AIM 2020 Challenge on Video Extreme Super-Resolution: Methods and Results

Dario Fuoli[1], Zhiwu Huang[1], Shuhang Gu[2], Radu Timofte[1] Arnau Raventos[3], Aryan Esfandiari[3], Salah Karout[3], Xuan Xu[3], Xin Li[3], Xin Xiong[3], Jinge Wang[3], Pablo Navarrete Michelini[3], Wenhao Zhang[3], Dongyang[3], Zhang[3], Hanwei Zhu[3], Dan Xia[3], Haoyu Chen[3], Jinjin Gu[3], Zhi Zhang[3], Tongtong Zhao[3], Shanshan Zhao[3], Kazutoshi Akita[3], Norimichi Ukita[3], Hrishikesh P S[3], Densen Puthussery[3], and Jiji C V[3]

[1] ETH Zürich, Switzerland
[2] The University of Sydney, Australia
Dario Fuoli (dario.fuoli@vision.ee.ethz.ch), Zhiwu Huang, Shuhang Gu and Radu Timofte are the AIM 2020 challenge organizers.
http://www.vision.ee.ethz.ch/aim20/
[3] Participants in the challenge. Appendix A contains the authors' teams and affiliations.

Abstract. This paper reviews the video extreme super-resolution challenge associated with the AIM 2020 workshop at ECCV 2020. Common scaling factors for learned video super-resolution (VSR) do not go beyond factor 4. Missing information can be restored quite well in this region, especially in high-resolution videos, where the high-frequency content mostly consists of texture details. The task in this challenge is to upscale videos with an extreme factor of 16, which results in more serious degradations that also affect the structural integrity of the videos. A single pixel in the low resolution domain corresponds to 256 pixels in the high-resolution domain. Due to this massive information loss, it is hard to accurately restore the missing information. Track 1 is set-up to gauge the state-of-the-art for such a demanding task, where fidelity to the ground truth is measured by PSNR and SSIM. Perceptually higher quality can be achieved in trade-off for fidelity by generating plausible high-frequency content. Track 2 therefore aims at generating visually pleasing results, which are ranked according to human perception, evaluated by a user study. In contrast to single image super-resolution (SISR), VSR can benefit from additional information in the temporal domain. However, this also imposes an additional requirement, as the generated frames need to be consistent along time.

Keywords: extreme super-resolution, video restoration, video enhancement, challenge

1 Introduction

Super-resolution (SR) aims at reconstructing a high-resolution (HR) output from a given low-resolution (LR) input. Single image SR (SISR) generally focuses on

Fig. 1. Downscaled crops with the extreme factor of ×16 (top) and corresponding 96×96 high-resolution crops (bottom).

restoring spatial details as the input consists of only one single image. By comparison, as the input of VSR is usually composed of consecutive frames, it is expected to concentrate on the exploitation of the additional temporal correlations, which can help improving restoration quality over SISR methods. Making full use of the temporal associations among multiple frames and keeping the temporal consistency for VSR remain non-trivial problems. Furthermore, when moving towards more extreme settings like higher scale factors that requires to restore a large amount of pixels from severely limited information, the VSR problem will get much more challenging.

Following our first AIM challenge [8], the goal of this challenge is to super resolve the given input videos with a extremely large zooming factor of 16, with searching for the current state-of-the-art and providing a standard benchmark protocol for future research in the field. Fig.1 presents a few downscaled crops and their corresponding HR crops. In this challenge, track 1 aims at probing the state-of-the-art for the extreme VSR task, where fidelity to the ground truth is measured by peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). As a trade-off for the fidelity measurement, track 2 is designed for the production of visually pleasing videos, which are ranked by human perception opinions with a user study.

2 Related Work

For SR, deep learning based methods [6, 17, 21, 36, 35, 4, 28] have proven their superiority over traditional shallow learning methods. For example, [6] introduces convolution neural networks (CNN) to address the SISR problem. In particular, it proposes a very shallow network to deeply learn LR features, which are subsequently leveraged to generate HR images via non-linear mapping. To reduce the time complexity of the network operations in the HR space, [38] proposes an effective sub-pixel convolution network to extract and map features from the LR space to the HR space using convolutional layers instead of classical interpolations (e.g., bilinear and bicubic). [47] exploits a residual dense network block

with direction connections for a more thorough extraction of local features from LR images. A comprehensive overview of SISR methods can be found in [41].

Compared with SISR, the VSR problem is considerably more complex due to the additional challenge of harnessing the temporal correlations among adjacent frames. To address this problem, a number of methods [22, 30, 5] are suggested to leverage temporal information by concatenating multiple LR frames to generate a single HR estimate. Following this strategy, [2] first warps consecutive frames towards the center frame, and then fuses the frames using a spatio-temporal network. [16] aggregates motion compensated adjacent frames, by computing optical flow and warping, followed by a few convolution layers for the processing on the fused frames. [25] calculates multiple HR estimates in parallel branches. In addition, it exploits an additional temporal modulation branch to balance the respective HR estimates for final aggregation. By contrast, [15] relies on implicit motion estimation. Dynamic upsampling filters and residuals are computed from adjacent LR frames with a single neural network. Finally, the dynamic upsampling filters are employed to process the center frame, which is then fused with the residuals. Similarly, [27] proposes a dynamic local filter network to perform implicit motion estimation and compensation. Besides, it suggests a global refinement neural network based on residual block and autoencoder structures to exploit non-local correlations and enhance the spatial consistency of the super-resolved frames.

In addition to the aggregation strategy, some other works suggest to make use of recurrent neural networks (RNN) for better VSR. Due to the better capacity of learning temporal information on input frames, they provide a potentially more powerful alternative to address the SR problem. For instance, [39] suggests an autoencoder style network as well as an intermediate convolutional long short-term memory (LSTM) layer. The whole network is capable of processing the preliminary HR estimate from a subpixel motion compensation layer, for better HR estimate. [13] proposes a bidirectional recurrent network, which exploits 2D and 3D convolutions with recurrent connections and combines a forward and a backward pass to produce the HR frames. To make use of temporal information, [37] designs a neural network that warps the previous HR output towards the current time step, by observing the optical flow in LR space. The warped output is concatenated with the current LR input frame and a SR network generates the HR estimate. [7] exploits a recurrent latent space propagation (RLSP) algorithm for more efficient VSR. Particularly, RLSP introduces high-dimensional latent states to propagate temporal information between frames in an implicit manner so that the efficiency can be highly improved.

