BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

VIDEO DENOISING USING DEEP LEARNING

POTLAČENÍ ŠUMU VE VIDEU POMOCÍ HLUBOKÝCH NEURONOVÝCH SÍTÍ

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR **MAKSIM NAUMENKO**
AUTOR PRÁCE

SUPERVISOR **Ing. MICHAL ŠPANĚL, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2024

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Bachelor's Thesis Assignment

154387

Institut:	Department of Computer Graphics and Multimedia (DCGM)
Student:	**Naumenko Maksim**
Programme:	Information Technology
Title:	**Video Denoising Using Deep Learning**
Category:	Image Processing
Academic year:	2023/24

Assignment:

1. Get familiar with the problem of noise removal in both image and video sequences, and principles of deep neural networks and their learning.
2. Find and study existing methods for image and video denoising using deep neural networks.
3. Create a dataset for your own experiments.
4. Choose appropriate methods and design a neural network architecture for video denoising.
5. Experiment with your implementation and possibly introduce your own modifications to the network design.
6. Compare the achieved results and discuss the possibilities of future development.
7. Create a brief poster, or a video, presenting your work, its goals and results.

Literature:

- Zhang, K., et. al., "Residual Learning of Deep CNN for Image Denoising," in *IEEE Transactions on Image Processing*, 2017, https://github.com/ocimakamboj/DnCNN, https://github.com/cszn/KAIR.
- Lehtinen, J. et al., "Noise2Noise: Learning Image Restoration without Clean Data", 2018, https://arxiv.org/abs/1803.04189.
- Zhao, H. et al., "Loss Functions for Image Restoration With Neural Networks," in *IEEE Transactions on Computational Imaging*, 2017, https://arxiv.org/pdf/1511.08861.pdf.

Requirements for the semestral defence:

- Fulfillment of the first three points of the assignment and partially the fourth point.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Španěl Michal, Ing., Ph.D.**
Head of Department:	Černocký Jan, prof. Dr. Ing.
Beginning of work:	1.11.2023
Submission deadline:	9.5.2024
Approval date:	9.11.2023

Abstract

In the era of digital multimedia, video content quality significantly impacts user experiences and system performance, particularly in domains such as entertainment, and video and image processing. This thesis addresses the persistent challenge of video noise, which degrades video quality, through the use of advanced deep learning techniques. Initially, traditional video denoising approaches are reviewed to establish a foundational understanding of denoising concepts. Subsequently, two state-of-the-art models, FastDVDNet and ViDeNN, are studied to familiarize with neural network architectures. The main product of this work is the development of a robust video denoising pipeline that utilizes a UNet architecture inspired by these state-of-the-art models. Throughout the thesis, the proposed UNet Baseline, ResUNet, and ResUNet Temporal models are explained, implemented, and evaluated to demonstrate their effectiveness in video denoising.

Abstrakt

V éře digitálních multimédií kvalita videoobsahu významně ovlivňuje uživatelský zážitek a výkon systému, zejména v oblastech, jako je zábava a zpracování videa a obrazu. Tato práce se zabývá přetrvávajícím problémem šumu ve videu, který zhoršuje jeho kvalitu, a to pomocí pokročilých technik hlubokého učení. Nejprve jsou přezkoumány tradiční přístupy k odstraňování šumu ve videu, aby bylo možné nastínit základní koncepty denoisingu. Následně jsou studovány dva referenční modely, FastDVDNet a ViDeNN, za účelem seznámení se s architekturami neuronových sítí. Hlavním výsledkem této práce je vývoj robustního systému pro odstraňování šumu ve videu, který je založen na architektuře UNet inspirované těmito referenčními modely. V průběhu práce jsou vysvětleny, implementovány a vyhodnoceny navrhované modely UNet Baseline, ResUNet a ResUNet Temporal, aby byla prokázána jejich účinnost v odstraňování šumu ve videu.

Keywords

deep learning, deep neural networks, convolutional neural networks, digital noise, video denoising, image denoising, UNet

Klíčová slova

hluboké učení, hluboké neuronové sítě, konvoluční neuronové sítě, digitální šum, denoising videa, denoising obrazu, UNet

Reference

NAUMENKO, Maksim. *Video Denoising Using Deep Learning.* Brno, 2024. Bachelor's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Ing. Michal Španěl, Ph.D.

Rozšířený abstrakt

V současné době digitální média mají obrovský vliv na život každého člověka, od zábavy po autonomní systémy. Stálý problém šumu v obrazech a videích nejen zhoršuje kvalitu, ale také ztěžuje funkci různých technik zpracování digitálních médií. Tato práce se zabývá odstraňováním šumu ve videu pomocí metod hlubokého učení, aby účinně řešila tyto výzvy.

Na začátku této práce je poskytnut základní přehled toho, co je šum, jak vzniká a jak ovlivňuje různá digitální média. Následně jsou prozkoumány různé tradiční techniky pro odstranění šumu ve videu, které ukazují svá omezení, jako je neschopnost zachovat jemné detaily a konzistenci v dynamických scénách. Proto nás tlačí k prozkoumání pokročilých řešení pomocí hlubokého učení, která jsou schopna překonat tyto omezení.

Následně práce zkoumá dva existující referenční modely, které jsou známé v oblasti odstraňování šumu ve videu: FastDVDNet a ViDeNN. Tyto modely byly použity k pochopení principů neuronových sítí a jejich architektury zaměřené na odstraňování šumu, včetně jejich silných a slabých stránek. Po prozkoumání těchto modelů jako základ pro vývoj vlastních modelů byla zvolena architektura UNet.

Před tím, než byly tyto modely vytvořeny pro testovací účely a experimenty, byl vytvořen vlastní dataset. Tento dataset byl vytvořen na základě existujícího datasetu Vimeo-90K, který poskytuje obrovské množství video sekvencí různých typů a druhů. Tento dataset byl následně zpracován pomocí existujících modelů pro generování šumu C2N. Tento model dokáže generovat šum, který je co nejvíce podobný šumu z reálného světa. Tím pádem na základě těchto dvou komponent vznikl nový dataset, který obsahuje velký rozsah různých scén ve video sekvencích a navíc obsahuje náročný a komplexní typ šumu.

Počáteční experiment zahrnoval trénování základního modelu pro vytvoření výchozího bodu pro odstraňování šumu ve videu a seznámení se s procesem trénování neuronové sítě. Základní model UNet Baseline byl trénován na navrženém datasetu. Validace navrhované neuronové sítě poskytla PSNR a SSIM skóre 33,4340 a 0,9440, což jsou vynikající výsledky pro základní model. Nicméně vizuální inspekce odhalila, že model má potíže s udržováním konzistence mezi snímky. Navíc produkoval vizuální artefakty, jako je blikání v některých video sekvencích, což vedlo ke zhoršení kvality videa ke konci sekvencí.

Druhý experiment se zaměřil na zdokonalení základního modelu UNet, což vedlo k vytvoření modelu ResUNet. Tato iterace využívá reziduálních bloků, což umožňuje hlubší architekturu modelu. Kromě toho byly provedeny úpravy v účelové funkci, která kombinuje SSIM a MAE pro zlepšení účinnosti. Validace modelu odhalila nižší výkon ve srovnání se základním modelem UNet Baseline, s PSNR a SSIM skóre 31,5674 a 0,9344. Nicméně vizuální inspekce ukázala významné zlepšení v udržování konzistence mezi snímky.

Poslední experiment vycházel z výsledků dvou předchozích modelů a vyvinul model ResUNet Temporal. Tento model klade důraz na temporální konzistenci ve video sekvencích tím, že mění způsob zpracování vstupních snímků a toho, co generuje. Na rozdíl od předchozích modelů, které generovaly jediný snímek bez šumu, model ResUNet Temporal generuje tři snímky. Tato změna umožnila integraci temporální účelové funkce, což zlepšilo schopnost modelu udržovat konzistenci mezi snímky. Výsledky validace ukázaly, že model úspěšně odstranil šum z video sekvencí a dosáhl PSNR a SSIM skóre 34,6109 a 0,9344. Navíc vizuální inspekce potvrdila, že tento model produkoval nejlépe odšuměná videa ze všech testovaných modelů, udržujíc plynulý pohyb bez výrazných artefaktů.

V závěrečném porovnání s předtrénovaným modelem ViDeNN bylo zjištěno, že navržený model ResUNet Temporal překonal tento referenční model na navrženém datasetu. Dosáhl lepších skóre validačních metrik a produkoval vyšší kvalitu videí, což dokazuje úspěch experimentů.

Závěrem lze konstatovat, že aplikace technik hlubokého učení ve sféře odstraňování šumu z videa měla významný dopad a poskytla nástroje, které mohou překonat tradiční metody. Modely navržené v této studii prokázaly vynikající výsledky na vybraném datasetu, zdůrazňující potenciál hlubokého učení efektivně zlepšovat kvalitu videa.

Video Denoising Using Deep Learning

Declaration

I hereby declare that this Bachelor's thesis was prepared as an original work by the author under the supervision of Mr. Ing. Michal Španěl Ph.D I have listed all the literary sources, publications and other sources, which were used during the preparation of this thesis.

.....
Maksim Naumenko
May 7, 2024

Acknowledgements

I would like to express my gratitude to my supervisor, Ing. Michal Španěl, Ph.D., for his guidance and valuable advice during my work on this thesis, which were essential in helping me understand my research topic more deeply. Many thanks to all my family members who supported me throughout my studies. Last but not least, I would like to thank my friends for your motivating words and support during this journey.

Contents

List of Figures

Chapter 1

Introduction

In the era of digital media, video content has become an essential part of our lives, affecting various aspects of it - from entertainment and social media to surveillance and autonomous systems. Since the demand for high-quality video remains constant, the problem associated with video noise - distortions that usually occur during the capture, processing, or transmission stages - is a serious problem. Video noise not only degrades quality but also affects the performance of various computer vision applications, including object recognition and tracking tasks.

Traditional video denoising methods, such as wavelet-based and filter-based techniques, have been extensively explored to solve this issue. However, these methods often struggle with the trade-off between noise reduction and the preservation of important details and temporal consistency. With a huge improvement in the sphere of machine learning and neural networks, a new approach has been integrated into the field of video processing. Deep learning models, characterized by their ability to learn complex patterns and dependencies from data, have shown noticeable success in image and video denoising tasks, often outperforming traditional algorithms in both effectiveness and efficiency.

The main goal of this thesis is to explore deep learning concepts and their application to video denoising through the development of a robust denoising pipeline. For this purpose, a unique dataset was created by combining the Vimeo-90k [29] dataset with a C2N [13] noise generation model, introducing a wide variety of scenes characterized by more realistic, real-world synthetic noise. During this investigation, two state-of-the-art models utilizing the UNet architecture, FastDVDNet [25] and ViDeNN [7], will be examined. These models will serve as a foundation for the proposed models in this work: UNet Baseline, ResUNet, and ResUNet Temporal.

