The Vid3oC and IntVID Datasets for Video Super Resolution and Quality Mapping

Sohyeong Kim  Guanju Li  Dario Fuoli  Martin Danelljan
Dario.fuoli@vision.ee.ethz.ch  martin.danelljan@vision.ee.ethz.ch
Zhiwu Huang  Shuhang Gu  Radu Timofte
zhiwu.huang@vision.ee.ethz.ch  shuhang.gu@vision.ee.ethz.ch  radu.timofte@vision.ee.ethz.ch
CVL, ETH Zürich, Switzerland

Abstract

The current rapid advancements of computational hardware has opened the door for deep networks to be applied for real-time video processing, even on consumer devices. Appealing tasks include video super-resolution, compression artifact removal, and quality enhancement. These problems require high-quality datasets that can be applied for training and benchmarking. In this work, we therefore introduce two video datasets, aimed for a variety of tasks. First, we propose the Vid3oC dataset, containing 82 simultaneous recordings of 3 camera sensors. It is recorded with a multi-camera rig, including a high-quality DSLR camera, a high-end smartphone, and a stereo camera sensor. Second, we introduce the IntVID dataset, containing over 150 high-quality videos crawled from the internet. The datasets were employed for the AIM 2019 challenges for video super-resolution and quality mapping.

1. Introduction

Video super-resolution is the problem of recovering rich details and high-frequency content from a given video. It finds application in restoration and enhancement of old video content, compressed video streamed over internet, and for enhancing video recorded by mobile devices. Due to these important applications and the gradual improvement of computational infrastructure, the field of video super-resolution is receiving a rapidly increasing amount of attention [15, 22, 3, 32, 29, 14, 26, 11]. It is also related to the emerging area of video quality mapping, which aims to transfer the quality of a video to that of a higher-quality camera or that from some websites where higher-quality videos are uploaded by people [35, 38, 39, 13].

Both of the aforementioned problems are ill-posed in general, since not all information needed reconstruct the high-resolution or higher quality setting is available in the given video data. Therefore, methods rely on learning powerful priors capable of guiding the process. This requires abundant high-quality data, that could be employed for learning. Moreover, the rising importance and popularity of these research problems calls for the introduction of standard benchmark procedures and datasets. In this work, we
therefore introduce two video datasets: Vid3oC and IntVID, that can be employed both for learning video models for a variety of tasks.

We first propose the multi-camera Vid3oC dataset. It consists of 82 simultaneous recordings, resulting in 328 videos in total. It is recorded with a rig containing three cameras: (i) a high quality DSLR camera, (ii) a high-end smartphone camera, and (iii) a stereo camera. This variety of cameras allows us to address multiple video enhancement and restoration tasks, including video super-resolution, deblurring and quality mapping. Moreover, by using the depth stream generated by the stereo camera, the dataset is also applicable for depth-guided video enhancement tasks, as well as novel viewpoint synthesis and monocular depth prediction.

As a second contribution, we introduce the IntVID dataset. It includes diverse videos from websites, as well as compressed videos which are generated using the standard video coding algorithm. The dataset allows for the video quality mapping application that aims at learning the map from low-quality compressed videos to high-quality uncompressed ones.

2. Proposed Vid3oC Dataset

We propose the Vid3oC dataset, consisting of videos recorded simultaneously by three different camera sensors. Our intention is to thereby capture different quality of video from the same scene to accommodate emerging applications, including video quality mapping and enhancement. The dataset is also designed to complement existing datasets for video super-resolution, by providing new challenges and diversity. In contrast to previous datasets of its kind, our Vid3oC includes data captured by a stereo camera, providing geometric information of the scenes, enabling the investigation of depth-guided video enhancement approaches.

