Cross-Modal Virtual Sensing for Combustion Instability Monitoring

Tryambak Gangopadhyay
Iowa State University
Ames, IA, USA
tryambak@iastate.edu

Vikram Ramanan
IIT Madras
Chennai, India
vikrambest@yahoo.co.in

Satyanarayanan R Chakravarthy
IIT Madras
Chennai, India
satyachakra@gmail.com

Soumik Sarkar
Iowa State University
Ames, IA, USA
soumiks@iastate.edu

Abstract

In many cyber-physical systems, imaging can be an important but expensive or 'difficult to deploy' sensing modality. One such example is detecting combustion instability using flame images, where deep learning frameworks have demonstrated state-of-the-art performance. The proposed frameworks are also shown to be quite trustworthy such that domain experts can have sufficient confidence to use these models in real systems to prevent unwanted incidents. However, flame imaging is not a common sensing modality in engine combustors today. Therefore, the current roadblock exists on the hardware side regarding the acquisition and processing of high-volume flame images. On the other hand, the acoustic pressure time series is a more feasible modality for data collection in real combustors. To utilize acoustic time series as a sensing modality, we propose a novel cross-modal encoder-decoder architecture that can reconstruct cross-modal visual features from acoustic pressure time series in combustion systems. With the "distillation" of cross-modal features, the results demonstrate that the detection accuracy can be enhanced using the virtual visual sensing modality. By providing the benefit of cross-modal reconstruction, our framework can prove to be useful in different domains well beyond the power generation and transportation industries.

1. Introduction

In aerospace and energy industries, ultra-lean premixed combustion is preferred to make gas turbine engines more fuel-efficient with lower cost and low NOx (nitrogen oxides) emissions. With this attempt to make engines efficient and environment-friendly, such operating regimes can make engines more prone to an undesirable phenomenon called combustion instability, which is caused by the establishment of a positive feedback loop between heat release rate fluctuations and fluctuating acoustic pressure [23]. In confined environments, fluctuating heat release rate leads to the generation of sound waves, which get reflected back to modify the heat release rate. A positive feedback loop can cause a growth of pressure fluctuations leading to large levels of vibration in an engine [4, 5]. This can result in huge revenue loss due to poor performance, reduced life, or catastrophic failure of engine [7]. Significant trade-offs are required in fuel efficiency and the design of combustion systems to prevent combustion instability ([17]). To avoid these trade-offs, it is important to develop an accurate and feasible framework for active detection and control of instability.

Previously, researchers have studied combustion instability using full-scale computational fluid dynamics [22], physics-based [2] and reduced order [28] modeling approaches. However, these approaches may require simplifying assumptions and face difficulty in achieving validation. An alternative is to implement data-driven methods utilizing acoustic pressure time-series [13, 20, 26]. These data-driven methods, based only on acoustic pressure time series, can sometimes be inaccurate due to interference from broadband background noise. Recently, researchers have started developing instability detection frameworks using machine learning [15, 27]. With the rapid development in the field of computer vision, the application of deep learning models has started in this domain to detect instability from flame images [25, 1, 8, 11, 9]. The effectiveness of model interpretability mechanisms such as 'attention' has been investigated from a domain knowledge perspective [11] and deep learning results have been verified from a physics-based understanding [9]. Therefore, the image-based deep learning frameworks have proved to be accurate, trustwor-

thy and can build the confidence of domain experts to implement these models in real systems to detect combustion instability. However, the current roadblock exists on the hardware side.

Acquisition and processing of high-volume flame image data may not be feasible to perform fast enough using existing commercial hardware. Also, flame imaging is not a common sensing modality in engines today. Therefore, the image-based deep learning frameworks can only become feasible with rapid improvement in the hardware sector. Acoustic pressure time series is a more feasible modality for data collection in real combustors. To circumvent the hardware roadblock and simultaneously ensure high detection accuracy, the optimal solution can be to utilize acoustic time series as a sensing modality and implement image-based deep learning models. In this work, we attempt to think in that direction by proposing a novel virtual sensing model (VSenseNet) to reconstruct cross-modal visual features from acoustic pressure time series in combustion systems. While researchers have proposed cross-modal reconstruction models for text-to-image [24, 32, 16], and speech-to-face [21, 6], there has been no work on the reconstruction of visual features from time series in any application domain.

Contributions. We summarize the contributions of this work as follows:

1. To the best of our knowledge, this is the first work on cross-modal reconstruction of visual features from time series in any domain. The proposed cross-modal encoder-decoder model VSenseNet is novel in the context of combustion systems to reconstruct flame images from acoustic pressure time series.

2. In VSenseNet, visual reconstruction from time series is achieved by training the encoder-decoder with "distillation" of cross-modal features from models pre-trained on images. During testing, the classification performance of synthetic images is compared against that of actual images.

3. With acoustic time series as the sensing modality, we demonstrate that instability detection accuracy can be enhanced using our proposed virtual sensing modeling approach. VSenseNet can prove to be a great resource in different sectors where imaging is an important but 'difficult to deploy' sensing modality.

2. Related Work

Researchers have proposed models involving cross-modal reconstruction for text, speech, and image datasets. Conditional Generative Adversarial Nets (GANs) [19] can direct the data generation process by conditioning the model on additional information. A training strategy involving GAN [12] architecture can enable text-to-image synthesis of bird and flower images from human-written descriptions [24]. Conditional GANs have been used to achieve the cross-modal audio-visual generation of musical performances [3]. AttnGAN [32] is an attention-driven model for text-to-image generation where the layered attentional structure can pay attention to the relevant words in the natural language description for generating different parts of the image. Obj-GAN [16] is an object-driven attentive generative network for synthesizing complex images from text descriptions utilizing an object-driven attentive generative network and an object-driven discriminator. Speech2Face [21] model has been proposed to study the task of reconstructing a facial image of a person from a short audio recording of that person speaking. Face images of a speaker can be generated with a self-supervised approach by exploiting the audio and visual signals naturally aligned in videos [6]. Cross-modal matching can be used to generate faces from voices that match several biometric characteristics of the speaker [31]. A model has been proposed to use both audio and a low-resolution image to perform extreme face super-resolution [18].

3. Virtual Sensing Model (VSenseNet)

To address the hardware roadblock of flame image acquisition, to use acoustic time series as sensing modality, and to simultaneously utilize an image-based detection framework for better accuracy, we propose a novel cross-modal encoder-decoder virtual sensing model VSenseNet. In this section, we demonstrate two versions of VSenseNet - VSenseNet I and VSenseNet II. The training framework of VSenseNet II additionally comprises of image classifier, while that of VSenseNet I only includes the time series encoder and the image decoder. We demonstrate the ablation studies of both versions in the supplementary materials.

3.1. Convolutional Autoencoder Pre-Training

Autoencoders can learn meaningful representations using a compression function (encoder) and a decompression function (decoder). The encoder compresses the input into a low dimensional embedding, and the decoder reconstructs the high dimensional information from that embedding. Without the requirement of explicit annotations, the weights of an autoencoder model can be learned with the objective of minimizing the reconstruction loss. The first step is to utilize the training dataset of flame images to pre-train a convolutional autoencoder which comprises an image encoder and an image decoder as demonstrated in Fig. 14. The encoder takes in a flame image (resolution 64 x 64) as input. The encoder model comprises a series of 2D convolutional and 2D max-pooling layers. After that, a fully connected layer is used to compute a 128-dimensional embedding. From the 128-dimensional embedding, the de-

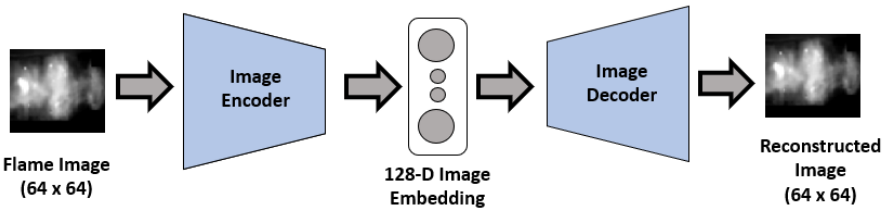


Figure 1. The encoder and decoder of the convolutional autoencoder model used for pre-training.

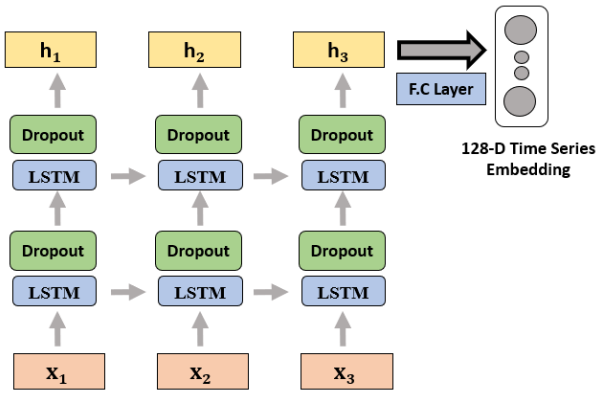


Figure 2. The details of time series encoder for training the VSenseNet.

coder attempts to reconstruct the original flame image as closely as possible using a series of 2D up-sampling and 2D convolutional transpose layers. The details of image encoder and decoder models are provided in the supplementary section.

