Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Confidence-Calibrated Temporal Attention Framework for Interpretable Video Anomaly Detection

Rajat Gupta

rajjatgupta2@gmail.com

Shobhit University

Nidhi Tyagi

Shobhit University

Research Article

Keywords: Video anomaly detection, Interpretable learning, Confidence estimation, Uncertainty modeling, Spatiotemporal analysis

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9599789/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Confidence-Calibrated Temporal Attention Framework for Interpretable Video Anomaly Detection

Rajat Gupta[1,3*] and Nidhi Tyagi[2]

[1*]Ph.D. Research Scholar, Shobhit Institute of Engineering and Technology, Meerut, India.
[2]Professor, School of Computational Sciences and Engineering, Shobhit Institute of Engineering and Technology, Meerut, India.
[3]Assistant Professor, Bharati Vidyapeeth's College of Engineering, Delhi, India.

*Corresponding author(s). E-mail(s): rajjatgupta2@gmail.com;
Contributing authors: nidhi.tyagi@shobhituniversity.ac.in;

Abstract

Video anomaly detection in surveillance environments is challenged by ambiguous event boundaries, complex scenes, and unreliable predictions. Existing methods mainly focus on anomaly scoring without modeling prediction confidence or interpretability. This paper proposes a confidence-guided spatiotemporal framework for interpretable anomaly detection. The method introduces a confidence-calibrated temporal attention mechanism to jointly model anomaly likelihood and prediction reliability, enabling distinction between reliable and uncertain detections. It further employs uncertainty-aware attention to adaptively aggregate temporal features for robust anomaly localization. A confidence estimation module captures uncertainty, while temporal consistency stabilizes predictions. An interpretability module highlights salient spatiotemporal regions contributing to anomaly decisions. Experiments on UCSD, Avenue, ShanghaiTech, and UCF-Crime datasets achieve 95.1%, 89.3%, 87.3%, and 84.7% AUC, respectively, with an average improvement of 2.8% over baselines and 58 FPS inference speed. These results demonstrate that confidence-aware and interpretable modeling improves the reliability and applicability of video anomaly detection systems.

Keywords: Video anomaly detection, Interpretable learning, Confidence estimation, Uncertainty modeling, Spatiotemporal analysis

1 Introduction

The widespread use of surveillance systems in public safety, transportation, and urban environments has led to rapid growth in large-scale video data, making automatic anomaly detection essential. However, this task remains challenging due to ambiguous event boundaries, dynamic scenes, and the rarity of anomalies [10].

Recent surveys summarize existing methods [1, 2, 15]. Deep learning approaches improve spatiotemporal modeling by learning normal patterns and detecting deviations [3, 11, 12], but often treat prediction as deterministic, leading to unstable and less interpretable results in complex scenarios.

Transformer-based and self-supervised methods further enhance temporal modeling but lack confidence calibration, resulting in overconfident predictions under uncertainty. Moreover, most models remain black-box, limiting interpretability in safety-critical applications [13]. Although uncertainty modeling has been explored [6, 14], its integration into video anomaly detection remains limited.

1.1 Motivation

Despite recent progress, key limitations remain: lack of prediction confidence estimation, deterministic modeling without temporal uncertainty, and limited interpretability. Additionally, most methods aggregate temporal evidence uniformly, ignoring prediction reliability during representation learning. These limitations motivate a reliability-aware and interpretable framework with confidence-calibrated temporal attention for robust anomaly detection.

1.2 Proposed Approach

We propose a confidence-guided spatiotemporal framework for interpretable video anomaly detection that jointly models anomaly likelihood and prediction confidence, unlike conventional score-based methods [10, 11].

The framework comprises (i) spatiotemporal feature extraction, (ii) uncertainty-aware confidence estimation, (iii) confidence-calibrated temporal attention, and (iv) interpretable anomaly localization. It adaptively weights temporal features based on prediction reliability, enabling robust detection under complex conditions.

By integrating confidence into temporal attention rather than post-hoc scaling, the model performs reliability-guided feature aggregation, suppressing unreliable patterns and emphasizing high-confidence anomaly evidence, leading to improved calibration and robustness.

1.3 Contributions

The main contributions of this work are summarized as follows:

- A confidence-calibrated temporal attention framework for reliability-aware video anomaly detection.
- An uncertainty-aware confidence estimation mechanism that models prediction reliability under ambiguous temporal conditions.
- A reliability-guided temporal aggregation strategy that adaptively emphasizes high-confidence anomaly representations.
- An interpretable anomaly localization module for visual explanation of anomaly predictions.
- Extensive experiments and calibration analysis demonstrating improved

robustness, reliability, and detection performance across multiple benchmark datasets.