3 AIM 2020 Challenge Setup

3.1 Data

We use the Vid3oC [18] dataset for this challenge, which has been part of previous challenges [8, 9]. The dataset is a collection of videos taken with three different cameras on a rig. This results in roughly aligned videos, which can be

Fig. 2. Track 1 Visual results for all competing teams. Additionally, we show the ground truth (GT) and bicubic interpolation (Bicubic) for reference. To present the details more clearly and to fit all methods on a single page, the frames are cropped to 800×1920. Highlights from yellow and blue boxes are shown at the top left.

used for weak supervision. In this challenge, only the high-quality DSLR camera (Canon 5D Mark IV) is used to serve as high-resolution ground truth. The corresponding low-resolution source data is obtained by downscaling the ground

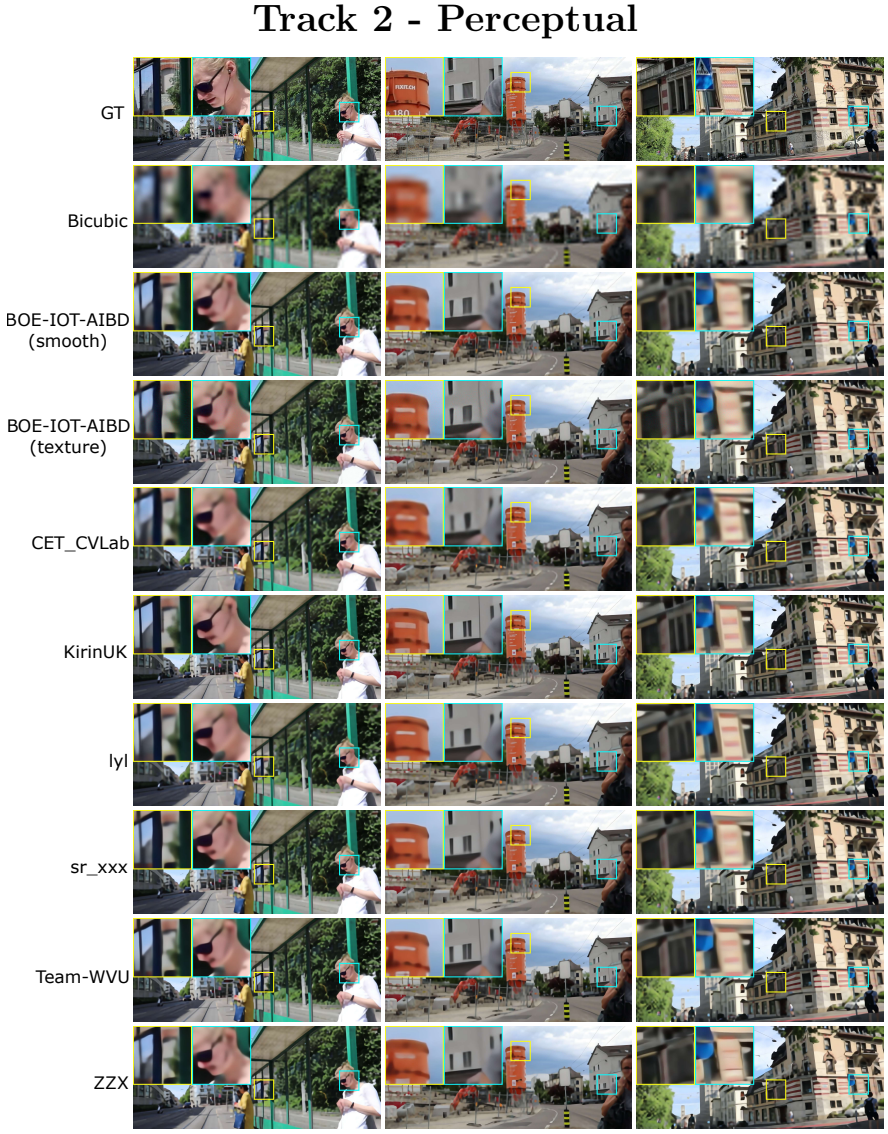


Fig. 3. Track 2 Visual results for all competing teams. Additionally, we show the ground truth (GT) and bicubic interpolation (Bicubic) for reference. To present the details more clearly and to fit all methods on a single page, the frames are cropped to 800×1920. Highlights from yellow and blue boxes are shown at the top left. Team BOE-IOT-AIBD provides two distinct solutions, which focus on high quality textures and temporal smoothness respectively.

Fig. 4. Low-resolution frame with extreme downscaling factor ×16 (left) and corresponding high-resolution frame (right).

truth by factor 16, using MATLAB's imresize function with standard settings, see Fig. 4. In order to retain proper pixel-alignment, the ground truth 1080×1920 FullHD frames are cropped to 1072 × 1920 before downscaling, to be dividable by 16. We provide 50 high-resolution sequences to be used for training. To save bandwidth, these videos are provided as MP4 files together with scripts to extract and generate the low-resolution source frames. Additionally, the dataset contains 16 paired sequences for validation and 16 paired sequences for testing, each consists of 120 frames in PNG format.

3.2 Challenge Phases

The challenge is hosted on CodaLab and is split up in a validation and a test phase. During the validation phase, only the validation source frames are provided and participants were asked to submit their super-resolved frames to the CodaLab servers to get feedback. Due to storage constraints on CodaLab, only a subset of frames could be submitted to the servers (every 20th frame in the sequence). In the following test phase, the final solutions had to be submitted to enter the challenge ranking. There was no feedback provided at this stage, in order to prevent overfitting to the test set. Additionally, the full set of frames had to be made accessible to the challenge organizers for the final rankings. After the submission deadline, the high-resolution validation ground truth was released, for public use of our dataset.

3.3 Track 1 - Fidelity

This track aims at high fidelity restoration. For each team, the restored frames are compared to the ground truth in terms of PSNR and SSIM and can be objectively quantified by these pixel-level metrics. The focus is on restoring the data faithfully to the underlying ground truth. Commonly, methods for this task are trained with a pixel-level L1-loss or L2-loss. The final ranking among teams is determined by PSNR/SSIM exclusively, without visual assessment of the produced frames.

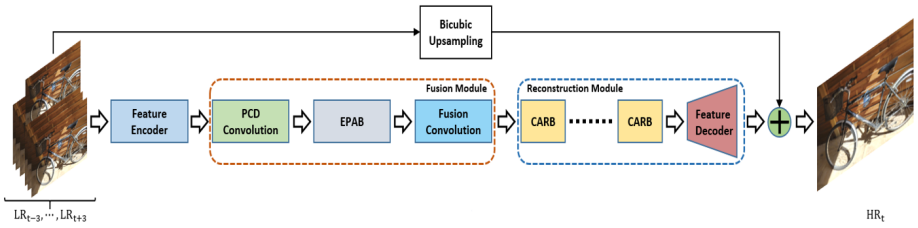


Fig. 5. Efficient Video Enhancement and Super-Resolution Net (EVESRNet) proposed by KirinUK.

3.4 Track 2 - Perceptual

Super-resolution methods optimized for PSNR tend to oversmooth and often fail to restore the highest frequencies. Also, PSNR does not correlate well with human perception of quality. Therefore, the focus in the field has shifted towards generation of perceptually more pleasing results in trade-off for fidelity to the ground truth. Since the extreme scaling factor 16 and its associated large information loss prohibits high fidelity results, the only possibility to achieve realistically looking high-resolution videos in this setting is by hallucinating plausible high frequencies. Track 2 is aimed at upscaling the videos for highest perceptual quality. Quantitative assessment of perceptual quality is difficult and remains largely an open problem. We therefore resort to a user study in track 2, which is still the most reliable benchmark for perceptual quality evaluation.

