Computational Visual Media
https://doi.org/10.1007/s41095-020-0175-7 Vol. 6, No. 3, September 2020, 307–317
Research Article

iSeeBetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks

Aman Chadha[1] (✉), **John Britto**[2], **and M. Mani Roja**[3]

Abstract Recently, learning-based models have enhanced the performance of single-image super-resolution (SISR). However, applying SISR successively to each video frame leads to a lack of temporal coherency. Convolutional neural networks (CNNs) outperform traditional approaches in terms of image quality metrics such as peak signal to noise ratio (PSNR) and structural similarity (SSIM). On the other hand, generative adversarial networks (GANs) offer a competitive advantage by being able to mitigate the issue of a lack of finer texture details, usually seen with CNNs when super-resolving at large upscaling factors. We present iSeeBetter, a novel GAN-based spatio-temporal approach to video super-resolution (VSR) that renders temporally consistent super-resolution videos. iSeeBetter extracts spatial and temporal information from the current and neighboring frames using the concept of recurrent back-projection networks as its generator. Furthermore, to improve the "naturality" of the super-resolved output while eliminating artifacts seen with traditional algorithms, we utilize the discriminator from super-resolution generative adversarial network. Although mean squared error (MSE) as a primary loss-minimization objective improves PSNR/SSIM, these metrics may not capture fine details in the image resulting in misrepresentation of perceptual quality. To address this, we use a four-fold (MSE, perceptual, adversarial, and total-variation loss function. Our results demonstrate that iSeeBetter offers superior VSR fidelity and surpasses state-of-the-art performance.

Keywords super resolution; video upscaling; frame recurrence; optical flow; generative adversarial networks; convolutional neural networks

1 Introduction

The goal of super-resolution (SR) is to enhance a low resolution (LR) image to a higher resolution (HR) image by filling in missing fine-grained details in the LR image. The domain of SR research can be divided into three main areas: single image SR (SISR) [1–4], multi image SR (MISR) [5, 6], and video SR (VSR) [7–11].

Consider an LR video source which consists of a sequence of LR video frames LR_{t-n}, ..., LR_t, ..., LR_{t+n}, where we super-resolve a target frame LR_t. The idea behind SISR is to super-resolve LR_t by utilizing spatial information inherent in the frame, independently of other frames in the video sequence. However, this technique fails to exploit the temporal details inherent in a video sequence resulting in temporal incoherence. MISR seeks to address just that—it utilizes the missing details available from the neighboring frames LR_{t-n}, ..., LR_t, ..., LR_{t+n} and fuses them for super-resolving LR_t. After spatially aligning frames, missing details are extracted by separating differences between the aligned frames from missing details observed only in one or some of the frames. However, in MISR, the alignment of the frames is done without any concern for temporal smoothness, which is in stark contrast to VSR where the frames are typically aligned in temporal smooth order.

1 Department of Computer Science, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA. E-mail: aman@amanchadha.com, amanc@stanford.edu (✉).
2 Department of Computer Science, University of Massachusetts Amherst, Amherst, MA 01003, USA. E-mail: jnadar@umass.edu.
3 Department of Electronics and Telecommunication Engineering, University of Mumbai, Mumbai, Maharashtra 400032, India. E-mail: maniroja@tsec.edu.
Manuscript received: 2020-02-21; accepted: 2020-04-23

Traditional VSR methods upscale based on a single degradation model (usually bicubic interpolation) followed by reconstruction. This is sub-optimal and adds computational complexity [12]. Recently, learning-based models that utilize convolutional neural networks (CNNs) have outperformed traditional approaches in terms of widely-accepted image reconstruction metrics such as peak signal to noise ratio (PSNR) and structural similarity (SSIM).

In some recent VSR methods that utilize CNNs, frames are concatenated [11] or fed into recurrent neural networks (RNNs) [13] in temporal order, without explicit alignment. In other methods, the frames are aligned explicitly, using motion cues between temporal frames with the alignment modules [7–9, 14]. The latter set of methods generally render temporally smoother results compared to the methods with no explicit spatial alignment [13, 15]. However, these VSR methods suffer from a number of problems. In the frame-concatenation approach [7, 11, 14], many frames are processed simultaneously in the network, resulting in significantly higher network training time. With methods that use RNNs [8, 9, 13], modeling both subtle and significant changes simultaneously (e.g., slow and quick motions of foreground objects) is a challenging task even if long short-term memory units (LSTMs) are deployed, which are designed for maintaining long-term temporal dependencies [16]. A crucial aspect of an effective VSR system is the ability to handle motion sequences, which are often integral components of videos [7, 17].

The proposed method, iSeeBetter, is inspired by recurrent back-projection networks (RBPNs) [10] which utilize "back-projection" as their underpinning approach, originally introduced in Refs. [18, 19] for MISR. The basic concept behind back-projection is to iteratively calculate residual images as reconstruction error between a target image and a set of neighboring images. The residuals are then back-projected to the target image for improving super-resolution accuracy. The multiple residuals enable representation of subtle and significant differences between the target frame and its adjacent frames, thus exploiting temporal relationships between adjacent frames as shown in Fig. 1. Deep back-projection networks (DBPNs) [2] use back-projection to perform SISR using learning-based methods by estimating the output frame SR_t using the corresponding LR_t frame. To this end, DBPN produces a high-resolution feature map that is iteratively refined through multiple up- and down-sampling layers. RBPN offers superior results by combining the benefits of the original MISR back-projection approach with DBPN. Specifically, RBPN uses the idea of iteratively refining HR feature maps from DBPN, but extracts missing details using neighboring video frames like the original back-projection technique [18, 19]. This results in superior SR accuracy.