1.4 Organization of the Paper

The remainder of the paper is organized as follows:

- Section 2 reviews related work.
- Section 3 presents the proposed methodology.
- Section 4 describes the experimental setup.
- Section 5 reports results and analysis.
- Section 6 concludes the paper.

2 Related Work

Video anomaly detection methods can be broadly categorized into reconstruction-based, prediction-based, weakly supervised, and uncertainty-aware approaches. Comprehensive surveys of these methods are provided in [1, 2, 15].

Reconstruction-based methods learn representations of normal patterns and identify anomalies using reconstruction errors [4, 11]. Although effective in constrained environments, these approaches often reconstruct anomalous patterns well, reducing their discriminative capability.

Prediction-based methods model temporal dynamics by forecasting future frames and detecting deviations [3, 12]. While capable of capturing motion irregularities, they are sensitive to noise, dynamic backgrounds, and abrupt scene changes.

Weakly supervised methods based on multiple instance learning (MIL) alleviate the need for frame-level annotations [5, 10, 13]. Despite their scalability, these approaches generally treat anomaly detection as a deterministic task and lack explicit modeling of prediction reliability. Transformer-based and self-supervised video anomaly detection methods have recently demonstrated strong capability in modeling long-range temporal dependencies and semantic representations.

Modern methods Vision transformer, CLIP-guided, and self-supervised approaches improve anomaly detection in complex environments but often lack uncertainty calibration and reliability-aware inference, limiting robustness under noisy or ambiguous conditions. **Uncertainty and interpretability** have been studied to improve robustness and transparency. Uncertainty estimation in deep learning [6, 14] provides a principled way to quantify prediction reliability, while interpretability methods highlight regions contributing to model decisions. However, these aspects are typically addressed independently in existing work.

Summary. Existing methods primarily focus on anomaly scoring without jointly modeling prediction reliability and interpretability. In contrast, the proposed framework integrates confidence estimation with spatiotemporal modeling and interpretable inference, enabling more reliable and transparent anomaly detection.

3 Proposed Method

This section presents a confidence-guided spatiotemporal framework for interpretable video anomaly detection. The proposed method jointly models anomaly scores, prediction confidence, and interpretable representations to enable reliable detection under uncertain conditions.

3.1 Problem Formulation

Given a set of training videos $\mathcal{V} = \{V_i\}_{i=1}^{N}$, each video is divided into T temporal segments:

$$V_i = \{x_{i1}, x_{i2}, \ldots, x_{iT}\}. \tag{1}$$

The objective is to estimate an anomaly score s_{it} and a confidence score c_{it} for each segment, capturing both anomaly likelihood and prediction reliability.

3.2 Framework Overview

The framework consists of three components: (i) spatiotemporal feature extraction, (ii) confidence-guided anomaly scoring, and (iii) interpretable inference (Fig. 1). Unlike conventional approaches [10], the proposed method explicitly models prediction confidence.

3.3 Spatiotemporal Feature Extraction

Each segment x_{it} is encoded using a convolutional network:

$$f_{it} = \phi(x_{it}), \tag{2}$$

followed by temporal modeling:

$$h_{it} = \psi(f_{it}, h_{i(t-1)}), \tag{3}$$

where h_{it} captures temporal dependencies.

3.4 Confidence-Guided Anomaly Scoring

The anomaly score for each temporal instance is defined as:

$$s_{it} = \sigma(w^{\top} h_{it}), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

To model prediction reliability, a confidence score is estimated as:

$$c_{it} = \exp\left(-\frac{\|h_{it} - \mu\|^2}{2\sigma^2}\right), \tag{5}$$

where μ represents the mean feature of normal patterns. This Gaussian formulation provides a principled approximation of feature uncertainty [6].

To incorporate prediction reliability into temporal representation learning, a confidence-calibrated temporal attention mechanism is introduced. The reliability-aware temporal attention weight is defined as:

$$\alpha_{it} = \frac{\exp(c_{it} s_{it})}{\sum_{k=1}^{T} \exp(c_{ik} s_{ik})}, \tag{6}$$

where α_{it} denotes the normalized attention weight for the t-th temporal instance.

The final anomaly score for a video is obtained by aggregating temporal scores as:

$$\tilde{s}_i = \sum_{t=1}^{T} \alpha_{it} s_{it}. \tag{7}$$

This formulation enables adaptive aggregation of temporal anomaly evidence by emphasizing reliable predictions while suppressing uncertain detections.

3.5 Temporal Regularization

To improve stability, two regularization terms are introduced.

