Anomaly Detection in Surveillance Video: A Comparative Study using PaDiM, SPADE, and CLIP

Tatyana Amugo
Department of Computer Science
University Name

August 4, 2025

Abstract

This capstone project investigates anomaly detection in surveillance video through three complementary perspectives: statistical distribution modeling (PaDiM), visual attention explainability (SPADE), and multimodal language-vision embedding similarity (CLIP). Each approach captures anomalies using distinct mechanisms—distribution modeling, saliency activation, and semantic proximity, respectively. Using the ShanghaiTech Campus dataset, we applied comprehensive data preprocessing, extracted thousands of image frames, and conducted systematic evaluations using ROC-AUC metrics, score distribution analysis, and visual interpretation tools. Our findings demonstrate that while PaDiM struggles with motion-less contexts, CLIP excels at semantic distinctions with an AUC of **0.797**. SPADE, designed primarily for saliency visualization, provides valuable interpretability but limited discrimination capability. Together, these models reveal complementary strengths essential for building robust anomaly detection pipelines in real-world surveillance applications.

1 Introduction

Anomaly detection in computer vision represents a fundamental challenge in intelligent surveillance systems, industrial inspection protocols, and safety-critical monitoring applications [1]. Unlike traditional supervised classification tasks, anomaly detection often operates under the constraint of limited or non-existent labeled anomalous data. Consequently, models must learn to generalize from normal data patterns and identify deviations in visual structure, semantic content, or temporal dynamics [2].

The complexity of real-world anomalies necessitates a multi-faceted approach to detection. This project explores anomaly detection through three distinct yet complementary model families, each offering unique perspectives on what constitutes an anomaly:

1. **PaDiM (Patch Distribution Modeling):** Leverages statistical distance measures in convolutional neural network feature spaces to model normal data distributions at the patch level.

2. **SPADE (Sub-Image Anomaly Detection with Deep Pyramid Analysis):** Employs gradient-based attention mechanisms using Grad-CAM to identify anomalous visual activation patterns.

3. **CLIP (Contrastive Language-Image Pre-training):** Utilizes multimodal language-vision embeddings to classify images based on semantic proximity to learned prototypes.

By comparing these fundamentally different approaches, we aim to understand their relative strengths, limitations, and potential for integration in comprehensive anomaly detection systems.

2 Dataset and Methodology

2.1 Dataset Selection: ShanghaiTech Campus

The ShanghaiTech Campus dataset serves as our evaluation benchmark, chosen for its realistic surveillance scenarios and comprehensive ground-truth annotations [6]. This dataset encompasses:

- **Training data:** Over 330 surveillance videos capturing normal campus activities
- **Testing data:** 107 labeled videos containing both normal and anomalous events
- **Scene diversity:** Multiple environments including walkways, plazas, building entrances, and outdoor spaces
- **Anomaly types:** Diverse unusual events such as cycling in pedestrian areas, sudden crowd formations, and atypical movement patterns

2.2 Frame Extraction and Preprocessing

To adapt video data for image-based anomaly detection models, we implemented a comprehensive frame extraction pipeline:

Algorithm 1 Frame Extraction and Labeling Process

1: Extract all frames from training and testing videos using OpenCV
2: Apply ground truth masks to categorize frames as "normal" or "abnormal"
3: Balance dataset by sampling equal numbers from each category
4: Resize frames to model-specific input dimensions
5: Normalize pixel values for neural network compatibility

This preprocessing yielded a total of 274,515 extracted frames, from which we selected 8,000 balanced frames (4,000 per class) for systematic evaluation. While this frame-based approach simplifies model training and inference, it inherently discards temporal context—a limitation that significantly impacts motion-dependent detection methods.

2.3 Model-Specific Data Utilization

Each model required different data preparation strategies:

Table 1: Data Utilization Summary by Model

Model	Training Frames	Test Frames	Purpose
PaDiM	2,000 (normal only)	8,000	Patch distribution modeling
SPADE	100 (normal only)	8,000	Reference CAM computation
CLIP	4,000 (2,000 per class)	8,000	Prototype generation (K-Means)

3 Model Analysis and Implementation

3.1 PaDiM: Patch Distribution Modeling

3.1.1 Theoretical Foundation

PaDiM operates on the principle that features extracted from normal image patches follow multivariate Gaussian distributions [3]. The model learns these distributions during training and identifies anomalies as patches exhibiting significant statistical deviations from the learned normal patterns.