Rig and Sensors The videos were simultaneously recorded by three cameras attached to a 3D-printed rig (figure 2). Firstly, we employ the Canon 5D Mark IV DSLR camera to serve as a high-quality reference. Secondly, since one of our main intentions is to cover smartphone video enhancement applications, we use a high-end smartphone camera, here the Huawei P20. Lastly, our setup includes a ZED stereo camera. While the ZED camera provides an estimated 3D depth of the scene, it also represents a separate video quality level, being significantly lower than that of the P20. As an important factor for image and video quality, the Canon 5D Mark IV has a sensor area that is 20 times larger than that of the P20 and 35 times larger than that of the ZED. Figure 3 shows sample images from the three cameras. More samples, including the sample generated depth images, are shown in figure 4.

Recording The dataset was manually recorded in the area in and around Zurich, Switzerland during the summer months. We used carefully chosen locations and scenes to ensure variety in content, appearance, and the dynamic nature. Videos were recorded with two different types of motion. In the first setting, the camera rig was attached to a tripod, with different panning motions applied. This setting ensures relatively smooth camera motion. In the second setting, the rig was hand-held and videos were recorded while walking, inducing non-smooth and sometimes fast motion. Sample snapshots from videos recorded at different scenes are shown in figure 4, demonstrating the variety of the data.

Video recording was started synchronously by a computer software. However, to allow more exact temporal alignment of the videos, a timer was also displayed in the first few seconds of the recording. The length of each recording was between 30 and 60 seconds. Videos were captured in 30 FPS, using the highest resolution available at that frame-rate. We used the auto/default video recording modes for the Canon and P20 cameras. The videos were stored as lossy but high-quality H.264 and H.265 compression for the Canon and P20 respectively, while the ZED video was saved as lossless PNG frames. Depth maps computed online by the ZED in the highest quality setting were saved with 16-bit quantization, along with the provided camera pose estimates. The settings for each camera are summarized in table 1.

<table>
<thead>
<tr>
<th>Camera</th>
<th>Video resolution</th>
<th>Frame-rate</th>
<th>Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon 5D Mark IV</td>
<td>1920 × 1080</td>
<td>29.97 FPS</td>
<td>H.264</td>
</tr>
<tr>
<td>Huawei P20</td>
<td>3840 × 2160</td>
<td>29.76 FPS</td>
<td>H.265</td>
</tr>
<tr>
<td>ZED stereo</td>
<td>1920 × 1080</td>
<td>30 FPS</td>
<td>Lossless PNG</td>
</tr>
</tbody>
</table>

Table 1. Camera settings used for the recording of Vid3oC dataset.
compressed with H.264. For each video in the training and validation sets we extract a carefully selected 4-second interval, containing 120 frames, which are stored as individual PNG image files. The dataset thus encompasses 82 recordings and 328 videos in total.

3. Proposed IntVID Dataset

We set up another dataset named IntVID, which includes videos crawled from the websites vimeo.com, mixkit.co and videvo.net. The downloaded videos cover 12 diverse categories: city, coffee, fashion, food, lifestyle, music/dance, narrative, nature, sports, talk, technique and transport. The resolutions of videos are mostly of 1920×1080. Their duration varies from 8 seconds to 360 seconds with frame rate being at the range of 23.98-25.00 FPS.

Processing In the dataset, most of the collected videos consist of changing scenes. Accordingly, we used a popular scene detection tool named PySceneDetect to split most of videos into three separate clips for training, validation and test respectively. In particular, we chose partitioned the resulting video clips such that the majority of the original video content is employed for training. For the validation and test video clips, we fix their length to 4 seconds containing 120 frames that are saved as PNG image files. Due to the bandwidth limit of Internet, video compression techniques are often applied to save the coding bitrate. Inspired by this, we further applied the standard video coding system H.264 to compress some of the collected videos. As a result, we generate a total of 60 unpaired compressed and uncompressed videos for training, 32 paired compressed/uncompressed clips for validation and testing respectively. Some sample frames from uncompressed and compressed videos are shown in Figure 5.

