IEEE journal paper format

Video Super-Resolution using GANs

1st Oleh Pomazan
Department of Software Engineering
University of Europe for Applied Sciences
oleh.pomazan@ue-germany.de

***Abstract*—Video super-resolution is used to produce high-resolution video frames from the given low-resolution video frames, which can be used in a variety of vision tasks, including video restoration and enhancement, within the entertainment industry and streaming services. generative adversarial networks (GANs) are widely used for the super-resolution problems and SRGAN [1] is the most popular GAN model for single-image super-resolution (SISR). The SRGAN model uses a complex loss function consisting of pixel-wise mean squared error (MSE), perception, adversarial and total-variation parts. The effects of different components of the loss function on the SRGAN performance were studied by many authors [1]–[3] but the role of total-variation loss is not well studied in the context of SISR.**

In this thesis we study the effect of total-variation (TV) loss on the performance of SRGAN model in the context of applying single-image super-resolution for video frames. The optimal range of the TV loss weights was found to be between 10^{-8} and 10^{-6}. Larger values of the TV loss weight causes significant degradation of performance metrics and visible smoothing of the super-resolved frames.

While the SRGAN model produces high-quality super-resolved images using a single low-resolution frame, it does not utilize additional information from the previous and next frames of the video. Therefore, we investigate the SRGAN model in terms of temporal consistency between the consecutive frames.

***Index Terms*—GANs, SRGAN, video super-resolution, total-variation loss**

I. Introduction

Image super-resolution (SR)is a process of recovering high-resolution (HR) images from low-resolution (LR) images. It is an important category of of image processing methods utilized in computer vision, where the goal is to generate one or more HR images from one or more LR images. The objective of the SR algorithm is to generate finer details in an image compared to the sampling grid of the imaging device by increasing the pixel density per unit area. SR is an ill-posed inverse problem, as several HR images can be valid for any given LR image due to many aspects like brightness and coloring. A LR image, possibly with noise, distortions and artifacts, is used to restore a HR image [4], [5].

SR finds applications across diverse fields, including satellite imaging and remote sensing, where multiple images of a single area are accessible, in security and surveillance where the need arises to magnify a specific point of interest in a scene (like zooming in on a criminal's face or license plate numbers), in computer vision to enhance pattern recognition performance, and in other domains such as facial image analysis, text image analysis, biometric identification, fingerprint image enhancement, and more [6].

Video plays an important role in our daily lives, making the enhancement of low-resolution videos through SR techniques a crucial endeavor. Video super-resolution originates from image super-resolution and has the objective of enhancing the quality of high-resolution videos by reconstructing them from several low-resolution frames. Nevertheless, it is crucial to note that video super-resolution differs notably from image super-resolution since it typically utilizes inter-frame data to achieve its results.

The progress in computing power and deep learning approaches has sparked a multitude of advancements in video-related problems, including frames interpolation, artifact removal, denoising, deblurring, and super-resolution. In this thesis, we focus on VSR problem applying GAN based approach that is a rapidly developing field.

Over the past few years, there has been a consistent increase in the resolution of consumer devices, including monitors, displays, virtual reality headsets, and other similar technologies. Content that was created in the past may not appear optimal on high-resolution screens for the users. Possibility of real-time VSR techniques could have significant implications for the entertainment industry, particularly in the context of streaming services. As the demand for streaming platforms continues to

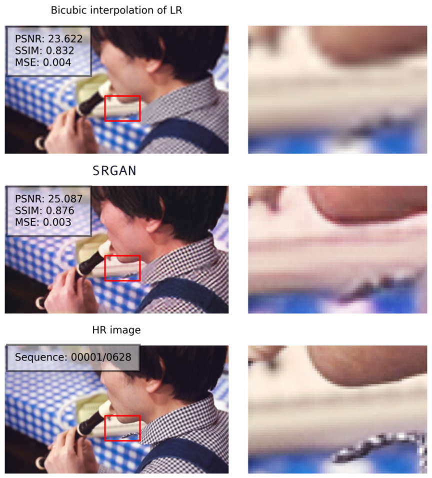

Fig. 1. Example of the super-resolved frame generated using the SRGAN model (sequence 00001/0628 from Vimeo90K dataset)

978-1-6654-5992-1/22/31.00 ©2023 IEEE

soar, applying post-processing VSR approaches on the client side can be feasible, thereby reducing the requirement for large-scale data transfers [7].

In this work, we investigate existing VSR approaches based on GANs and selected the SRGAN [1] model for the research. We study whether single-image super-resolution methods based on GANs can be applied to video super-resolution problem and if it can achieve super-resolution performance comparable to VSR models. We trained several SRGAN models with different parameters on Vimeo90K dataset, see an example of the super-resolved video frame on Fig. 1.

First, we study if the SRGAN model can be used to the VSR problem and to what temporal inconsistencies and artifacts it can lead. Second, the loss function components are analyzed with emphasis on the total-variation loss. At last, we study the effect of different patch size values on the model's training and performance.

II. Literature Review

Three - four paragraphs (or more) with one paragraph per important literature. A table summarizing the characteristics of the existing literature along with the novelty of your proposed work (as shown in Table **??**).

A. *Generative Adversarial Networks*

Generative adversarial networks (GANs) belong to the class of generative models used in unsupervised machine learning. They consist of two networks, namely the generator and the discriminator, engaged in a competitive zero-sum game framework. GANs employ a latent code that encapsulates all aspects of the generated output.

In the GAN framework, we have two models engaged in a competitive scenario akin to game theory. The setup involves a game with defined payoff functions, where each player strives to maximize their respective payoffs. In this game, one of the networks serves as the generator, which is our main focus and is responsible for producing samples (also known as generated or fake samples) with the goal of imitating those originating from the real training distribution (real samples). The other competing model is the discriminator, which examines the samples and determines whether they are real or fake [8].