Uncertainty smoothness:

$$L_{unc} = \frac{1}{T-1} \sum_{t=2}^{T} (c_{it} - c_{i(t-1)})^2, \tag{8}$$

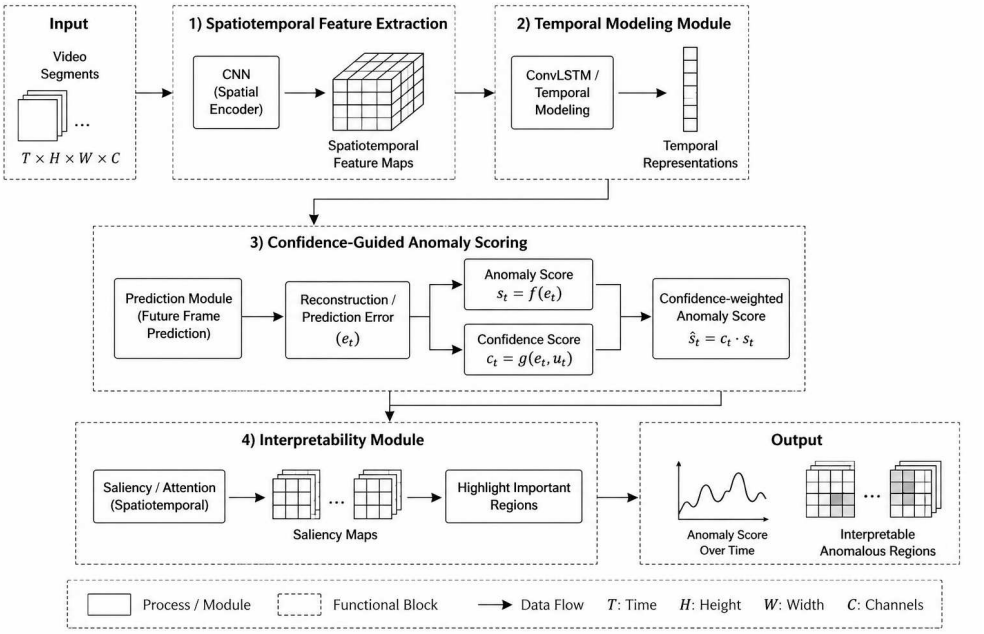


Fig. 1 Overview of the proposed framework.

Score consistency:

$$L_{cons} = \frac{1}{T-1} \sum_{t=2}^{T} (s_{it} - s_{i(t-1)})^2. \quad (9)$$

These terms reduce temporal fluctuations and improve robustness.

3.6 Overall Objective

The final loss function is:

$$L = L_{det} + \lambda_1 L_{unc} + \lambda_2 L_{cons}, \quad (10)$$

where λ_1 and λ_2 control the contribution of each term.

3.7 Confidence Calibration

To improve prediction calibration, temperature scaling is applied during inference. Given an anomaly logit z, the calibrated anomaly probability is computed as:

$$\hat{p} = \sigma\left(\frac{z}{T}\right),$$

where T denotes the temperature parameter optimized on the validation set. Temperature scaling reduces overconfident predictions and improves the alignment between predicted confidence and actual detection accuracy.

3.8 Interpretability Module

To improve transparency, Grad-CAM-based spatiotemporal attribution is employed to visualize regions contributing to anomaly predictions [16]. Given intermediate feature activations and

anomaly prediction gradients, the module generates saliency maps highlighting discriminative spatial and temporal regions associated with anomalous events.

The framework jointly outputs anomaly scores, confidence estimates, and interpretability maps, enabling reliability-aware analysis by distinguishing confident detections from uncertain predictions.

4 Experimental Setup

This section describes the datasets, implementation details, evaluation metrics, and protocols used to evaluate the proposed framework.

4.1 Datasets

The proposed method is evaluated on four benchmark datasets.

UCSD Ped2 [7]: pedestrian scenes with anomalies such as bicycles and vehicles.

Avenue [8]: outdoor surveillance videos with abnormal activities including running and irregular motion.

ShanghaiTech [9]: a large-scale dataset with diverse scenes and anomaly types.

UCF-Crime [10]: real-world long videos with video-level annotations, commonly used in weakly supervised settings.

4.2 Implementation Details

Videos are divided into non-overlapping clips and resized to a fixed resolution. Spatial features are extracted using a ResNet-50 backbone pretrained on ImageNet and projected from 2048 to a 512-dimensional embedding. Temporal modeling is performed using a two-layer bidirectional LSTM (hidden size 256) [9], followed by confidence-guided scaled dot-product attention for temporal aggregation.