4. Applications and Challenges

AIM 2019 Video Extreme Super-Resolution Challenge

Video Super-Resolution (VSR) is the task of reconstructing a high-resolution (HR) video from the low-resolution (LR) input. Super-resolution is an ill-posed problem, as high-frequency information is inherently lost when downsampling an image or video, because of the lower Nyquist frequency in LR space. Single image super-resolution (SISR) methods usually restore this information by learning image priors through paired examples. For video super-resolution (VSR), additional information is present in the temporal domain, which can help significantly improving restoration quality over SISR methods. SISR has been an active research for a long time [4, 16, 19, 20, 9, 30, 33, 28], while VSR has gained traction in recent years [25, 12, 29, 14, 5, 2, 37, 32, 27, 15, 36], also due to the availability of more and faster computing resources. While there exists a lot of prior work on super-resolution factors ×2, ×3 and ×4 with impressive results, attempts at higher factors are less common in the field [18].

Restoring such a large amount of pixels from very limited information is a very challenging task. The aim of this challenge [7] is therefore to investigate if super-resolution with such high downsampling ratios is still possible with acceptable performance. Two tracks are provided in this challenge. Track 1 is set up for fully supervised example-based VSR. The restoration quality is evaluated with standard metrics in the field, Peak Signal-to-Noise Ratio (PSNR) and structural similarity index (SSIM). Because PSNR and SSIM are not always well correlated with human perception of quality, track 2 is aimed at judging the outputs according to how humans perceive quality. Track 2 is also example-based, however, the final scores are determined by a mean opinion score (MOS).

For the video extreme super-resolution challenge [7], the high-quality DSLR videos were used to serve as high quality ground truth. The whole training set (encoded HR videos) was provided to the participants, in order to generate their training data according to their needs and to reduce data traffic. For the validation and test phase, two separate sets are provided of 16 sequences each, all composed of 120 LR frames in PNG format, to be processed and evaluated on the CodaLab servers. To generate the LR data from the FullHD (1920×1080) HR videos for factor ×16, the HR sequences are first cropped to 1920×1072 to be dividable by 16. The sequences were then downscaled by Matlab’s imresize method with factor ×16 using the standard settings, resulting in sequences of 120×67.
**Video Quality Mapping Application** Recorded and transmitted videos often suffer from various quality issues. For instance, despite the incredible development of current smartphone and depth cameras, compact sensors and lenses still limit the quality of their captured videos. To address such issue, the Vid3oC dataset allows for learning mappings for either P20 or ZED videos to the quality of Canon captured ones. This includes tasks such as enhancing the perceived quality of videos, which includes enhancements like increasing color vividness, boosting contrast, sharpening up textures, etc.

In addition, the bandwidth limit over internet requires that videos have to be compressed for easier transmission. However, the compressed videos inevitably suffer from compression artifacts. Therefore, we suggest our IntVID dataset to learn the quality mapping from compressed videos to uncompressed ones, which will be helpful to restore the quality of compressed videos.
5. Video Super Resolution Datasets

In contrast to SISR [1, 31, 8, 41, 23], there are no standardized benchmarks that are widely accepted in the field for higher resolutions, longer duration and more realistic video material. The most frequently used benchmark for VSR is vid4 [21]. Vid4 is a dataset composed of 4 short sequences with a resolution between 480×704 and 576×704. The content is therefore not representative of modern video material. There are other, less popular datasets, which are occasionally used for validation, but suffer from similar limitations.

A new dataset is introduced in the NTIRE 2019 challenge [24], called REDS. It includes 30000 HR frames (1280×720) of various content. However, since the dataset has been collected mainly for video deblurring, the videos are intentionally not stabilized and therefore expose large displacements between frames, which are not representative for general video content. Vid3oC has been created to mitigate these issues and to provide realistic, up-to-date videos. The dataset contains videos of 3 different HR cameras with different levels of quality, including diverse content and varying motion patterns. All videos are aligned in the temporal domain, even depth information is recorded. The dataset is therefore not limited to super-resolution and can be used for many other tasks.