To mitigate the issue of a lack of finer texture details when super-resolving at large upscaling factors that is usually seen with CNNs [20], iSeeBetter utilizes GANs with a loss function that weighs adversarial loss, perceptual loss [20], mean square error (MSE)-based loss, and total-variation (TV) loss [21]. Our approach combines the merits of RBPN and SRGAN [20]—it is based on RBPN as its generator and is complemented by SRGAN's discriminator architecture, which is trained to differentiate between super-resolved images and original photo-realistic images. Blending these techniques yields iSeeBetter, a state-of-the-art system that is able to recover precise photo-realistic textures and motion-based scenes from heavily down-sampled videos.

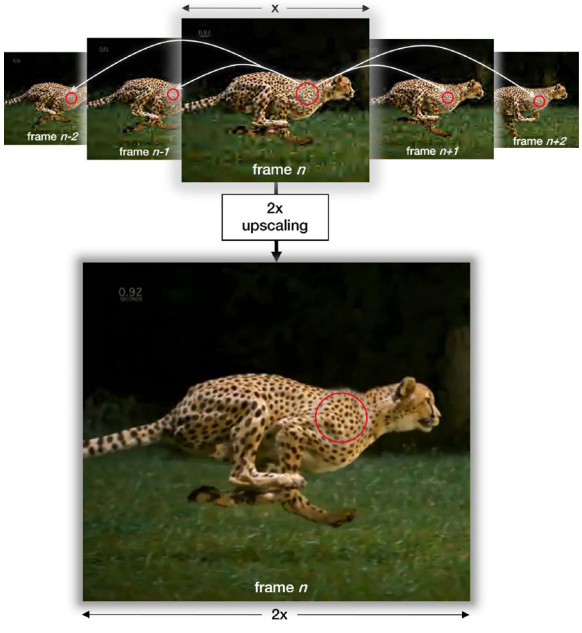


Fig. 1 Temporal relationships between adjacent frames.

Our contributions include the following key innovations.

Combining the state-of-the-art in SR: We propose a model that leverages two superior SR techniques—(i) RBPN, which is based on the idea of integrating SISR and MISR in a unified VSR framework using back-projection and, (ii) SRGAN, which is a framework capable of inferring photo-realistic natural images. RBPN enables iSeeBetter to extract details from neighboring frames, complemented by the generator–discriminator architecture in GANs which pushes iSeeBetter to generate more realistic and appealing frames while eliminating artifacts seen with traditional algorithms [22]. iSeeBetter thus yields more than the sum of the benefits of RBPN and SRGAN.

"Optimizing" the loss function: Pixel-wise loss functions such as L1 loss, used in RBPN [10], struggle to handle the uncertainty inherent in recovering lost high-frequency details such as complex textures that commonly exist in many videos. Minimizing MSE encourages finding pixel-wise averages of plausible solutions that are typically overly-smooth and thus have poor perceptual quality [23–26]. To address this, we adopt a four-fold (MSE, perceptual, adversarial, and TV) loss function for superior results. Similar to SRGAN [20], we utilize a loss function that optimizes perceptual quality by minimizing adversarial loss and content loss. Adversarial loss helps improve the "naturality" associated with the output image using the discriminator. On the other hand, the content loss focuses on optimizing perceptual similarity instead of similarity in pixel space. Furthermore, we use the MSE loss and a de-noising loss function called TV loss [21]. We carried out experiments comparing L1 loss with our four-fold loss and found significant improvements with the latter.

Extended evaluation protocol: To evaluate iSeeBetter, we used standard datasets: Vimeo90K [27], Vid4 [28], and SPMCS [8]. Since Vid4 and SPMCS lack significant motion sequences, we included Vimeo90K, a dataset containing various types of motion. This enabled us to conduct a more holistic evaluation of the strengths and weaknesses of iSeeBetter. To make iSeeBetter more robust and enable it to handle real-world videos, we expanded the spectrum of data diversity and wrote scripts to collect additional data from YouTube. As a result, we augmented our dataset to about 170,000 clips.

User-friendly infrastructure: We built several useful tools to download and structure datasets, visualize temporal profiles of intermediate blocks and the output, and run predefined benchmark sequences on a trained model to be able to iterate on different models quickly. In addition, we built a video-to-frames tool to directly input videos to iSeeBetter, rather than frames. We also ensured our script infrastructure is flexible (such that it supports a myriad of options) and can be easily leveraged for extending this work. The code and pre-trained models are available at `https://iseebetter.amanchadha.com`.

2 Related work

Since the seminal work by Tsai on image registration [29] two decades ago, many SR techniques based on various underlying principles have been proposed. Initial methods included spatial or frequency domain signal processing, statistical models, and interpolation approaches [30]. In this section, we focus our discussion on learning-based methods which have emerged as superior VSR techniques compared to traditional statistical methods.

2.1 Deep SISR

First introduced by SRCNN [1], deep SISR required a predefined up-sampling operator. Further improvements in this field include better up-sampling layers [12], residual learning [31], back-projection [2], recursive layers [32], and progressive up-sampling [33]. A significant milestone in SR research was the introduction of a GAN-powered SR approach [20], which achieved state-of-the-art performance.

2.2 Deep VSR

Deep VSR can be primarily divided into five types based on the approach to preserving temporal information.

(a) Temporal concatenation. The most popular approach to retain temporal information in VSR is concatenating multiple frames [7, 11, 15, 34]. This approach can be seen as an extension of SISR to accept multiple input images. VSR-DUF [11] proposed a mechanism to construct up-sampling filters and residual images. However, this approach fails to represent multiple motion regimes within a single input sequence since the input frames are

concatenated together.

(b) Temporal aggregation. To address the dynamic motion problem in VSR, Ref. [14] proposed multiple SR inferences which work on different motion regimes. The final layer aggregates the outputs of all branches to construct SR frame. However, this approach still concatenates many input frames, resulting in lengthy convergence during global optimization.