3.2. Time Series Encoder

The training frameworks of both versions of VSenseNet consist of the time series encoder to compute an embedding from the time series. Long Short Term Memory (LSTM) networks can effectively capture long-term temporal dependencies [14], and LSTM networks are effective in different applications involving time series data [10]. We develop the time series encoder model consisting of two LSTM layers with dropout added after each layer to prevent over-fitting. The time series encoder is shown in Fig. 2. The first LSTM layer takes in the multivariate time series as input. The hidden state outputs of the first LSTM layer act as inputs to the second LSTM layer. The last hidden state of the second LSTM layer is considered the compressed information for the time series. A fully connected layer is used to get a 128-dimensional time series embedding.

3.3. VSenseNet I

The VSenseNet I modeling approach is illustrated in Fig. 9. The trainable parts of the VSenseNet I framework are the time series encoder and image decoder, and the pre-trained (and fixed) part is the image encoder. The weights of VSenseNet I are learned with two learning objectives - minimizing the embedding loss and minimizing the reconstruction loss.

The pre-trained image encoder is utilized to compute the 128-dimensional image embedding. The trainable time series encoder (Fig. 2) generates the 128-dimensional time series embedding. The embedding loss is the mean squared error (MSE) computed between the time series embedding and image embedding. In the training process, the time series encoder learns to compute an embedding that can match the image embedding as closely as possible.

The image decoder in Fig. 9 is trained from scratch alongside the time series encoder in the training loop. The model architecture of the decoder is the same as that in the autoencoder model (Fig. 14). The input to the image decoder is the time series embedding, from which it attempts to reconstruct the image corresponding to the input time series. The learning objective of the image decoder is to minimize the reconstruction loss between the actual flame image and the reconstructed image.

3.4. VSenseNet II

Compared to VSenseNet I, VSenseNet II has an image classifier model in the training framework apart from the time series encoder and the image decoder. We demonstrate the training framework of VSenseNet II in Fig. 12. The trainable parts are the time series encoder, image decoder, and image classifier. Similar to VSenseNet I, in VsenseNet II also, the pre-trained image encoder is utilized. The training framework consists of two steps.

In the first step, the time series encoder is trained to regress to the image embedding computed from the image encoder. With the time series as input, the time series encoder computes a 128-dimensional time series embedding. The learning objective is to minimize the embedding loss (MSE) between the time series and image embeddings.

In the next step, the image decoder and image classifier

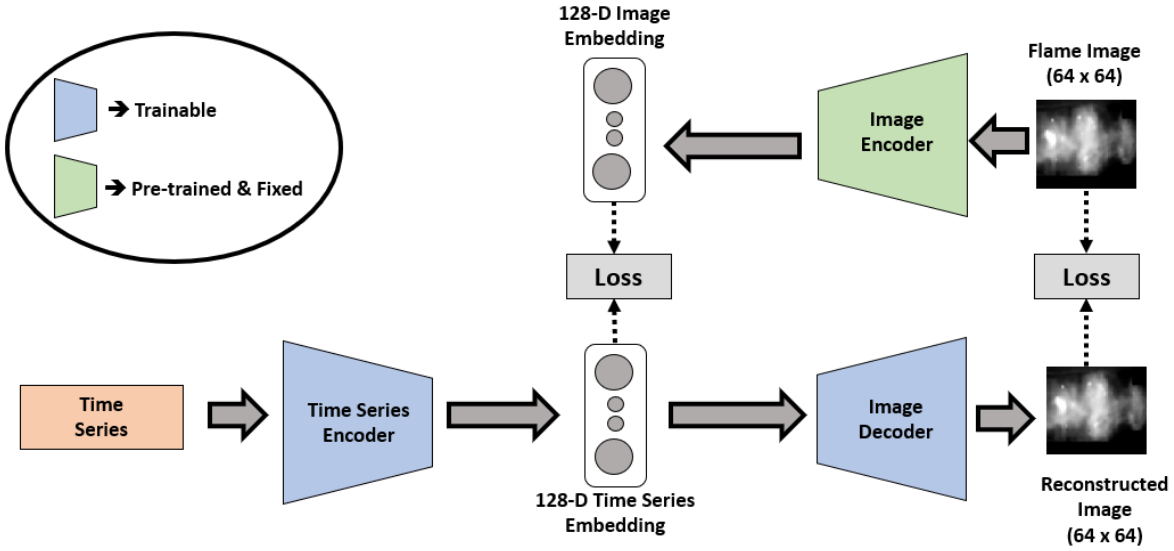


Figure 3. Training framework of VSenseNet I. It utilizes the pre-trained image encoder to compute the image embedding. The time series encoder and image decoder are trained with two loss functions - embedding loss and reconstruction loss.

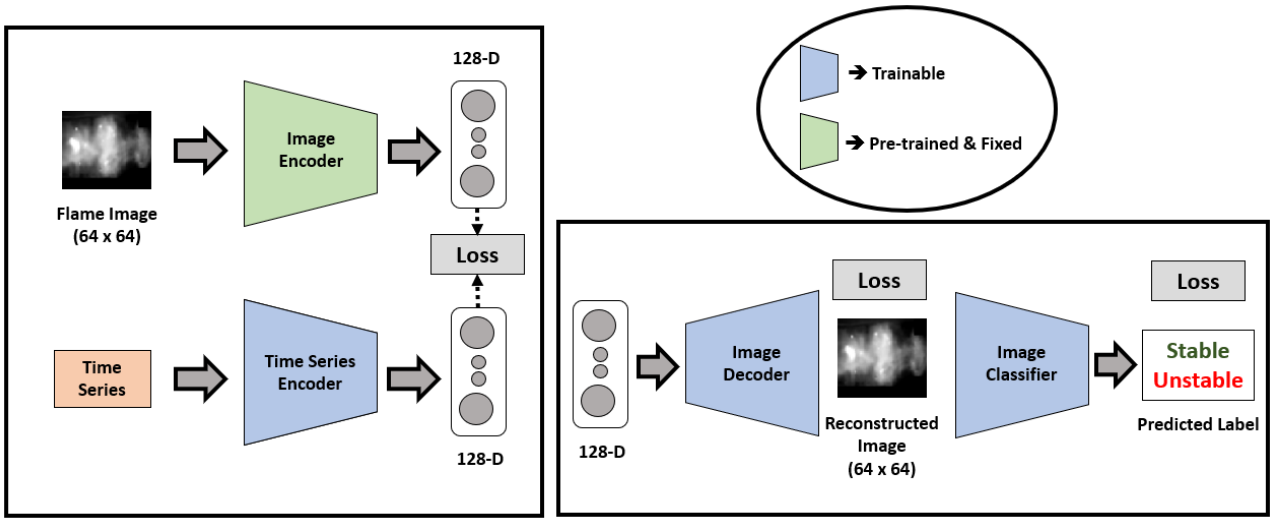


Figure 4. Training framework of VSenseNet II. It comprises two steps. The first step is to train the time series encoder. The next step is to train the image decoder and image classifier utilizing reconstruction loss and classification loss.

models are trained. The model architecture of the decoder is the same as that used in Fig. 14. The image classifier model comprises 2D convolutional, 2D max-pooling, and fully connected layers. It is a binary classification model to predict a flame image as stable or unstable. The model architecture of the image classifier is provided in supplementary materials. The generated time series embeddings are utilized for training this part of the framework. From an embedding, the image decoder learns to reconstruct the corresponding flame image as closely as possible. With the reconstructed flame image as input, the image classifier model predicts it as stable or unstable.

3.5. Test Framework for VSenseNet

The overall test framework for VSenseNet is shown in Fig. 5. It consists of three steps - time series encoder, image decoder, and image classifier.

For VSenseNet I, the time series encoder and the image decoder are trained as shown in Fig. 9. The image classifier is trained separately using actual flame images of the training dataset. For VSenseNet II, the time series encoder, the image decoder, and the image classifier are trained as

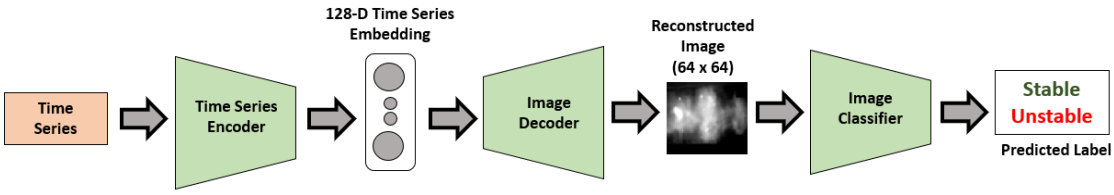


Figure 5. Test framework of VSenseNet comprising time series encoder, image decoder and image classifier.

shown in Fig. 12.