4 Challenge Methods and Teams

4.1 KirinUK

Recent video super-resolution approaches [3] propose splitting the spatio-temporal attention operation in several dimensions, their aim is to reduce the computational cost of a traditional 3D non-local block. Nevertheless, these methods still need to store the HWxHW attention matrices, which is challenging, especially when dealing with GPUs with limited amount of memory or when upscaling high resolution videos. To tackle this, the KirinUK team proposes to extend the VESRNet [3] architecture by replacing the Separate Non Local (SNL) module with an Efficient Point-Wise Temporal Attention Block (EPAB). This block aggregates the spatio-temporal information with less operations and memory consumption, while still keeping the same performance. The team names this new architecture Efficient Video Enhancement and Super-Resolution Net (EVESRNet) and an overview of it can be seen in Fig. 5. It is mainly composed of Pyramid, Cascading and Deformable Convolutions (PCDs) [42], the EPAB, and Channel Attention Residual Blocks (CARBs) [3] [46].

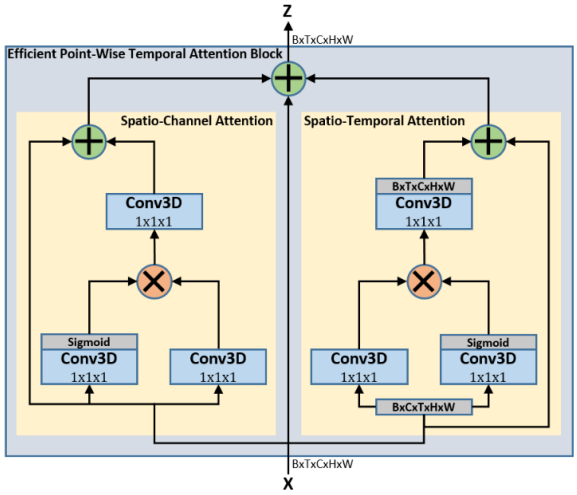


Fig. 6. Efficient Point-Wise Temporal Attention Block proposed by KirinUK.

The EPAB module is illustrated in Fig.6 and can be divided into two sub-blocks, the Spatio-Channel Attention (SCA) and the Spatio-Temporal Attention (STA). Both of them share the same structure, the only difference is the permutation operation at the beginning and at the end of the STA sub-block.

To perform Extreme Video SR the team employs two 4x stages in cascade mode. Each stage was trained independently with an EVESRNet architecture. Moreover, their training does not start from scratch, they first pretrain a model with the REDS dataset [31] to initialize the networks. This helps preventing overfitting in the first stage where the amount of spatial data is limited. The REDS dataset was only utilized for pretraining. As a reference, the initialization model achieves 31.19 PSNR in the internal REDS validation set defined in [42], which corresponds to a +0.1 dB improvement with respect to EDVR [42].

Each track uses the same pipeline, the only differences are: 1) For the fidelity track, the two stages were trained using L2 loss. 2) For the perceptual track, the second stage was trained using the following combined loss:

$$L = \lambda_1 * L_{L1} + \lambda_2 * L_{VGG} + \lambda_3 * L_{RaGAN} \tag{1}$$

where λ_1 is 1e-3, λ_2 is 1 and λ_3 is 5e-3. They used a patch discriminator as in [14]. The rest of the hyperparameters are the same as in [43].

4.2 Team-WVU

Recently, deformable convolution [48] has been received increasingly more attention to solve low-level vision tasks such as video super-resolution. EDVR [42] and TDAN [40] have already successfully implemented deformable convolution

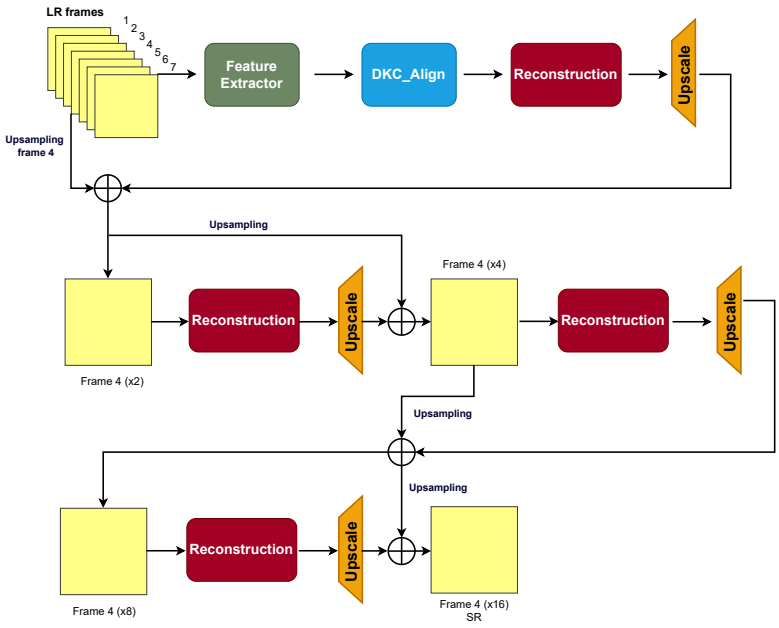
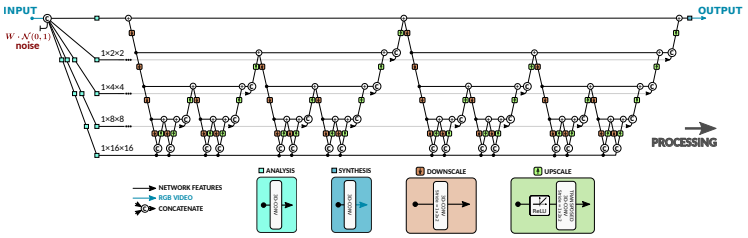


Fig. 7. Network proposed by WVU.

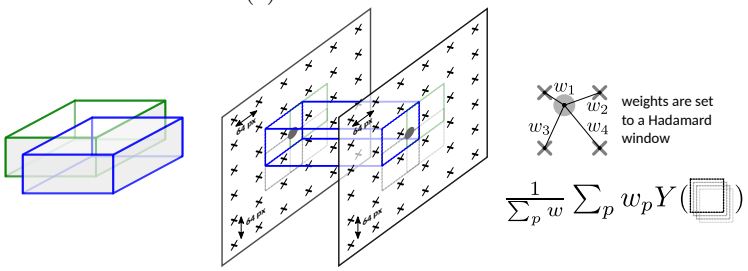
to temporally align reference frame and its neighboring frames which can let networks better utilize both spatial and temporal information to enhance the final results.

Inspired by state-of-the-art video SR method EDVR [42], the team develops the novel Multi-Frame based Deformable Kernel Convolution Networks to temporally align the non-reference and reference frames with deformable kernel [10] convolution alignment module and enhance the edge and texture features via deformable kernel spatial attention module.