All proposed models will be evaluated and compared on this newly created dataset and benchmarked against the pre-trained, state-of-the-art ViDeNN model. The final evaluations will show that the ResUNet Temporal model significantly outperforms the ViDeNN model on the custom dataset, achieving scores of PSNR (34.6109) and SSIM (0.9344), compared to ViDeNN's PSNR (31.9800) and SSIM (0.8381). Furthermore, ResUNet Temporal produces denoised video sequences with enhanced temporal consistency, ensuring smooth motion and sharp quality in each frame without any highly noticeable artifacts. In contrast, some video sequences denoised by the ViDeNN model contain visual artifacts.

Chapter 2

Principles of Denoising

This chapter covers the fundamentals of video denoising, starting with a brief overview of video noise and its impact on video clarity. It then delves into the principles of denoising, highlighting some traditional techniques used to improve video quality. Finally, the chapter discusses the metrics for evaluating the effectiveness of denoising methods and explores the challenges faced in the field of video processing.

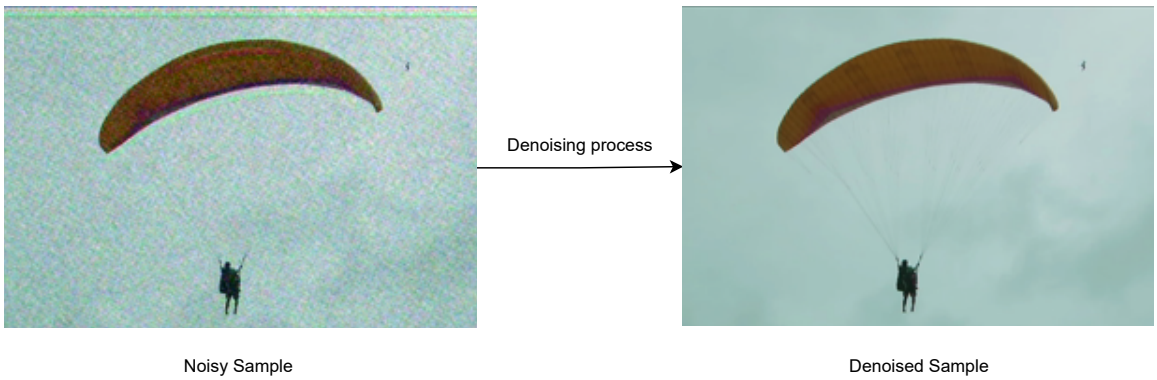


Figure 2.1: Example of the denoising.

2.1 Nature of Noise

In digital images and videos, noise manifests as unwanted or random variations in visual data, deviating from the intended scene. This phenomenon, common in digital imaging, appears as grainy or speckled disturbances, diminishing overall quality. Sources of noise are varied, ranging from sensor limitations during capture, and environmental conditions, to processing techniques. Noise, typically viewed as disruptive, obscures the true content and reduces image or video fidelity.

The characteristics of noise in digital media include:

- **Graininess** [15]: A primary indicator of noise is a grainy texture, causing images or videos to appear rough or granular, especially in areas that should be smooth. This graininess affects the aesthetic and clarity of the media.

- **Color Distortions** [15]: Noise can lead to random color shifts, resulting in unnatural hues and saturation levels. These distortions impact the visual appeal and interpretability of the media.
- **Pixel Variations** [15]: At its essence, noise is seen as random fluctuations in pixel values. These inconsistencies in brightness and color may manifest as flickering or shifting in videos.

Understanding these characteristics is crucial for identifying and addressing noise. Recognizing its disruptive nature and common attributes aids in the development of effective noise mitigation techniques.

2.1.1 Types of Noise

To better understand the complexity of noise in digital media, it's important to explore the different types of noise we often encounter:

- **Gaussian Noise** [5]: This common type of noise is caused by electronic interference in the camera sensor. It results in a normal distribution of pixel values the image and is easily visible in areas with the same color or brightness. Gaussian noise is a fundamental problem in digital media processing that affects its quality.
- **Salt-and-Pepper Noise** [5]: This type of noise, characterized by random black and white pixels that appear scattered across the image, is usually caused by sudden interference in the signal, such as data transmission errors or defective pixels in the camera sensor.
- **Speckle Noise** [5]: Especially common in radar and ultrasonic images, occurs as a result of coherent processing of signals reflected from multiple scatterers. This type of noise gives images a grainy appearance that can mask smaller details, making them difficult to recognize.
- **Poisson Noise** [5]: Also known as photon capture noise, it arises from inherent randomness in the way the image sensor detects photons. This is more noticeable in low-light conditions when changes in the number of photons affect image clarity.
- **Quantization Noise** [5]: This occurs when a continuous signal is quantized into a discrete signal, as in the analog-to-digital conversion process. Rounding the actual values to the nearest available digital values results in this kind of noise.

Understanding the different types of noise is crucial for developing effective noise reduction techniques. Advances in technology and scientific research are constantly improving these methods, increasing the clarity and accuracy of digital media. Developers can significantly improve the quality of images and videos by applying sophisticated noise reduction algorithms.

2.2 Overview of Denoising

The desire to achieve a clean image, free from any noise, goes hand in hand with the development of the image processing field. Fundamentally, denoising is a restoration process,

that aims to return an image or video to a state as close as possible to the original signal, before it is distorted by noise (Figure 2.1). Achieving this goal while preserving the consistency and clarity of visual data requires an understanding of both the nature of noise and signal characteristics.

Historically, denoising approaches were based on hardware improvements and signal processing techniques, that were aimed at filtering noise while preserving the signal. However, with the advent of digital imagery and video, noise reduction evolved into a computational problem, requiring multiple algorithms to solve it. The complexity of applied methods ranges from simple linear filters [2], such as Gaussian blur, which softens noise by losing detail, to more complex nonlinear methods like anisotropic diffusion [23], that preserves edges while smoothing noise.

As the quality of images and videos improved, traditional denoising techniques began to experience difficulties, especially in the context of maintaining consistency between frames in video recordings. The introduction of motion and real-time processing requirements has contributed to the development in this area.

With the development of machine learning and deep learning, the classical algorithm-based approach to noise reduction transitioned to data-driven methods. Using large datasets and powerful computational models, deep learning methods have demonstrated the ability to distinguish noise from signal, learning from examples to predict the appearance of a denoised frame.

Modern noise removal techniques often include concepts such as spatiotemporal coherence [28] and motion compensation [4]. They use redundant information from neighboring frames to determine and restore the quality of the noisy frame. These methods are especially effective in handling random and inconsistent noise, which classical approaches struggled with [8].

2.3 Traditional Denoising Techniques

Before machine learning became standard practice in image and video processing, traditional methods built a solid foundation for noise reduction. These techniques are based on a deep understanding of signal processing and statistical and mathematical theories. They served as the primary tools against image and video degradation due to noise.

- **Spatial Filtering [20]:** Among the earliest strategies, it operates on a frame-by-frame basis, applying filters such as mean and median filters to reduce noise. These filters replace each pixel's value with an average or median of the intensities in neighboring pixels. Effective for certain noise types like salt and pepper noise, however, they often cause image blurring and detail loss.

- **Temporal Filtering [19]:** Utilizes temporal redundancy in videos. By averaging corresponding pixels across successive frames, it reduces inconsistent noise over time. Its effectiveness is limited by motion in the video, potentially resulting in ghosting artifacts and loss of temporal detail without accurate motion tracking.

- **Frequency Domain Methods [20]:** Methods like the Wiener filter transform frames into the frequency domain using the Fourier Transform, attenuating frequencies where noise is dominant. These methods can be powerful but often require knowledge or assumptions about the noise.

- **Wavelet-based Denoising [20]:** Represents a significant advancement by representing image data at multiple scales. Decomposing a video frame into wavelet coefficients allows for denoising based on signal and noise characteristics at different resolution levels, improving edge and detail preservation. However, it faces challenges in tuning for different noise types and artifacts.
- **Anisotropic Diffusion [9]:** Developed to overcome the limitations of linear filters, it aims to reduce noise without removing significant image content, such as edges, by diffusing the image in areas with no significant intensity variation. As an early non-linear method, it shows promise in preserving important features but requires careful parameter tuning and is computationally intensive.

As we can see all of these traditional methods provide valuable insights into the characteristics of noise and lay the foundation for noise reduction, but also each of them has different limitations. These limitations range from an inability to handle complex noise structures to the trade-off between noise reduction and detail preservation. The deterministic nature of these techniques means that they cannot adaptively respond to the variability of noise in different videos

2.4 Metrics for Denoising

Evaluating the performance of denoising algorithms is crucial to determine their effectiveness and applicability in real-world scenarios. The metrics used to assess denoising techniques can be broadly categorized into objective metrics, which provide quantitative assessments, and subjective metrics, which are based on human perception.

2.4.1 Objective Metrics

Objective metrics are indispensable in the comparison and evaluation of denoising algorithms. These metrics offer a quantifiable measure of the quality of denoised images or videos, facilitating a direct comparison between different denoising approaches.

- **Peak Signal-to-Noise Ratio (PSNR) [26]:** One of the most widely used objective metrics for image and video quality assessment. PSNR measures the ratio between the maximum possible power of a signal (in this case, the original image or video) and the power of corrupting noise that affects its representation. Higher PSNR values indicate better denoising performance, as they signify a greater degree of similarity between the denoised and original signals.

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \tag{2.1}$$

  Where:

  - MAX_I is the maximum possible pixel value of the image.
  - MSE is the mean squared error between the original and denoised images.

- **Structural Similarity Index (SSIM) [26]:** SSIM is a more sophisticated metric that considers changes in structural information, luminance, and contrast between the original and denoised images. Unlike PSNR, which measures absolute errors, SSIM

is a perception-based model that reflects the visual impact of three characteristics of the human visual system: luminance, contrast, and structure. An SSIM value of 1 indicates perfect similarity, making it an effective tool for evaluating the perceptual quality of denoised images.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2.2}$$

Where:

- μ_x and μ_y are the averages of x and y respectively.
- σ_x^2 and σ_y^2 are the variances of x and y respectively.
- σ_{xy} is the covariance of x and y.
- $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are two constants to stabilize the division with a weak denominator; L is the dynamic range of the pixel-values (typically this is $2^{\text{bits per pixel}} - 1$), $k_1 = 0.01$ and $k_2 = 0.03$ by default.

- **Mean Squared Error (MSE) [26]:** MSE calculates the average squared difference between pixels of the original and denoised images. It provides a straightforward measure of the energy of the noise removed but does not always correlate well with perceived visual quality.

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (I(i, j) - K(i, j))^2 \tag{2.3}$$

Where:

- m and n are the dimensions of the image.
- $I(i, j)$ is the pixel value of the original image at position (i, j).
- $K(i, j)$ is the pixel value of the denoised image at position (i, j).

2.4.2 Subjective Metrics

Subjective metrics involve human evaluators who rate the quality of denoised images or videos based on their perception. Although subjective evaluation can be time-consuming and less consistent due to personal biases, it remains a crucial aspect of performance assessment because it directly reflects the viewer's experience.