During training, images or other samples are provided to the discriminator. The discriminator is typically a differentiable function, often implemented as a deep neural network, whose parameters can be learned through gradient descent. When the discriminator is presented with samples/images from the training set (real samples), its objective is to output a value close to one, indicating a high probability that the input is real rather than fake [8].

The discriminator is also used to evaluate samples generated by the generator (fake samples), and in this case, the objective of the discriminator is to produce an output as close to zero as possible, indicating that the sample is fake. The generator, on the other hand, is a differentiable function, often implemented as a deep neural network, and its parameters can be learned through gradient descent [8].

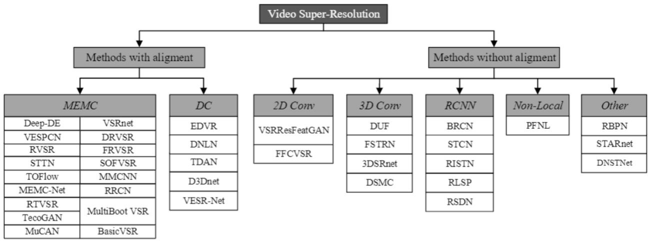

Fig. 2. A classification scheme for the current state-of-the-art video super-resolution methods [7]

The generator function operates on a sampled latent vector 'z', which serves as initial noise and acts as a source of randomness to aid the generator in producing a diverse range of outputs. The images generated by the generator are then evaluated by the discriminator, and the generator strives to trick the discriminator into outputting a value of one, making it believe that the generated image is real when, in fact, it is not. For more detailed technical information on GANs, readers can refer to [8].

B. *Video Super-Resolution Overview*

Over the past few years, numerous video super-resolution techniques have emerged, dividing mainly into two categories: traditional approaches and deep learning-based approaches. Traditional super-resolution methods are categorized into three groups: interpolation-based, reconstruction-based, and frequency-based approaches.

So far, many VSR algorithms have been proposed. Different deep learning models have demonstrated their effectiveness in video super-resolution tasks. Video Super-Resolution typically adopts the multi-input-single-output approach, as it involves providing multiple low-resolution frames to the model to predict a single reference frame. The key emphasis in VSR is on capturing the spatial and temporal relationships between frames.

In Liu et al. [7] the existing VSR methods are categorized into two main categories: methods with alignment and methods without alignment, according to whether the video frames are explicitly aligned.

A classification of the current VSR methods is presented at Fig. 2. There, MEMC stands for Motion Estimation and Compensation, DC is Deformable Convolution, 3D Conv means 3D Convolution and RCNN is Recurrent Convolutional Neural Network [7].

C. *Frame-Reccurent Video Super-Resolution*

Frame recurrent video super-resolution (FRVSR), as proposed by Sajjadi et al. [9], focuses on using the previously inferred high-resolution (HR) estimate to super-resolve the subsequent frame. This approach aims to achieve temporally consistent results and reduce computational costs. The architecture of FRVSR is presented in Figure **??**.

The detailed implementation involves the use of an optical estimation network to compute the optical flow between the

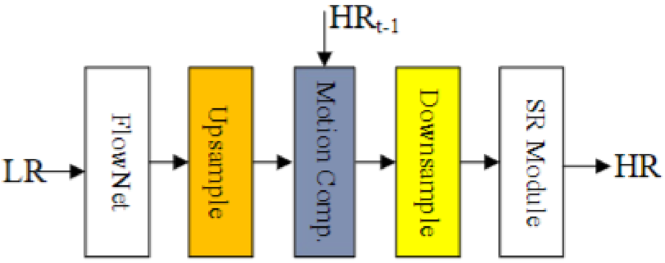

Fig. 3. The network architecture of FRVSR [7]

previous frame and the target frame. Subsequently, the low-resolution optical flow is upsampled to match the size of the high-resolution video using bilinear interpolation. The HR variant of the previous frame is then warped using the upsampled LR optical flow, and the warped HR frame is downsampled using space-to-depth transformation to obtain the LR version. Finally, the LR variant of the warped HR frame and the target frame are fed into the subsequent super-resolution network to produce the result for the target frame [7].

D. TecoGAN

The Temporally coherent GAN (TecoGAN) by Chu et al. [10] introduces a spatio-temporal discriminator to achieve realistic and coherent video super-resolution. To address recurrent artifacts, they propose a new ”Ping-Pong” loss. Similar to GANs, TecoGAN comprises a generator and a discriminator, and its architecture is illustrated in Fig. 4.

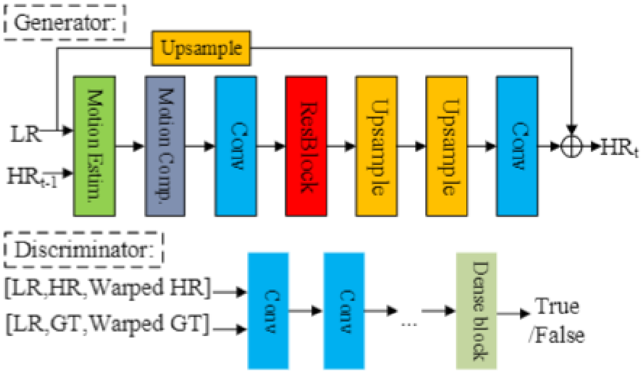

Fig. 4. The network architecture of TecoGAN [7]

E. VSRResFeatGAN

Rather than employing alignment operations like motion estimation and motion compensation between frames, the input frames are directly fed into a 2D convolutional network for spatial feature extraction, fusion, and super-resolution operations. This approach simplifies the video super-resolution problem by allowing the network to independently learn the correlation information within frames. Representative methods following this approach include VSRResFeatGAN (Lucas et al. [11]) and FFCVSR (Yan et al. [12]).