3.1.2 Architecture and Implementation

Our PaDiM implementation utilized the following configuration:

- **Backbone network:** ResNet-18 pre-trained on ImageNet
- **Feature extraction layers:** `layer1`, `layer2`, and `layer3`
- **Dimensionality reduction:** Features compressed to 300 dimensions using random projection
- **Statistical modeling:** Multivariate Gaussian distributions computed per spatial location

3.1.3 Results and Analysis

PaDiM achieved an ROC-AUC score of **0.603**, indicating moderate discrimination capability. The model generated interpretable heatmaps highlighting regions of statistical deviation, but showed limited sensitivity to subtle, localized anomalies. The absence of temporal information particularly impacted performance on motion-based anomalies.

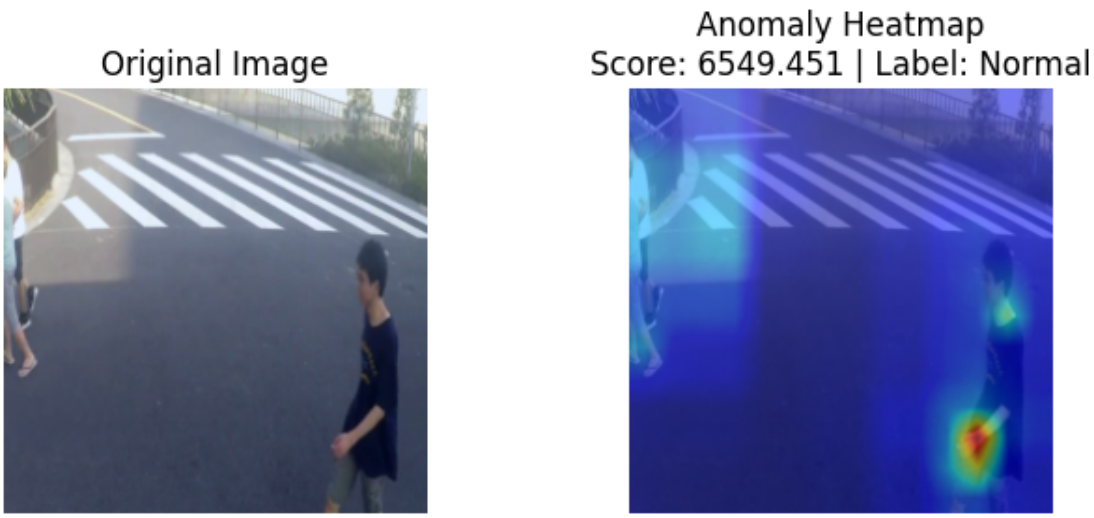


Figure 1: PaDiM anomaly heatmap demonstration. Red regions indicate higher statistical deviation from learned normal distributions. The model successfully identifies structural anomalies but struggles with subtle contextual deviations.

3.2 SPADE: Gradient-Based Attention Analysis

3.2.1 Methodological Approach

SPADE leverages the hypothesis that normal images produce consistent neural activation patterns. By analyzing Class Activation Maps (CAMs) generated through Grad-CAM [8], the model identifies anomalies as deviations from typical attention patterns.

3.2.2 Implementation Pipeline

The SPADE pipeline consists of several key components:

1. **CAM Generation:** Extract Grad-CAM visualizations from ResNet-18 layers 2–4
2. **Reference Creation:** Average normal CAMs to establish baseline attention patterns
3. **Similarity Measurement:** Compute Structural Similarity Index (SSIM) between test CAMs and reference
4. **Anomaly Scoring:** Lower SSIM scores indicate higher anomaly likelihood

3.2.3 Performance and Limitations

SPADE obtained an ROC-AUC of **0.509**, suggesting near-random performance for our dataset. The primary challenges included:

- **Static image limitations:** Absence of temporal dynamics reduced saliency effectiveness

- **SSIM sensitivity:** High sensitivity to minor spatial shifts caused excessive false positives
- **Attention consistency:** Normal surveillance scenes showed high variability in attention patterns