(c) Recurrent networks. RNNs deal with temporal inputs and/or outputs and have been deployed for a myriad of applications ranging from video captioning [35–37], video summarization [38, 39], and VSR [8, 9, 13]. Two types of RNN have been used for VSR. A many-to-one architecture is used in Refs. [8, 13] where a sequence of LR frames is mapped to a single target HR frame. A many-to-many RNN has recently been used by Ref. [9] to map the last and the current LR frame to an optical flow network, which feeds an SR network along with the previous HR estimate. This approach was first proposed by Ref. [13] using bidirectional RNNs. However, the network has a small network capacity and has no frame alignment step. Further improvement is proposed by Ref. [8] using a motion compensation module and a ConvLSTM layer [40].

(d) Optical flow-based methods. The above methods estimate a single HR frame by combining a batch of LR frames and are thus computationally expensive. They often result in unwanted flickering artifacts in the output frames [21]. To address this, Ref. [9] proposed a method that utilizes a network trained on estimating the optical flow along with the SR network. Optical flow methods allow estimation of the trajectories of moving objects, thereby assisting in VSR. Ref. [34] warps video frames LR_{t-1} and LR_{t+1} onto LR_t using the optical flow method of Ref. [41], concatenates the three frames, and passes them through a CNN that produces the output frame SR_{t+1}. Ref. [7] follows the same approach but replaces the optical flow model with a trainable motion compensation network.

(e) Pre-training then fine-tuning v/s end-to-end training. While most of the above-mentioned methods are end-to-end trainable, certain approaches first pre-train each component before fine-tuning the system as a whole in a final step [7, 8, 14].

Our approach is a combination of (i) an RNN-based optical flow method that preserves spatio-temporal information in the current and adjacent frames as the generator and, (ii) a discriminator that is adept at ensuring the generated SR frame offers superior fidelity.

3 Methods

3.1 Datasets

To train iSeeBetter, we amalgamated diverse datasets with differing video lengths, resolutions, motion sequences, and number of clips. Table 1 presents a summary of the datasets used. When training our model, we generated the corresponding LR frame for each HR input frame by performing 4× down-sampling using bicubic interpolation. We thus perform self-supervised learning by automatically generating the input-output pairs for training without any human intervention. We also applied data augmentation techniques such as rotation, flipping, and random cropping. To further extend our dataset, we wrote scripts to collect additional data from YouTube. The dataset was shuffled for training and testing. Our training/validation/test split was 80%/10%/10%.

3.2 Network architecture

Figure 2 shows the iSeeBetter architecture that consists of RBPN [10] and SRGAN [20] as its generator and discriminator respectively. Table 2 shows the adopted notation. RBPN has two approaches that extract missing details from different sources, namely SISR and MISR. Figure 3 shows the

Table 1 Datasets used for training and evaluation

Dataset	Resolution	# of clips	# of frames/clip	# of frames
Vimeo90K	448 × 256	13,100	7	91,701
SPMCS	240 × 135	30	31	930
Vid4	(720 × 576 × 3) × 2, (704 × 576 × 3) × 2	4	41, 34, 49, 47	684
Augmented	960 × 720	7000	110	77,000
Total	—	**46,034**	—	**170,315**

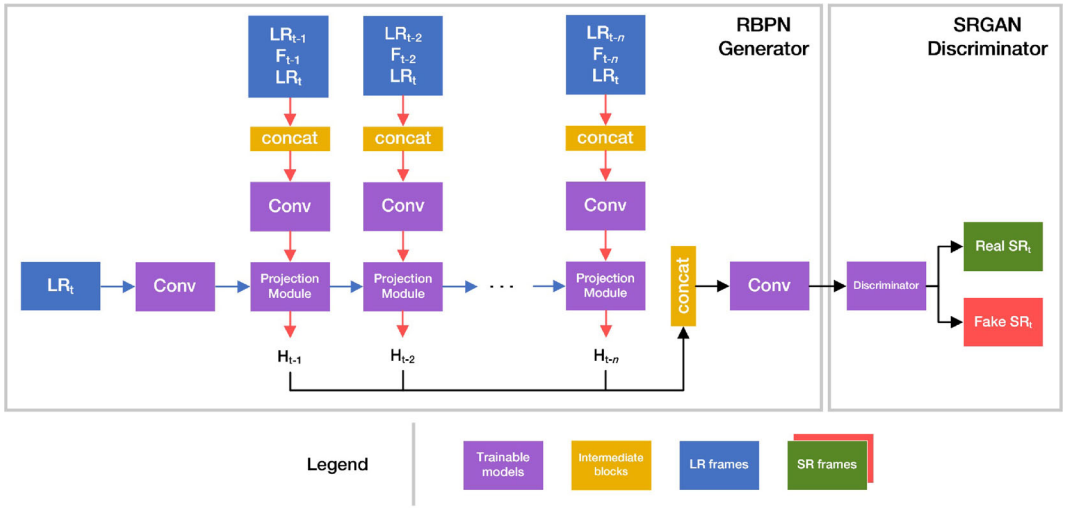


Fig. 2 Overview of iSeeBetter.

Table 2 Adopted notation

HR_t	input high resolution image
LR_t	low resolution image (derived from HR_t)
F_t	optical flow output
H_{t-1}	residual features extracted from (LR_{t-1}, F_{t-1}, LR_t)
SR_t	estimated HR output

horizontal flow (represented by blue arrows in Fig. 2) that enlarges LR_t using SISR. Figure 4 shows the vertical flow (represented by red arrows in Fig. 2) which is based on MISR that computes residual features from a pair of LR_t and its neighboring frames (LR_{t-1}, ..., LR_{t-n}) coupled with the pre-computed dense motion flow maps (F_{t-1}, ..., F_{t-n}). At each projection step, RBPN observes the missing details from LR_t and extracts residual features from neighboring frames to recover details. Within the projection models, RBPN utilizes a recurrent encoder–decoder mechanism for fusing details extracted from adjacent frames in SISR and MISR and incorporates them into the estimated frame SR_t through back-projection. Once an SR frame is synthesized, it is sent over to the discriminator (shown in Fig. 5) to validate its "authenticity".