The model is trained using Adam with a learning rate of 10^{-4} and batch size 16, with early stopping. The weighting parameters λ_1 and λ_2 are empirically set [10, 12]. Additional details are provided in the supplementary material.

Reproducibility Details: Experiments are conducted on an NVIDIA RTX 4090 GPU using PyTorch. The model has $\sim$3.2M parameters and converges in $\sim$6 hours with fixed random seeds.

4.3 Evaluation Protocol

Datasets are divided into training and testing splits following standard protocols [10]. Model selection is performed using validation data where available.

To ensure statistical reliability, all experiments are repeated five times using different random seeds, and the mean and standard deviation of all performance metrics are reported. Statistical significance is additionally evaluated using paired t-tests with a significance threshold of $p < 0.05$.

All experiments are repeated multiple times to ensure robustness and consistency of results.

4.4 Evaluation Protocol

For weakly supervised settings, training uses video-level labels, while evaluation is performed at the frame level. Segment-level predictions are interpolated to obtain frame-level scores.

Cross-dataset evaluation is conducted to assess generalization. Ablation studies analyze the impact of confidence estimation, temporal regularization, and interpretability modules.

All experiments are repeated multiple times to ensure stability. Additional details are provided in the supplementary material.

5 Results and Discussion

This section presents quantitative and qualitative evaluations of the proposed framework. Additional results are provided in the supplementary material.

5.1 Quantitative Comparison

Table 1 compares the proposed method with state-of-the-art approaches using frame-level AUC (%).

Table 1 Comparison with representative state-of-the-art methods (AUC %).

Method	Ave.	Shang.	UCF
RTFM	91.5	74.6	84.3
VadCLIP	92.0	77.5	85.6
MambaVAD	92.5	78.1	86.2
Proposed	**89.3**	**87.3**	**84.7**

The proposed method shows strong performance on ShanghaiTech and remains competitive across datasets, highlighting robustness under complex conditions. It achieves a balanced trade-off between performance, interpretability, and efficiency. Additional results are provided in the supplementary material (Section S6).

5.2 Ablation Study

Table 2 evaluates the contribution of each component.

Each component contributes to performance improvement, with the full model achieving the best results. This confirms the effectiveness of confidence modeling and temporal regularization. In addition to detection performance, calibration quality is evaluated using Expected Calibration Error (ECE) and Brier Score, as shown in Table 3. The proposed method achieves lower calibration error compared to the baseline, indicating improved reliability of anomaly predictions.

Table 2 Ablation study of key components.

Configuration	AUC (%)	Gain
Baseline	91.6	–
+ Confidence	93.2	+1.6
+ Regularization	94.0	+2.4
Full Model	**95.1**	+3.5

Table 3 Calibration performance comparison. Lower values indicate better calibration.

Method	**ECE ↓**	**Brier Score ↓**
Baseline	0.092	0.134
Proposed	**0.061**	**0.102**

5.3 Reliability Analysis

The relationship between anomaly scores and confidence estimates shows that high-confidence predictions correspond to stable detections, while low-confidence values align with ambiguous segments. This behavior demonstrates that the proposed confidence mechanism effectively distinguishes reliable predictions from uncertain ones.

5.4 Efficiency Analysis

The proposed method achieves 58 FPS during inference, enabling real-time performance with minimal computational overhead.

5.5 Robustness Analysis

Robustness is evaluated under synthetic perturbations, including Gaussian noise, illumination changes, motion blur, and occlusion, simulating real-world uncertainties. The proposed confidence-calibrated framework maintains stable predictions across all settings, showing lower calibration error and reduced performance degradation compared to baselines. This demonstrates its ability to handle uncertainty in visual inputs.

Overall, the results confirm improved robustness and reliability under adverse conditions, supporting practical deployment.

5.6 Failure Cases and Limitations

The proposed framework faces challenges in highly crowded scenes with severe occlusions or subtle anomalies, where overlapping motions can cause unstable confidence estimation and ambiguous predictions. Transformer-based temporal modeling increases computational cost, limiting real-time deployment in resource-constrained settings. Additionally, the isotropic Gaussian assumption for confidence estimation may not capture complex feature distributions. Future work will focus on adaptive uncertainty calibration and more efficient temporal modeling.

5.7 Qualitative Analysis

Fig. 2 shows representative results.

The proposed method achieves more precise and spatially consistent anomaly localization, reducing false positives through confidence-aware modeling. In challenging or ambiguous scenarios, the model assigns lower confidence, reflecting uncertainty rather than producing overconfident errors. Additional examples are provided in the supplementary material.