- **Mean Opinion Score (MOS) [16]:** In denoising applications, MOS is obtained by averaging the scores from a panel of viewers rating the quality of denoised images or videos on a predefined scale, such as 1 to 5. MOS provides insight into the perceived quality improvement or degradation resulting from denoising processes.

- **Double Stimulus Categorical Rating [16]:** In this method, both the original and denoised images are displayed to observers, who then rate the quality of the denoised image compared to the original.

- **Forced-Choice Pair-Wise Comparison [16]:** In this approach, two images of the same scene are shown to observers. They are asked to select the image they believe has better quality. This method is useful for directly comparing the effectiveness of various denoising techniques under identical conditions.

Balancing Objective and Subjective Metrics: The ultimate goal of denoising is not only to achieve high scores in objective metrics but also to ensure that the processed images or videos are pleasing to the human eye. Therefore, an effective denoising algorithm should optimize both objective and subjective metrics, maintaining a balance between removing noise and preserving the naturalness and detail of the original content.

2.5 Challenges in Video Denoising

Video denoising introduces a set of challenges that extend beyond those encountered in still image denoising. The addition of the temporal dimension in videos necessitates sophisticated approaches to effectively reduce noise without compromising the quality and dynamics of the video content. Below are key challenges faced in video denoising:

- **Temporal Consistency:** Maintaining temporal consistency across frames is critical, as noise levels can vary significantly from one frame to the next. Achieving a balance between noise reduction and temporal coherence is essential to prevent flickering or ghosting effects that can degrade the viewing experience.
- **Motion Estimation and Compensation:** The presence of motion requires accurate estimation and compensation to ensure that denoising does not blur or distort moving objects. Incorrect motion estimation can lead to artifacts, such as smearing or unnatural patterns, especially noticeable in fast-moving scenes.
- **Real-Time Processing:** For applications requiring real-time video denoising, such as live broadcasting or video conferencing, the computational complexity of denoising algorithms poses a significant challenge. Efficient algorithms are needed to process video frames quickly without sacrificing output quality.
- **High-Resolution and High-Dynamic Range Content:** The advent of high-resolution and high-dynamic range video content complicates the denoising process by increasing the data volume and highlighting the need for algorithms capable of handling nuanced variations in brightness and color without losing detail or introducing artifacts.
- **Adaptive Noise Modeling:** Effective video denoising algorithms must adapt their noise model to different types of noise (e.g., Gaussian, speckle, salt-and-pepper) and varying noise levels, crucial for preserving the authenticity and quality of the denoised video.

Overcoming these challenges involves a blend of advanced signal processing techniques, machine learning models, and hardware optimizations. As the role of video content continues to expand, the development of efficient and effective video denoising methods remains a pivotal area of research in digital imaging and computer vision.

Chapter 3

Review of Existing Neural Networks for Video and Image Denoising

Advances in denoising technology have led to the development of various algorithms tailored to reduce noise in digital images and videos effectively. These methods range from spatial and temporal filtering in the early days of digital imaging to sophisticated machine-learning models that leverage large datasets to learn noise patterns and how to remove them.

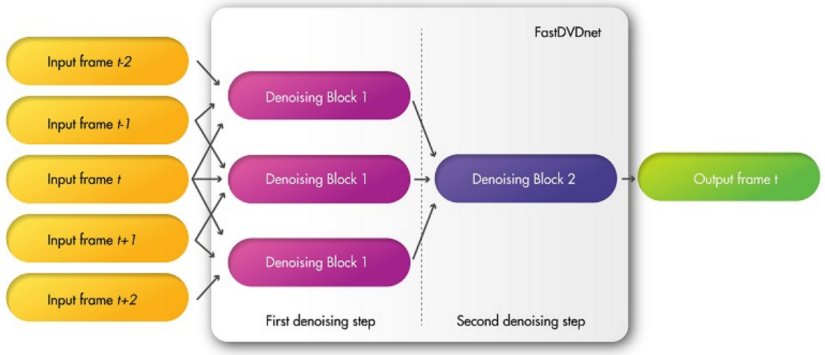
This chapter looks at how video and image denoising techniques have evolved, moving from older methods to the latest ones based on neural networks. We will focus on two of the most advanced methods available today: **FastDVDNet** [25] and **ViDeNN** [7]. These methods are examined for how they work, what makes them strong, and where they could be better, especially when dealing with the complex issue of denoising.

3.1 FastDVDNet

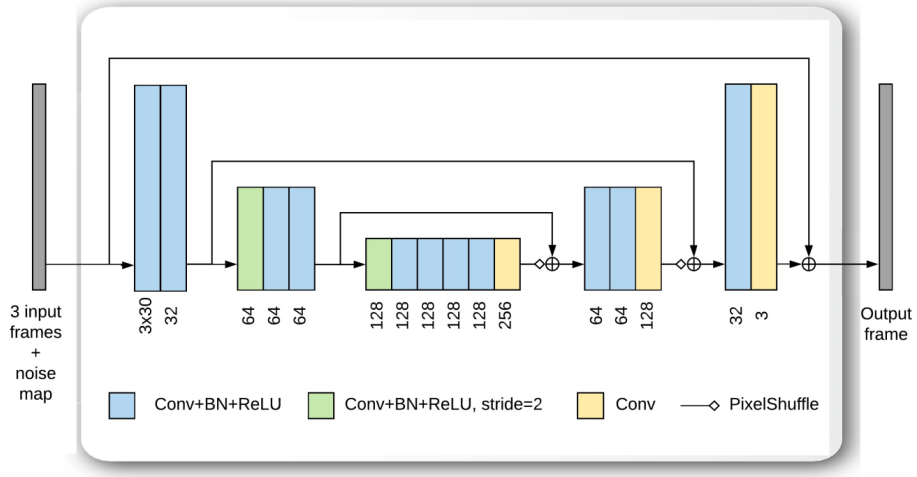
FastDVDNet is a notable advancement in the field of video denoising, developed to offer both rapid processing and exceptional quality. Its design allows it to efficiently address a broad spectrum of noise intensities, facilitating its application in real-time scenarios without sacrificing performance. This capability positions FastDVDNet as a versatile and powerful tool in modern video processing technologies.

3.1.1 Architecture and Approach

FastDVDNet uses a unique approach to video denoising by incorporating traditional motion estimation techniques, such as calculating optical flow. This choice aims to simplify the denoising process while improving its speed and maintaining high output quality. The algorithm uses advanced methods to accurately understand and use motion information in videos, which leads to better noise reduction.



(a)



(b)

Figure 3.1: FastDVDNet model architecture. (a) A high-level diagram of the architecture. (b) The denoising blocks of FastDVDnet Adopted from [25].

Two-step Denoising Process FastDVDNet's effectiveness relies on its two-step process (Figure 3.1), a technique also used in other denoising frameworks like DVDnet [24] and ViDeNN [7]. This method is designed to make the most of temporal information, which relates to how frames interact over time. The value of this dual-phase approach was proven through a modified version of FastDVDNet, which handles five frames at once instead of the usual three. In tests, this standard two-step setup showed better results in keeping video smooth and minimizing flickering, confirming its importance in effective video denoising [25].

Multi-scale Architecture and End-to-end Training Additionally, FastDVDNet's design includes multi-scale denoising blocks, which help the algorithm tackle noise at different scales or detail levels within the video. Experiments that modified FastDVDNet to use single-scale denoising blocks, similar to those in DVDnet [24], showed that multi-scale blocks are more effective at denoising. Originally, parts of FastDVDNet were trained in separate phases. However, switching to an end-to-end training approach greatly reduced temporal artifacts, highlighting the benefits of this more integrated training method [25].

Efficient Motion Handling Among FastDVDNet's significant advancements is its proficient handling of motion without relying on explicit motion estimation—a task that is notably prone to inaccuracies in scenarios characterized by significant noise or complex motion patterns. Through the synergistic application of multi-scale analysis, the cascaded two-step denoising process, and end-to-end training, FastDVDNet navigates video motion with exceptional efficiency. Including residual learning within its denoising blocks further augments the quality of denoising. These methods not only enhance the noise reduction capabilities of FastDVDNet but also make it faster and more efficient. This makes it especially suitable for real-time video denoising tasks [25].

3.1.2 Strengths and Weaknesses

Strengths FastDVDNet showcases several strengths that set it apart in the field of video denoising. Primarily, its ability to maintain temporal coherence across denoised sequences is a significant advantage. This characteristic ensures minimal flickering, especially in flat areas where patch-based methods may leave low-frequency residual noise, an aspect critically important for viewer satisfaction. FastDVDNet's performance is commendable across different datasets, including DAVIS [18] and Set8 [25], indicating its robustness to various content types and noise levels.

Additionally, the algorithm performs well in processing non-repetitive textures, maintaining clarity in elements such as text and vegetation, which can be challenging for patch-based algorithms. This ability indicates an advanced approach to managing texture and detail that is essential for preserving visual quality. Another key advantage of FastDVDNet is its computational efficiency. It operates much faster than many other advanced methods, being up to 80 times quicker than some competitors. This speed makes FastDVDNet especially suitable for real-time applications, which is impressive given the complexity of video-denoising tasks [25].

Weaknesses While FastDVDNet demonstrates outstanding capabilities, certain limitations are noteworthy. The algorithm, like any, is not without its challenges in specific scenarios. For instance, in sequences with a large portion of repetitive structures, patch-based methods might outperform FastDVDNet, as these methods leverage the non-local similarity prior effectively. This limitation indicates a potential area for improvement in how FastDVDNet processes and denoises repetitive patterns and structures within video sequences.

Furthermore, the current implementation and evaluations focus primarily on Gaussian noise. Although this focus encompasses a wide range of practical applications, the adaptability of FastDVDNet to other noise types, such as speckle or Poisson noise, remains less clear. Extending the algorithm to efficiently tackle a broader spectrum of noise types could enhance its applicability and robustness, making it even more versatile in real-world denoising tasks.

In conclusion, FastDVDNet stands as a powerful tool for video denoising, offering a unique balance between quality and efficiency. Its strengths in temporal coherence, detail preservation, and computational speed are complemented by opportunities for further enhancement, particularly in the processing of repetitive structures and the handling of various noise types.

3.2 ViDeNN

ViDeNN, a novel convolutional neural network (CNN) tailored for the task of video denoising without prior knowledge of the noise type, represents a significant leap in addressing the complex challenge of video noise reduction under blind conditions. Its design cleverly merges spatial and temporal denoising into a singular, streamlined process, showcasing its unique ability to adapt to various noise distributions and video dynamics.

3.2.1 Architecture and Approach

The design of ViDeNN is carefully developed (Figure 3.2), consisting of two complementary parts that focus on reducing spatial noise first and then decreasing temporal noise. This two-step process not only improves the quality of each frame but also maintains consistency and coherence over time [7].

