively, regions containing bright colors and clear details are more appealing to human eyes. Thus, similar to [24], we choose to locate best exposed regions based on some quantized visual perception quality measures, namely contrast, saturation and exposedness.

Visual Perception Quality Measures In general, we apply a Laplacian filter on the input intensity image, and select the absolute value of the filtering result as the quantized contrast. Note that, since Laplacian filter is sensitive to noise, we usually mildly smooth the intensity image using a Gaussian filter before applying the Laplacian filter. This measure tends to assign high values to important features such as edges and textures that human eyes usually pay more attention. Commonly, saturated colors are desired since it can provide us vivid look. We thus include a saturation measure, which is roughly estimated by the standard deviation with the R, G and B channels. As for the exposedness, we exploit it to evaluate the friendliness of local brightness. As human beings usually show little interest to pixels that are under- or overexposed, we thus measure the exposedness by computing the Gaussian-modeled distance between the input intensity \( I_i \) and the average normalized range value (0.5):

\[
E_p = \exp\left\{ -\frac{(I_{i,p} - 0.5)^2}{2\sigma_e^2} \right\}
\]

(3)

where \( I_{i,p} \) denotes the intensity of the pixel \( p \), and the standard deviation \( \sigma_e \) is set to 0.25 in the experiments.

By combining these three measurements, we can determine the visual friendliness of a pixel. For an input image, we define the visual friendliness map as \( W = C^n \times E^p \times S^n \), where \( C, E \) and \( S \) denote quantized contrast, exposedness, and saturation, respectively. \( \eta, \varphi \) and \( \gamma \) are exponents that controls the contribution of each measure. If an exponent is set to zero, the corresponding measure will not be taken into account.

By default, these three exponents are set to 1.0. In Fig. 4, we validate the effectiveness of our visual friendliness map in identifying best exposed regions from a multi-exposure image sequence. The top row shows five images with different exposure, and the bottom row shows the corresponding visual friendliness map. As we have seen, compared with well-exposed regions, those under- or overexposed regions are assigned lower visual friendliness values.

4.3 Perception-driven Progressive Fusion

In this section, we introduce the perception-driven progressive fusion framework. The proposed framework is built upon a global-to-local strategy. In particular, the perception-driven progressive fusion computes the desired output frame by keeping only the best exposed regions in its multi-exposure image sequence. Note that the temporal consistency is guaranteed by considering temporal information during the fusion.

Fig. 5. Pipeline of the proposed perception-driven progressive fusion framework.

In Fig. 5, we illustrate the pipeline of the proposed perception-driven progressive fusion framework. As an example, we introduce how our framework expands the dynamic range of frame \( I_t \) with the assistance of its multi-exposure image sequence \( \{ I_{M}^{m} \}_{m=1}^{M} \). Considering the original video frame may contain some well-exposed regions, and following a coarse-to-fine strategy, we begin
by fusing $I_i$ with $I_i^M$ to integrate their best exposed regions into a fusion result $I_i'$. To ensure the temporal consistency, we actually incorporate two additional temporally adjacent virtual exposed frames $I_{i-1}^M$ and $I_{i+1}^M$ into the fusion process between $I_i$ and $I_i^M$. In a similar manner, $I_i'$ is then fused with $I_i^M-1$ to obtain an updated version of $I_i'$. We iteratively implement the fusion, until $I_i'$ is fused with $I_1^M$ to obtain the final fusion result.

With the visual friendliness map indicating best exposed regions, we now proceed to fulfill the fusion task. Simple linear fusion is the first choice since it enables us to conveniently gather best exposed regions. However, as shown in Fig. 6b, naive linearly using the straightforward linear fusion produces an unsatisfactory result that suffers from unpleasing artifacts. It turns out that these disturbing artifacts will appear in areas where the visual friendliness values vary quickly. To resolve this, we need to avoid sharp weights transition. Edge aware nonlinear filters [25], [26] seem to be good choices for the problem. However, these methods may induce additional computation burden or tedious parameter setting, and a blending technique producing good results directly is preferable in this case. We thus adopt the multiscale fusion framework, which is proved quite effective at avoiding unpleasing blending artifacts.