F. iSeeBetter

iSeeBetter [13] is a GAN-based spatio-temporal method used for video super-resolution that renders temporally consistent super-resolution videos. iSeeBetter’s generator utilizes recurrent back-projection networks to extract spatial and temporal information from both the current and neighboring frames. To enhance the natural appearance of the super-resolved image and eliminate artifacts associated with conventional methods, they incorporate the discriminator from the super-resolution generative adversarial network (SRGAN).

G. Loss function

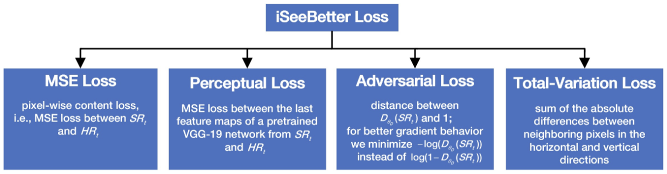
iSeeBetter Loss
MSE Loss
pixel-wise content loss, i.e., MSE loss between SR_t and HR_t
Perceptual Loss
MSE loss between the last feature maps of a pretrained VGG-19 network from SR_t and HR_t
Adversarial Loss
distance between $D_{\theta_D}(SR_t)$ and 1; for better gradient behavior we minimize $-\log(D_{\theta_D}(SR_t))$ instead of $\log(1-D_{\theta_D}(SR_t))$
Total-Variation Loss
sum of the absolute differences between neighboring pixels in the horizontal and vertical directions

Fig. 5. The MSE, perceptual, adversarial and TV loss components [13].

The 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-of-the-art SR approaches, which aims to improve the PSNR of an image. Despite optimizing MSE makes better PSNR and SSIM numbers, these metrics does not take into account fine details in the image, potentially misrepresenting the perceptual quality [13]. MSE’s capability to capture intricate texture details by analyzing individual pixel differences in consecutive frames is highly restricted, leading to the potentially excessively smooth video frames [14]. Through a sequence of experiments, it was discovered that even when images were deliberately distorted manually, their MSE scores were similar to those of the original, undistorted image [15].

To address this, [13] uses a four-fold (mean squared error, perceptual, adversarial, and total-variation) loss instead of solely relying on pixel-wise MSE loss. Fig. 5 shows the components of the loss function used in this work.

1) Mean-Squared Error Loss: In this work we use MSE loss (or, alternatively, content loss) for the super-resolved frame I^{SR} against the ground truth I^{HR}:

$$L_{MSE} = \frac{1}{WH} \sum_{x=0}^{W} \sum_{y=0}^{H} (I^{HR}_{x,y} - G_{\theta_G}(I^{LR})_{x,y})^2 \qquad (1)$$

where, $G_{\theta_G}(I^{LR})$ is the estimated frame SR, W and H are width and height of the image frames correspondingly [13].

2) Perceptual Loss : Instead of utilizing low-level measures that focus on pixel-wise errors, perceptual loss relies on features extracted from the activation layers of the pre-trained VGG network in the reference paper [16]. Perceptual loss, which was introduced in [17], [18], focuses on perceptual similarity instead of similarity in pixel space.

Here perceptual loss is defined the Euclidean distance between the feature representations of the estimated SR image $G_{\theta_G}(I^{LR})$ and the ground truth image I^{HR}:

$$L_{perceptual} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=0}^{W_{i,j}} \sum_{y=0}^{H_{i,j}} (VGG_{i,j}(I^{HR}_{x,y}) \\ -VGG_{i,j}(G_{\theta_G}(I^{LR})_{x,y}))^2$$

where, $VGG_{i,j}$ is the feature map which is the result of the j^{th} convolution after activation before the maxpooling layer in the VGG network, $W_{i,j}$ and $H_{i,j}$ are the widths and heights of the corresponding feature maps in VGG model [13].

3) Adversarial Loss: Adversarial loss is defined as:

$$L_{adversarial} = -\log(D_{\theta_D}(G_{\theta_G}(I^{LR}))) \tag{2}$$

where D_{θ_D} is the output of the discriminator which is the probability that the reconstructed image $G_{\theta_G}(I^{LR}$ is a real HR image. It was shown in [19] that it is better for the gradient behavior to minimize $-\log(D_{\theta_D}(G_{\theta_G}(I^{LR})))$ instead of $\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))$

4) Total-Variation Loss: TV loss was proposed in [20] as a loss function for super-resolution problems. It is defined as the sum of the absolute differences between neighboring pixels in the horizontal and vertical directions. By incorporating TV loss into our overall loss objective, we aim to minimize noise in the input, leading to a denoised output super-resolution image that promotes spatial smoothness.

TV loss is defined as in [21]:

$$L_{TV} = \frac{1}{WH} \sum_{x=0}^{W} \sum_{y=0}^{H} ((G_{\theta_G}(I^{LR})_{i,j+1} - G_{\theta_G}(I^{LR})_{i,j})^2 + \\ (G_{\theta_G}(I^{LR})_{i+1,j} - G_{\theta_G}(I^{LR})_{i,j})^2)^{\frac{1}{2}}$$

5) Loss formulation: The overall loss objective for each frame is defined as a combination of the MSE, adversarial, perceptual, and TV loss components, with appropriate weights assigned to each component:

$$\begin{aligned} L_{G_{\theta_G}} = \alpha \times L_{MSE}(I^{SR}, I^{HR}) + \\ \beta \times L_{perceptual}(I^{SR}, I^{HR}) + \\ \gamma \times L_{adversarial} + \\ \delta \times L_{TV}(I^{SR}, I^{HR}) \end{aligned} \tag{3}$$

where $\alpha, \beta, \gamma, \delta$ are weights.