3.3 Loss functions

Perceptual image quality of the resulting SR image is dependent on the choice of the loss function. To evaluate the quality of an image, MSE is the most commonly used loss function in a wide variety of state-

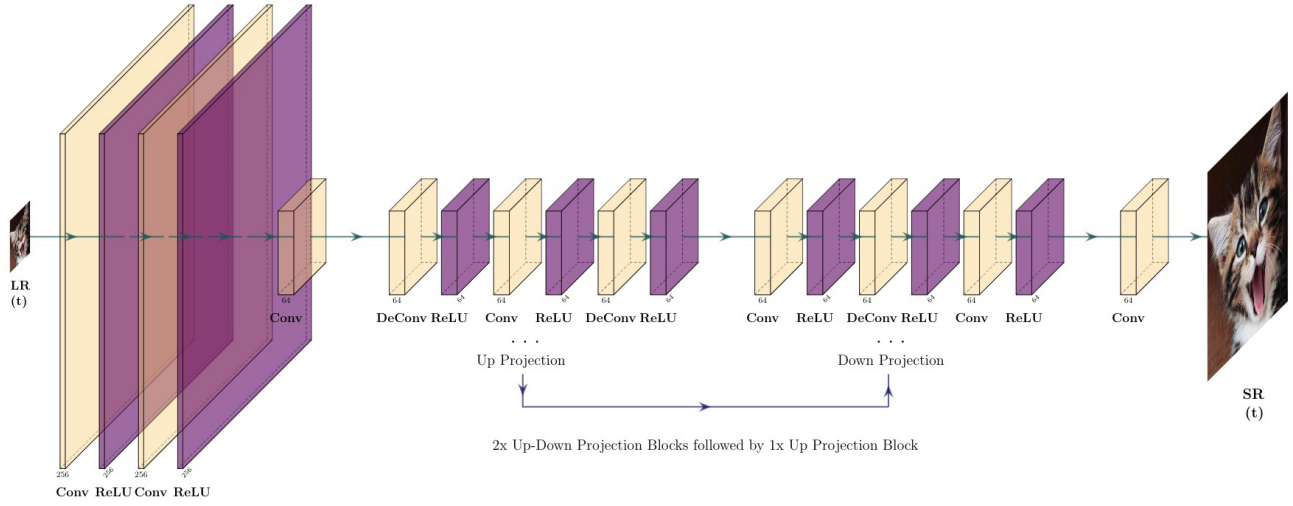


Fig. 3 DBPN [2] architecture for SISR, where we perform up–down–up sampling using 8 × 8 kernels with a stride of 4 and padding of 2. Similar to the ResNet architecture above, the DBPN network also uses Parametric ReLUs [42] as its activation functions.

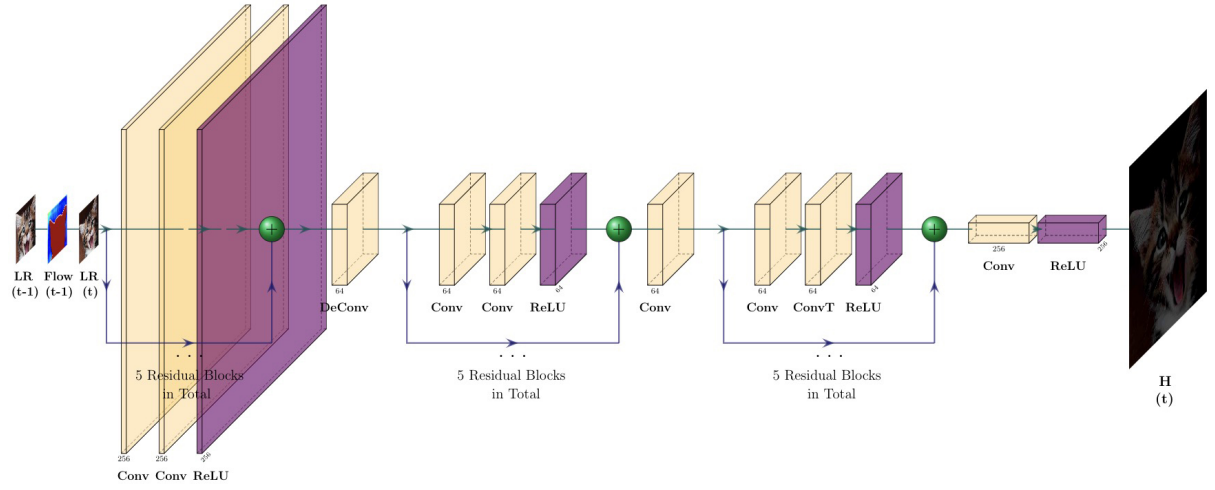


Fig. 4 ResNet architecture for MISR that is composed of three tiles of five blocks where each block consists of two convolutional layers with 3×3 kernels, a stride of 1 and padding of 1. The network uses Parametric ReLUs [42] for its activations.