Multiscale Fusion To implement the fusion between $I_i$ and $I_i^M$ in a multiscale manner, we first align the two temporal adjacent virtual exposed frames $I_{i-1}^M$ and $I_{i+1}^M$ to the central frame $I^M_i$ using the bidirectional correspondence. Then, $I_i$, $I_i^M$, $I_{i-1}^M$ and $I_{i+1}^M$ are decomposed into Laplacian pyramids $\{L[I_i]\}$, $\{L[I_i^M]\}$, $\{L[I_{i-1}^M]\}$ and $\{L[I_{i+1}^M]\}$ consisted of low pass filtered versions of different scales. Meanwhile, corresponding Gaussian pyramids $\{G[W_i]\}$, $\{G[W_i^M]\}$, $\{G[W_{i-1}^M]\}$ and $\{G[W_{i+1}^M]\}$ are constructed, where $W_i$, $W_i^M$, $W_{i-1}^M$ and $W_{i+1}^M$ are normalized visual friendliness maps of $I_i$, $I_i^M$, $I_{i-1}^M$ and $I_{i+1}^M$, respectively. By mixing the coefficients between the Laplacian pyramid and Gaussian pyramid at each level independently, we can construct the output Laplacian pyramid $L[I_i']$. Specifically, we define the $l$-th level of the output Laplacian pyramid as follows:

$$L_i[I_i']_p = G_i[W_i]_p L_i[I_i]_p + \sum_{i=i-1}^{l+1} G_i[W_i^M]_p L_i[I_i^M]_p$$

where $p$ denotes the pixel coordinate. After all coefficients of the output Laplacian pyramid have been computed, we collapse the output pyramid to get the fusion result $I_i'$. As demonstrated in Fig. 6c, under our perception-driven progressive fusion framework, we achieve the desired visually pleasing result with the multiscale fusion.

![Fig. 6. Comparative fusion results between linear fusion and multiscale fusion. (a) One frame of the input video. (b) Straightforward linear fusion yields unpleasing artifacts. (c) Using the multiscale fusion, we avoid unpleasing artifacts and produce a visually friendly result.](image)

![Fig. 7. Bidirectional correspondence correction. To correct the bidirectional correspondence of pixel $p$ in $I_i^M$, We first exploit the initial correspondence $u_p^b$ (shown in red arrow) and $u_p^f$ (shown in blue arrow) to get two corresponding pixels of $p$, namely $p'$ and $p''$. We then perform a patch searching to correct $u_p^b$ and $u_p^f$. For the patch centered at $p$, we search the closest patches inside $N \times N$ windows (shown in the dashed line) centered at $p'$ and $p''$, respectively. In this case, patches in purple and green are those closest patches. Thereafter, we update the initial bidirectional correspondence by replacing it with the two new correspondences.](image)
To handle color, it is possible to treat only the intensity component and reintroduce chrominance after the perception-driven progressive fusion, just as what we do in the section 4.1. However, our approach extends naturally to color frames as well. By applying the multiscale fusion on each color channel separately, our approach produces good results.

In Fig. 8, we carefully evaluate both the proposed tone mapping operator and progressive fusion scheme. To validate the effectiveness of the tone mapping operator in Eq. (2), we first sample a family of tone mapping curves from the gamma curve model. Then, we replace all the tone mapping curves with the sampled gamma curves and finally receive the gamma curves based results shown in Fig. 8b. Compared with the gamma curves, our tone mapping curves have better performance in revealing details and noise control.

Because of the aggressive smoothing in the temporal domain, the tone mapping component of [4] leads to clear ghosting artifacts around the basketball. Moreover, their tone mapped frames exhibit visually unpleasant appearance, such as the whitish floor and severely degraded textures, as shown in Fig. 8c. In contrast, our method produces visually pleasing result without introducing those visual artifacts.

Considering that [7] also introduced an effective exposure correction method, we further compare our approach with [7] in Fig. 8. As we can see that both our method and [7] successfully avoid inducing distortion appearance and ghosting artifacts. Benefiting from the contrast-aware optimal zones estimation, [7] outperforms our method in preserving relative contrast between different regions. However, our method has better performance in improving the visibility of severely underexposed regions, and tends to produce results with better overall appearance, such as more reasonable brightness and vivid colors.

Implementation We limit the depth of pyramid to at most 5 levels when performing the multiscale fusion, and exploit an efficient implementation of Lucas-Kanade optical flow [28] in OpenCV to quickly obtain the initial bidirectional correspondence. To accelerate the bidirectional correspondence correction, we progressively shrink the patch searching window along with the progressive fusion. Besides, we find that 2-3 bidirectional correspondence corrections are usually robust enough for most videos. In most cases, the resulting videos by the proposed implementation scheme are visually indistinguishable from the complete implementation while the performance has been improved.