H. TV loss weight

Total-variation (TV) denoising technique is known in image processing [22]. The total-variation based loss component was introduced in the original SRGAN paper [1] for training of SRResNet-VGG22 model. The TV loss weight was set to 2×10^{-8}. The same value of TV loss $\delta = 2 \times 10^{-8}$ is used in [13] and [23].

The effect of TV loss on the performance of the SRGAN model for the face super-resolution was studied in [24] where the TV loss weight was also set to 2×10^{-8} with reference to the SRGAN implementation by Hao Ren [1].

TV loss was used in the research related to the style transfer and super-resolution [25] where they set total-variation regularization "with a strength of between 1×10^{-6} and 1×10^{-8}, chosen via cross-validation per style target" [25].

Following an extensive literature review, we encountered a lack of sufficient explanation for the selected TV loss weight values utilized in the aforementioned research works. Consequently, we aim to conduct an in-depth investigation into this topic to offer more profound insights into the optimal range of TV loss weight values for the SRGAN model in the context of the VSR problem. The summarized review of different TV vaues is shown in Table I.

III. Methodology

A. Dataset

To train the model in this thesis, we use a subset of the original test set of Vimeo90K dataset (not downsampled or downgraded by noise, ∼15GB of video frames) [28]. The septuplet dataset consists of 91,701 7-frame sequences with fixed resolution 448 × 256, extracted from 39K selected video clips from Vimeo-90K. This dataset is designed to video denoising, deblocking, and super-resolution. Apart from Vimeo90K dataset, we use other datasets, that are known in the super-resolution research, for the evaluation of the trained models and comparison of our results with other authors.

B. Detailed Methodology

Three paragraphs for explaining the pipeline and workflow of your study. A figure depicting the workflow of your work (shown in Fig 6 or in Fig 7).

C. Evaluation Metrics

Video quality refers to the visual attributes of videos, which can be evaluated using either quantitative metrics or perceptual assessments conducted by human viewers. While human assessments offer initial insights, quantitative metrics are essential for benchmarking results and gaining widespread acceptance within the scientific community. This study primarily concentrates on achieving results based on reconstruction accuracy and perceptual naturalness, thereby considering metrics such as mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM). These metrics are classified as full-reference metrics, meaning

[1] Github page: https://github.com/leftthomas/SRGAN

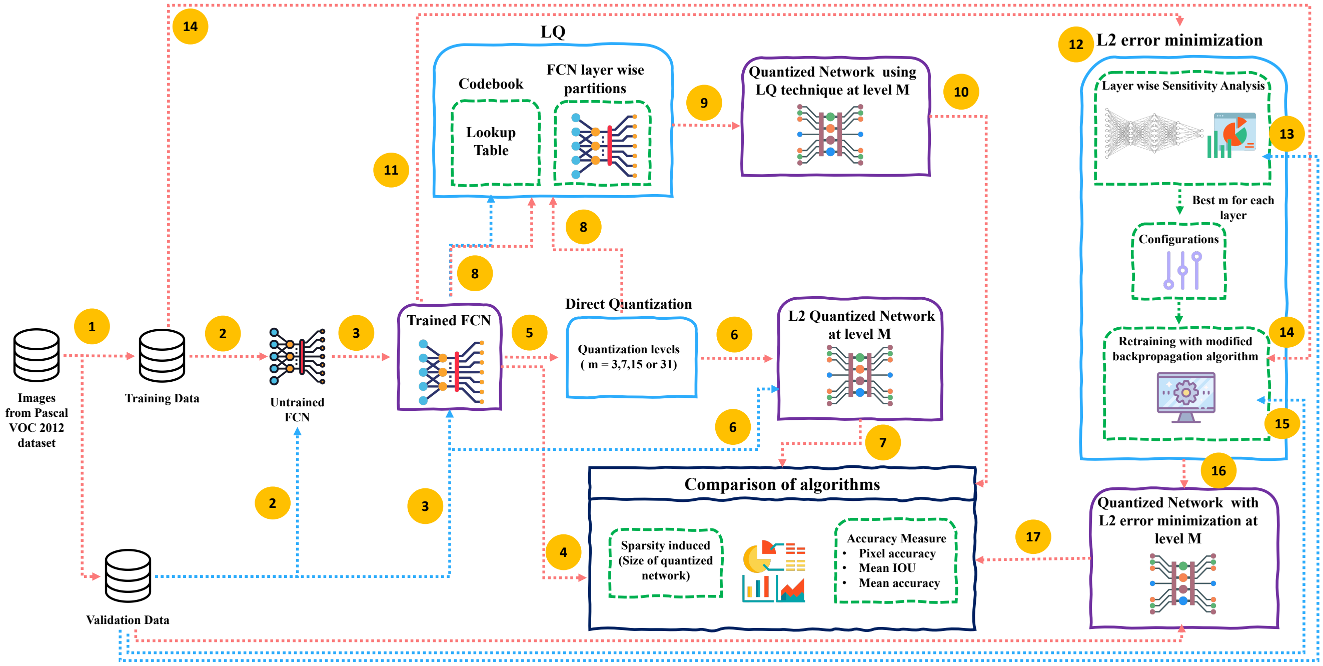
LQ
Codebook
Lookup Table
FCN layer wise partitions
Quantized Network using LQ technique at level M
L2 error minimization
Layer wise Sensitivity Analysis
Best m for each layer
Configurations
Retraining with modified backpropagation algorithm
Images from Pascal VOC 2012 dataset
Training Data
Untrained FCN
Trained FCN
Direct Quantization
Quantization levels (m = 3,7,15 or 31)
L2 Quantized Network at level M
Comparison of algorithms
Sparsity induced (Size of quantized network)
Accuracy Measure
Pixel accuracy
Mean IOU
Mean accuracy
Quantized Network with L2 error minimization at level M
Validation Data
1
2
2
3
3
4
5
6
6
7
8
8
9
10
11
12
13
14
14
15
16
17

Fig. 6. Sample workflow image. Make it in Powerpoint with svg images and save as pdf. Sanity check: Zoom in and pixels should not break.