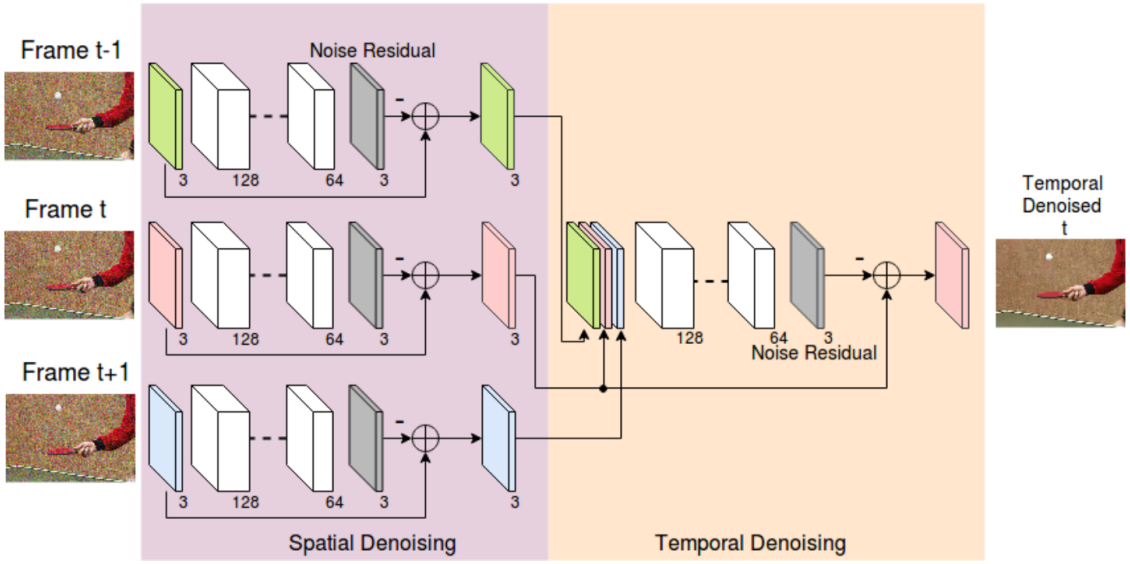


Figure 3.2: ViDeNN model architecture. Adopted from [7].

Spatial Denoising Component Drawing inspiration from the advancements in CNN-based image restoration, ViDeNN employs a deep network for the initial spatial denoising phase. This network layering is optimized to address multiple noise models through a comprehensive approach involving both traditional and novel techniques. Key features include the adoption of residual learning to focus on noise patterns and the strategic use of activation functions to improve learning efficiency [7].

Temporal Denoising Component (Temp3-CNN) After the frames are denoised spatially, the Temp3-CNN component of ViDeNN uses these frames to integrate temporal information by processing them in sequences. This element of ViDeNN's architecture merges the spatially enhanced frames, using their temporal sequence to effectively reduce inconsistencies and motion-related artifacts [7].

3.2.2 Strengths and Weaknesses

Strengths ViDeNN's architecture showcases remarkable flexibility in handling diverse noise conditions and video content, setting a new standard for blind video denoising. Its comprehensive approach to spatial and temporal noise reduction is particularly effective in challenging conditions, such as low-light environments, where it demonstrates exceptional noise reduction capabilities. Moreover, ViDeNN's ability to preserve motion integrity and reduce temporal artifacts offers a clear advantage over traditional denoising methods that often struggle with dynamic content.

Weaknesses Despite its creative approach and significant advantages, ViDeNN has some drawbacks, particularly in terms of computational efficiency. The high demand for resources in its current form might restrict its use in real-time applications or on less powerful devices. Moreover, the division into separate spatial and temporal denoising stages, though effective, indicates that there could be room for more integration and optimization to better utilize the interaction between spatial and temporal elements of video denoising.

In conclusion, ViDeNN stands out as an innovative solution in the field of video denoising, effectively handling the challenges of blind noise reduction under various conditions and types of content. Future improvements and research are likely to boost its performance and broaden its use, offering promising prospects for advancements in video processing technology.

Chapter 4

Dataset for Experiments

This chapter offers an overview of the dataset that will be used to achieve the objectives of this thesis. The first section introduces the **Vimeo-90k**[1] dataset, which forms the foundation of this research. Subsequent sections discuss the pre-trained noise generation model, **C2N**[2], used for synthesizing noise in the video sequences. Finally, the chapter discusses the preprocessing methods to structure the final dataset.

4.1 Vimeo-90k

The Vimeo-90k dataset [29] is a comprehensive video collection specifically designed for deep learning applications in video processing, including tasks such as super-resolution, denoising, and frame interpolation. The dataset compromises a diverse range of video content sourced from the Vimeo[3] platform (Figure 4.1).

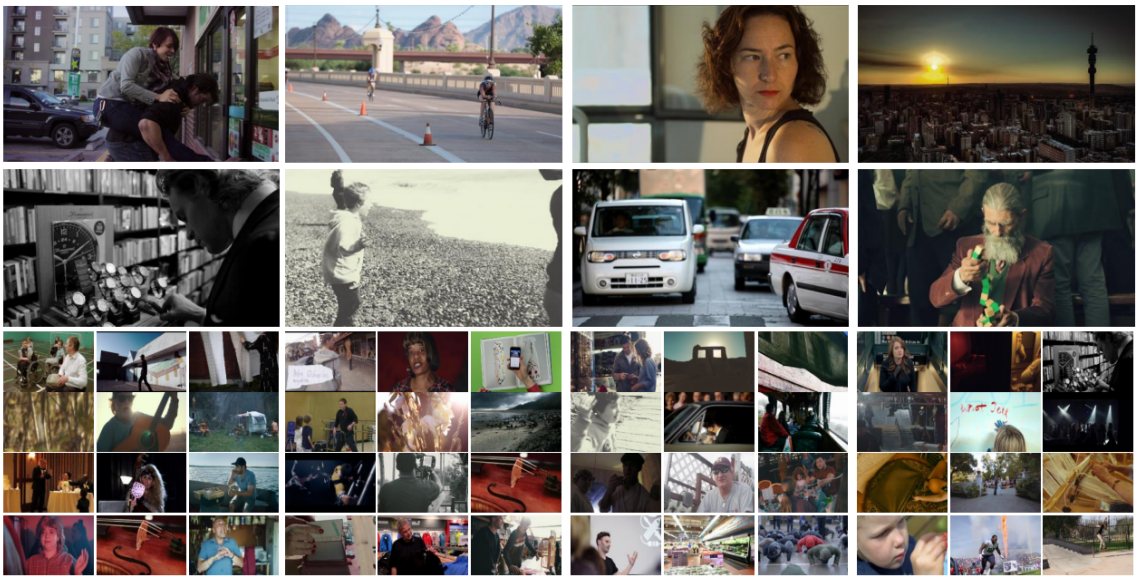


Figure 4.1: Sampled frames from the Vimeo-90K dataset. Adopted from [29].

[1] http://toflow.csail.mit.edu/
[2] https://github.com/onwn/C2N
[3] https://vimeo.com/

Dataset Size and Structure: Vimeo-90k consists of approximately 90,000 video sequences, each containing 7 frames with a resolution of 448x256 pixels. This structure is particularly suited for applications requiring temporal consistency analysis, such as video denoising, where multiple consecutive frames are essential for evaluating the performance across time.

4.2 C2N Noise Generation Model

The C2N (Clean to Noisy) Noise Generation Model [13] is an innovative approach to creating synthetic noise for video sequences. It stands out in its field for not requiring paired noisy and clean images to learn noise distributions, utilizing an unsupervised learning strategy instead.

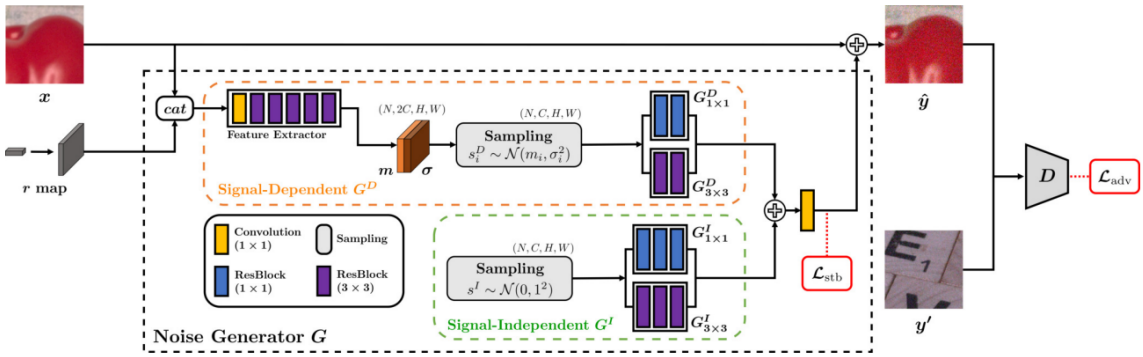


Figure 4.2: Architecture of the C2N framework. Adopted from [13].

Noise Types Generated: C2N is adept at simulating real-world noise, including both signal-dependent noise, which varies with the image content, and signal-independent noise, which is random. This allows it to mimic the complex noise typically introduced by various factors in real-world scenarios, such as sensor noise, lighting conditions, and electronic interference [13].

Functioning Mechanism: The model operates through a convolutional neural network that predicts noise patterns from clean images. This noise is then superimposed onto the clean images to generate their noisy counterparts. The architecture of C2N consists of a noise generator that synthesizes both the signal-dependent and independent noise components and a discriminator that works in tandem to ensure the generated noise is indistinguishable from genuine noise found in actual noisy images [13] (Figure 4.2).

In essence, the C2N model facilitates the generation of realistic noisy videos required for the development and training of video-denoising algorithms, providing a valuable tool for researchers in the field of video processing.

4.3 Dataset Preparation

The preparation of our dataset involved several key steps to ensure its suitability for training and evaluating video-denoising models:

Cleaning Invalid Sequences: Initial preprocessing involved identifying and removing any corrupt or invalid video sequences from the Vimeo-90K dataset. This step was crucial

to ensure the integrity and quality of the training and testing data. Furthermore, the dataset size was trimmed to the 20.000 video sequences.

Generating Noise/Clean Video Pairs: Using the C2N model [13], we generated synthetic noise for each of the clean video sequences in the dataset. This process resulted in pairs of videos, where each clean sequence had a corresponding noisy version that closely resembled real-world noisy footage.

Final Dataset Structure: The final dataset consists of pairs of clean and noisy video sequences, maintaining the original structure of 7 frames per sequence with a resolution of 448x256 (Figure 4.3). The dataset was then partitioned into training, validation, and testing sets.

```
Dataset
|-- 1
|   |-- noisy
|   |     |-- im1.png
|   |     |-- ...
|   |     |-- im7.png
|   |-- clean
|         |-- im1.png
|         |-- ...
|         |-- im7.png
|-- 20000
    |-- noisy
    |-- clean
```

Figure 4.3: Tree structure of the proposed dataset.

The dataset created through this process serves as a crucial resource for exploring and advancing video denoising techniques. By closely mimicking real-world conditions, it provides a challenging and diverse testing ground for evaluating the effectiveness of different denoising approaches.

Chapter 5

Proposed Solution for Video Denoising

The main goal of this thesis is to get acquainted with video denoising and deep learning techniques. This includes a study and experimentation with various deep learning models, strategies, and architectures to understand their effectiveness in reducing noise from video sequences. The product of this work is a video denoising pipeline using deep learning methods.