4.4 Texture-preserving Spatio-temporal Filtering

To remove noise interference in the well-exposed video \( \{I_1', I_2', ..., I_T'\} \) while avoiding degrading textures, we introduce a texture-preserving spatio-temporal filtering in this section. We build a filtering framework upon the patch-based sparsity formulation [29]. For each patch in a video, supporting patches are gathered from current frame and temporal adjacent frames, and then aggregated together with weights based on patch similarities. In addition, we integrate the temporal bidirectional correspondence into a noise model to estimate the noise level of every frame. As a consequence, we can adapt the smoothness to different frames based on the estimated noise level.

We use \( v = (x, y, t) \) to index the input space-time volume \( \{I_1', I_2', ..., I_T'\} \), and \( P(v) \) to denote a 2D patch centered at pixel \( v \). We then focus on searching for supporting patches of patch \( P(v) \) within frame \( I'_t \) and temporally adjacent frames of \( I'_t \). To ensure both the spatial and temporal sparsity, we demand all the supporting patches to share as similar as possible structures. This demand is met by employing an approximate \( k \) nearest neighbor matching for current frame and reliable temporal correspondence to determine temporally supporting patches. Thanks to the bidirectional correspondence correction introduced in section 4.3, we can conveniently access reliable temporal correspondence here. To speed up the patch matching, we resort to an efficient approximate \( k \) nearest neighbors (\( knn \)) algorithm [30].

To determine all supporting patches for an example patch \( P(v) \), we first gather \( k \) nearest patches (including \( P(v) \) itself) of \( P(v) \) within frame \( I'_t \) by using the \( knn \) algorithm in [30]. Considering the temporally bidirectional
 correspondences, we access to two temporal patches \( P(v') \) and \( P(v'') \) corresponding to patch \( P(v) \) in frame \( I'_{t-1} \) and \( I'_{t+1} \), respectively. Then, the \( knn \) of patch \( P(v') (P(v'')) \) within frame \( I'_{t-1} (I'_{t+1}) \) are then treated as temporally supporting patches of \( P(v) \). To form the entire temporally supporting patches, we actually gather all the \( knn \) along the bidirectional motion trajectory of \( P(v) \) with a temporal radius between 3 and 5 in the experiments.

With all these spatio-temporal supporting patches, we then follow a patch-based approach, much like [22], and compute the smoothed pixel \( v \) as:

\[
\hat{I}'_{t,v} = \frac{1}{Z_v} \sum_{i=1}^{4\tau} \sum_{j=1}^{k} \omega_{v,v_{i,j}} I'_{t,v_{i,j}}
\]

where \( Z_v = \sum_{i=1}^{4\tau} \sum_{j=1}^{k} \omega_{v,v_{i,j}} \) is a normalization factor, \( \tau \) is the temporal radius. We set \( \lambda = 0.9 \) to act as a temporal decay factor that adapts to the reduced accuracy of temporal correspondence away from frame \( I'_t \). \( \omega_{v,v_{i,j}} \) measures the similarity between two patches centered at pixel \( v \) and \( v_{i,j} \), where \( v_{i,j} \) denotes a patch center coordinate indicating the \( j \)-th nearest neighbor of the \( j \)-th spatio-temporal pixel. Specifically, \( \omega_{v,v_{i,j}} \) is formulated as:

\[
\omega_{v,v_{i,j}} = \exp\left(-\frac{\text{dist}(P(v), P(v_{i,j}))^2}{2\sigma_t^2}\right)
\]

where \( \text{dist}(P(v), P(v_{i,j})) \) measures the SSD between patch \( P(v) \) and \( P(v_{i,j}) \), which is weighted by a pixelwise Gaussian-weight on the spatial offset relative to the patch center. In particular, the patch’s central pixel receives the heaviest weight, and neighboring pixels receive smaller weights as their distance to the central pixel increases. \( \sigma_t \) is a vital parameter that controls the smoothness of frame \( I'_t \). Since we take into account complete patches instead of single pixel intensities in the similarity measuring, the proposed approach is capable of removing noise from textured frames while preserving textures.