The overall diagram of proposed network is shown in Fig. 7. It mainly includes four parts, feature extractor, DKC_Align module (deformable kernel convolution alignment module), reconstruction module and upscale module. Different from PCD alignment module from EDVR, the team implemented stacked deformable kernel convolution layers instead of traditional convolution layer to extract offset. Deformable kernel can better adapt effective receptive fields than normal convolution [10] which can better enhance the offset extraction compared with the normal convolution. On the step of reconstruction, to calibrate reconstructed feature maps before feeding into each upscaling module, the team proposes a Deformable Kernel Spatial Attention (DKSA) module (integrated to reconstruction module) to enhance the textures that can help the proposed network to reconstruct SR frames sharper and clearer. Because this challenge aims to super-resolve extreme low-resolution videos with the scale factor of 16, to avoid generating undesired blurring and artifacts, the low-resolution frames



(a) 3D–MGBP Network



(b) Overlapping strategy

Fig. 8. 3D – MultiGrid BackProjection network (3D–MGBP) proposed by BOE-IOT-AIBD.

are super-resolved with a scale factor of 2 each time (see Fig. 7). Finally, the low-resolution frames are super-resolved four times in total to upscale the LR frames with a magnification factor $\times 16$. Charbonnier Loss [42, 20] is used as the loss function for both track 1 and track 2 training, the loss can be expressed as follows:

$$L = \sqrt{||\hat{X}_r - X_r||^2 + \xi^2} \tag{2}$$

where $\xi = 1 \times 10^{-3}$, $\hat{X}_r$ is super-resolved frame and X_r is target frame (ground-truth).

4.3 BOE-IOT-AIBD

The team proposes 3D-MGBP, a fully 3D–convolutional architecture designed to scale efficiently for the difficult task of extreme video SR. 3D–MGBP is based on the Multi–Grid Back–Projection network introduced and studied in [34, 32, 33]. In particular, they extend the MGBPv2 network [32] that was designed to scale efficiently for the task of extreme image SR and was successfully used in the 2019–AIM Extreme Image SR competition [29] to win the Perceptual track of that challenge. For this challenge they redesigned the MGBPv2 network to use 3D–convolutions strided in space. The network works as a video enhancer that, ignoring memory constrains, can take a whole video stream and outputs a whole video stream with the same resolution and framerate. They input a $16\times$ Bicubic upscaled video and the network enhances the quality of the video stream. The receptive field of the network extends in space as well as time by

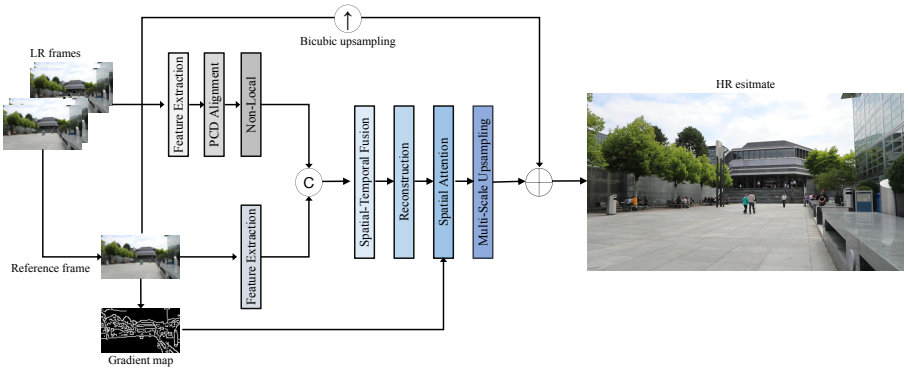


Fig. 9. Network proposed by ZZX.

using 3D–convolutional kernels of size $3 \times 3 \times 3$. The overall architecture uses only 3D–convolutions and ReLU units. This is in contrast to general trends in video processing networks that often include attention, deformable convolutions, warping or other non–linear modules.

Figure 8 displays the diagram of the 3D–MGBP network used in the competition. In inference it is impossible for 3D–MGBP to process the whole video stream and so they extend the idea of overlapped patches used in MGBPv2 by using overlapped spatio–temporal patches (overlapping in space and time). More precisely, to upscale arbitrarily long video sequences they propose a patch based approach in which they average the output of overlapping video patches produced by the Bicubic upscaled input. First, they divide input streams into overlapping patches (of same size as training patches) as shown in Figure 8; second, they multiply each output by weights set to a Hadamard window; and third, they average the results.

They trained the 3D–MGBP network starting from random parameters (no pre–trained models were used). For the Fidelity track they trained the model using L2 loss on the output spatio–temporal patch. For the Perceptual track they submitted the output of two different configurations of the same architecture. The first submission, labeled *Smooth*, was trained with L2 loss as they noticed better time–consistency and smooth edges. In their second submission, labeled *Texture*, they followed the loss and training strategy of G–MGBP [33], adding a noise input to activate and deactivate the generation of artificial details. The noise input consists of one channel of Gaussian noise concatenated to the Bicubic upscaled input. In this solution, although more noisy and farther from ground truth due to the perception–distortion trade–off [1], they noticed better perception of textures.

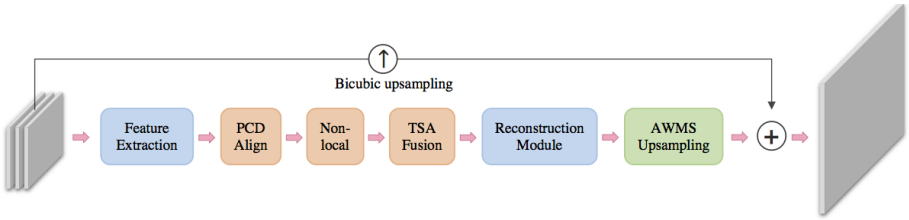


Fig. 10. Network proposed by sr_xxx.

4.4 ZZX

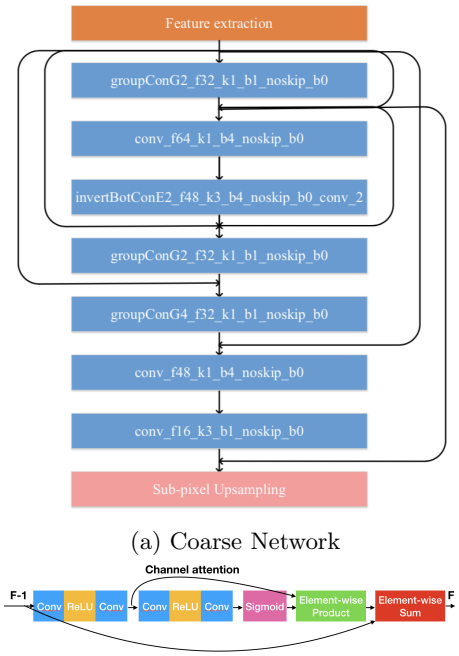
In order to restore the high frequency information of the extreme video, the team designed the multi-scale aggregated upsampling based on high frequency attention (MAHA) network. The framework is illustrated in Fig. 9. The team inputs seven low-resolution frames into the feature extraction module, and inspired by the EDVR [42], the Pyramid, Cascading and Deformable (PCD) alignment module was applied to address global motion. The non-local module was used to select valid inter-frame information. Then, the extracted reference frame feature concatenated with the alignment features that were utilized to perform spatial-temporal fusion in a progressive strategy, which helps to aggregate spatial-temporal information. Next, the team also proposed an attention-guided multi-level residual feature reconstruction module to fully improve feature representation. Finally, to generate a sharp structure high-resolution video, the team computed the gradient map of the low-resolution image to guide the spatial attention module.