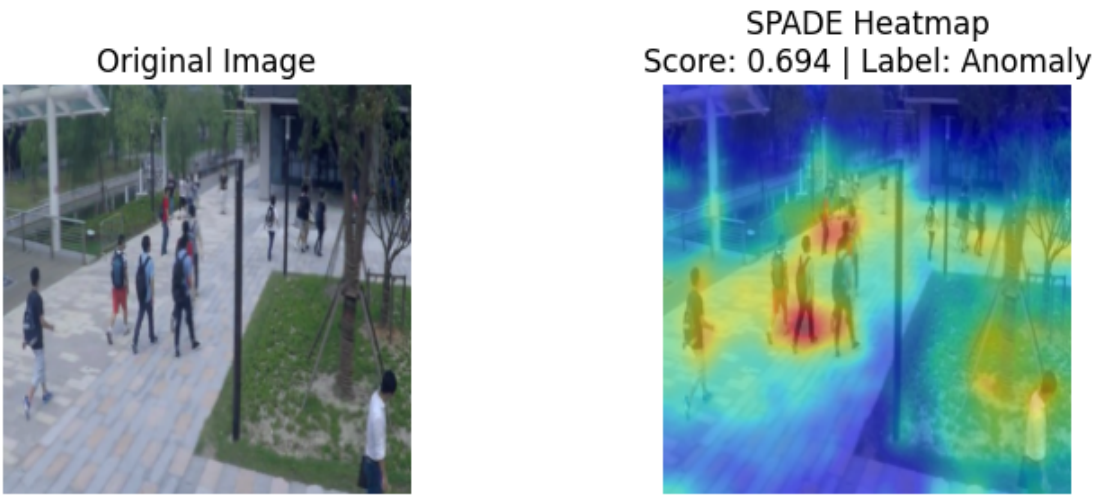


Figure 2: SPADE Class Activation Map overlay. The heatmap shows regions of high neural attention, with warmer colors indicating stronger activation. Inconsistencies with reference patterns suggest potential anomalies.

3.3 CLIP: Multimodal Semantic Analysis

3.3.1 Conceptual Framework

CLIP's approach to anomaly detection leverages its pre-trained multimodal understanding to map images into a semantic embedding space [5]. We adapted this capability for visual anomaly detection by comparing test images to learned class-specific prototype vectors.

3.3.2 Prototype-Based Detection Strategy

Our CLIP implementation employed multiple similarity computation methods:

- **Prototype Generation:** K-Means clustering (k=100) applied to image embeddings from each class
- **Max Similarity:** Highest cosine similarity to any prototype
- **Top-k Average:** Mean similarity to k most similar prototypes
- **Distance-Based Scoring:** Euclidean distance measurements in embedding space
- **Weighted Ensemble:** Combination of multiple similarity metrics

3.3.3 Superior Performance Results

CLIP demonstrated exceptional performance with an ensemble ROC-AUC of **0.797**, significantly outperforming the other models. Key strengths included:

- **Semantic understanding:** Effective detection of contextually inappropriate objects (e.g., bicycles in pedestrian areas)
- **Robust feature representation:** Pre-trained embeddings captured relevant visual semantics
- **Scalable architecture:** Prototype-based approach enables efficient inference

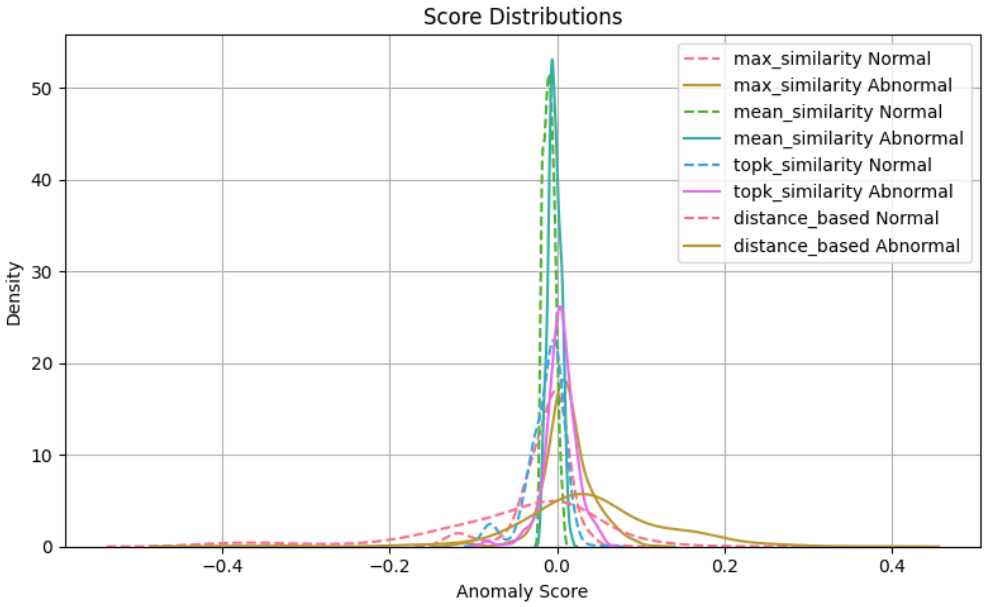


Figure 3: CLIP anomaly score distributions for normal (blue) and abnormal (red) samples. Clear separation between distributions demonstrates the model's strong discrimination capability, with minimal overlap in the decision boundary region.