In the above filtering framework, a suitable \( \sigma_t \) is indeed vital to control the smoothness. Intuitively, if the noise level of the input video is low, we should set a relative small \( \sigma_t \) to avoid over-smoothing, and if the noise level is high, we should set a relative large \( \sigma_t \) to strengthen the smoothing. To avoid the noise level estimation, most existing denoising algorithms [31], [22] simply assume the noise level is known, which in turn greatly prevents them from practical use. Thus, we believe that the noise level should be adaptively tuned according to different frames. Inspired by [32], [33], we then introduce a frame-wise noise level estimation method to automatically adapt the value of \( \sigma_t \) to each frame.

**Frame-wise Noise Level Estimation** Theoretically, we mix a pixel-wise Gaussian distribution and a pixel-wise uniform distribution to simulate the noise statistical distribution of each video frame, and we assume that the differences between neighboring aligned frames are identical to the difference of independent noise. Then, we formulate the noise model as follows:

\[
I'_t,v - \frac{1}{2}(I'_{t-1,v'} + I'_{t+1,v''}) = \xi_v \hat{n}_v + (1 - \xi_v) \hat{u}_v
\]

where \( \hat{n}_v \) denotes a pixel-wise Gaussian random variable with \( \hat{n}_v \sim N(0, \sigma_n) \), and \( \hat{u}_v \sim U[-1,1] \) is a pixel-wise uniform random variable. The backward and forward corresponding pixels of \( v \) are denoted by \( v' \) and \( v'' \), respectively. We use a parameter \( \xi_v \) to trade off the two random variables. Let \( X_{t,v} = |I'_{t,v} - \frac{1}{2}(I'_{t-1,v'} + I'_{t+1,v''})| \) and \( \vartheta_v = -\frac{X_{t,v}}{2\sigma_n} \), we can estimate the noise level \( \sigma_n \) using an Expectance-Maximization algorithm as illustrated in Algorithm 1.

**Algorithm 1 Noise Level Estimation**

**Input:** pixelwise noise variable \( X_{t,v} \) \((v \in I'_t)\) of frame \( I'_t \)

**Output:** noise level standard variance \( \sigma_n \)

1: Initialization. Set \( \sigma_n = 10 \).
2: E-Step. Update \( \xi_v \) by \( \xi_v = 2e^{\vartheta_v} / (2e^{\vartheta_v} + \sqrt{2\pi}\sigma_n) \).
3: M-Step. Update \( \sigma_n = \sqrt{\sum_{v \in I'_t} X_{t,v}^2 / \left( \sum_{v \in I'_t} \xi_v \right)} \).
4: Iteration. Iterate step 2 and 3 until convergence.
5: Output estimated \( \sigma_n \).

![Fig. 9. Noise reduction. (a) The input well-exposed frame. (b) Noise reduction result of our approach.](image-url)

In the experiments, we use \( 7 \times 7 \) patches, and \( k = 7 \) as the number of the nearest neighbors in each frame, and we typically set \( \tau = 3 \) to integrate 7 temporal frames into the spatio-temporal filtering framework. In practice, we estimate the noise level of each RGB channel, and we have found that our approach produce visually satisfactory filtering results when we set \( \sigma_t = \sigma_n \) in default. As we can see in Fig. 9, our approach successfully reduce the noise level of the input well-exposed video without degrading the textures.

**5 Applications**

We now present some practical video editing applications to demonstrate the scalability of our framework.
Further video results can be found in the supplemental material.

**Video Dehazing** Our perception-driven progressive fusion framework is well suited to video dehazing. As introduced in [34], [35], [36], estimating the scene’s depth map is indeed vital to recover the haze-free image. To avoid directly estimating depth, we turn to sample a series of possible depth values within a specified range. Then, we can rewrite the hazy image formation model as follows:

$$J_m(x) = I(x) + (I(x) - A)\frac{1 - e^{-\beta d_m(x)}}{e^{-\beta d_m(x)}} \quad (8)$$

where $I(x)$ denotes the pixel value at the location $x$, $J_m$ denotes the haze-free scene radiance, $A$ denotes the global atmospheric light, $d_m = m \Delta d$ denotes the discretely sampled scenes depth value and $\beta$ denotes the scattering coefficient of the atmosphere. For more detail, please refer to [34].