5.1 Proposed Architectures

This section provides an overview of the iterative process of how the video denoising model was developed.

- **UNet Baseline:** Establishes the foundational architecture, adapting the conventional UNet structure for video frame processing. It introduces an encoder-decoder design tailored for capturing and denoising temporal video data.
- **ResUNet:** Enhances the baseline model by incorporating residual blocks and attention mechanisms. This design aims to improve the handling of complex noise patterns and ensure deeper network training without the risk of vanishing gradients.
- **ResUNet Temporal:** Extends the capabilities of the ResUNet by adding ConvLSTM2D layers to capture temporal dependencies between frames effectively. This model leverages both spatial and temporal data to achieve consistent and superior denoising across video sequences.

5.1.1 UNet Baseline

The basic UNet model is a fundamental component for studying video noise reduction techniques based on deep learning. Using the UNet architecture [21], known for its efficiency in image segmentation tasks, this model is adapted to solve video denoising problems.

The Baseline UNet Model is structured into an encoder-decoder architecture (Figure 5.1), typical of UNets, but tailored for processing video frames. The model operates on input data shaped as $(256, 448, 9)$, representing a sequence of three consecutive RGB video frames ($Frame_{t-1}$, $Frame_t$, and $Frame_{t+1}$). This input structure is pivotal for capturing temporal information crucial for effective denoising.

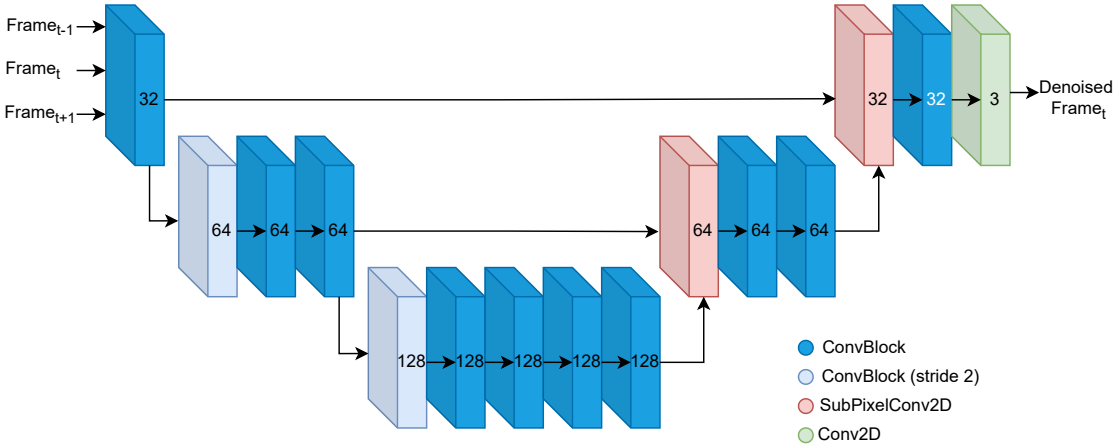


Figure 5.1: UNet Baseline model architecture.

Encoder:

- The encoder begins with a feature extraction convolutional block [17] (Figure 5.2) employing 32 filters of size 3x3, essential for initial feature identification from the input frames.
- The subsequent layers contain two downsampling blocks:
  - The first block increases the filter count to 64 and incorporates strides of 2, effectively reducing the spatial dimensions while enhancing feature detection. This block includes two convolutional blocks, each using the same number of filters.
  - The second downsampling block further increases the filter count to 128 and maintains the stride to continue spatial reduction. It consists of four consecutive convolutional blocks that maintain the same number of filters.

Decoder:

- The decoder reverses the process through two upsampling blocks:
  - The first utilizes sub-pixel convolution [22] (SubPixelConv2D) with 64 filters and a scaling factor of 2, increasing the resolution of feature maps. This stage also includes a concatenation with the output of the corresponding downsampling block, thereby preserving high-frequency details essential for accurate reconstruction.
  - The second block continues with the upsampling, linking back to the outputs from the initial feature extraction phase and reducing the filter count to 32.
- The final output is constructed using a convolutional layer with 3 filters, which synthesizes the denoised video frame by refining and integrating the extracted features.

Functional Overview of Key Components

- **ConvBlock:** Central to the model, each ConvBlock includes a convolutional layer, batch normalization [12], and ReLU activation [1]. This configuration not only aids

in feature enhancement but also stabilizes the learning process through normalization and non-linear activation.

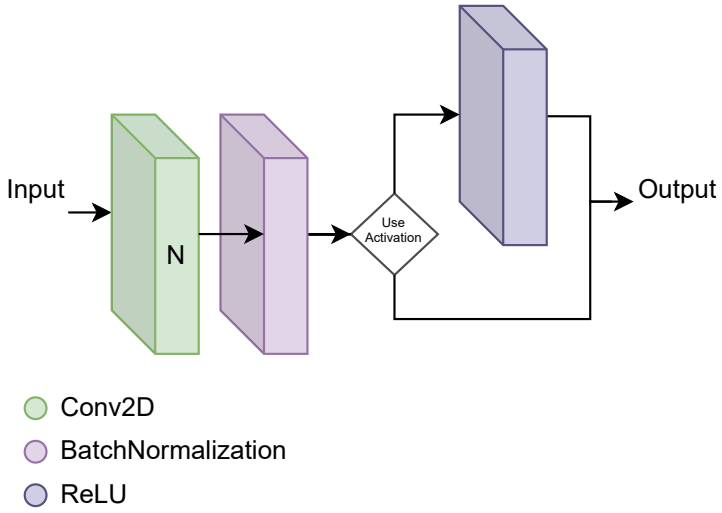


Figure 5.2: ConvBlock architecture.

- **SubPixelConv2D:** This layer plays a critical role in the decoder for effectively upsampling the feature maps. It operates by first expanding the channel depth using a 3×3 convolution. This is followed by a spatial rearrangement through the `tf.nn.depth_to_space`[1] operation, which increases the spatial resolution while reducing the depth. This method enhances resolution efficiently, enabling detailed and precise reconstruction of the denoised frame.

This baseline model, with its straightforward yet robust architecture, lays the foundation for future advancements in our video-denoising research. Its design promotes easy training and adaptability, making it an excellent starting point for iterative improvements and experiments with more sophisticated deep-learning strategies for video denoising.

5.1.2 ResUNet

The ResUNet architecture extends the capabilities of the UNet Baseline by incorporating residual blocks [10] and integrating attention mechanisms [6], aiming to improve the effectiveness of video denoising. These modifications are designed to deepen the network while ensuring effective training and information flow, which is critical for handling detailed noise characteristics in videos (Figure 5.3).

[1] https://www.tensorflow.org/api_docs/python/tf/nn/depth_to_space

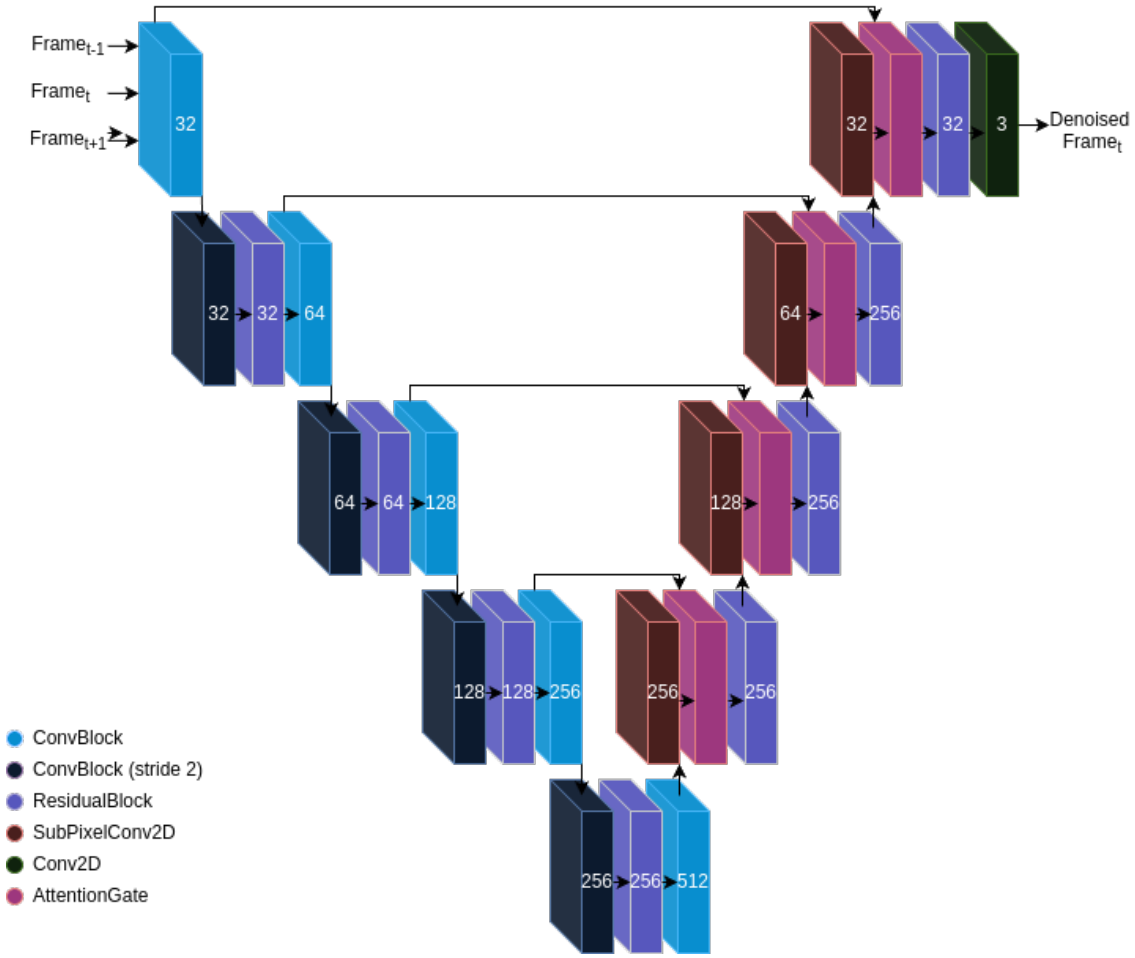


Figure 5.3: ResUNet architecture.

Encoder:

- The encoder features a series of downsampling blocks that increase the depth of feature maps while reducing spatial dimensions:
  - Each block consists of a convolutional operation for feature extraction followed by another convolution with an increased stride for dimensionality reduction.
  - A residual block is then applied, which includes a standard convolutional block followed by another convolutional operation without activation. A shortcut connection adds the input of the block to its output, followed by a ReLU activation. This structure helps in maintaining the flow of gradients and reduces the risk of information loss across deeper layers (Figure 5.4).