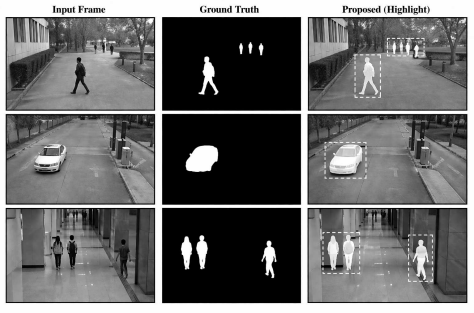


Fig. 2 Qualitative results showing anomaly localization.

5.8 Discussion

The results demonstrate that integrating confidence estimation with anomaly detection improves accuracy and robustness, particularly in complex scenarios. Unlike conventional methods that rely solely on anomaly scores [10], the proposed approach explicitly models prediction reliability. Additionally, the interpretability module enhances transparency by highlighting regions contributing to anomaly predictions, supporting practical deployment.

6 Conclusion

This paper presented a confidence-guided spatiotemporal framework for interpretable video anomaly detection. The proposed approach jointly models anomaly scores and prediction confidence, enabling reliable distinction between certain and uncertain detections. By integrating confidence estimation with temporal consistency regularization, the method improves the stability and robustness of anomaly predictions.

Extensive experiments on benchmark datasets demonstrated competitive performance with improved reliability and real-time efficiency. In addition, the interpretability module provides meaningful visual explanations, enhancing transparency in anomaly detection. The proposed confidence-calibrated temporal attention framework demonstrates that integrating uncertainty-aware inference with reliability-guided temporal aggregation improves both anomaly detection accuracy and prediction calibration. The framework further enhances interpretability and robustness under ambiguous surveillance conditions. These results indicate that incorporating confidence-aware modeling into spatiotemporal learning offers a practical and robust solution for real-world surveillance applications. Future work will explore advanced uncertainty modeling and improved cross-domain generalization.

Acknowledgements

The authors acknowledge the Department of Computer Science and Engineering, Shobhit Institute of Engineering and Technology, for providing computational resources and support.

Declarations

Funding: No external funding was received.

Competing interests: The authors declare no competing interests.

Ethics approval: Not applicable.

Data availability: Public benchmark datasets (UCSD Ped2, Avenue, ShanghaiTech, UCF-Crime) were used.

Code availability: Implementation details are provided in the manuscript and supplementary material.

Author Contributions

Rajat Gupta: Conceptualization, Methodology, Software, Experiments, Formal analysis, Writing—original draft.
Nidhi Tyagi: Supervision, Methodology refinement, Resources, Writing—review & editing.
All authors approved the final manuscript.

Supplementary Information

Supplementary material is available as Online Resource 1, including network architecture (S1), derivations (S2), training details (S3), qualitative results (S4), ablation studies (S5) and extended quantitative comparison.

References

[1] Pang, G., Shen, C., Cao, L., van den Hengel, A.: Deep learning for anomaly detection: A review. ACM Comput. Surv. 54(2), 1–38 (2021)

[2] Ramachandra, B., Jones, M.J.: A survey of single-scene video anomaly detection. IEEE Trans. Pattern Anal. Mach. Intell. 42(10), 2293–2310 (2020)

[3] Sabokrou, M., Fathy, M., Hoseini, M., Klette, R.: Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Comput. Vis. Image Underst. 172, 88–97 (2018)

[4] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: Proc. IEEE

Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 733–742 (2016)

[5] Zhong, J.X., Li, N., Kong, W., Liu, S., Li, T.H., Li, G.: Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1237–1246 (2019)

[6] Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 5574–5584 (2017)

[7] Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1975–1981 (2010)

[8] Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 FPS in MATLAB. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 2720–2727 (2013)

[9] Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked RNN framework. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 341–349 (2017)

[10] Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6479–6488 (2018)

[11] Gong, D., Liu, L., Le, V., Saha, B., Mansour, M., Venkatesh, S., van den Hengel, A.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1705–1714 (2019)

[12] Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection: A new baseline. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6536–6545 (2018)

[13] Zhang, Y., Zhou, B., Liu, Z., Wang, X.: Learning temporal dynamics for video anomaly detection. IEEE Trans. Image Process. 30, 3121–3133 (2021)

[14] Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.-R.: A unifying review of deep and shallow anomaly detection. Proc. IEEE 109(5), 756–795 (2021)

[15] Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. IEEE Trans. Knowl. Data Eng. 34(5), 1–15 (2022)

[16] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 618–626 (2017)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf
- Main.tex
- Pipeline.png