![Fig. 10. Video dehazing. Top: Original frame. Middle: Result of [34]. Bottom: Our result. Videos courtesy of YouTube user MD881212 (left) and ©Tapes Up Productions (right).](image)

With a proper sampling interval $\Delta t$ and sampling numbers $M$, we can obtain a haze-free image sequence $\{J_m(x)\}_{M=1}^M$ for each input hazy frame. We then perform our perception-driven progressive fusion to fuse all these sequences to achieve a final haze-free video. Two video examples using our method are presented in Fig. 10. In the middle row of Fig. 10, we show the result of applying the recent method by Lang et al. [37] over the individual haze-free frame returned by He’s method [34]. As our constructed haze-free image sequence contains almost all possible pixel values, we can customize the haze-free video according to different needs. For instance, we exhibit a haze-free video with stronger contrast by assigning higher exponent to the contrast measurement, as shown in the bottom row of Fig. 10.

**Video Denoising** Our patch-based texture-preserving spatio-temporal filtering can be directly applied for video denoising. We run our filtering step with parameters $k = 11$ and $\tau = 5$ on synthetically generated noisy sequences (standard deviation $\sigma = 0.2$), and compare with an effective method [31], as illustrated in Fig. 11. We can notice that the visual difference between ours and [31] is subtle. However, since our approach relies on a robust noise estimation and a more abundant spatio-temporal patch collection, our denoising result outperforms [31] in texture preserving. As a result, our result exhibits slightly clearer textures.

**HDR Video Reconstruction** High dynamic range (HDR) imaging [38] is now popular and becoming more widespread. Our method can also be applied to reconstruct HDR video from the alternating exposed video. To do this, we first construct bidirectional correspondence between adjacent frames. Since most current optical flow algorithms rely on the brightness constancy assumption, we choose to adjust the exposure of adjacent frames to match the one with the highest exposure. Based on the bidirectional correspondence, we warp all adjacent frames to the central frame. With all warped adjacent frames and the central frame, we then run our perception-driven progressive fusion to reconstruct uniform exposed video frame. Finally, these uniform exposed frames are combined using the HDR merge process presented in [39] to form HDR frames.

Fig. 12 shows an example, where we constructed a plausible HDR video from frames with alternate exposure. As we can see, our result successfully utilizes the information from neighboring frames to restore all under- and overexposed regions without losing details.

### 6 Results and Discussion

Our underexposed video enhancement approach is implemented using C++ on a desktop PC with an Intel Core 2 Duo 2.4 GHz CPU. In this section, we first validate our approach on a variety of underexposed videos and compare with the previous methods [4], [5] in both visual appearance and runtime. We then demonstrate the effectiveness of our approach by a user study. Finally, we give the limitations of our approach.

In Fig. 13, 14, 15, 16 and 17, we present video enhancement results for various types of underexposed videos and compare our approach with [4] and [5]. Note that, all our results are produced automatically. The two previous approaches are powerful in enhancing contrast but may present visually unpleasant appearance due to the disturbing visual artifacts, such as uneven exposure, distorted appearance, colored noise, etc. Experimental results in Fig. 13 to Fig. 17 showed that our approach produces visually friendly enhancement results without introducing uneven exposure and other visual artifacts.

In Table 1, we give the time consumption comparison with Bennett and McMillan [4] and Malm et al. [5], where Stage 2 and Stage 3 refer to the two main time-consuming steps, namely perception-driven progressive
Fig. 11. Video denoising. (a) Original frame. (b) Synthesized noisy frame. (c) Result of [31]. (d) Our result. Videos courtesy of ©Smoky Cat Productions (top) and ©Universal Pictures (Bottom).

Fig. 12. HDR video reconstruction. Top: Input frames with alternate exposure (The 1th, 26th, 51th and 160th frames are shown). Bottom: Our reconstructed frames. Video courtesy of Nima Khademi Kalantari.

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Time consumption comparison with Bennett and McMillan [4] and Malm et al. [5].</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 13</td>
<td>960 × 544 × 78</td>
</tr>
<tr>
<td>Fig. 14</td>
<td>960 × 544 × 68</td>
</tr>
<tr>
<td>Fig. 15</td>
<td>960 × 544 × 87</td>
</tr>
<tr>
<td>Fig. 16</td>
<td>1280 × 720 × 113</td>
</tr>
<tr>
<td>Fig. 17</td>
<td>1280 × 720 × 157</td>
</tr>
</tbody>
</table>

Fig. 13. Comparison on the evening parking lot video clip. The first column: The 15th frame of the input video. The second column: Result of [4]. The third column: Result of [5]. The fourth column: Our result.
Fig. 14. Comparison on the rolling basketball video clip. The first column: The 28th frame of the input video. The second column: Result of [4]. The third column: Result of [5]. The fourth column: Our result.