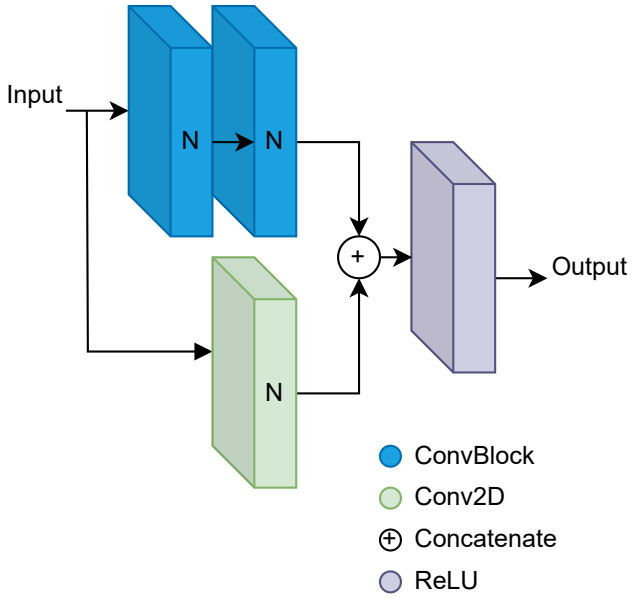


Figure 5.4: Residual block architecture.

Bridge:

- The bridge between the encoder and decoder features a convolution block with a high number of filters (512), serving as a critical transition point that synthesizes the compressed features before they are upsampled in the decoder.

Decoder:

- Compared to the baseline UNet, the ResUNet decoder enhances the upsampling process with the integration of SubPixelConv2D, attention gates, and residual blocks:
  - **SubPixelConv2D** is utilized for efficient spatial expansion of feature maps, as detailed in the baseline section.
  - **Attention Gates** are a key enhancement, selectively refining the upsampled feature maps to focus the model's recovery capabilities on areas with significant noise or important details, thereby enhancing the quality and precision of the denoising process (Figure 5.5).

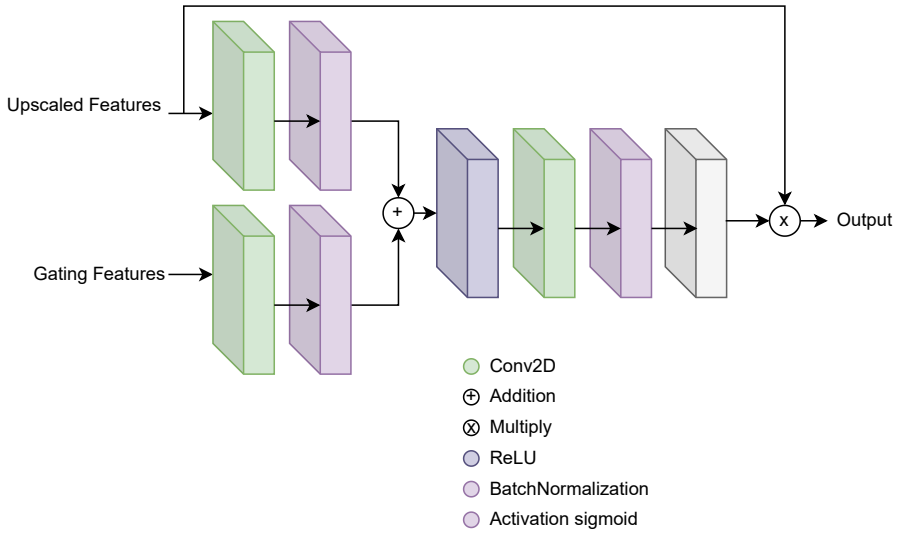


Figure 5.5: Attention Gate architecture.

– **Residual Blocks** in the decoder further stabilize the feature transformation during the upsampling, ensuring that even deep layers retain essential information without degradation, critical for maintaining the fidelity of the reconstructed frames.

The ResUNet model, with its robust and advanced architecture, significantly advances the baseline capabilities by tackling more complex noise patterns and maintaining high-quality video output. Its comprehensive design ensures effective training and adaptability, making it highly effective for video denoising tasks.

5.1.3 ResUNet Temporal

ResUNet Temporal extends the capabilities of the ResUNet by integrating temporal dynamics directly into the video denoising process. This model variation uses ConvLSTM2D layers [11] to capture temporal dependencies between frames effectively, enhancing the ability to address variations in noise patterns over time (Figure 5.6).

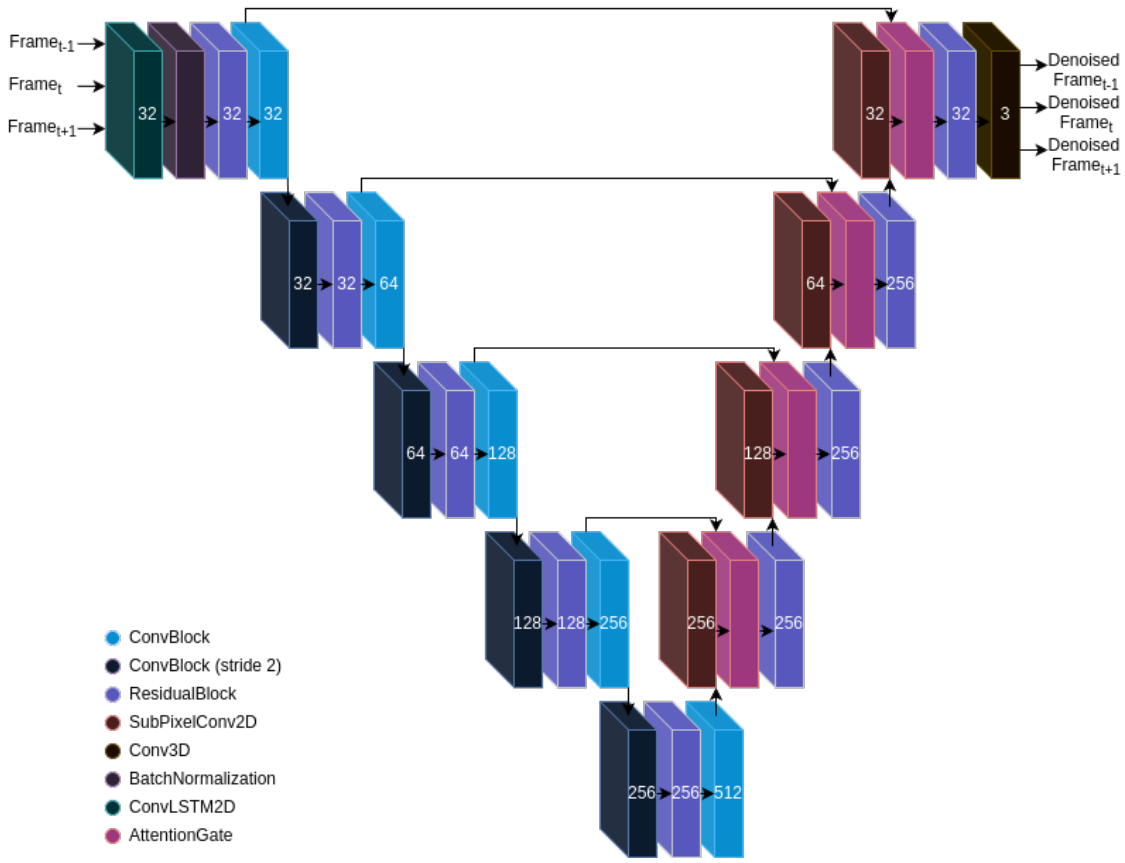


Figure 5.6: ResUNet Temporal architecture.

Temporal Feature Processing:

- The input sequence undergoes initial processing by a ConvLSTM2D layer, which is adept at capturing temporal correlations within video data. This layer recurrently applies convolutional operations, maintaining state across frames to adaptively respond to changes in video noise.
- This temporal information is then enhanced through the integration of a residual block, which refines the features while maintaining temporal continuity.

Encoder:

- Following the temporal processing layer, the encoder features successive downsampling blocks, similar to those in the ResUNet. Each block deepens feature extraction while compressing spatial dimensions, incorporating residual blocks to preserve essential information and ensure robust feature propagation.

Decoder:

- The decoder employs SubPixelConv2D for efficient spatial upsampling, combined with attention gates that selectively refine features based on their relevance to the denoising task. This selective attention is critical in focusing the model's capabilities on regions with significant noise or detail.

- Skip connections, enhanced with attention gates, reintroduce and integrate detailed features from earlier in the network, ensuring the preservation of important textural information.

Output:

- Uniquely, the ResUNet Temporal model outputs a sequence of denoised frames using a Conv3D layer [27]. This approach allows the model to maintain temporal consistency in the output, reducing flickering and improving the visual quality of the denoised video.

The ResUNet Temporal model represents a significant advancement in video-denoising technology by addressing both spatial and temporal aspects of noise. This sophisticated architecture not only reduces noise effectively but also maintains the natural appearance and dynamic consistency of the video, making it particularly suitable for high-quality video applications.

5.2 Denoising Pipeline

This section provides an overview of the final denoising pipeline and its components.

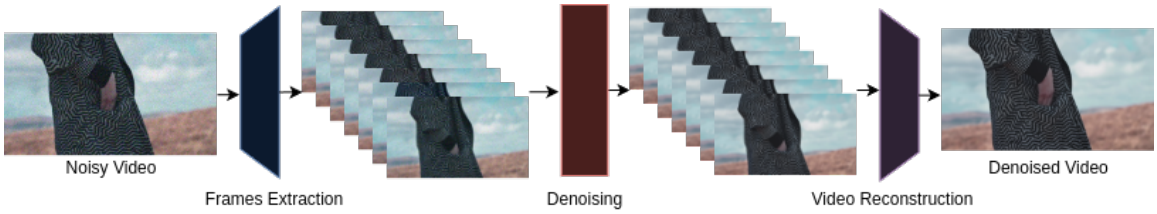


Figure 5.7: Video Denosing pipeline.

Figure 5.7 illustrates a schematic view of the proposed video denoising pipeline. The pipeline begins with a noisy video, which is the input requiring denoising. Initially, the preprocessing block extracts individual frames from the video. These frames are then formatted to fit the input requirements of the ResUNet Temporal model and are processed through a sliding window technique. This technique processes sets of three frames at a time and saves the middle frame as the denoised result. The final step reconstructs the video from the denoised frames providing the final denoised video as output.

Chapter 6

Experiments and Results

This chapter provides a detailed overview of the conducted experiments and their corresponding results. Initially, the chapter outlines the training methodologies and outcomes of the proposed deep learning models as discussed in Chapter 5. Subsequent sections will compare these models both among themselves and against the established ViDeNN model [7]. Finally, the chapter will provide an overview of the resulting complete video denoising pipeline.

6.1 Models Training and Validation

Training neural networks is a complex and demanding process. To achieve a high degree of accuracy, it is essential to use large and robust datasets. Additionally, the architecture of the models and the selection of hyperparameters play a crucial role in the training process.

6.1.1 Baseline UNet model

The input sample X for the model is 3 consecutive noisy frames $Frame_{t-1}$, $Frame_t$, $Frame_{t+1}$ that are concatenated along the channel dimension resulting in the input having shape $(256, 448, 9)$. Concatenating the frames in this manner allows the model to access information from the immediate past and future states of the current frame, enhancing its ability to infer the noise characteristics and underlying clean signal.