After taking in the multivariate time series as input, the trained time series encoder model computes the time series embedding. From this embedding, the trained image decoder computes the reconstructed image. This image is fed into the trained image classifier model to classify it as stable or unstable. Therefore, the virtual sensing framework utilizes acoustic time series as the sensing modality from which it reconstructs the cross-modal visual features. The reconstructed image is then used to identify the presence of combustion instability. With this approach of using a feasible sensing modality, the hardware roadblock of flame image acquisition can be avoided, and simultaneously, better detection accuracy can be achieved.

4. Experiments

4.1. Dataset

For dataset collection, we induce combustion instability in a laboratory-scale swirl combustor (details provided in the supplementary section). The fuel is injected co-axially with air at selected upstream distances. For dataset collection, the chosen upstream distances are 90 mm and 120 mm. For the upstream distance of 90 mm, partial premixing of the fuel with air occurs, while the distance of 120 mm facilitates full premixing of the fuel and air.

The ground truth labels (stable, unstable) for the hi-speed flame image sequences are provided by the domain experts. The conditions are defined as stable or unstable based on the dominant frequency, and root mean square (RMS) value of pressure fluctuations (between initial and final instants). The stable conditions demonstrate broadband frequency (estimated from fast Fourier transform) and RMS pressure values less than 100 Pa. For unstable conditions, the frequency of oscillations corresponds to sharp values in the range of 130-150 Hz, and the RMS pressure exceeds 500 Pa. Hence, the images are labeled using the corresponding pressure modality and not using image features.

We identify the conditions by upstream distance (premixing length), airflow rate (AFR), and fuel flow rate (FFR). Both AFR and FFR are expressed in lpm (liters per minute). The hi-speed images are captured at 3000 Hz (with a resolution of 1024 x 1024) for 3 seconds at each condition. Simultaneously, the pressure data is recorded at 4 locations of the experimental setup with a frequency of 9000 Hz. Therefore, for each condition, we have 9000 frames and 27000 time steps of multivariate pressure data. From a total of six conditions, we use four conditions for training our proposed models and keep two conditions for testing the performance of the models.

The stable and unstable conditions in the training set are:

1. $\mathbf{Stable_{120/60/600}}$: Condition has Premixing Length = 120 mm, FFR = 60 and AFR = 600.
2. $\mathbf{Stable_{90/45/450}}$: Condition has Premixing Length = 90 mm, FFR = 45 and AFR = 450.
3. $\mathbf{Unstable_{120/45/900}}$: Condition has Premixing Length = 120 mm, FFR = 45 and AFR = 900.
4. $\mathbf{Unstable_{90/28/600}}$: Condition has Premixing Length = 90 mm, FFR = 28 and AFR = 600.

The stable and unstable conditions in the test set are:

1. $\mathbf{Stable_{120/45/450}}$: Condition has Premixing Length = 120 mm, FFR = 45 and AFR = 450.
2. $\mathbf{Unstable_{90/45/900}}$: Condition has Premixing Length = 90 mm, FFR = 45 and AFR = 900.

4.2. Baseline Models

For comparison of results, we use three baseline models.

1. **Image Classifier:** The image classifier model is the same as that used in the test framework of VSenseNet (Fig. 5). This image-based baseline model takes a flame image as input and classifies it as stable or unstable. It is trained on actual training set images and tested on actual test set images.
2. **Time Series Classifier:** This is an entirely time series based model with no cross-modal reconstruction of visual features. The time series classifier model is developed by augmenting the time series encoder model (Fig. 2) - two fully connected layers are added after the computation of the 128-dimensional time series embedding. The entire model architecture of the time series classifier has been provided in the supplementary

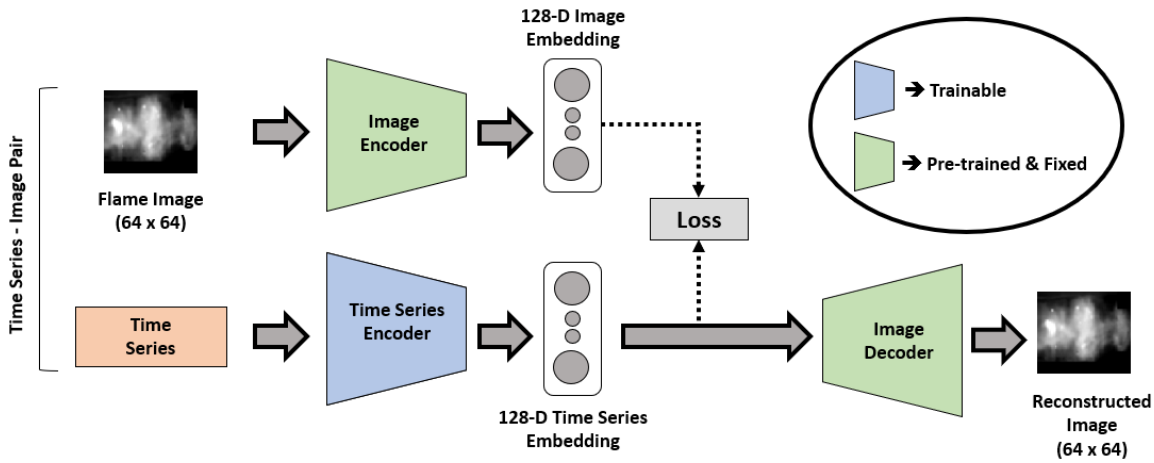


Figure 6. Framework to train the time series encoder of the baseline Cross-Modal Model which is inspired from Speech2Face [21].

section. With the multivariate time series as input, the model predicts it as stable or unstable. The time series classifier model is trained using time series only, and it is tested using time series of the test set.

3. **Cross-Modal Model:** We develop this cross-modal baseline model inspired from the Speech2Face [21] model. Speech2Face was proposed to reconstruct a human face from an audio recording of a person speaking. Speech2Face model consists of a voice encoder that takes spectrogram as input - the spectrogram is computed from the audio recording. The voice encoder is trained utilizing a pre-trained face encoder network. The voice encoder computes an embedding that is fed to a pre-trained face decoder model to reconstruct the human face. We implement a framework similar to that of Speech2Face - time series encoder is used instead of voice encoder to encode the time series, pre-trained image encoder, and image decoder models (Fig. 14) are used as replacements of pre-trained face encoder and face decoder models, respectively. The framework to train the time series encoder is shown in Fig. 6. The test framework of the Cross-Modal model is the same as that of VSenseNet, as demonstrated in Fig. 5.

4.3. Results

In this section, we discuss the classification and reconstruction performance of the proposed approach.

We use three evaluation metrics for classification performance: Accuracy, F1 Score, and False Negative Rate. F1 Score, also known as F-score or F-measure, is the harmonic mean of precision and recall. False negative rate (FNR) refers to falsely predicting negative (stable) when it is actually positive (unstable). FNR is the ratio between the predicted number of false negative samples and the actual number of positive samples. It summarizes how often stable is predicted when the actual is unstable. For detection of combustion instability, it is highly significant to have a low model FNR. We use Adam optimizer with a learning rate of 0.001 and a batch size of 32. The models are trained using NVIDIA Titan RTX GPU.

For reconstruction performance, we use two evaluation metrics - Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM). MSE computed between the actual and reconstructed images may not always be highly indicative of the structural similarity. SSIM [29, 30] addresses this issue by considering texture and also including luminance masking and contrast masking terms. Therefore, SSIM can be efficient in estimating the perceived similarity between the actual and reconstructed images.

Table 1 presents the empirical results for the test set conditions. The Image Classifier Model, which is tested using actual flame images, shows an average accuracy of 99.38%. The Time Series Classifier Model, which doesn't involve any cross-modal reconstruction, shows an average accuracy of 98.50%. Therefore, the image-based model is more accurate than the time series based framework. From Table 1, we observe that our proposed models (VSenseNet I and VSenseNet II) demonstrate better accuracy than the time series model. Both of our proposed models also outperform the Cross-Modal baseline model in terms of accuracy, F1 Score, and FNR. Using time series as the input modality with virtual sensing approach, we can enhance the average classification accuracy from 98.50% to 99.01% and approach closer towards the 99.38% accuracy achieved by the imaging modality-based model. Therefore, by adopting the proposed training approach of distillation of cross-modal features and reconstruction of synthetic images, we enhance the instability detection performance with acoustic time series as the sensing modality.

Model	Reconstruction Performance		Classification Performance		
	SSIM	Mean Squared Error	Accuracy	F1 Score	FNR
Image Classifier	NA	NA	**0.9938 ± 0.0064**	**0.9938 ± 0.0065**	**0.0122 ± 0.0129**
Time Series Classifier	NA	NA	**0.9850 ± 0.0021**	**0.9847 ± 0.0022**	**0.0300 ± 0.0044**
Cross-Modal	0.6655	0.0072	0.9888 ± 0.0005	0.9887 ± 0.0005	0.0222 ± 0.0010
VSenseNet I	**0.6980**	**0.0062**	0.9895 ± 0.0003	0.9894 ± 0.0002	0.0209 ± 0.0006
VSenseNet II	0.6876	0.0070	**0.9901 ± 0.0003**	**0.9900 ± 0.0003**	**0.0197 ± 0.0007**

Table 1. Empirical Results for the test set. For classification performance, average and standard deviation of the evaluation metrics (Accuracy, F1 Score, FNR) are reported after training each model five times. For reconstruction performance, average values of SSIM and MSE are reported.