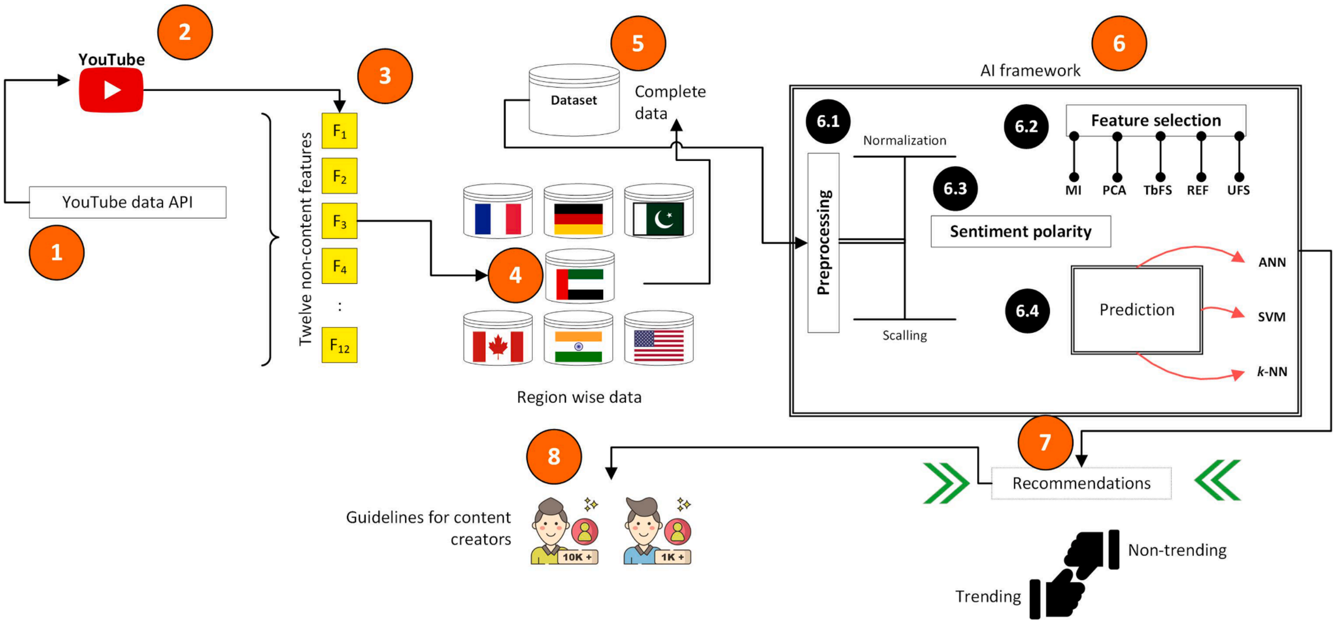

Fig. 7. Sample workflow image. Make it in Powerpoint with svg images and save as pdf. Sanity check: Zoom in and pixels should not break.

TABLE I
LITERATURE REVIEW TABLE SHOWING THE CONTRIBUTIONS OF VARIOUS AUTHORS FOR QUANTIZATION OF NETWORKS.

Paper	TV loss weight	Reasoning
SRGAN paper [1]	2×10^{-8}	References to [20], [25]
Faces SR paper [24]	2×10^{-8}	Reference to implementation (based on [1])
Style transfer paper [25]	$10^{-6} - 10^{-8}$	”chosen via cross-validation per style target”
iSeeBetter paper [13]	2×10^{-8}	References to [26]
iSeeBetter milestone paper [23]	2×10^{-8}	References to [27]

TABLE II
DATASETS SUMMARY.

Dataset	Amount	Average resolution	Format	Keywords
Set5	5	313 × 336	PNG	baby, bird, butterfly, head, woman
Set14	14	492 × 446	PNG	humans, animals, insects, etc.
Urban100	100	984 × 797	PNG	architecture, city, structure, urban, etc.
BSD100	100	435 × 367	PNG	animal, building, food, etc.
Vid4	171	720 × 480	PNG	calendar, city, foliage, walk

they require both the target (original) and the reconstructed (super-resolved) output to calculate the image/video quality [21].

1) Mean Squared Error: One widely-used metric is mean squared error (MSE), which is defined by the Equation 4:

$$MSE = \frac{1}{N}\sum_{i}^{N}(\hat{I}_i - \tilde{I}_i)^2 \tag{4}$$

where N is the total number of pixels in the frame, $\hat{I}$ is the ground-truth HR image, and $\tilde{I}$ is the super-resolved frame.

2) Peak Signal to Noise Ratio: Peak signal-to-noise ratio (PSNR) is one of the most popular reconstruction quality measurement of lossy transformation such as image compression. For image super-resolution, PSNR is defined via the maximum pixel value and the MSE between images. PSNR of one SR frame is defined as [7]:

$$PSNR = 10\log_{10}(\frac{L^2}{MSE}) \tag{5}$$

where L is the maximum range of the pixel color value (usually 255). In general, a higher value of PSNR means superior quality of the image. Because PSNR relies solely on pixel-level MSE, it focuses solely on pixel differences without considering visual perception. Consequently, it often yields unsatisfactory results in representing the true reconstruction quality in real-world scenarios, where human perception matters more. Nevertheless, as a necessity to compare with existing literature and due to the absence of entirely accurate perceptual metrics, PSNR remains the most commonly used evaluation criterion for SR models [21].