The target for each input sample X is the clean version of the current frame $Frame_t$. This frame is chosen because it directly corresponds to the middle frame of the three noisy frames used in X. The goal of the model is to learn to predict this clean frame from the noisy input sequence provided.

The loss function for this task was the Structural Similarity Index (SSIM) 2.4, which computes the difference between the predicted denoised frame and target. The Adam optimizer [14] was chosen for its efficiency and effectiveness, with a learning rate set to 0.0001. The batch size was set to 16. The number of epochs was set to 100.

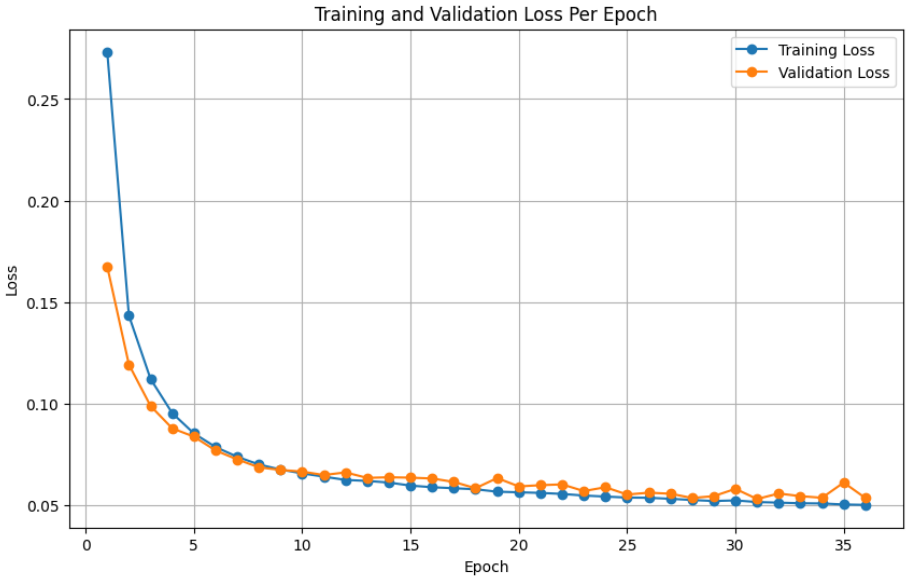


Figure 6.1: Training and validation losses for the UNet Baseline model over epochs.

As depicted in Figure 6.1, both the training and validation loss consistently decreased over the iterations. The training process ended earlier due to the implementation of the Early Stop regularisation [3]. Early Stop is a regularization element, which stops the training if the validation loss stops decreasing or begins to increase, thereby preventing overfitting. The patience parameter was set to 5 epochs for this mechanism.

Figure 6.2: Example of frame denoising using the UNet Baseline model.

Figure 6.2 demonstrates that although the baseline UNet model is capable of effectively reducing noise in individual frames, it tends to blur the images slightly. This blurring effect is a common challenge in denoising, where the process of reducing noise can sometimes obscure finer details in the image.

Another significant issue observed with the baseline model is its inability to maintain consistency between consecutive frames when applied to entire video sequences. This results in noticeable flickering or blinking effects, which can degrade the overall viewing experience by introducing visual disturbances that are particularly evident in dynamic scenes. The examples shown in Figure 6.3 highlight how these inconsistencies appear throughout a video sequence.

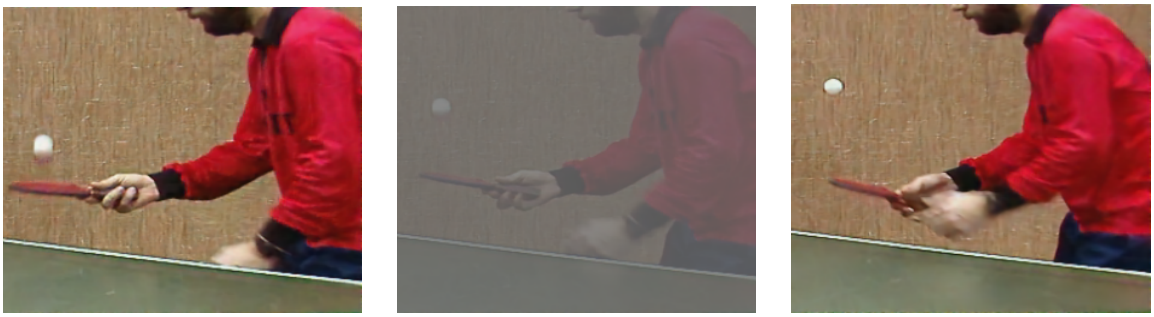


Figure 6.3: Example of video sequence denoising using the UNet Baseline model. Original clean video sequence was adopted from [7].

Despite these limitations, the baseline UNet model provides a solid foundation for further exploration in the field of video denoising. It demonstrates the applicability of deep learning techniques to complex video processing tasks but also underscores the need for improvements in maintaining image detail and enhancing frame-to-frame consistency.

6.1.2 ResUNet model

The input configuration for the ResUNet model mirrors that of the baseline model 6.1.1, utilizing three consecutive noisy frames concatenated along the channel dimension. The target remains the clean version of the current frame ($Frame_t$), which directly corresponds to the middle frame of the input sequence.

This model introduces an improved loss function that combines the Structural Similarity Index (SSIM) and Mean Absolute Error (MAE) to effectively address both perceptual quality and pixel-wise errors. The combined loss function is defined as:

$$\text{Combined Loss} = \alpha \cdot SSIM\_LOSS + (1 - \alpha) \cdot MAE\_LOSS \tag{6.1}$$

Where:

- α is set to 0.84.
- $SSIM\_LOSS$ is the loss based on the Structural Similarity Index Mesure (SSIM) references in Equation 2.2.
- MAE_LOSS is based on the Mean Absolute Error (MAE), similar to the Mean Squared Error (MSE) referenced in Equation 2.3. Unlike MSE, which squares the differences, MAE uses the absolute values, providing a measure less sensitive to outliers.

This configuration aims to balance the trade-offs between maintaining structural integrity and minimizing average error, providing a more robust approach to video denoising.

The Adam optimizer was retained for its proven efficiency and effectiveness. The learning rate was set to 0.0001. Due to the deeper architecture of the ResUNet model the batch size was adjusted to 8. The number of epochs was set to 100.

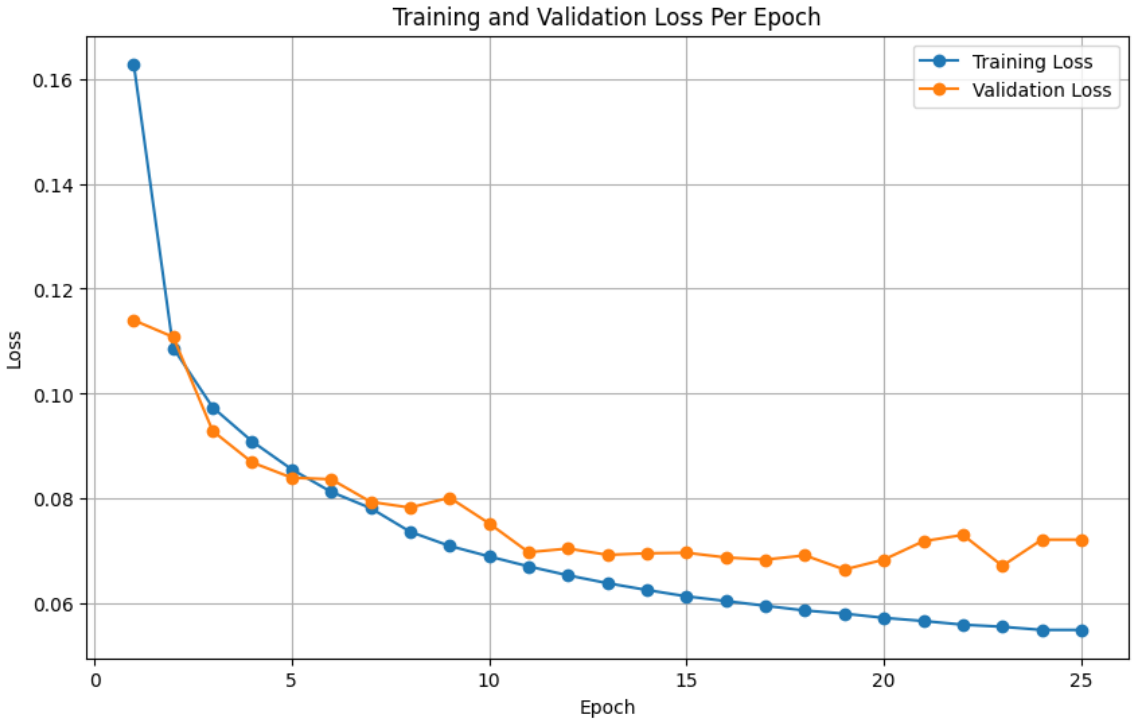


Figure 6.4: Training and validation losses for the ResUNet model over epochs.

Figure 6.4 demonstrates that both training and validation loss decreased over the iterations, but the validation loss had struggles with not decreasing smoothly. The training process again ended earlier due to the Early Stop regularization. The patience parameter was set to 5 epochs for this mechanism.

Figure 6.5: Example of frame denoising using the ResUNet model.

Figure 6.5 provides an example of a frame denoised using the Enhanced ResUNet model. Comparing this with the baseline UNet model shown in Figure 6.2, it is evident that while both models are effective at reducing noise, some blurring remains in the denoised frames.

A significant improvement of the Enhanced ResUNet model over the baseline is its performance in maintaining consistency between consecutive frames. The denoised video exhibits far fewer inconsistencies, and the visual quality does not degrade over time as noticeably. However, some artifacts related to frame consistency, such as blinking, are still present. Figure 6.6 displays denoised frames where such inconsistencies are minimally visible in static frames but become more noticeable during playback.

Figure 6.6: Example of video sequence denoising using the ResUNet model. Original clean video sequence was adopted from [7].

Despite these enhancements, the Enhanced ResUNet model, while superior to the baseline in several aspects, still exhibits room for improvement, particularly in eliminating subtle artifacts such as blinking. It has demonstrated significant progress in addressing the challenges of video denoising, particularly in maintaining frame consistency and reducing temporal artifacts. This model provides a robust foundation for further research, especially in optimizing frame-to-frame transitions and enhancing overall video quality.

6.1.3 ResUNet Temporal

ResUNet Temporal adopts a distinct approach for handling inputs and outputs. Unlike previous models where consecutive noisy frames were concatenated along channel dimensions, this model processes the frames as a list of three separate frames. Furthermore, the model predicts three clean frames as the target output instead of a single clean middle frame. These modifications enhance the model's ability to leverage temporal information between frames more effectively.