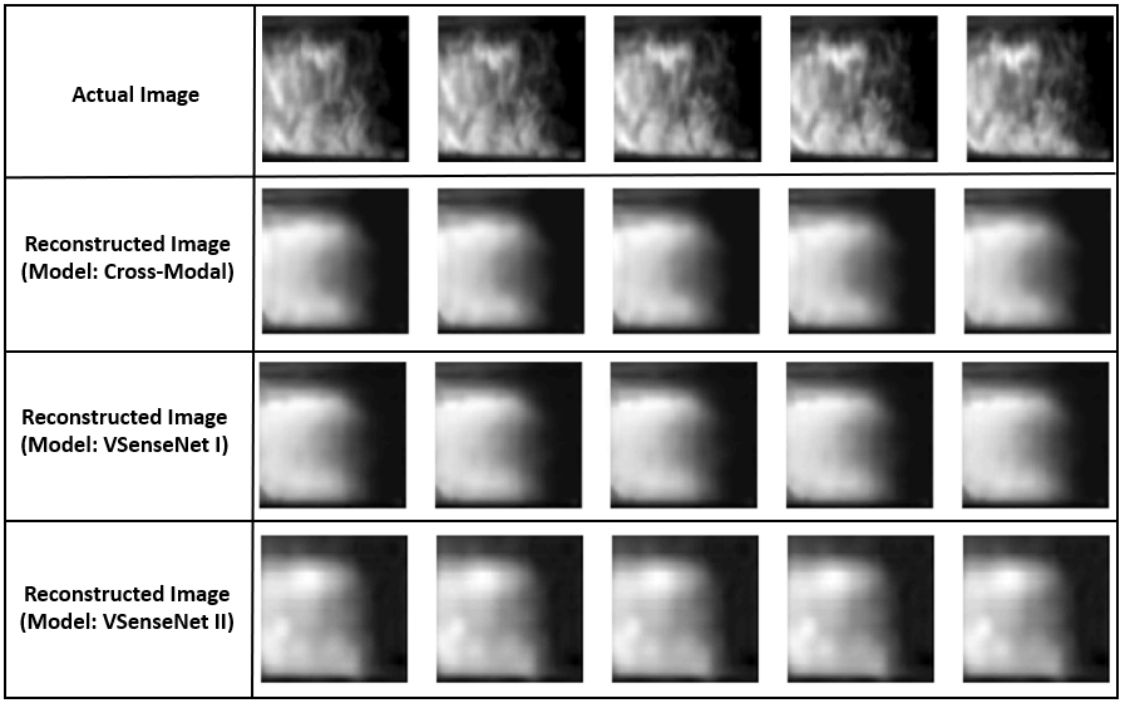


Figure 7. The reconstructions obtained from the proposed VSenseNet models are compared against the reconstructions obtained from the Cross-Modal baseline model, and the actual flame images for the test set $\mathbf{Stable_{120/45/450}}$ condition.

In terms of reconstruction performance, both VSenseNet I and VSenseNet II outperform the Cross-Modal baseline model in terms of SSIM and MSE as demonstrated in Table 1. Sample reconstruction results are shown in Fig. 7 and Fig. 8. The reconstructed images are generally a bit more smooth than the actual images. From Fig. 7, we observe that all the models can reliably reconstruct the flame images for the test set $\mathbf{Stable_{120/45/450}}$ condition. For the other test set condition $\mathbf{Unstable_{90/45/900}}$, the proposed VSenseNet models reconstruct the flame structures better than the baseline Cross-Modal model as highlighted by red boxes in Fig. 8.

While stable-unstable flame classification can help in designing active combustion control mechanisms to mitigate the instability, it does not provide any scientific insight to the domain experts in terms of the coherent structures responsible for triggering the instability, and hence, the overall approach may lack interpretability. Therefore, we stress upon the need for flame image reconstruction that can provide valuable insights during offline analysis as well as build sufficient trust of the domain experts via necessary interpretability.

5. Conclusion

Deep learning frameworks have demonstrated state-of-the-art performance in detecting combustion instability from flame images. Such frameworks have also proved to be trustworthy to build the confidence of domain experts. But the current roadblock exists in the acquisition

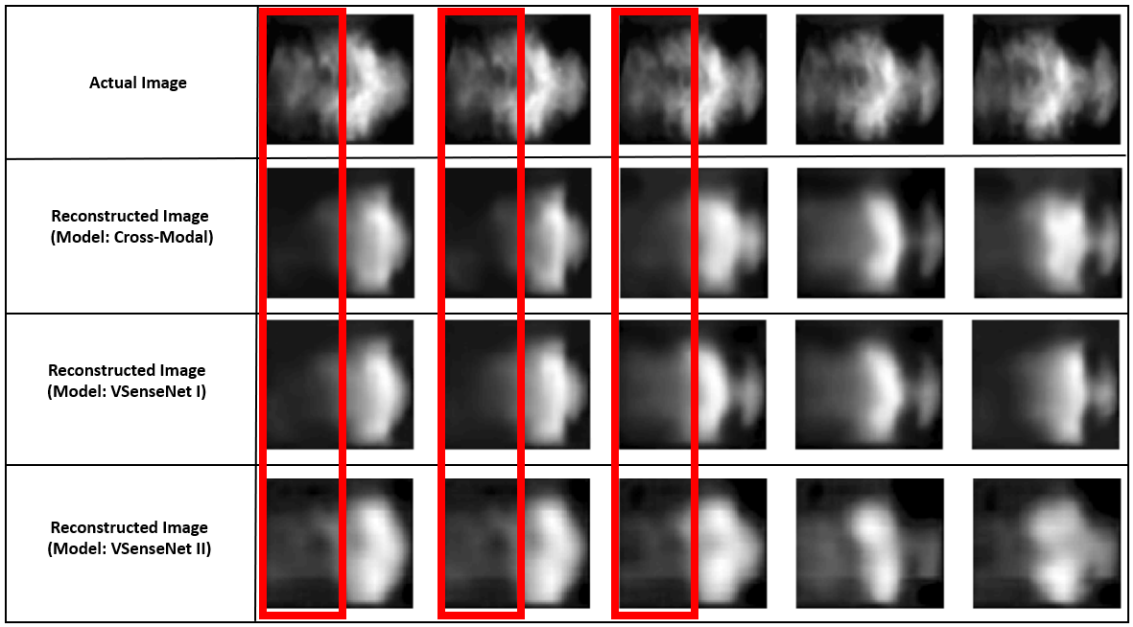


Figure 8. The reconstructions obtained from the proposed VSenseNet models are compared against the reconstructions obtained from the Cross-Modal baseline model, and the actual flame images for the test set $\mathbf{Unstable}_{\mathbf{90/45/900}}$ condition.

of high-volume flame images within the confines of an engine having high temperatures. From the hardware side, capturing acoustic pressure time series in real combustors is a more feasible modality. To utilize acoustic time series as a sensing modality and, at the same time, to simultaneously ensure high detection accuracy, we propose a novel cross-modal encoder-decoder virtual sensing model that can reconstruct cross-modal visual features from acoustic pressure time series in combustion systems.

Our proposed VSenseNet approach demonstrates effectiveness in reconstructing the flame images. By choosing different conditions in the test set, we demonstrate the robustness of our model. We demonstrate that by cross-modal reconstruction of synthetic images, the classification performance is better than that from time series alone. Therefore we are enhancing the accuracy of combustion instability detection using time series data with our virtual sensing modeling approach. Domain experts can also gain valuable insights during an offline analysis of the reconstructed flame images.

VSenseNet provides the unique benefit of generating synthetic visual features corresponding to time series information. We believe that our proposed approach has the potential to impact different sectors dealing with cyber-physical systems where imaging is an important but 'difficult to deploy' sensing modality. Our VSenseNet approach can fill up the gap by providing the benefit of cross-modal reconstruction, and we envision that this can bring a transformative advancement in different application domains. In the future, we plan to extend VSenseNet to capture transitions in a combustion system. We would also like to apply our modeling approach to other cyber-physical systems.

References

[1] Adedotun Akintayo, Kin Gwn Lore, Soumalya Sarkar, and Soumik Sarkar. Prognostics of combustion instabilities from hi-speed flame video using a deep convolutional selective autoencoder. *International Journal of Prognostics and Health Management*, 7(023):1–14, 2016.

[2] Benjamin D Bellows, Mohan K Bobba, Annalisa Forte, Jerry M Seitzman, and Tim Lieuwen. Flame transfer function saturation mechanisms in a swirl-stabilized combustor. *Proceedings of the combustion institute*, 31(2):3181–3188, 2007.

[3] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357, 2017.

[4] FE Culick and Paul Kuentzmann. Unsteady motions in combustion chambers for propulsion systems. Technical report, NATO RESEARCH AND TECHNOLOGY ORGANIZATION NEUILLY-SUR-SEINE (FRANCE), 2006.