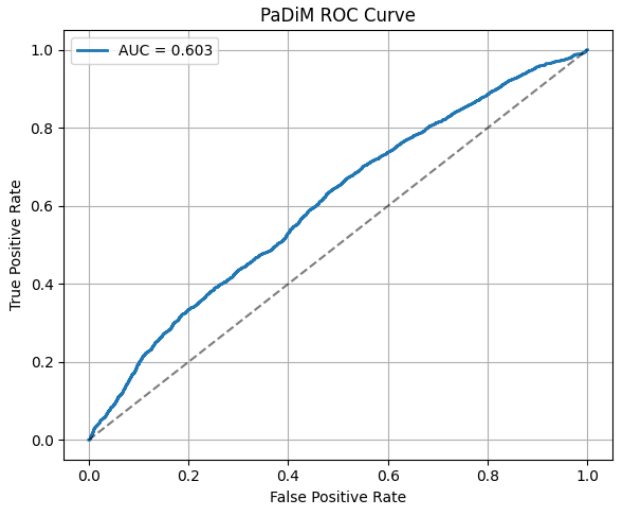
4 Comparative Analysis

4.1 Performance Comparison

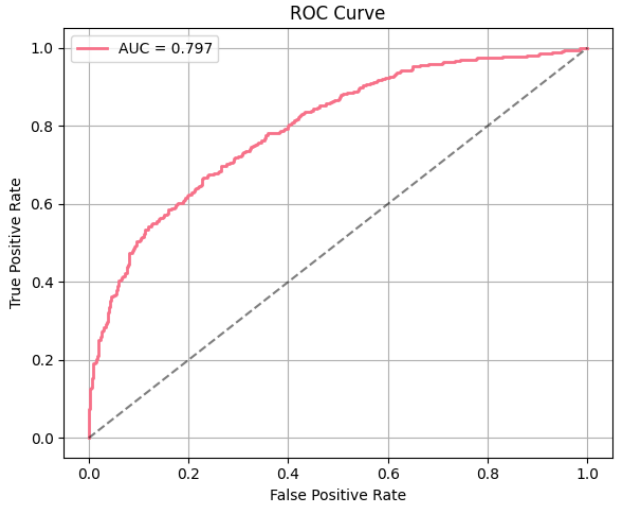
Table 2: Quantitative Performance Summary

Model	ROC-AUC	Computational Cost	Interpretability
PaDiM	0.603	Medium	High (Heatmaps)
SPADE	0.509	Low	High (Attention Maps)
CLIP	**0.797**	High	Medium (Embeddings)

4.2 ROC Curve Analysis



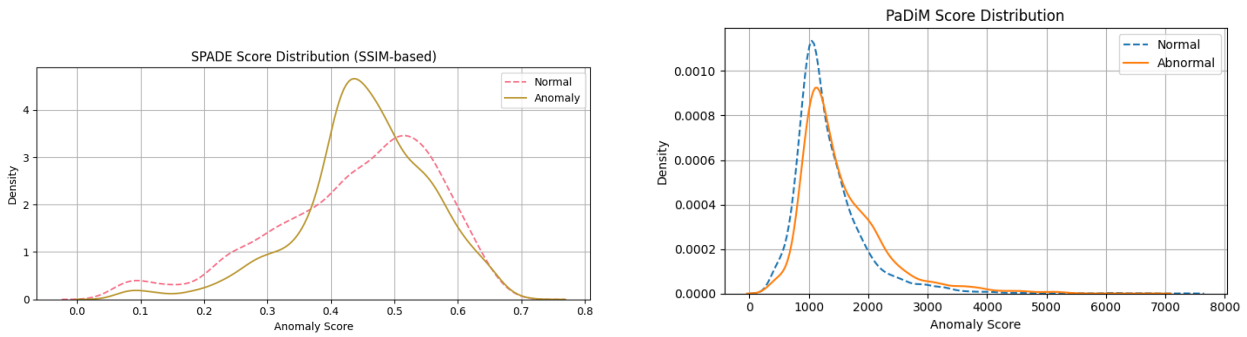
(a) PaDiM ROC curve



(b) CLIP ROC curve

Figure 4: Individual ROC curves for each model. (a) PaDiM shows moderate discrimination with AUC = 0.603, and (b) CLIP demonstrates superior performance with AUC = 0.797.