Fig. 15. Comparison on the underexposed office video clip. The first column: The 77th frame of the input video. The second column: Result of [4]. The third column: Result of [5]. The fourth column: Our result.

Fig. 16. Comparison on the playing scooter video clip. The first column: The 23rd frame of the input video. The second column: Result of [4]. The third column: Result of [5]. The fourth column: Our result. Video courtesy of Seb Bodman.

Fig. 17. Comparison on the baby crawling video clip. The first column: The 113th frame of the input video. The second column: Result of [4]. The third column: Result of [5]. The fourth column: Our result. Video courtesy of Damien Newton.
fusion and the texture-preserving spatio-temporal filtering. Though our current C++ implementation is not optimized, our method is still much faster than both [4] and [5]. Our perception-driven progressive fusion typically takes less than 15 minutes to process a video stream with size of 640×360×100. The most time-consuming step in our method is the texture-preserving spatio-temporal filtering. It usually takes about 30 minutes to smooth the aforementioned size of video. However, we believe that the runtime performance of our approach can be significantly accelerated by a parallel GPU implementation or by using some efficient sampling [40] and hashing strategies [41].

In Fig. 13, the input video shows an evening parking lot scene with a moving car. Because of the aggressive smoothing in the temporal domain, the moving car is degraded in result of [4]. Since their subsequent tone mapping may bring out noise, their final result also suffers from clear noise interference. Due to the over enhancement in contrast, [5] produces visually unnatural result suffering from black stain. Besides, as shown in the close-up, [5] also introduces colored noise. Compared with [4] and [5], our approach produces visually result without degrading the moving car or introducing visual artifacts.

Fig. 14 shows a basketball rolling toward the umbrella. As the umbrella occludes some light, the inner side of the umbrella is barely noticeable. Though [4] successfully restores the original underexposed region, it induces clear ghosting artifacts around the basketball. Moreover, result of [4] suffers from uneaven exposure problem. The ground is obviously overexposed and exhibits unpleasant appearance. As [5] relies on a powerful local histogram equalization to stretch the dynamic range, their result exhibits the strongest local contrast among all of the three results. However, their aggressive contrast enhancement in turn leads to widespread black stain artifacts on the ground. Besides, [5] produces over-blurred textures of the umbrella. In this case, our method does not produce those aforementioned artifacts.

In Fig. 15, the input video shows an underexposed office scene captured by a moving camera where only a few areas are lit by sunshine. Due to the wide range of camera motion, [4] fails to integrate the relevant temporal pixels into their spatio-temporal filtering framework, which leads to ghosting artifacts and over-blurred appearance. Besides, [4] introduces clear halo and aliasing artifacts around edges of the trash can. Subject to the byproduct from the local histogram based tone mapping, result of [5] presents unnatural appearance even we ignore all visual artifacts on the leather chair. As we can see, our result exhibits visually pleasing browsing.

In Fig. 16, we show an evening scene in which a boy is playing scooter. Our method successfully light-s all underexposed regions without introducing visual artifacts. [4] produces a competitive result on this case. However, our approach restores clearer features. Though [5] effectively enhances the contrast, even for the dark clouds, it produces over-blurred features and introduces some disturbing artifacts on the house and the body of the boy. As shown in the close-up, our method does not induce these disturbing artifacts.

Fig. 17 shows an indoor underexposed video in which a baby is crawling. Among all the three results, [4] compensates the light most and exhibits the overall brightest appearance. However, their heuristic tone mapping leads to overall oversaturated color. Specifically, the keyboard in result of [4] exhibits distorted appearance. As shown in the close-up, [5] seriously degrade the whole scene and induces additional disturbing black stain. Compared with [4] and [5], our method produces visually pleasing result.

**User Study** We performed a user study with 50 random volunteers to validate the effectiveness of the proposed method. For each volunteer, we randomly show them enhancement results of our approach and other two previous methods [4], [5]. In fact, all the results are labeled to avoid potential unfair comparison. Results of [4] and [5] are labeled by $R_1$ and $R_2$, respectively. Our results are labeled by $R_3$.

Once a volunteer has finished browsing the labeled video enhancement results corresponding to a video example shown in Fig. 13 to Fig. 17, a survey is conducted to collect the feedbacks on following questions: 1) Which one do you think exhibits the best overall visual appearance? 2) Which one do you think best restores the original underexposed regions? 3) Which one do you think introduces the least visual artifacts? 4) Which one do you think has the least amount of temporal fluctuation? 5) Which one do you think preserves the clearest textures?

For each question, the volunteer is asked to vote for only one result. We repeat the survey until we have finished showing a volunteer all the enhancement results in Fig. 13 to Fig. 17. After all volunteers have finished the survey, we count the number of votes on each question for each label. Let $V_{ij}$ denotes the total votes of $R_i$ on $j$-th question. To evaluate each method on the individual question, we compute the percentage of votes ($PoV$) obtained by $R_i$ on the $j$-th question as follows:

$$PoV = \frac{V_{ij}}{250} \times 100\% \quad (9)$$

To provide an overall evaluation of different methods, we further calculate the percentage of votes obtained by $R_i$ in all by

$$PoV = \frac{\sum_{j=1}^{5} V_{ij}}{1250} \times 100\% \quad (10)$$

In Table 2, we give the percentage of votes obtained by different methods in the survey, where Qu. $x$ denotes the x-th question. From Table 2, we can see that results of our method get the favorites of most volunteers. In other words, most volunteers think that our results are visually better than that of [4] and [5].
We believe that our work is complementary to the video enhancement and editing since it relies on human visual perception to adaptively customize the final visual appearance instead of a single heuristic transform function. Our method opens new perspective on types of underexposed videos to confirm the method’s high-quality video enhancement results over various wide range of practical applications. We demonstrate capable of consistently producing perceptually friendly enhanced video on the Internet [42].

**ACKNOWLEDGMENTS**

The authors would like to thank the anonymous reviewers for their valuable comments and insightful suggestions. This work was partly supported by the National Basic Research Program of China (No. 2012CB725303), the NSFC (No.61472288), NCET (NCET-13-0441) and the Key Grant Project of Hubei province (2013AAA02). Chunxia Xiao is the corresponding author.

**REFERENCES**


**TABLE 2**

<table>
<thead>
<tr>
<th>Labels</th>
<th>Question ID</th>
<th>Qu. 1</th>
<th>Qu. 2</th>
<th>Qu. 3</th>
<th>Qu. 4</th>
<th>Qu. 5</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td></td>
<td>9.2%</td>
<td>13.2%</td>
<td>5.6%</td>
<td>10.8%</td>
<td>13.2%</td>
<td>10.4%</td>
</tr>
<tr>
<td>R2</td>
<td></td>
<td>6.4%</td>
<td>17.6%</td>
<td>3.6%</td>
<td>15.6%</td>
<td>10.4%</td>
<td>10.72%</td>
</tr>
<tr>
<td>R3</td>
<td></td>
<td>84.4%</td>
<td>69.2%</td>
<td>90.8%</td>
<td>73.6%</td>
<td>76.4%</td>
<td>78.88%</td>
</tr>
</tbody>
</table>

**Fig. 18. Limitations. (a) One frame of the input video. (b) Our result.**

**7 Conclusion and Future Work**

Video enhancement is a useful technique for improving the visual appearance of underexposed videos. In this paper, we have presented a novel method for underexposed video enhancement. It is conceptually simple and capable of consistently producing perceptually friendly results without introducing artifacts, and allows for a wide range of practical applications. We demonstrate high-quality video enhancement results over various types of underexposed videos to confirm the method’s effectiveness. Our method opens new perspective on video enhancement and editing since it relies on human visual perception to adaptively customize the final visual appearance instead of a single heuristic transform function. We believe that our work is complementary to the video enhancement community and can have a broad impact in the domain of perception-aware video editing and its related applications.

In the future, we will investigate how to integrate the dynamic range expansion and the noise reduction into a unified framework. We will also borrow more visual measures to better imitate human vision. Currently, our method just allows the user to customize the visual appearance of the dynamic range expanded videos. We will extend our method to synthesize a plausible scene at a different time of day instead of simply enhancing the underexposed videos, e.g., from sunset to night or from night to the noon. Another promising extension is to apply our method to large collections of images and video on the Internet [42].