3) Structural Similarity Index Metric: The structural similarity index measure (SSIM) is defined for measuring the structural similarity between images, based on independent comparisons of luminance, contrast, and structures [29]. For an image I with N pixels, the luminance $\mu_I = \frac{1}{N}\sum_{i=1}^{N} I(i)$ and contrast $\sigma_I = (\frac{1}{N-1}\sum_{i=1}^{N}(I(i) - \mu_I)^2)^{\frac{1}{2}}$ are calculated as the mean and standard deviation of the image intensity I.

The comparisons of luminance and contrast are denoted as $C_l(I, \hat{I})$ and $C_c(I, \hat{I})$ correspondingly, are defined by:

$$C_l(I, \hat{I}) = \frac{2\mu_I\mu_{\hat{I}} + C_1}{\mu_I^2 + \mu_{\hat{I}}^2 + C_1} \tag{6}$$

$$C_c(I, \hat{I}) = \frac{2\sigma_I\sigma_{\hat{I}} + C_2}{\sigma_I^2 + \sigma_{\hat{I}}^2 + C_1} \tag{7}$$

where $C_l = (k_1L)^2$ and $C_2 = (k_2L)^2$ are constants preventing instabilities, $k_1 << 1$ and $k_2 << 1$ [21].

Structural similarity index measure (SSIM) is defined as:

$$SSIM(I, \hat{I}) = [C_l(I, \hat{I})]^{\alpha}[C_c(I, \hat{I})]^{\beta}[C_s(I, \hat{I})]^{\gamma} \tag{8}$$

where α, β and γ are control parameters for relative importance, C_s is the structure comparison function, defined by:

$$C_s(I, \hat{I}) = \frac{\sigma_{I\hat{I}} + C_3}{\sigma_I\sigma_{\hat{I}} + C_3} \tag{9}$$

where $\sigma_{I,\hat{I}}$ is the covariance between I and $\hat{I}$ and C_3 is a constant [21].

Structural similarity index measure (SSIM) is designed as an enhancement over conventional metrics like peak PSNR and MSE. Its purpose is to provide improved performance in assessing the similarity between images.

D. Experimental settings

One paragraph for experimental settings of your and competing methods (if any). (Optional) One paragraph for hyper-parameter settings and network architecture

A table with hyper-parameter settings (shown in Table III) and a figure for network architecture (shown in Fig 8).

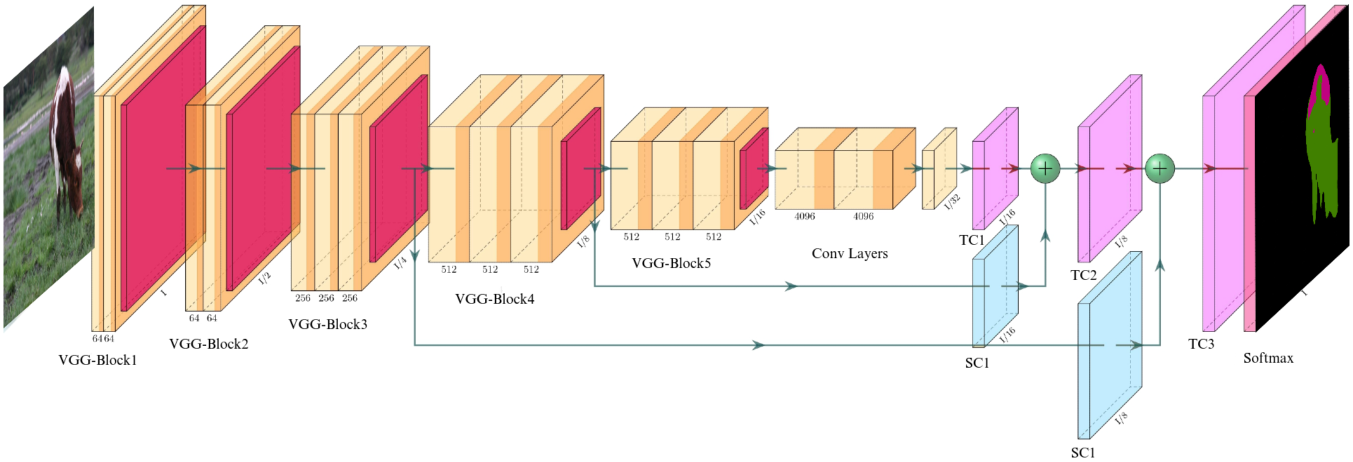

Fig. 8. Sample network architecture image. Make it in Powerpoint with svg images and save as pdf. Sanity check: Zoom in and pixels should not break.

TABLE III
CONFIGURATION TABLE SHOWING THE NETWORK CONFIGURATION OF FCN USED IN THIS STUDY. THE TABLE SHOWS THE VARIOUS CONFIGURATION SETTINGS USED FOR FCN8.

Network Configuration	
Epochs	50
Learning rate	0.0001
Mini batch size	20
Optimizer	SGD
Momentum	0.9
Weight decay	0.0002
L_2 Regularization	None
Samples in training set	8498
Samples in validation set	786

IV. RESULTS

Three (or more) paragraph(s) explaining your results. [30]. At least one paragraph targeting one research question with at least one figure (preferably) or table (where figure is not possible). Note that each figure must be drawn keeping in mind the black and white print (different patterns are prefered alongwith different intensity colors). This section must contain only results and nothing else (not your own opinion or any sort of discussion on quality of results). Also include comparison with existing contemporary methods in this section (sample shown in Figure 9 and in Table IV).

V. DISCUSSION

Five to six paragraphs discussing the results (at least one paragraph for each research question). Your opinion on how good/bad the results are. Draw inferences from the results here. Explain novelty of your contributions and what was missing that you have explored here. Discuss how results from your proposed method compare with other existing contemporary methods in this section. Any other point you would like to discuss related to this study. One paragraph for what are the future directions in your opinion for continuing this study. Limitations of your work, if any? Any assumptions that effect your analysis?

VI. CONCLUSION

One paragraph related to conclusions drawn from your whole experimentation.

References will be added automatically by using the following lines. Add the relevant citations in the attached bibliogrpahy.bib file. Get help from me where you want to work on citations.

REFERENCES

[1] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” *CoRR*, vol. abs/1609.04802, 2016. [Online]. Available: http://arxiv.org/abs/1609.04802

[2] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, “ESRGAN: enhanced super-resolution generative adversarial networks,” *CoRR*, vol. abs/1809.00219, 2018. [Online]. Available: http://arxiv.org/abs/1809.00219

[3] H. Cao and S. Mi, “Weighted SRGAN and Reconstruction Loss Analysis for Accurate Image Super Resolution,” *Journal of Physics: Conference Series*, vol. 1903, no. 1, p. 012050, Apr. 2021, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1903/1/012050

[4] B. B. Moser, F. Raue, S. Frolov, S. Palacio, J. Hees, and A. Dengel, “Hitchhiker's guide to super-resolution: Introduction and recent advances,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9862–9882, aug 2023. [Online]. Available: https://doi.org/10.1109%2Ftpami.2023.3243794

[5] Y. Jo, S. W. Oh, P. Vajda, and S. J. Kim, “Tackling the ill-posedness of super-resolution through adaptive target generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 236–16 245.

[6] S. C. Park, M. K. Park, and M. G. Kang, “Super-resolution image reconstruction: a technical overview,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[7] H. Liu, Z. Ruan, P. Zhao, F. Shang, L. Yang, and Y. Liu, “Video super resolution based on deep learning: A comprehensive survey,” *CoRR*, vol. abs/2007.12928, 2020. [Online]. Available: https://arxiv.org/abs/2007.12928

[8] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[9] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, “Frame-recurrent video super-resolution,” *CoRR*, vol. abs/1801.04590, 2018. [Online]. Available: http://arxiv.org/abs/1801.04590

TABLE IV

PERFORMANCE COMPARISON TABLE SHOWING THE PERFORMANCE OF VARIOUS QUANTIZATION TECHNIQUES APPLIED ON VARIOUS NETWORKS. THE RESULTS ARE DIRECTLY TAKEN FROM THE PAPERS AFTER APPLYING QUANTIZATION ON THE MENTIONED DATASETS.

Paper Name	Dataset Used	Performance Comparison							
		PA	MA	IoU	Time	Test Error	Sparsity Induced	size reduction	memory reduction
Vanhoucke et al. [26]	×				10x speedup				
Courbariaux et al. [27]	MNIST, SVHN, CIFAR-10					0.59%, 14.82%, 4.95%			
Gupta et al. [28]	MNIST, CIFAR-10					0.90%			
Denton et al. [29]	ImageNet 2012					0.83%	13%		
Lin et al. [30]	MNIST, SVHN, CIFAR-10					1.29%, 12.08%, 2.48%			
Hwang et al. [31]	MNIST					1.08%			
Anwar et al. [32]	MNIST					0.92%	17.1%		
Shin et al. [33]	TIMIT Corpus, MNIST					2.11%			
Xu et al. [39]	MICCAI Gland 2015	90.1%						6,4x	
Hubara et al. [46]	MNIST, SVHN, CIFAR-10, ImageNet 2012, Penn-tree bank					1.40% 2.53% 10.15%			
Lin et al. [47]	CIFAR-10					6.74%			20%
Proposed Approach	Pascal VOC 2012	89.30%	75%	56%		1.72%	40%	6.35x	84%

[10] M. Chu, Y. Xie, L. Leal-Taixé, and N. Thuerey, "Temporally coherent gans for video super-resolution (tecogan)," *CoRR*, vol. abs/1811.09393, 2018. [Online]. Available: http://arxiv.org/abs/1811.09393

[11] A. Lucas, S. L. Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *CoRR*, vol. abs/1806.05764, 2018. [Online]. Available: http://arxiv.org/abs/1806.05764

[12] B. Yan, C. Lin, and W. Tan, "Frame and feature-context video super-resolution," 09 2019.

[13] A. Chadha, J. Britto, and M. M. Roja, "iseebetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks," *CoRR*, vol. abs/2006.11161, 2020. [Online]. Available: https://arxiv.org/abs/2006.11161

[14] M.-H. Cheng, N.-W. Lin, K.-S. Hwang, and J.-H. Jeng, "Fast video super-resolution using artificial neural networks," in *2012 8th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP)*, 2012, pp. 1–4.

[15] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," *arXiv preprint arXiv:1511.05666*, 2015.

[18] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Advances in neural information processing systems*, vol. 28, 2015.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[20] H. Aly and E. Dubois, "Image up-sampling using total-variation regularization with a new observation model," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1647–1659, 2005.

[21] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[22] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016727899290242F

[23] A. Chadha, "iseebetter: A novel approach to video super-resolution using adaptive frame recurrence and generative adversarial networks," 2019, system Performance and Architecture, Apple Inc., CS230: Deep Learning, Project Milestone.