4.3 Score Distribution Analysis



(a) SPADE Score Distribution

(b) PaDiM Score Distribution

Figure 5: Kernel Density Estimation (KDE) plots comparing anomaly scores for normal and abnormal frames. Both SPADE and PaDiM exhibit considerable overlap between the two distributions, suggesting limited discriminative power. The lack of clear separation makes reliable thresholding challenging.

5 Technical Challenges and Limitations

Our experimental methodology encountered several significant challenges that influenced model performance:

5.1 Temporal Information Loss

The conversion from video sequences to static frames fundamentally eliminated temporal dynamics, severely impacting models designed to leverage motion information. This particularly affected:

- SPADE's saliency detection, which relies on temporal attention consistency
- PaDiM's ability to detect motion-based anomalies
- Overall system capability to identify behavioral anomalies requiring temporal context

5.2 Computational Resource Requirements

All models demanded substantial GPU resources, with training and inference times extending to several hours. This limitation constrained:

- Extensive hyperparameter optimization
- Large-scale dataset experimentation
- Real-time deployment feasibility

5.3 Dataset-Specific Limitations

The ShanghaiTech dataset's characteristics introduced specific challenges:

- Variable lighting conditions affecting visual consistency
- Diverse scene layouts reducing model generalization
- Subjective anomaly definitions complicating ground truth reliability

6 Future Research Directions

Building upon our findings, several promising research avenues emerge:

6.1 Temporal Integration

- **3D CNN Integration:** Implement I3D or similar architectures for motion-aware anomaly detection
- **Sequence Modeling:** Utilize Transformers or LSTM networks for temporal pattern recognition
- **Optical Flow Analysis:** Incorporate motion vectors for enhanced temporal understanding

6.2 Multimodal Enhancement

- **Advanced Vision-Language Models:** Explore BLIP [14] or GPT-4V for improved semantic understanding
- **Segmentation Integration:** Incorporate Segment Anything Model (SAM) [15] for object-level analysis
- **Audio-Visual Fusion:** Include acoustic information for comprehensive scene understanding [16]

6.3 System Integration

- **Ensemble Methods:** Develop sophisticated fusion strategies combining all three approaches
- **Contextual Metadata:** Integrate temporal, spatial, and environmental context
- **Adaptive Thresholding:** Implement dynamic decision boundaries based on scene characteristics

7 Conclusion

This comprehensive study evaluated three distinct paradigms for anomaly detection in surveillance video: statistical distribution modeling (PaDiM), visual attention analysis (SPADE), and semantic embedding similarity (CLIP). Each method demonstrated unique strengths and revealed specific limitations when applied to real-world surveillance data.

PaDiM provided valuable interpretability through statistical heatmaps but showed moderate discrimination performance (AUC = 0.603) due to its reliance on spatial features without temporal context. SPADE offered insights into neural attention patterns but struggled with the static nature of our frame-based approach (AUC = 0.509). CLIP emerged as the clear performance leader (AUC = 0.797), leveraging its pre-trained semantic understanding to effectively distinguish between normal and anomalous visual content.

The complementary nature of these approaches suggests significant potential for integrated anomaly detection systems. Future research should focus on temporal information integration, multimodal enhancement, and sophisticated ensemble methods that leverage the unique strengths of each paradigm. Such integrated approaches could provide the robust, interpretable, and accurate anomaly detection capabilities required for real-world surveillance applications.

Our findings contribute to the growing understanding of anomaly detection methodologies and provide a foundation for developing more sophisticated, multi-faceted detection systems capable of handling the complexity and variability inherent in real-world surveillance scenarios [17, 18].

8 References

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[2] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021.

[3] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: a patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*, 2021, pp. 475–489.

[4] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," *arXiv preprint arXiv:2005.02357*, 2021.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Oskamp, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[6] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 341–349.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, 1967, pp. 281–297.

[11] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022, pp. 12888–12900.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[16] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *European Conference on Computer Vision*, 2016, pp. 801–816.

[17] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[18] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6536–6545.