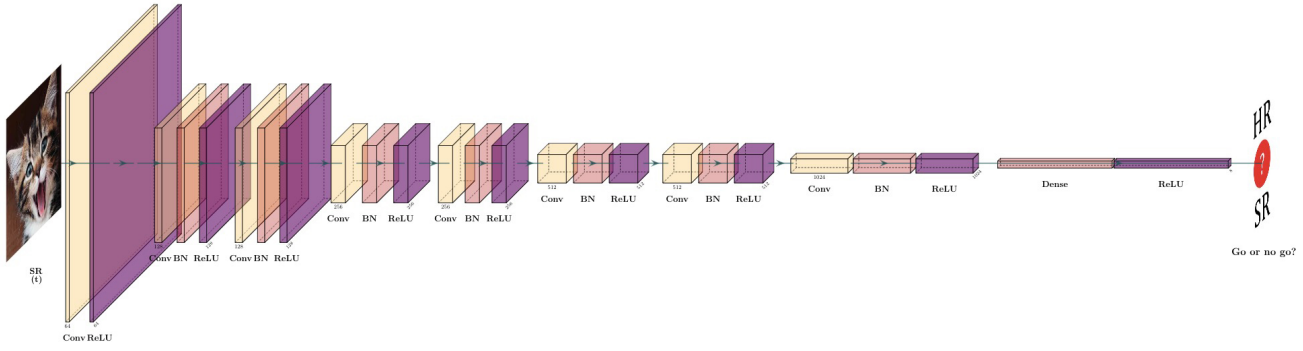


Fig. 5 Discriminator architecture from SRGAN [20]. The discriminator uses Leaky ReLUs for computing its activations.

of-the-art SR approaches, which aims to improve the PSNR of an image [43]. While optimizing MSE during training improves PSNR and SSIM, these metrics may not capture fine details in the image leading to misrepresentation of perceptual quality [20]. The ability of MSE to capture intricate texture details based on pixel-wise frame differences is very limited, and can cause the resulting video frames to be overly-smooth [44]. In a series of experiments, it was found that even manually distorted images had an MSE score comparable to the original image [45]. To address this, iSeeBetter uses a four-fold (MSE, perceptual, adversarial, and TV) loss instead of solely relying on pixel-wise MSE loss. We weigh these losses together as a final evaluation standard for training iSeeBetter, thus taking into account both pixel-wise similarities and high-level features when judging the quality of the SR images. Figure 6 shows the individual components of the iSeeBetter loss function.

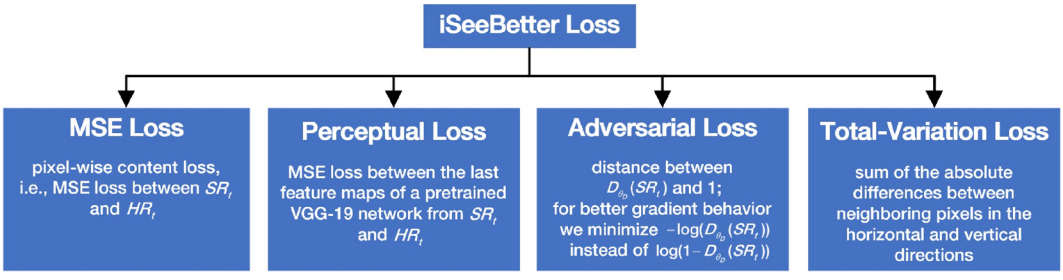


Fig. 6 MSE, perceptual, adversarial, and TV loss components of the iSeeBetter loss function.

3.3.1 MSE loss

We use pixel-wise MSE loss (also called content loss) for the estimated frame $\mathrm{SR_t}$ against the ground truth $\mathrm{HR_t}$.

$$MSE_{\mathrm{t}} = \frac{1}{WH}\sum_{x=0}^{W}\sum_{y=0}^{H}((HR_{\mathrm{t}})_{x,y} - G_{\theta_G}(LR_{\mathrm{t}})_{x,y})^2 \quad (1)$$

where $G_{\theta_G}(LR_{\mathrm{t}})$ is the estimated frame $\mathrm{SR_t}$. W and H represent the width and height of the frames respectively.

3.3.2 Perceptual loss

Refs. [26, 46] introduced a new loss function called perceptual loss, also used in Refs. [20, 24], which focuses on perceptual similarity instead of similarity in pixel space. Perceptual loss relies on features extracted from the activation layers of the pre-trained VGG-19 network in Ref. [47], instead of low-level pixel-wise error measures. We define perceptual loss as the euclidean distance between the feature representations of the estimated SR image $G_{\theta_G}(LR_{\mathrm{t}})$ and the ground truth $\mathrm{HR_t}$.

$$PerceptualLoss_{\mathrm{t}} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}\left(\begin{array}{l} VGG_{i,j}(HR_{\mathrm{t}})_{x,y} - \\ VGG_{i,j}(G_{\theta_G}(LR_{\mathrm{t}}))_{x,y}\end{array}\right)^2 \quad (2)$$

where $VGG_{i,j}$ denotes the feature map obtained by the j^{th} convolution (after activation) before the i^{th} maxpooling layer in the VGG-19 network. $W_{i,j}$ and $H_{i,j}$ are the dimensions of the respective feature maps in the VGG-19 network.

3.3.3 Adversarial loss

Similar to Ref. [20], we use the generative component of iSeeBetter as the adversarial loss to limit model "fantasy", thus improving the "naturality" associated with the super-resolved image. Adversarial loss is defined as

$$AdversarialLoss_{\mathrm{t}} = -\log(D_{\theta_D}(G_{\theta_G}(LR_{\mathrm{t}})) \quad (3)$$

where $D_{\theta_D}(G_{\theta_G}(LR_{\mathrm{t}}))$ is the discriminator's output probability that the reconstructed image $G_{\theta_G}(LR_{\mathrm{t}})$ is a real HR image. We minimize $-\log(D_{\theta_D}(G_{\theta_G}(LR_{\mathrm{t}}))$ instead of $\log(1-D_{\theta_D}(G_{\theta_G}(LR_{\mathrm{t}}))$ for better gradient behavior [48].