[24] H. Nguyen-Truong, K. N. A. Nguyen, and S. Cao, "Srgan with total variation loss in face super-resolution," in *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, 2020, pp. 292–297.

[25] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016. [Online]. Available: http://arxiv.org/abs/1603.08155

[26] J. Hany and G. Walters, *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.

[27] S. López-Tapia, A. Lucas, R. Molina, and A. K. Katsaggelos, "A single video super-resolution GAN for multiple downsampling operators based on pseudo-inverse image formation models," *CoRR*, vol. abs/1907.01399, 2019. [Online]. Available: http://arxiv.org/abs/1907.01399

[28] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision (IJCV)*, vol. 127, no. 8, pp. 1106–1125, 2019.

[29] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, pp. 600 – 612, 05 2004.

[30] C. Liu and D. Sun, "On bayesian adaptive video super resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 346–360, 2014.

[31] U. Nepal and H. Eslamiat, "Comparing yolov3, yolov4 and yolov5 for

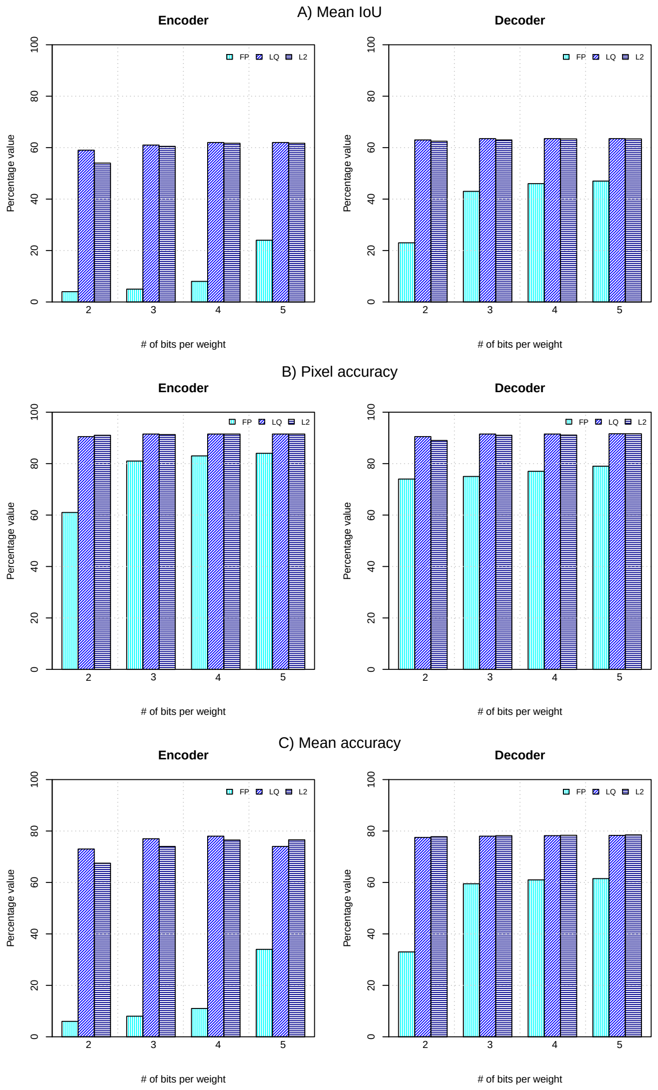

Fig. 9. Sample Figure comparing the three quantization techniques Fixed Point (FP), Lloyd's quantizer (LQ) and L_2 error minimization (L_2) on the three performance metrics divided into encoder and decoder layers. Mean IoU is shown for the three techniques in Panel A), pixel accuracy in Panel B), and mean accuracy in Panel C) respectively. Note that FP is consistently worse than both LQ and L_2, while L_2 and LQ are of comparable accuracy. Also, FP is most sensitive to number of bits in all metrics while L_2 and LQ are relatively insensitive.

autonomous landing spot detection in faulty uavs," *Sensors*, vol. 22, no. 2, p. 464, 2022.

VII. Other Headings and Reference Material

A. Ease of Use

1) Maintaining the Integrity of the Specifications: The IEEEtran class file is used to format your Report and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your Report as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

B. Prepare Your Report Before Styling

Before you begin to format your Report, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections VII-B1–VII-B5 below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

1) Abbreviations and Acronyms: Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

2) Units:

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

3) Equations: Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{10}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(10)", not "Eq. (10)" or "equation (10)", except at the beginning of a sentence: "Equation (10) is . . ."

4) LaTeX-Specific Advice: Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

5) Some Common Mistakes:

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your Report title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [30].

6) Authors and Affiliations: **The class file is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

7) Identify the Headings: Headings, or heads, are organizational devices that guide the reader through your Report. There are two types: component heads and text heads.

Component heads identify the different components of your Report and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the Report title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

8) Figures and Tables:

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 10", even at the beginning of a sentence.

TABLE V
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of

Fig. 10. Example of a figure caption.

quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks . . .”. Instead, try “R. B. G. thanks. . .”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Please number citations consecutively within brackets [30]. The sentence punctuation follows the bracket [31]. Refer simply to the reference number, as in [30]—do not use “Ref. [30]” or “reference [30]” except at the beginning of a sentence: “Reference [30] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Reports that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [30]. Reports that have been accepted for publication should be cited as “in press” [30]. Capitalize only the first word in a Report title, except for proper nouns and element symbols.

For Reports published in translation journals, please give the English citation first, followed by the original foreign-language citation [30].

IEEE conference templates contain guidance text for composing and formatting conference Reports. Please ensure that all template text is removed from your conference Report prior to submission to the conference. Failure to remove the template text from your Report may result in your Report not being published.