[5] Ann P Dowling. Nonlinear self-excited oscillations of a ducted flame. *Journal of fluid mechanics*, 346:271–290, 1997.

[6] Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i Nieto. Wav2pix: Speech-conditioned face generation using generative adversarial networks. In *ICASSP*, pages 8633–8637, 2019.

[7] Steven C Fisher and Shamim A Rahman. Remembering the giants: Apollo rocket propulsion development. 2009.

[8] Tryambak Gangopadhyay, Anthony Locurto, James B Michael, and Soumik Sarkar. Deep learning algorithms for detecting combustion instabilities. In *Dynamics and Control of Energy Systems*, pages 283–300. Springer, 2020.

[9] Tryambak Gangopadhyay, Vikram Ramanan, Adedotun Akintayo, Paige K Boor, Soumalya Sarkar, Satyanarayanan R Chakravarthy, and Soumik Sarkar. 3d convolutional selective autoencoder for instability detection in combustion systems. *Energy and AI*, 4:100067, 2021.

[10] Tryambak Gangopadhyay, Sin Yong Tan, Zhanhong Jiang, Rui Meng, and Soumik Sarkar. Spatiotemporal attention for multivariate time series prediction and interpretation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3560–3564. IEEE, 2021.

[11] Tryambak Gangopadhyay, Sin Yong Tan, Anthony LoCurto, James B Michael, and Soumik Sarkar. Interpretable deep learning for monitoring combustion instability. *IFAC-PapersOnLine*, 53(2):832–837, 2020.

[12] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[13] Hiroshi Gotoda, Hiroyuki Nikimoto, Takaya Miyano, and Shigeru Tachibana. Dynamic properties of combustion instability in a lean premixed gas-turbine combustor. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(1):013124, 2011.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Tsubasa Kobayashi, Shogo Murayama, Takayoshi Hachijo, and Hiroshi Gotoda. Early detection of thermoacoustic combustion instability using a methodology combining complex networks and machine learning. *Physical Review Applied*, 11(6):064034, 2019.

[16] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[17] Tim C Lieuwen. *Unsteady combustor physics*. Cambridge University Press, 2012.

[18] Givi Meishvili, Simon Jenni, and Paolo Favaro. Learning to have an ear for face super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1364–1374, 2020.

[19] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[20] Vineeth Nair and RI Sujith. Multifractality in combustion noise: predicting an impending combustion instability. *Journal of Fluid Mechanics*, 747:635–655, 2014.

[21] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7539–7548, 2019.

[22] P Palies, T Schuller, D Durox, and S Candel. Modeling of premixed swirling flames transfer functions. *Proceedings of the combustion institute*, 33(2):2967–2974, 2011.

[23] John William Strutt Rayleigh. The explanation of certain acoustical phenomena. *Nature*, 18(455):319–321, 1878.

[24] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[25] Soumalya Sarkar, Kin Gwn Lore, and Soumik Sarkar. Early detection of combustion instability by neural-symbolic analysis on hi-speed video. In *CoCo@ NIPS*, 2015.

[26] Uddalok Sen, Tryambak Gangopadhyay, Chandrachur Bhattacharya, Achintya Mukhopadhyay, and Swarnendu Sen. Dynamic characterization of a ducted inverse diffusion flame using recurrence analysis. *Combustion Science and Technology*, 190(1):32–56, 2018.

[27] Ushnish Sengupta, Carl Rasmussen, and Matthew Juniper. Bayesian machine learning for the prognosis of combustion instabilities from noise. 2020.

[28] Simon R Stow and Ann P Dowling. Low-order modelling of thermoacoustic limit cycles. In *ASME turbo expo 2004: power for land, sea, and air*, pages 775–786. American Society of Mechanical Engineers Digital Collection, 2004.

[29] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[31] Yandong Wen, Bhiksha Raj, and Rita Singh. Face reconstruction from voice using generative adversarial networks. 2019.

[32] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

Supplementary Materials

S.1 Ablation Study

The training framework of VSenseNet II additionally comprises of image classifier, while that of VSenseNet I only includes the time series encoder and the image decoder. For the ablation study of VSenseNet I and VSenseNet II, we conduct multiple experiments.

S.1.1 VSenseNet I

The VSenseNet I modeling approach is illustrated in Fig. 9. The trainable parts of the VSenseNet I framework are the time series encoder and image decoder, and the pre-trained (and fixed) part is the image encoder. The weights of VSenseNet I are learned with two learning objectives - minimizing the embedding loss and minimizing the reconstruction loss. The pre-trained image encoder is utilized to compute the 128-dimensional image embedding. The trainable time series encoder generates the 128-dimensional time series embedding. The embedding loss is the mean squared error (MSE) computed between the time series embedding and image embedding. In the training process, the time series encoder learns to compute an embedding that can match the image embedding as closely as possible. The input to the image decoder is the time series embedding, from which it attempts to reconstruct the image corresponding to the input time series. The learning objective of the image decoder is to minimize the reconstruction loss between the actual flame image and the reconstructed image.

As part of ablation study for VSenseNet I, we develop the models VSenseNet I(A) and VSenseNet I(B), demonstrated in Fig. 10 and Fig. 11 respectively. In VSenseNet I(A), we remove the embedding loss - the time series encoder and the image decoder are trained with only reconstruction loss. In VSenseNet I(B), we add another loss function in addition to the embedding loss and reconstruction loss. The additional loss is added after the second convolutional transpose layer of the image decoder to facilitate knowledge distillation from the image decoder pre-trained on images. From Table 2, we observe that while VSenseNet I(A) is better than VSenseNet I(B) in terms of reconstruction performance, VSenseNet I(B) outperforms VSenseNet I(A) in terms of classification performance. Overall, VSenseNet I performs better in terms of both reconstruction and classification performance.

S.1.2 VSenseNet II

Compared to VSenseNet I, VSenseNet II has the image classifier model in the training framework apart from the time series encoder and the image decoder. We demonstrate the training framework of VSenseNet II in Fig. 12. The trainable parts are the time series encoder, image decoder, and image classifier. With the time series as input, the time series encoder computes a 128-dimensional time series embedding. From a time series embedding, the image decoder learns to reconstruct the corresponding flame image as closely as possible. With the reconstructed flame image as input, the image classifier model predicts it as stable or unstable.

As part of the ablation study for VSenseNet II, we develop the model VSenseNet II(A) demonstrated in Fig. 13. While VSenseNet II consists of two steps for training, VSenseNet II(A) consists of a single step to train the time series encoder, image decoder, and image classifier. From Table 3, we observe that in terms of classification performance VSenseNet II is better than VSenseNet II(A). For reconstruction, VSenseNet II(A) performs better, but the performance is not as good as VSenseNet I.

S.2 Convolutional Autoencoder

Autoencoders can learn meaningful representations using a compression function (encoder) and a decompression function (decoder). The encoder compresses the input into a low dimensional embedding, and the decoder reconstructs the high dimensional information from that embedding. Without the requirement of explicit annotations, the weights of an autoencoder model can be learned with the objective of minimizing the reconstruction loss. The first step of developing our proposed framework is to utilize the training dataset of flame images to pre-train a convolutional autoencoder which comprises an image encoder and an image decoder as demonstrated in Fig. 14. The encoder takes in a flame image (resolution 64 x 64) as input. The encoder model comprises a series of 2D convolutional and 2D max-pooling layers. After that, a fully connected layer is used to compute a 128-dimensional embedding. From the 128-dimensional embedding, the decoder attempts to reconstruct the original flame image as closely as possible using a series of 2D up-sampling and 2D convolutional transpose layers. The details of the image encoder and decoder model are shown in Fig. 15 and Fig. 16 respectively.

S.3 Baseline Models

S.3.1 Image Classifier

This image-based model is a binary classification model which takes a flame image as input and classifies it as stable or unstable. It is trained on actual training set images and tested on actual test set images. The image classifier model comprises 2D convolutional, 2D max-pooling, and fully connected layers. The model architecture of the image classifier is shown in Fig. 17.