3.3.4 Total-variation loss

TV loss was introduced as a loss function in the domain of SR by Ref. [49]. It is defined as the sum of the absolute differences between neighboring pixels in the horizontal and vertical directions [22]. Since TV loss measures noise in the input, minimizing it as part of our overall loss objective helps de-noise the output SR image and thus encourages spatial smoothness. TV loss is defined as follows:

$$TVLoss_{\mathrm{t}} = \frac{1}{WH}\sum_{i=0}^{W}\sum_{j=0}^{H}\sqrt{\begin{array}{l}(G_{\theta_G}(LR_{\mathrm{t}})_{i,j+1,k} - G_{\theta_G}(LR_{\mathrm{t}})_{i,j,k})^2 + \\ (G_{\theta_G}(LR_{\mathrm{t}})_{i+1,j,k} - G_{\theta_G}(LR_{\mathrm{t}})_{i,j,k})^2\end{array}} \quad (4)$$

3.3.5 Loss formulation

We define our overall loss objective for each frame as the weighted sum of the MSE, adversarial, perceptual, and TV loss components:

$$\begin{aligned} Loss_{G_{\theta_G}}(SR_{\mathrm{t}}) = {} & \alpha \times MSE(SR_{\mathrm{t}}, HR_{\mathrm{t}}) \\ & - \beta \times AdversarialLoss(SR_{\mathrm{t}}) \\ & + \gamma \times PerceptualLoss(SR_{\mathrm{t}}, HR_{\mathrm{t}}) \\ & + \delta \times TVLoss(SR_{\mathrm{t}}, HR_{\mathrm{t}}) \end{aligned} \quad (5)$$

where α, β, γ, δ are weights set as 1, 10^{-3}, 6×10^{-3}, and 2×10^{-8} respectively [50].

The discriminator loss for each frame is as follows:

$$Loss_{D_{\theta_D}}(SR_{\mathrm{t}}) = 1 - D_{\theta_D}(HR_{\mathrm{t}}) + D_{\theta_D}(SR_{\mathrm{t}}) \quad (6)$$

The total loss of an input sample is the average loss of all frames.

$$\begin{cases} Loss_{G_{\theta_G}} = \frac{1}{N}\sum_{t=1}^{N}(Loss_{G_{\theta_G}}(SR_{\mathrm{t}})) \\ Loss_{D_{\theta_D}} = \frac{1}{N}\sum_{t=1}^{N}(Loss_{D_{\theta_D}}(SR_{\mathrm{t}})) \end{cases} \quad (7)$$

4 Experimental evaluation

To train the model, we used an Amazon EC2 P3.2xLarge instance with an NVIDIA Tesla V100 GPU with 16 GB VRAM, 8 vCPUs, and 64 GB of host memory. We used the hyperparameters from RBPN and SRGAN. Table 3 compares iSeeBetter with six state-of-the-art VSR algorithms: DBPN [2], $\mathrm{B_{123}}$ + T [14], DRDVSR [8], FRVSR [9], VSR-DUF [11], and RBPN/6-PF [10]. Table 4 offers a visual analysis of VSR-DUF and iSeeBetter. Table 5 shows ablation studies to assess the impact of using a generator discriminator architecture and the four-fold loss as design decisions.

5 Conclusions and future work

We proposed iSeeBetter, a novel spatio-temporal approach to VSR that uses recurrent-generative back-

Table 3 SNR/SSIM evaluation of state-of-the-art VSR algorithms using Vid4 and Vimeo90K for 4× upscaling. Bold numbers indicate best performance

Dataset	Clip name	Flow	Bicubic	DBPN [2]	$B_{123}+T$ [14]	DRDVSR [8]	FRVSR [9]	RBPN/6-PF [10]	VSR-DUF [11]	**iSeeBetter**
Vid4	Calendar	1.14	19.82/0.554	22.19/0.714	21.66/0.704	22.18/0.746	—	23.99/0.807	24.09/0.813	**24.13/0.817**
	City	1.63	24.93/0.586	26.01/0.684	26.45/0.720	26.98/0.755	—	27.73/0.803	28.26/0.833	**28.34/0.841**
	Foliage	1.48	23.42/0.575	24.67/0.662	24.98/0.698	25.42/0.720	—	26.22/0.757	26.38/0.771	**26.57/0.773**
	Walk	1.44	26.03/0.802	28.61/0.870	28.26/0.859	28.92/0.875	—	**30.70/0.909**	30.50/0.912	30.68/0.908
Average		1.42	23.53/0.629	25.37/0.737	25.34/0.745	25.88/0.774	26.69/0.822	27.12/0.818	27.31/0.832	**27.43/0.835**
Vimeo90K	Fast Motion	8.30	34.05/0.902	37.46/0.944	—	—	—	40.03/0.960	37.49/0.949	**40.17/0.971**

Table 4 Visually inspecting examples from Vid4, SPMCS, and Vimeo-90k comparing VSR-DUF and iSeeBetter. We chose VSR-DUF for comparison because it was the state-of-the-art at the time of publication. Top row: fine-grained textual features that help with readability; middle row: intricate high-frequency image details; bottom row: camera panning motion