The team divided the network training into two stages for both track 1 and track 2. For stage 1, the $L1$ loss was used. For stage 2, the team fine-tuned the results of stage 1 and using $0.15 * L_{L1} + 0.85 * L_{SSIM}$ to train the network. However, the team applied different testing strategies for track 1 & 2 . For Track 1, they employed model fusion testing and test enhancement strategies to improve the PSNR value. For the track 2, the best validation performance was used to directly generate the test results.

4.5 sr_xxx

The team employs the high-level architecture design of EDVR [1], with improvements to accommodate large upscaling factors which up to 16. The used network is illustrated in Fig.10. To explain the framework, they first start with EDVR baseline. EDVR is a unified framework which can achieve good alignment and fusion quality in video restoration tasks. It proposed a Pyramid, Cascading and Deformable (PCD) alignment module, or the first time uses deformable convolutions to align temporal frames. Besides, EDVR includes a Temporal and Spatial Attention (TSA) fusion module to emphasize important features.

Their proposed network takes 5 low-resolution frames as input and generates one high-resolution output image frame. They first conduct feature extraction,



(a) Coarse Network

(b) Fine Network

Fig. 11. Network proposed by lyl.

followed with PCD alignment module, Non-local module and TSA module to align and fusion multiple frames. Right after reconstruction module, they use adaptive weighted multi-scale (AWMS) module as our upsampling Layer. In the last module, they add the learned residual to a direct bicubic upsampled image to obtain the final high-resolution outputs.

They trained two different reconstruction models and ensemble their results to obtain more stable texture reconstructions. To incorporate finer detail ensembling, they combine residual feature aggregation block [26] and residual channel attention block [46].

4.6 lyl

As shown in Fig.11, the team proposes a coarse to fine network for progressive super-resolution reconstruction. By using the suggested FineNet:lightweight upsampling module (LUM), they achieve competitive results with a modest number of parameters. Two requirements are contained in a coarse to fine network(CFN): (1) progressiveness and (2) merge the output of the LUM to correct the input in each level. Such progressive cause-and-effect process helps to achieve the principle for image SR: high-level information can guide an LR image to recover a better SR image. In the proposed network, there are three indispensable parts

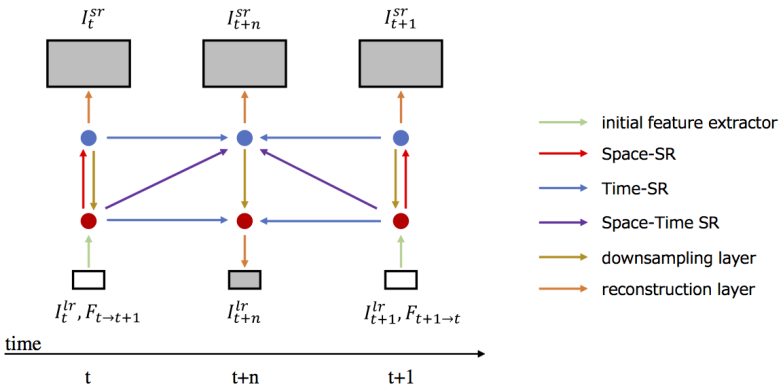


Fig. 12. Network proposed by TTI.

to enforce the suggested CFN: (1) tying the loss at each level (2) using LUM structure and (3) providing a lower level extracted feature input to ensure the availability of low-level information.

They propose to construct their network based on the Laplacian pyramid framework, as shown in Fig.11. Their model takes an LR image as input and progressively predicts residual images at $S_1, S_2...S_n$ levels where S is the scale factor, $S = S_1 \times S_2... \times S_n$. Normally, the $n = log_2^S$. For example, the network consists of 4 sub-networks for super-reconstructing an LR image at a scale factor of 16, if the scale factor is 3, $S = S_1 \times S_2$, $S_1 = 1.5, S2 = 2$. Their model has three branches: (1) feature extraction and (2) image reconstruction (3) loss function.

4.7 TTI

The team used base network for x16 video SR is STARnet [11] shown in Fig.12. With the idea that space and time are related, STARnet jointly optimizes three tasks (i.e., space SR, time SR, and space-time SR). In the experiments, STARnet was initially trained using three losses (i.e., space, time, and space-time losses, which evaluate the errors of images reconstructed through space SR paths (red arrows in the figure), time SR paths (blue arrows), and space-time SR paths (purple arrows), respectively), and then fine-tuned using only the space loss for optimizing the network model specialized for space SR. While space SR in STARnet is basically based on RBPN [12], this fine-tuning strategy allows them to superior to RBPN trained only with the space loss. While the original STARnet employs pyflow [24] for optical flow computation, pyflow almost cannot estimate optical flows in LR frames in this challenge (i.e., 120 $\times$ 67 pixels) because the optical flows are significantly small between subsequent frames. Based on an extensive survey, we chose sift-flow [23] that shows the better performance on the LR images used in this challenge.

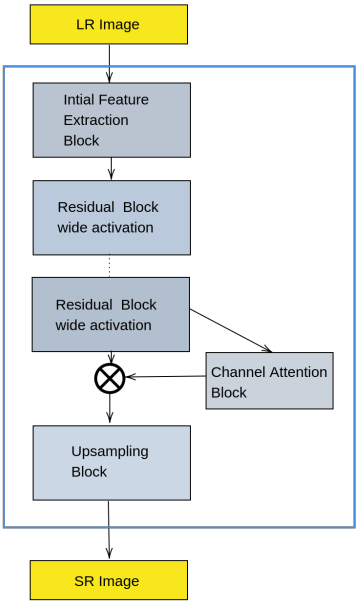


Fig. 13. Network proposed by CET-CVLab.

4.8 CET_CVLab

The architecture used by the team is inspired from wide activation based network by [44] and channel attention network by [46]. As shown in Fig.13, the network mainly consists of 3 blocks, a feature extraction block, series of wide activation residual blocks and a set of progressive upsampling blocks($\times 2$). Charbonnier loss is used for training the network as it better captures the edge information than with mean squared errorloss (MSE).