S.3.2 Time Series Classifier

This is an entirely time series based model with no cross-modal reconstruction of visual features. The model archi-

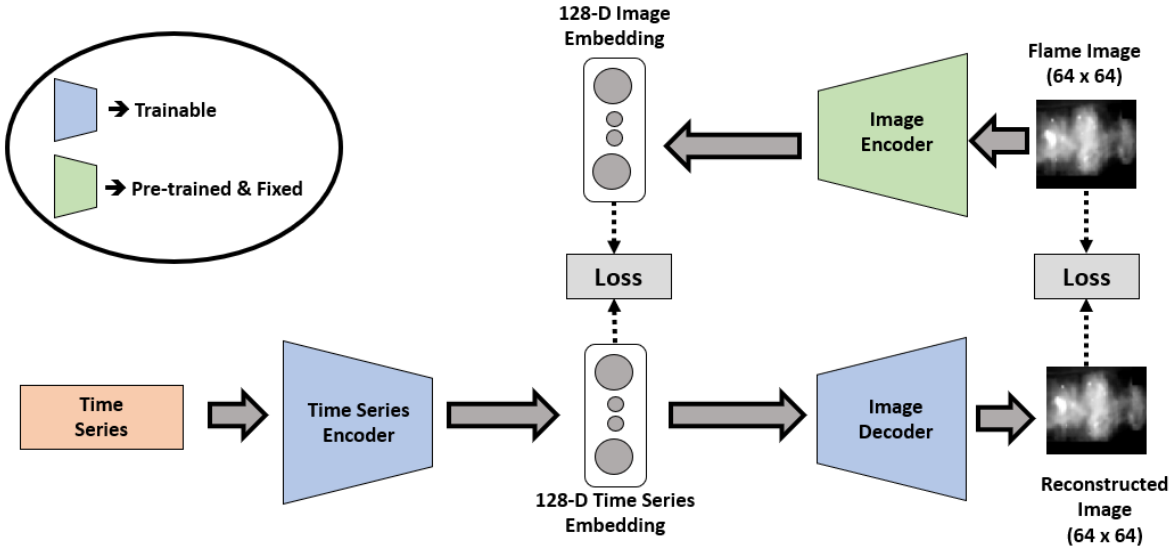


Figure 9. Training framework of VSenseNet I. It utilizes the pre-trained image encoder to compute the image embedding. The time series encoder and image decoder are trained with two loss functions - embedding loss and reconstruction loss.

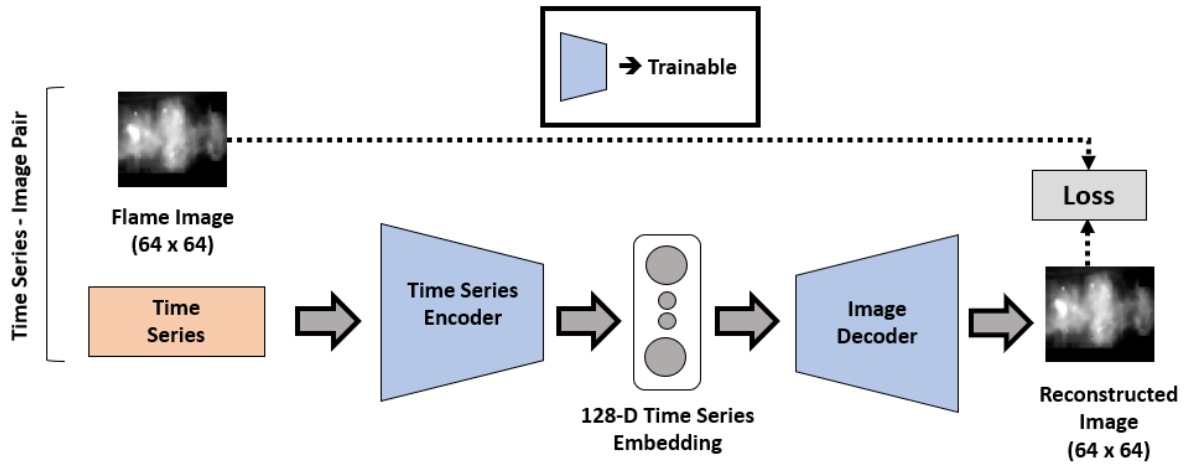


Figure 10. Training framework of VSenseNet I(A). The time series encoder and image decoder are trained with reconstruction loss.

tecture of the time series classifier has been demonstrated in Fig. 18. With the multivariate time series as input, the model predicts it as stable or unstable. The time series classifier model is trained using time series only, and it is tested using time series of the test set.

S.4 Dataset Collection

For dataset collection, we induce combustion instability in a laboratory-scale swirl combustor (Fig. 19), which has a swirler of diameter 30 mm and vane angles of 60 degrees. Air is provided to the combustor through a settling chamber of diameter 28 cm and thereafter through a square cross-section of side 6 cm. The experimental setup includes an inlet section, an inlet optical access module (IOAM), a primary combustion chamber, and a secondary duct. The IOAM facilitates optical access to the fuel tube. The fuel is injected co-axially with air through a fuel injection tube, having slots on the surface at selected distances upstream of the swirler.

The chosen upstream distances are 90 mm and 120 mm. For the upstream distance of 90 mm, partial premixing of the fuel with air occurs, while the distance of 120 mm facilitates full premixing of the fuel and air. Based on the swirler diameter, different airflow rates are chosen for a fixed fuel flow rate. In another way, the inlet air flow rate can be kept fixed for different fuel flow rates.

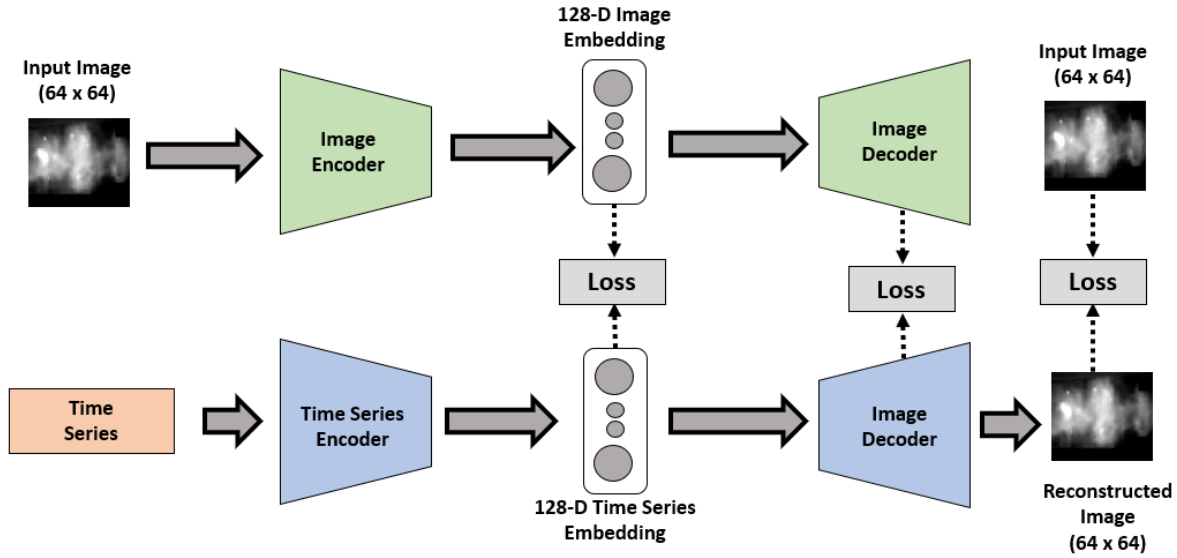


Figure 11. Training framework of VSenseNet I(B). The time series encoder and image decoder are trained with three loss functions.

Model	Reconstruction Performance		Classification Performance		
	SSIM	Mean Squared Error	Accuracy	F1 Score	FNR
Image Classifier	NA	NA	**0.9938 ± 0.0064**	**0.9938 ± 0.0065**	**0.0122 ± 0.0129**
Time Series Classifier	NA	NA	**0.9850 ± 0.0021**	**0.9847 ± 0.0022**	**0.0300 ± 0.0044**
Cross-Modal	0.6655	0.0072	0.9888 ± 0.0005	0.9887 ± 0.0005	0.0222 ± 0.0010
VSenseNet I	**0.6980**	0.0062	**0.9895 ± 0.0003**	**0.9894 ± 0.0002**	**0.0209 ± 0.0006**
VSenseNet I(A)	0.6965	**0.0061**	0.9859 ± 0.0011	0.9857 ± 0.0012	0.0280 ± 0.0023
VSenseNet I(B)	0.6894	0.0063	0.9890 ± 0.0002	0.9889 ± 0.0002	0.0218 ± 0.0005

Table 2. Ablation study with VSenseNet I for the test set. For classification performance, average and standard deviation of the evaluation metrics (Accuracy, F1 Score, FNR) are reported after training each model five times. For reconstruction performance, average values of SSIM and MSE are reported.