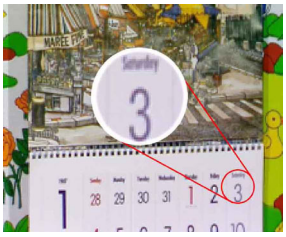
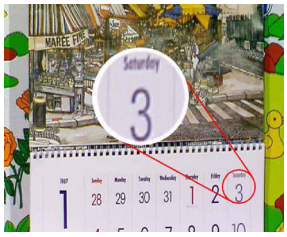
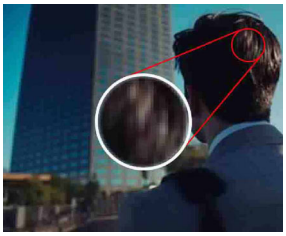
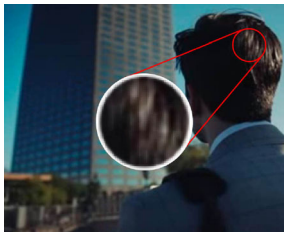
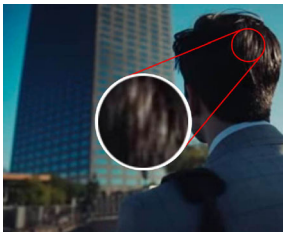
Dataset	Clip name	VSR-DUF [11]	**iSeeBetter**	Ground truth
Vid4	Calendar			
SPMCS	Pagoda			
Vimeo90K	Motion			

Table 5 Ablation analysis for iSeeBetter using the "City" clip from Vid4

iSeeBetter Config	PSNR
RBPN baseline with L1 loss	27.73
RBPN baseline with MSE loss	27.77
RBPN generator + SRGAN discriminator with adversarial loss	28.08
RBPN generator + SRGAN discriminator with adversarial + MSE loss	28.12
RBPN generator + SRGAN discriminator with adversarial + MSE + perceptual loss	28.27
RBPN generator + SRGAN discriminator with adversarial + MSE + perceptual + TV loss	28.34

projection networks. iSeeBetter couples the virtues of RBPN and SRGAN. RBPN enables iSeeBetter to generate superior SR images by combining spatial and temporal information from the input and neighboring frames. In addition, SRGAN's discriminator architecture fosters generation of photo-realistic frames. We used a four-fold loss function that emphasizes perceptual quality. Furthermore, we proposed a new evaluation protocol for video SR by collating diverse datasets. With extensive experiments, we assessed the role played by various design choices in the ultimate performance of

iSeeBetter, and demonstrated that on a vast majority of test video sequences, iSeeBetter advances the state-of-the-art.

To improve iSeeBetter, a couple of ideas could be explored. In visual imagery the foreground receives much more attention than the background since it typically includes subjects such as humans. To improve perceptual quality, we can segment the foreground and background, and make iSeeBetter perform "adaptive VSR" by utilizing different policies for the foreground and background. For instance, we could adopt a wider span of the number of frames to extract details from for the foreground compared to the background. Another idea is to decompose a video sequence into scenes on the basis of frame-similarity and make iSeeBetter assign weights to adjacent frames based on which scene they belong to. Adjacent frames from a different scene can be weighed lower compared to frames from the same scene, thereby making iSeeBetter focus on extracting details from frames within the same scene—a la the concept of attention applied to VSR.

Acknowledgements

The author would like to thank Andrew Ng's lab at Stanford University for their guidance on this project. In particular, the authors express their gratitude to Mohamed El-Geish for the idea-inducing brainstorming sessions throughout the project.

References

[1] Dong, C.; Loy, C. C.; He, K. M.; Tang, X. O. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 38, No. 2, 295–307, 2016.

[2] Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1664–1673, 2018.

[3] Haris, M.; Widyanto, M. R.; Nobuhara, H. Inception learning super-resolution. *Applied Optics* Vol. 56, No. 22, 6043, 2017.

[4] Kim, J.; Lee, J. K.; Lee, K. M. Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1646–1654, 2016.

[5] Faramarzi, E.; Rajan, D.; Christensen, M. P. Unified blind method for multi-image super-resolution and single/multi-image blur deconvolution. *IEEE Transactions on Image Processing* Vol. 22, No. 6, 2101–2114, 2013.

[6] Garcia, D. C.; Dorea, C.; de Queiroz, R. L. Super resolution for multiview images using depth information. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 22, No. 9, 1249–1256, 2012.

[7] Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z. H.; Shi, W. Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4778–4787, 2017.

[8] Tao, X.; Gao, H. Y.; Liao, R. J.; Wang, J.; Jia, J. Y. Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, 4472–4480, 2017.

[9] Sajjadi, M. S. M.; Vemulapalli, R.; Brown, M. Frame-recurrent video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6626–6634, 2018.

[10] Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3897–3906, 2019.

[11] Jo, Y.; Oh, S. W.; Kang, J.; Kim, S. J. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3224–3232, 2018.

[12] Shi, W. Z.; Caballero, J.; Huszar, 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1874–1883, 2016.

[13] Huang, Y.; Wang, W.; Wang, L. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In: Proceedings of the Advances in Neural Information Processing Systems 28, 235–243, 2015.

[14] Liu, D.; Wang, Z. W.; Fan, Y. C.; Liu, X. M.; Wang, Z. Y.; Chang, S. Y.; Huang, T. Robust video super-resolution with learned temporal dynamics. In: Proceedings of the IEEE International Conference on Computer Vision, 2507–2515, 2017.

[15] Liao, R. J.; Tao, X.; Li, R. Y.; Ma, Z. Y.; Jia, J. Y. Video super-resolution via deep draft-ensemble learning. In: Proceedings of the IEEE International Conference on Computer Vision, 531–539, 2015.

[16] Gers, F. A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Computation* Vol. 12, No. 10, 2451–2471, 2000.

[17] Makansi, O.; Ilg, E.; Brox, T. End-to-end learning of video super-resolution with motion compensation. In: *Pattern Recognition. Lecture Notes in Computer Science, Vol. 10496*. Roth, V.; Vetter, T. Eds. Springer Cham, 203–214, 2017.

[18] Irani, M.; Peleg, S. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing* Vol. 53, No. 3, 231–239, 1991.

[19] Irani, M.; Peleg, S. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation* Vol. 4, No. 4, 324–335, 1993.

[20] Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4681–4690, 2017.

[21] Ren, H.; Fang, X. Recurrent back-projection network for video super-resolution. In: Final Project for MIT 6.819 Advances in Computer Vision, 1–6, 2018.

[22] Wang, Z. H.; Chen, J.; Hoi, S. C. H. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* DOI: 10.1109/TPAMI.2020.2982166, 2020.

[23] Mathieu, M.; Couprie, C.; LeCun, Y. Deep multi-scale video prediction beyond mean square error. *arXiv preprint* arXiv:1511.05440, 2015.

[24] Johnson, J.; Alahi, A.; Li, F. F. Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9906*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 694–711, 2016.

[25] Dosovitskiy, A.; Brox, T. Generating images with perceptual similarity metrics based on deep networks. In: Proceedings of the Advances in Neural Information Processing Systems 29, 658–666, 2016.

[26] Bruna, J.; Sprechmann, P.; LeCun, Y. Super-resolution with deep convolutional sufficient statistics. In: Proceedings of the 4th International Conference on Learning Representations, 2016.

[27] Xue, T. F.; Chen, B. A.; Wu, J. J.; Wei, D. L.; Freeman, W. T. Video enhancement with task-oriented flow. *International Journal of Computer Vision* Vol. 127, No. 8, 1106–1125, 2019.

[28] Liu, C.; Sun, D. Q. A Bayesian approach to adaptive video super resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 209–216, 2011.

[29] Tsai, R. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing* Vol. 1, 317–339, 1984.

[30] Yang, J. C.; Huang, T. Image super-resolution: Historical overview and future challenges. In: *Super-Resolution Imaging*. Milanfar, P. Ed. CRC Press, 1–34, 2017.

[31] Tai, Y.; Yang, J.; Liu, X. M. Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3147–3155, 2017.

[32] Kim, J.; Lee, J. K.; Lee, K. M. Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1637–1645, 2016.

[33] 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, 624–632, 2017.

[34] Kappeler, A.; Yoo, S.; Dai, Q. Q.; Katsaggelos, A. K. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging* Vol. 2, No. 2, 109–122, 2016.

[35] Johnson, J.; Karpathy, A.; Li, F. F. DenseCap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4565–4574, 2016.

[36] Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint* arXiv:1412.6632, 2014.

[37] Yu, H. N.; Wang, J.; Huang, Z. H.; Yang, Y.; Xu, W. Video paragraph captioning using hierarchical recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4584–4593, 2016.

[38] Donahue, J.; Hendricks, L. A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Darrell, T.; Saenko, K. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2625–2634, 2015.

[39] Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; Saenko, K. Translating videos to natural language using deep recurrent neural networks In: Proceedings of the Annual Conference of the North American Chapter of the ACL, 1494–1504, 2015.

[40] Shi, X.; Chen, Z.; Wang, H.; Yeung, D.; Wong, W.; Woo, W. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In:

Proceedings of the Advances in Neural Information Processing Systems 28, 1–9, 2015.

[41] Drulea, M.; Nedevschi, S. Total variation regularization of local-global optical flow. In: Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems, 318–323, 2011.

[42] He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, 1026–1034, 2015.

[43] Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In: Proceedings of the 20th International Conference on Pattern Recognition, 2366–2369, 2010.

[44] Cheng, M.-H.; Lin, N.-W.; Hwang, K.-S.; Jeng, J.-H. Fast video super-resolution using artificial neural networks. In: Proceedings of the 8th International Symposium on Communication Systems, Networks & Digital Signal Processing, 1–4, 2012.

[45] Wang, Z.; Bovik, A. C. A universal image quality index. *IEEE Signal Processing Letters* Vol. 9, No. 3, 81–84, 2002.

[46] Gatys, L.; Ecker, A. S.; Bethge, M. Texture synthesis using convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems 28, 262–270, 2015.

[47] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. a*rXiv preprint* arXiv:1409.1556, 2014.

[48] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In: Proceedings of the Advances in Neural Information Processing Systems 27, 2672–2680, 2014.

[49] Aly, H. A.; Dubois, E. Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing* Vol. 14, No. 10, 1647–1659, 2005.

[50] Hany, J.; Walters, G. Hands-On Generative Adversarial Networks with PyTorch 1. x: Implement next-generation neural networks to build powerful GAN models using Python. Packt Publishing Ltd., 2019.

Aman Chadha has held positions at some of the world's leading semiconductor/product companies. He is currently based out of Cupertino (Silicon Valley), California and is currently pursuing his graduate studies in artificial intelligence from Stanford University. He has published in prestigious international journals and conferences, and has authored two books. His publications have garnered about 200 citations. He currently serves on the editorial boards of several international journals including IJATCA, IJLTET, IJCET, IJEACS, and IJRTER. He has served as a reviewer for IJEST, IJCST, IJCSEIT, and JESTEC. Aman graduated with an M.S. degree from the University of Wisconsin-Madison with an outstanding graduate student award in 2014 and his B.E. degree with distinction from the University of Mumbai in 2012. His research interests include computer vision (particularly, pattern recognition), artificial intelligence, machine learning and computer architecture. Aman has 18 publications to his credit.

John Britto is pursuing his M.S. degree in computer science from the University of Massachusetts, Amherst. He completed his B.E. degree in computer engineering from the University of Mumbai in 2018. His research interests lie in machine learning, natural language processing, and artificial intelligence.

M. Mani Roja is a full professor in the Electronics and Telecommunication Department at the University of Mumbai since the past 30 years. She received her Ph.D. degree in electronics and telecommunication engineering from Sant Gadge Baba Amravati University and her master degree in electronics and telecommunication engineering from the University of Mumbai. She has collaborated across the years in the fields of image processing, speech processing, and biometric recognition.