5 Challenge Results

5.1 Track 1 - Fidelity

This track aims at restoring the missing high frequencies that were lost during downsampling with the highest fidelity to the underlying ground truth. Challenge track 1 has 65 registered participants, from which 12 submitted solutions to the validation server and 8 teams entered the final ranking. The ranking, along with details about the training and testing are summarized in Table 1. Most provided solutions make use of large networks, as super-resolution with factor 16 is highly challenging and requires high-complexity networks in order to restore the details from learned priors. The top teams mostly employ window based approaches, attention modules and 3D-convolutions to additionally aggregate the temporal information. Team lyl and CET_CVLab do not process temporal information

	Method	↑PSNR	↑SSIM	Train Req	Train Time	Test Req	Test Time	Params	Extra Data
Participants	1. KirinUK	**22.83**	**0.6450**	4×V100	10d	1×2080Ti	6.1s	45.29M	Yes
	2. Team-WVU	22.48	0.6378	4×TitanXp	4d	1×TitanXp	4.90s	29.51M	No
	BOE-IOT-AIBD	22.48	0.6304	1×V100	> 30d	1× 1080	4.83s	53M	No
	4. sr_xxx	22.43	0.6353	8×V100	2d	1×V100	4s	n/a	No
	5. ZZX	22.28	0.6321	6×1080Ti	4d	1×1080Ti	4s	31,14M	No
	6. lyl	22.08	0.6256	1 × V100	2d	n/a	13s	n/a	No
	7. TTI	21.91	0.6165	V100	n/a	n/a	0.249s	n/a	No
	8. CET_CVLab	21.77	0.6112	1×P100	6d	1×P100	0.04s	n/a	Yes
	Bicubic (baseline)	20.69	0.5770						

Table 1. Quantitative results for track 1. Train Time: days per model, Test Time: seconds per frame.

into their networks and instead rely only on a single frame for the upscaling process. They can not compete with the top teams, which shows the importance of temporal information for high-quality restoration. The winner in track 1 is team KirinUK with a PSNR score of 22.83dB, followed by team Team-WVU and BOE-IOT-AIBD, which share the second place due to their identical PSNR scores.

Metrics Since this track is about high fidelity restoration, we rank the teams according to PSNR, which is a pixel-level metric. Additionally, we compute SSIM scores which is a metric based on patch statistics and is considered to correlate better with human perception of image quality. PSNR does not explicitly enforce to retain smooth temporal dynamics. It is therefore possible, that a method can generate high quality image quality on frame level, but introduces temporal artifacts like flickering. Most window based and 3D-convolution approaches however manage to produce frames with only minimal flickering artifacts, as they have access to adjacent frames. On the other hand, the flickering is very prominent for the single frame enhancers in this challenge.

Visual Results In addition to the metrics, we also provide visual examples in Fig. 2 for all competing methods in the challenge. We also show the Bicubic baseline (MATLAB's *imresize*) together with the ground truth frames for reference. All methods manage to clearly outperform the Bicubic baseline, which is also reflected in the PSNR and SSIM metrics in Table 1. The methods improve PSNR and SSIM by 1.08dB to 2.14dB and 0.0342 to 0.0680 respectively. As expected for such a challenging task, no method is capable of restoring all the fine details presented in the ground truth. Predominantly, only sharp edges and smooth textures can be recovered even by the top teams, since the information loss is so extreme. Interestingly, teams KirinUK, Team-WVU, BOE-IOT-AIBD, sr_xxx, ZZX and TTI restore two windows (blue box highlight, second column) instead of a single one as presented in the ground truth frame. This could indicate, the method's upscaling for such a large factor is highly dependent on image priors and having access to temporal information is not sufficient to achieve high restoration quality. Still, a top ranking is only achieved by teams that leverage temporal information, which shows its importance for video super-resolution even in such extreme settings.

	Method	K	T	Z	B (t)	s	B (s)	l	C	Wins (tot)	Wins (%)	↑PSNR	↑SSIM	↓LPIPS
Frame Level	1. KirinUK	-	115	126	116	113	117	126	144	**857**	**76.52**	**22.79**	**0.6474**	**0.447**
	2. Team-WVU	45	-	85	97	97	110	109	132	675	60.27	22.48	0.6378	0.507
	3. ZZX	34	75	-	78	97	101	104	130	619	55.27	22.09	0.6268	0.505
	4. BOE-IOT-AIBD (t)	44	63	82	-	68	87	95	118	557	49.73	21.18	0.3633	0.514
	5. sr_xxx	47	63	63	92	-	95	109	131	600	53.57	22.43	0.6353	0.509
	BOE-IOT-AIBD (s)	43	50	59	73	65	-	85	113	488	43.57	22.48	0.6304	0.550
	7. lyl	34	51	56	65	51	75	-	119	451	40.27	22.08	0.6256	0.535
	8. CET_CVLab	16	28	30	42	29	47	41	-	233	20.80	21.77	0.6112	0.602
												Final Scores		
												Frame	Video	Total
Video Level	1. KirinUK	-	7	7	8	7	8	8	6	**51**	**72.86**	76.52	72.86	**149.38**
	2. Team-WVU	3	-	7	6	8	2	8	8	42	60.00	60.27	60.00	120.27
	3. ZZX	3	3	-	5	5	4	9	8	37	52.86	55.27	52.86	108.13
	4. BOE-IOT-AIBD (t)	2	4	5	-	8	8	5	8	40	57.14	49.73	57.14	106.87
	5. sr_xxx	3	2	5	2	-	4	7	6	29	41.43	53.57	41.43	95.00
	BOE-IOT-AIBD (s)	2	8	6	2	6	-	5	7	36	51.43	43.57	51.43	95.00
	7. lyl	2	2	1	5	3	5	-	7	25	35.71	40.27	35.71	75.98
	8. CET_CVLab	4	2	2	2	4	3	3	-	20	28.57	20.80	28.57	49.37

Table 2. User study results for track 2. The results are obtained by a one vs. one user study on frame level and video level. Wins(tot) indicates absolute wins in all comparisons. Wins (%) reflects relative wins, which are normalized by the number of comparisons with other teams. Compared to absolute wins, the relative wins allow direct comparison between frame level and video level performance. The aggregated relative wins of both studies on frame and video levels led to the final ranking. Additionally, we provide PSNR, SSIM and LPIPS scores for reference. Note, these metrics are not considered for ranking.

5.2 Track 2 - Perceptual

Due to the extreme information loss, it is hard to accurately restore the high frequency content with respect to the ground truth. If deviations from the ground truth can be accepted and more visually pleasing results are desired, perceptual quality can be traded-off for fidelity. The results may not entirely reflect the underlying ground truth, but instead boost the perceived quality considerably. We therefore do not rely on PSNR and SSIM for evaluation in track 2, but instead conduct a user study to asses human perceived quality. Challenge track 2 has 54 registered participants, from which 7 submitted solutions to the validation server and 7 teams entered the final ranking

Metrics Assessing perceptual quality quantitatively is difficult in such a setting and remains largely an open problem. Attempts for such metrics have been made in the past and one of the most promising metrics is called Learned Perceptual Image Patch Similarity (LPIPS), which is proposed in [45]. This metric measures similarity to the ground truth in feature space of popular architectures [19]. While this metric is widely adopted for perceptual quality assessment, especially on images, it still fails in some cases to reflect human perception. Like PSNR and SSIM, it does not discriminate on temporal dynamics, which are crucial for high quality videos.

User Study Since quantitative metrics are not reliable, we resort to a user study to rank the participating teams in track 2. For that matter we split the

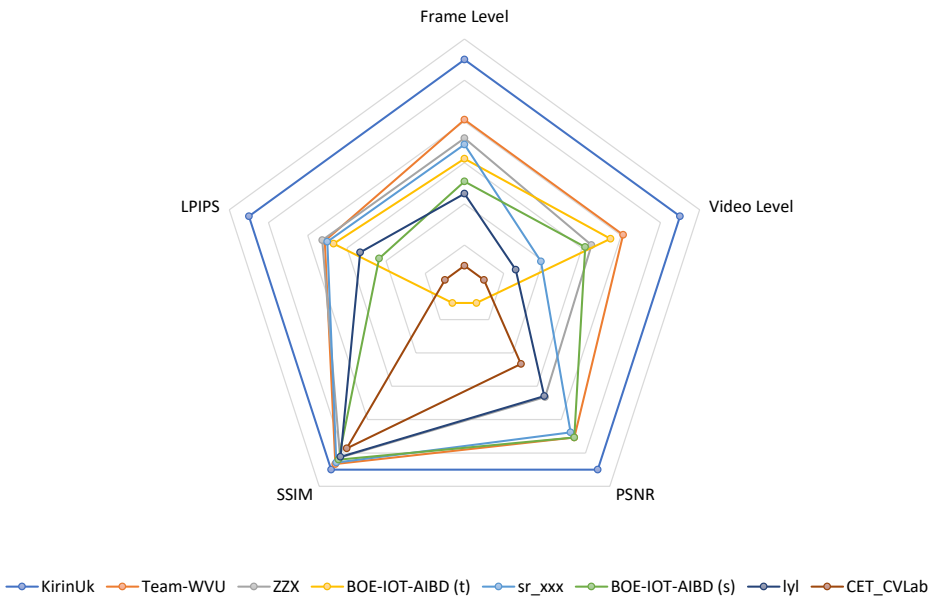


Fig. 14. Radar plot for track 2. All values per axis have been normalized to lie in [0,1]. The range for LPIPS has been reversed to indicate better values towards the outside in accordance with the other metrics.

evaluation in two separate user studies, a frame level study and a video level study. The frame level study is meant to judge the image level quality and is performed on randomly subsampled frames from all 16 sequences in the test set. The competing method's frames are compared side-by-side in a one-to-one setting, resulting in 28 comparisons per frame. We asked 10 users to judge the frame level quality. which results in $16 \times 28 \times 10 = 4480$ total ratings. The detailed results are shown in a confusion matrix in Table 2. Each row shows the preference of the method in the first column against all other methods. Additionally, we show the total number of preferences (Wins (tot)) plus the relative preferences, which are normalized to the number of total comparisons ($16 \times 7 \times 10 = 1120$) with all teams (Wins (%)). The video level study is meant for evaluating the temporal dynamics in the videos and the overall perceived quality when watching the videos. Again, we generated side-by-side videos between all methods for comparison in a one-to-one setting. The videos are generated by compiling all 1920 frames from 16 sequences into a single video, showing two competing methods. This results in 28 short videos of ≈ 1 minute. We also ask 10 users to perform the video level user study and get a total number of $28 \times 10 = 280$ ratings. In order to directly compare with the frame level study, we normalize the total wins by the total comparisons with all the teams ($10 \times 7 = 70$) to get the relative scores. We derive the ranking from the combined relative scores of both frame level and video level user studies (see Table 2, lower right). Team KirinUk is the clear winner in track 2, followed by Team-WVU and ZZX on the

second and third place. We also provide PSNR, SSIM and LPIPS metrics for reference. A qualitative illustration of all track 2 results is presented in Fig. 14, including the user study results. Note, the final ranking is only derived from the user study.

Visual Results To allow direct comparison to track 1, we provide the visual results on the same frames for track 2 in Fig. 3. Note that teams Team-WVU, sr_xxx, lyl and CET_CVLab submitted the same set of frames to both tracks, while KirinUK, BOE-IOT-AIBD and ZZX adapted their solutions to the specific requirements in both tracks. BOE-IOT-AIBD even provided two distinct solutions for track 2. One is optimized with emphasis on textures, the other is designed for temporal smoothness, abbreviated with (t) and (s) respectively in Table 2. Surprisingly, the texture based solution of BOE-IOT-AIBD also performs better in the video level user study. According to the users, the sharper texture details obviously have a higher impact on the quality than the flickering artifacts. The winning team KirinUK manages to not only outperform all other teams in both user studies, but also in the provided metrics PSNR, SSIM and LPIPS. However, it has to be considered, that the solution is optimized with L1 and VGG-loss, which are both closely related to these metrics. BOE-IOT-AIBD and KirinUK are the only teams that incorporate a GAN loss into their training strategy. Some other teams like Team-WVU trains its network only on the pixel-based Charbonnier Loss, and ZZX trains their perceptual solution on L1 and SSIM. Nevertheless, they outperform BOE-IOT-AIBD, which employs a GAN loss. Therefore, strong guidance from a pixel-based loss might be important for such an extreme scaling factor.

6 Conclusions

This paper presents the AIM 2020 challenge on Video Extreme Super-Resolution setup and results. We evaluate the performance in this challenging setting for both high fidelity restoration (track 1) and perceptual quality (track 2). The overall winner KirinUK manages to strike the best balance between restoration and perceptual quality. The participating teams provided innovative and diverse solutions to deal with the extreme upscaling factor of 16. Further improvements could be achieved by reducing and ideally removing the notorious flickering artifacts associated with video enhancement in general. On top of that, a more powerful generative setting could be designed for higher perceptual quality in track 2. Quantitative evaluation for perceptual quality still requires more research, especially in the video domain, where temporal consistency is important. We hope this challenge attracts more researchers to enter the area of extreme video super-resolution as it offers great opportunities for innovation.

Acknowledgements

We thank the AIM 2020 sponsors: Huawei, MediaTek and Google, and Computer Vision Lab (CVL), ETH Zurich.

Appendix A: Teams and Affiliations

AIM 2020 team
Title: AIM 2020 Challenge on Video Extreme Super-Resolution
Members: Dario Fuoli, Zhiwu Huang, Shuhang Gu, Radu Timofte
Affiliations:
Computer Vision Lab, ETH Zurich, Switzerland; The University of Sydney, Australia

KirinUK
Title: Efficient Video Enhancement and Super-Resolution Net (EVESRNet)
Members: Arnau Raventos, Aryan Esfandiari, Salah Karout
Affiliations:
Huawei Technologies R&D UK

Team-WVU
Title: Multi-Frame based Deformable Kernel Convolution Networks
Members: Xuan Xu, Xin Li, Xin Xiong, Jinge Wang
Affiliations:
West Virginia University, USA; Huazhong University of Science and Technology, China

BOE-IOT-AIBD
Title: Fully 3D–Convolutional MultiGrid–BackProjection Network
Members: Pablo Navarrete Michelini, Wenhao Zhang
Affiliations:
BOE Technology Group Co., Ltd

ZZX
Title: Multi-Scale Aggregated Upsampling Extreme Video Based on High Frequency Attention
Members: Dongyang Zhang, Hanwei Zhu, Dan Xia
Affiliations:
Jiangxi University of Finance and Economics; National Key Laboratory for Remanufacturing, Army Academy of Armored Forces

sr_xxx
Title: Residual Receptive Attention for Video Super-Resolution
Members: Haoyu Chen, Jinjin Gu, Zhi Zhang
Affiliations:
Amazon Web Services; The Chinese University of Hong Kong, Shenzhen

lyl
Title: Coarse to Fine Pyramid Networks for Progressive Image Super-Resolution
Members: Tongtong Zhao, Shanshan Zhao
Affiliations:

Dalian Maritime University; China Everbright Bank Co., Ltd

TTI

Title: STARnet
Members: Kazutoshi Akita, Norimichi Ukita
Affiliations:
Toyota Technological Institute (TTI)

CET_CVLab

Title: Video Extreme Super-Resolution using Progressive Wide Activation Net
Members: Hrishikesh P S, Densen Puthussery, Jiji C V
Affiliations:
College of Engineering, Trivandrum, India

References

1. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
2. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
3. Chen, J., Tan, X., Shan, C., Liu, S., Chen, Z.: Vesr-net: The winning solution to youku video enhancement and super-resolution challenge. arXiv preprint arXiv:2003.02115 (2020)
4. Dahl, R., Norouzi, M., Shlens, J.: Pixel recursive super resolution. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
5. Dai, Q., Yoo, S., Kappeler, A., Katsaggelos, A.K.: Sparse representation-based multiple frame video super-resolution. IEEE Transactions on Image Processing **26**(2), 765–781 (Feb 2017). https://doi.org/10.1109/TIP.2016.2631339
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (Feb 2016). https://doi.org/10.1109/TPAMI.2015.2439281
7. Fuoli, D., Gu, S., Timofte, R.: Efficient video super-resolution through recurrent latent space propagation. In: ICCV Workshops (2019)
8. Fuoli, D., Gu, S., Timofte, R., et al.: Aim 2019 challenge on video extreme super-resolution: Methods and results. In: ICCV Workshops (2019)
9. Fuoli, D., Huang, Z., Danelljan, M., Timofte, R., et al.: Ntire 2020 challenge on video quality mapping: Methods and results. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020)
10. Gao, H., Zhu, X., Lin, S., Dai, J.: Deformable kernels: Adapting effective receptive fields for object deformation. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SkxSv6VFvS
11. Haris, M., Shakhnarovich, G., Ukita, N.: Space-time-aware multi-resolution video enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2859–2868 (2020)
12. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3897–3906 (2019)
13. Huang, Y., Wang, W., Wang, L.: Bidirectional recurrent convolutional networks for multi-frame super-resolution. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. pp. 235–243. NIPS'15, MIT Press, Cambridge, MA, USA (2015), http://dl.acm.org/citation.cfm?id=2969239.2969266
14. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 466–467 (2020)
15. Jo, Y., Wug Oh, S., Kang, J., Joo Kim, S.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
16. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.: Video super-resolution with convolutional neural networks. In: IEEE Transactions on Computational Imaging (2016)

17. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
18. Kim, S., Li, G., Fuoli, D., Danelljan, M., Huang, Z., Gu, S., Timofte, R.: The vid3oc and intvid datasets for video super resolution and quality mapping. In: ICCV Workshops (2019)
19. Krizhevsky, A., Sutskever, I., E. Hinton, G.: Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems **25** (01 2012). https://doi.org/10.1145/3065386
20. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017)
21. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
22. Liao, R., Tao, X., Li, R., Ma, Z., Jia, J.: Video super-resolution via deep draft-ensemble learning. In: The IEEE International Conference on Computer Vision (ICCV) (December 2015)
23. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE transactions on pattern analysis and machine intelligence **33**(5), 978–994 (2010)
24. Liu, C., et al.: Beyond pixels: exploring new representations and applications for motion analysis. Ph.D. thesis, Massachusetts Institute of Technology (2009)
25. Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., Huang, T.: Robust video super-resolution with learned temporal dynamics. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
26. Liu, J., Zhang, W., Tang, Y., Tang, J., Wu, G.: Residual feature aggregation network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2359–2368 (2020)
27. Liu, X., Kong, L., Zhou, Y., Zhao, J., Chen, J.: End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 2416–2425 (2020)
28. Lucas, A., Lopez Tapia, S., Molina, R., Katsaggelos, A.K.: Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution. arXiv e-prints (Jun 2018)
29. Lugmayr, A., Danelljan, M., Timofte, R., Fritsche, M., Gu, S., Purohit, K., Kandula, P., Suin, M., Rajagopalan, A., Joon, N.H., et al.: AIM 2019 challenge on real-world image super-resolution: Methods and results. arXiv preprint arXiv:1911.07783 (2019)
30. Makansi, O., Ilg, E., Brox, T.: End-to-End Learning of Video Super-Resolution with Motion Compensation. arXiv e-prints (Jul 2017)
31. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Lee, K.M.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
32. Navarrete Michelini, P., Chen, W., Liu, H., Zhu, D.: MGBPv2: Scaling up multi–grid back–projection networks. In: The IEEE International Conference on Computer Vision Workshops (ICCVW) (October 2019), https://arxiv.org/abs/1909.12983

33. Navarrete Michelini, P., Liu, H., Zhu, D.: Multi–scale recursive and perception–distortion controllable image super–resolution. In: The European Conference on Computer Vision Workshops (ECCVW) (September 2018), http://arxiv.org/abs/1809.10711
34. Navarrete Michelini, P., Liu, H., Zhu, D.: Multigrid backprojection super–resolution and deep filter visualization. In: Proceedings of the Thirty–Third AAAI Conference on Artificial Intelligence (AAAI 2019). AAAI (2019)
35. Pérez-Pellitero, E., Sajjadi, M.S.M., Hirsch, M., Schölkopf, B.: Photorealistic Video Super Resolution. arXiv e-prints (Jul 2018)
36. Sajjadi, M.S.M., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
37. Sajjadi, M.S.M., Vemulapalli, R., Brown, M.: Frame-Recurrent Video Super-Resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
38. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR (2016)
39. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
40. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3360–3369 (2020)
41. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
42. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
43. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
44. Yu, J., Fan, Y., Yang, J., Xu, N., Wang, Z., Wang, X., Huang, T.: Wide activation for efficient and accurate image super-resolution. arXiv preprint arXiv:1808.08718 (2018)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
46. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018)
47. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2472–2481 (2018)
48. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9308–9316 (2019)