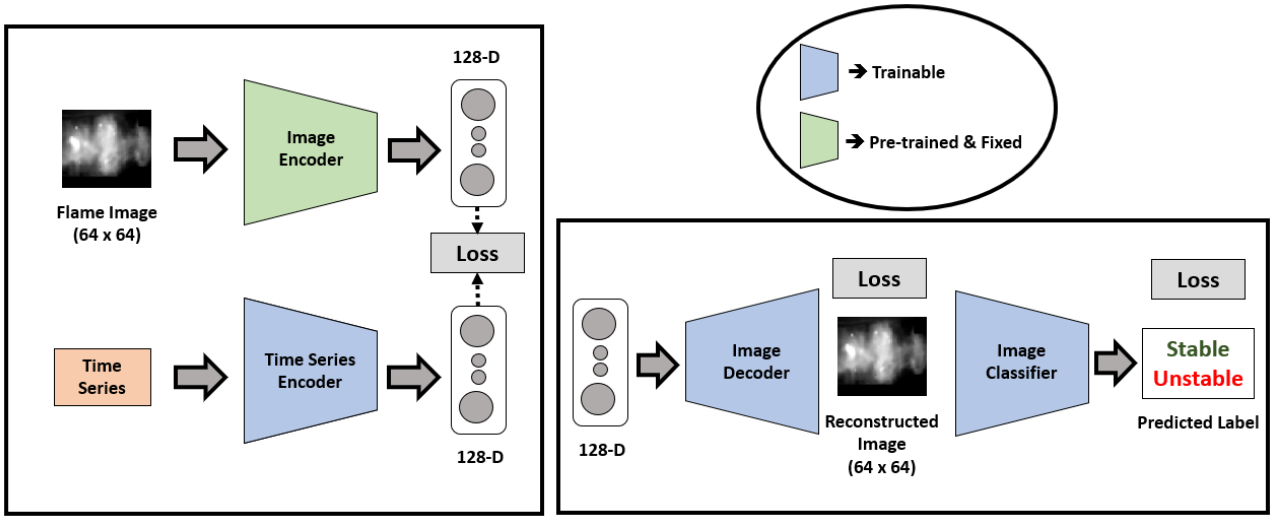


Figure 12. Training framework of VSenseNet II. It comprises two steps. The first step is to train the time series encoder. The next step is to train the image decoder and image classifier utilizing reconstruction loss and classification loss.

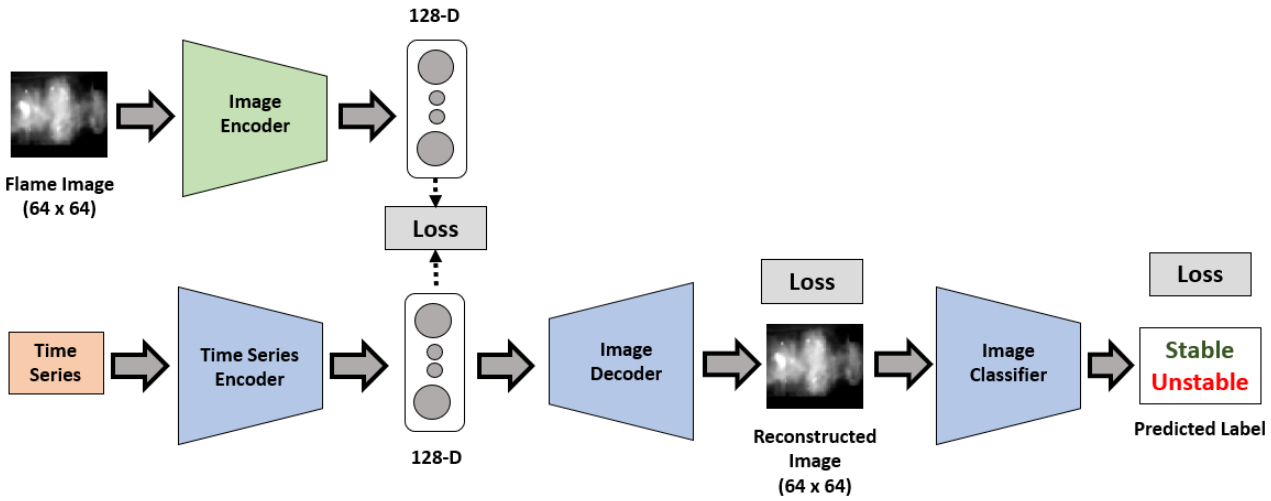


Figure 13. Training framework of VSenseNet II(A). It consists of a single step to train the time series encoder, image decoder and image classifier.

Model	Reconstruction Performance		Classification Performance		
	SSIM	Mean Squared Error	Accuracy	F1 Score	FNR
Image Classifier	NA	NA	**0.9938 ± 0.0064**	**0.9938 ± 0.0065**	**0.0122 ± 0.0129**
Time Series Classifier	NA	NA	**0.9850 ± 0.0021**	**0.9847 ± 0.0022**	**0.0300 ± 0.0044**
Cross-Modal	0.6655	0.0072	0.9888 ± 0.0005	0.9887 ± 0.0005	0.0222 ± 0.0010
VSenseNet II	0.6876	0.0070	**0.9901 ± 0.0003**	**0.9900 ± 0.0003**	**0.0197 ± 0.0007**
VSenseNet II(A)	**0.6978**	**0.0065**	0.9898 ± 0.0006	0.9897 ± 0.0006	0.0202 ± 0.0012

Table 3. Ablation study with VSenseNet II for the test set. For classification performance, average and standard deviation of the evaluation metrics (Accuracy, F1 Score, FNR) are reported after training each model five times. For reconstruction performance, average values of SSIM and MSE are reported.

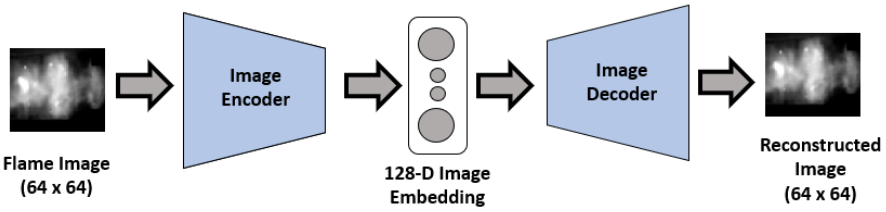


Figure 14. The encoder and decoder of the convolutional autoencoder model used for pre-training.

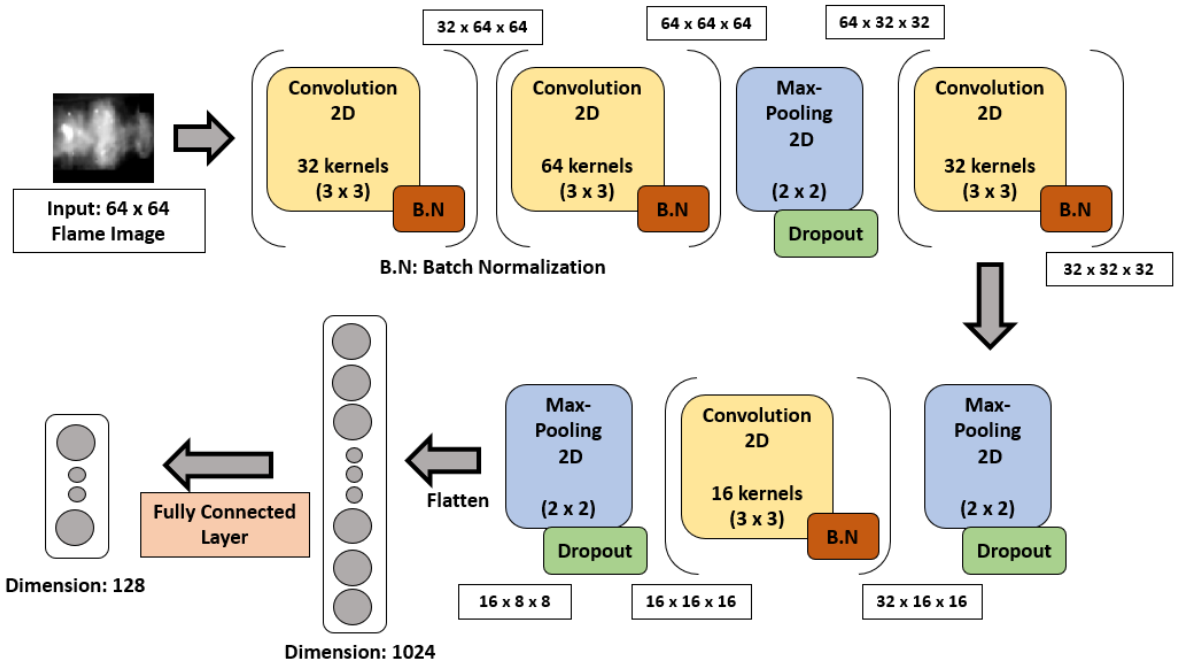


Figure 15. Encoder model for pre-training of convolutional autoencoder.