6. AIM 2019 VSR Challenge Results

This section presents results of the AIM 2019 VSR challenge, which employs the proposed Vid3oC dataset. From initially 39 and 30 registered participants in track 1 and 2 respectively, three teams (fenglinglwb, NERCMS, and HIT-XLab) entered the final ranking and submitted their results, codes, executables and factsheets. Team fenglinglwb and HIT-XLab provided the same solutions to both tracks, while NERCMS submitted a solution to the first track. According to PSNR/SSIM and MOS, the winner of both challenge tracks is fenglinglwb with +0.17 dB over team NERCMS and +1.84 dB over the Bicubic baseline. The final ranking, PSNR, SSIM, MOS, runtimes, platform and type of hardware is shown in table 2.

7. AIM 2019 Video Super Resolution Methods

This section gives brief descriptions of the participating methods in the AIM 2019 VSR challenge. Further details can be found in the challenge report [7].

7.1. fenglinglwb team

The winner of both tracks presents a solution based on the network EDVR [36]. They extend their method by applying an edge mask guided model during learning, which encourages the network to learn more refined edges. The method also incorporates a non-local module to make use of global information. For extreme VSR, it is essential to take the context information into full play. The final performance in the challenge is improved by using ensembling. The final ensemble model (see figure 6) is composed of two edge mask models of different sizes and a non-local model. The individual outputs are combined produce the HR estimate. The team submitted identical results for both tracks 1 & 2.

![Figure 6. Architecture of team fenglinglwb [7].](image)

7.2. NERCMS team

The NERCMS team follows a progressive fusion network by Yi et al. [40], which is composed of a series of progressive fusion residual blocks (PFRBs). These blocks leverage both spatial (intra-frame) and temporal (inter-frame) correlations between frames. Figure 7 illustrates the case of adopting 5 frames as input. The method gathers a batch of 7 consecutive frames to produce a single HR estimate. A non-local block computes correlations between a single pixel and all other pixels in the batch to enable global feature extraction. A residual learning scheme is adopted to improve training. Finally, the residuals are added to the input center frame, which is upsampled with Bicubic interpolation. The team submitted results for track 1.

7.3. HIT-XLab team

The team’s method uses EDSR [20] as the backbone but replaces 2D convolutions with 3D convolutions [34]. Memory usage is lowered by reducing the number of residual blocks [10] to 8. The method is trained with 64 LR patches of size 5×32×32×3 per iteration. The network is trained with L1 loss and gradient loss with Adam optimizer [17]. The team chose to use no temporal padding needed for 3D convolutions, instead the missing frames at the tempo-
Table 2. Results for the participating teams. We evaluated PSNR and SSIM for track 1 and conducted a MOS for track 2.

Figure 7. Architecture of team NERCMS [40].

Figure 8. Architecture of RLSP [6].

7.4. RLSP (baseline)

In addition to Bicubic interpolation, we provide another method as a baseline called RLSP [6]. The method proposes a fully convolutional RNN for VSR with upsampling factor 4. RLSP introduces high dimensional hidden states to enable implicit information propagation along time. No optical flow and/or warping is used, accumulation and processing of temporal information is handled implicitly by the hidden states. Due to the recurrent nature, the method is extremely efficient at runtime. The upsampling is divided into 4 stages of factor $\times 2$, to handle the extreme super-resolution factor.

8. Conclusions

This paper introduces two video datasets: Vid3oC and IntVID. The former consists of 82 video recordings collected with a multi-camera rig which includes a high-quality DSLR camera, a high-end smartphone, and a stereo camera sensor. The IntVID contains over 150 videos crawled from the internet. These datasets are designed for a variety of tasks, including video super-resolution and quality mapping. In particular, we demonstrate its use for the former, in the AIM 2019 Video Super-Resolution Challenge. Future work includes adapting these datasets for other video tasks.

Acknowledgments. This work was partly supported by ETH General Fund, by Nvidia through a GPU grant, and by Huawei.

References


[28] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan