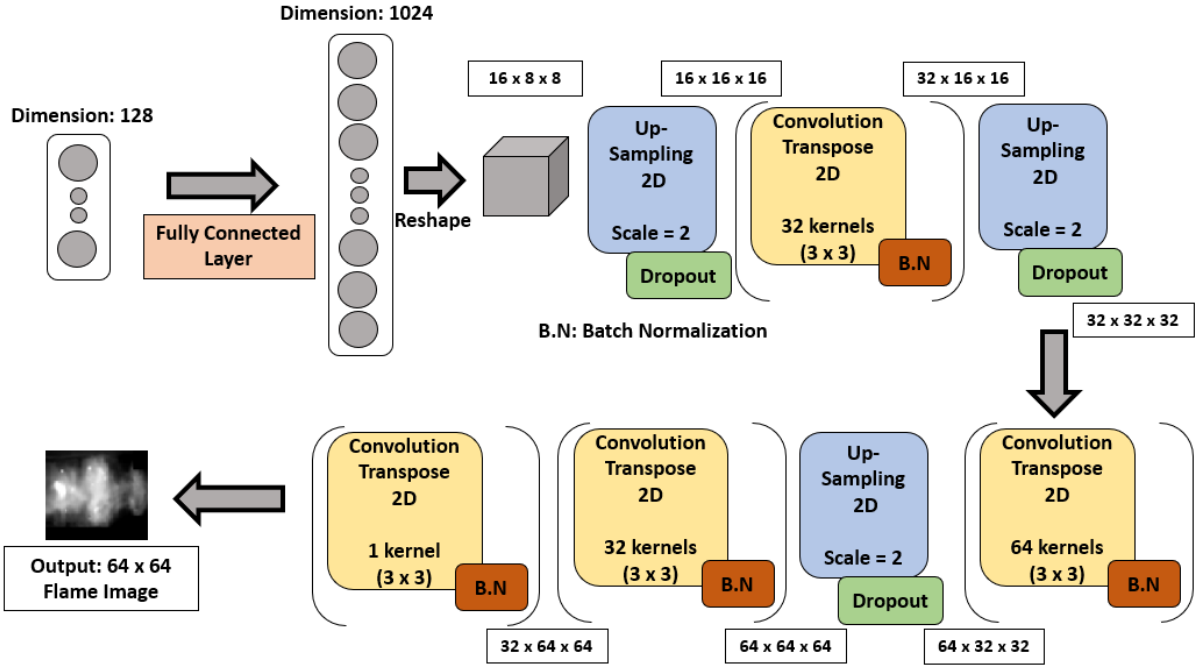


Figure 16. Decoder model for pre-training of convolutional autoencoder.

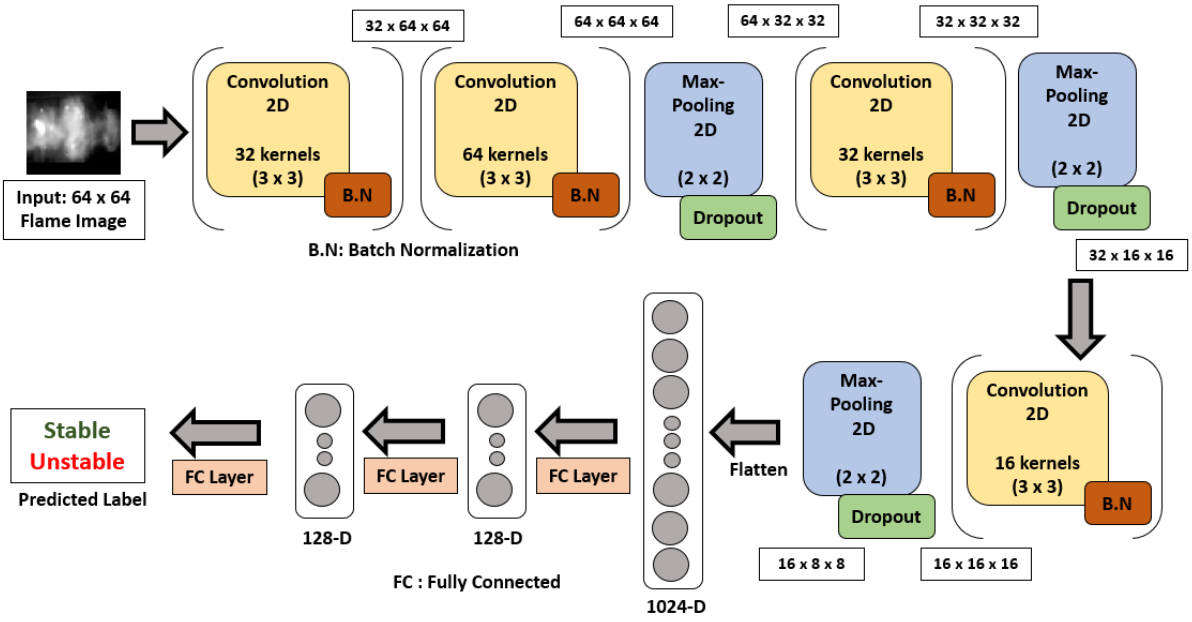


Figure 17. Image Classifier model classifies a flame image as stable or unstable.

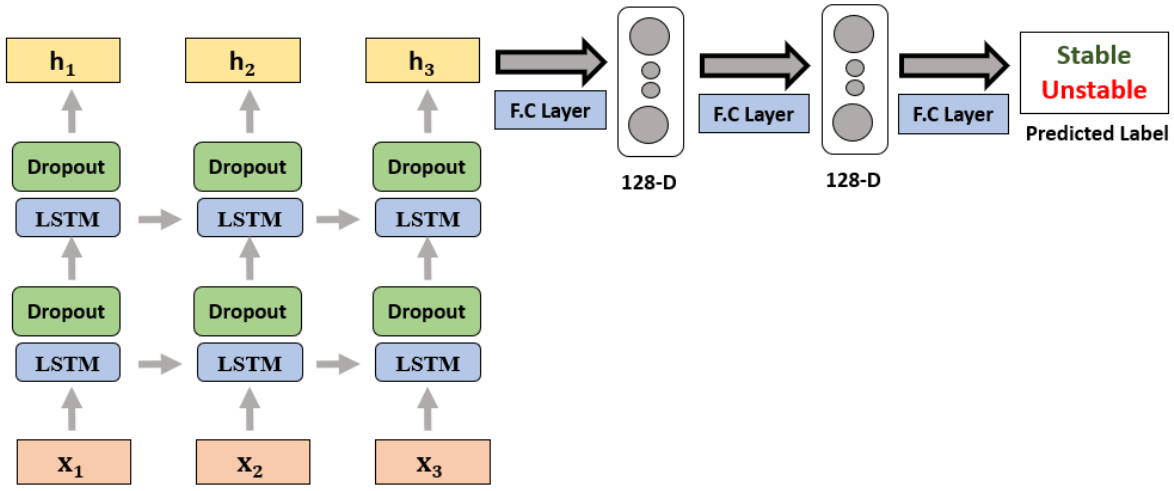
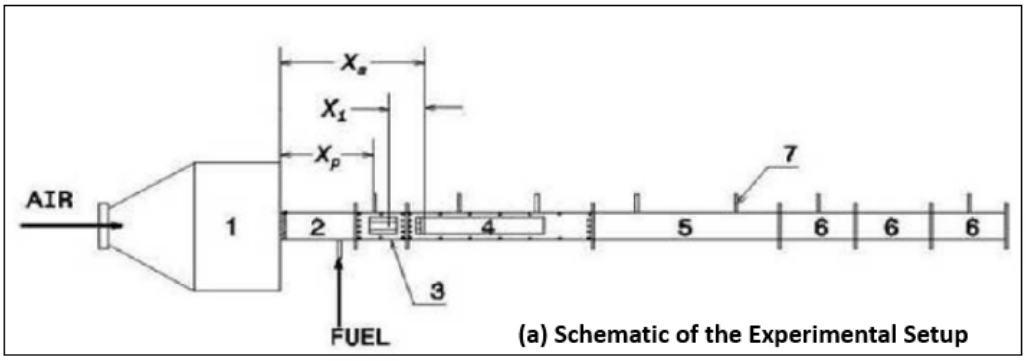
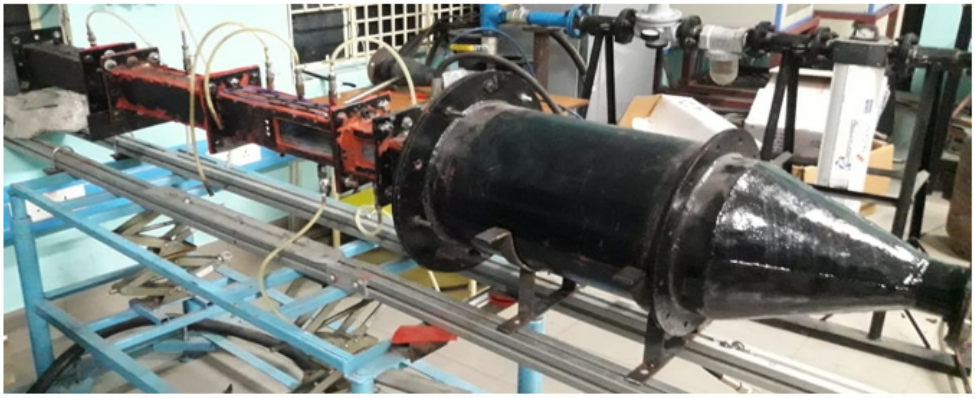


Figure 18. Time Series Classifier model classifies an acoustic pressure time series as stable or unstable.





(b) Image of the Laboratory Scale Combustor

Figure 19. The laboratory-scale combustor used for data collection. (a) Schematic of the experimental setup. 1 - settling chamber, 2 - inlet duct, 3 - IOAM, 4 - test section, 5 - big extension duct, 6 – small extension ducts, 7 - pressure transducers, Xs - swirler location measured downstream from settling chamber exit, Xp - transducer port location measured downstream from settling chamber exit, Xi - fuel injection location measured upstream from swirler exit. (b) Image of the laboratory scale combustor.