Moreover, with the capacity to handle multiple input and target frames, we can further refine the loss function. The model employs a combination of three loss functions: Structural Similarity Index (SSIM), Mean Absolute Error (MAE), and a custom Temporal loss. The temporal loss is computed as follows:

$$\text{TEMP\_LOSS}(Y_\text{t}, Y_\text{p}) = \frac{1}{N-1} \sum_{i=1}^{N-1} \|(Y_{\text{t},i+1} - Y_{\text{t},i}) - (Y_{\text{p},i+1} - Y_{\text{p},i})\| \tag{6.2}$$

Where:

- N is the number of frames.
- $\|\cdot\|$ denotes the L1 norm (absolute difference).

The combined loss function is then defined as:

$$\text{Combined Loss} = (1 - \alpha - \beta) \cdot \text{MAE\_LOSS} + \alpha \cdot \text{SSIM\_LOSS} + \beta \cdot \text{TEMP\_LOSS} \tag{6.3}$$

Where:

- α is set to 0.8.

- β is set to 0.1.
- $SSIM\_LOSS$ is the loss based on the Structural Similarity Index Mesure (SSIM) references in Equation 2.2.
- MAE_LOSS is based on the Mean Absolute Error (MAE), similar to the Mean Squared Error (MSE) referenced in Equation 2.3. Unlike MSE, which squares the differences, MAE uses the absolute values, providing a measure less sensitive to outliers.

The optimizer remains unchanged, we continue to use Adam due to its proven effectiveness in prior training. The learning rate is set to 0.0001. The batch size was reduced to 2 to accommodate the deeper architecture of the model, and the number of epochs was set to 100.

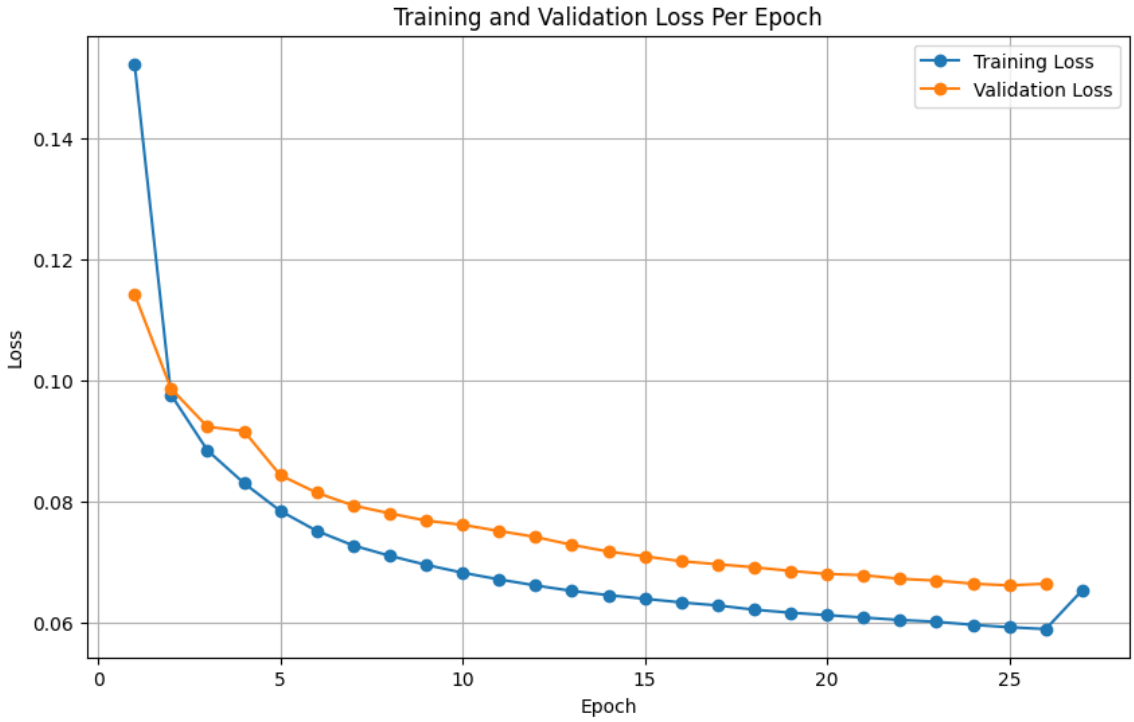


Figure 6.7: Training and validation losses for the ResUNet Temporal model over epochs.

Figure 6.7 shows that the training did not run for the 100 epochs and stopped earlier at 27. epoch due to the implementation of the Early Stop regularization. The patience parameter for this mechanism was set again to 5 epochs. With this model, both training and validation losses decreased much smoother than with the ResUNet model.

Figure 6.8: Example of frame denoising using the ResUNet Temporal model.

Figure 6.8 illustrates an example of a frame denoised using the ResUNet Temporal. Compared to the two previous models, UNet baseline and ResUNet, there is a noticeable improvement in image sharpness. However, a minor blurring effect persists on some fine details.

The ResUNet Temporal's major advancement is evident when observing the denoised video sequence, particularly in maintaining consistency between frames. The processed video displays minimal inconsistencies and artifacts such as blinking, and the visual quality remains stable throughout the sequence. Nonetheless, specific artifacts are observed when the video undergoes rapid transitions, such as switching between multiple recording sources or changing viewing angles. These artifacts manifest as residual content from a previous frame briefly lingering in the subsequent frame, slightly muddying the visual clarity.

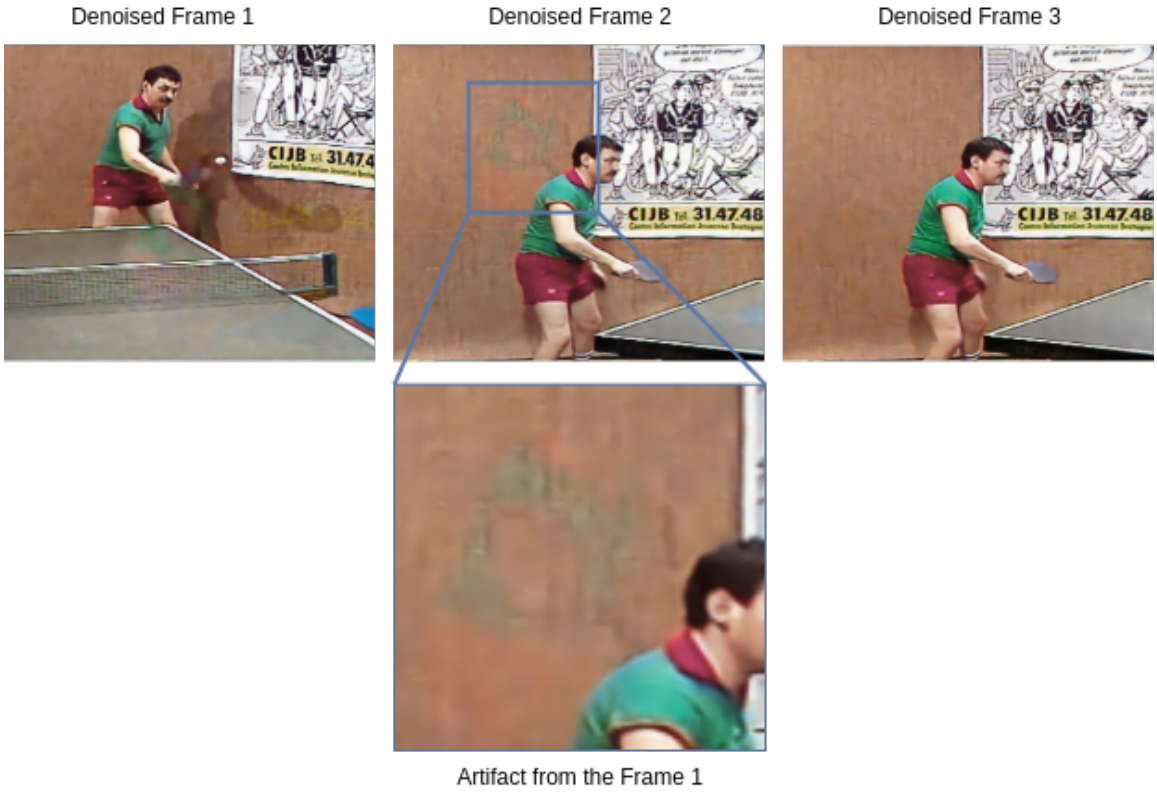


Figure 6.9: Example of video sequence denoising using the ResUNet Temporal model. Original clean video sequence was adopted from [7].

As depicted in Figure 6.9, there is a camera switch between *Frame* 1 and *Frame* 2. The zoomed fragment from *Frame* 2 reveals a residual green area, located in the same position as in *Frame* 1, thus creating an artifact. This artifact is barely noticeable during video playback since it persists for only one frame. By *Frame* 3, the issue is resolved, and the artifact is no longer visible.

In general, the ResUNet Temporal model demonstrated significant improvements in maintaining consistency between frames, effectively preserving video dynamics, and reducing artifacts that degrade video quality. Additionally, the model shows enhanced performance in denoising individual frames, achieving slightly better preservation of image quality and reducing blurriness.

6.2 Results Comparations

This section sums up the results of each model that was obtained during the training and validation process. Furthermore, the results will be compared with a pre-trained version of the ViDeNN model that was described in Section 3.2.

6.2.1 Performance Metrics of Developed Models

Since all models were trained using the same dataset, a consistent baseline is established, allowing for an effective comparison of their performance on the test set. The validation metrics employed are the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). Additionally, the performance of the developed models was benchmarked against a pre-trained version of the ViDeNN model, a recognized state-of-the-art approach in video denoising. This comparison aims to highlight the enhancements or limitations of the newly developed models relative to existing advanced technologies.

Model	**PSNR (dB)**	**SSIM**
UNet Baseline	33.4340	**0.9440**
ResUNet	31.5674	0.9237
ResUNet Temporal	**34.6109**	0.9344
ViDeNN	31.9800	0.8381

Table 6.1: Comparison of PSNR and SSIM metrics for each model.

According to Table 6.1, the UNet Baseline model scores better in PSNR and SSIM metrics compared to the ResUNet and ViDeNN models and achieves a higher SSIM value than the ResUNet Temporal model. Despite these metrics, visual inspections have revealed that the UNet Baseline model struggles with maintaining temporal consistency across different video sequences.

The ResUNet model slightly outperforms the ViDeNN model but scores significantly lower than the UNet Baseline and ResUNet Temporal models. The visual investigation revealed that the model handles temporal consistency much better than the UNet Baseline model. However, it also displays certain anomalies and struggles to effectively denoise individual frames.

The ResUNet Temporal model achieves the best results in the PSNR metric and a slightly lower SSIM score compared to the UNet Baseline. This illustrates a typical trade-off encountered in neural network training. Overall, this model provides the most consistent visual quality, preserving both the integrity of individual frames and temporal consistency across sequences.

The ViDeNN model shows nearly the lowest performance metrics, which may be attributed to its training primarily on Gaussian noise, while the dataset used for the proposed models contains a more complex type of noise. Additionally, various video sequences denoised with ViDeNN contain artifacts, underscoring the challenges it faces with complex noise types (Figure 6.10).

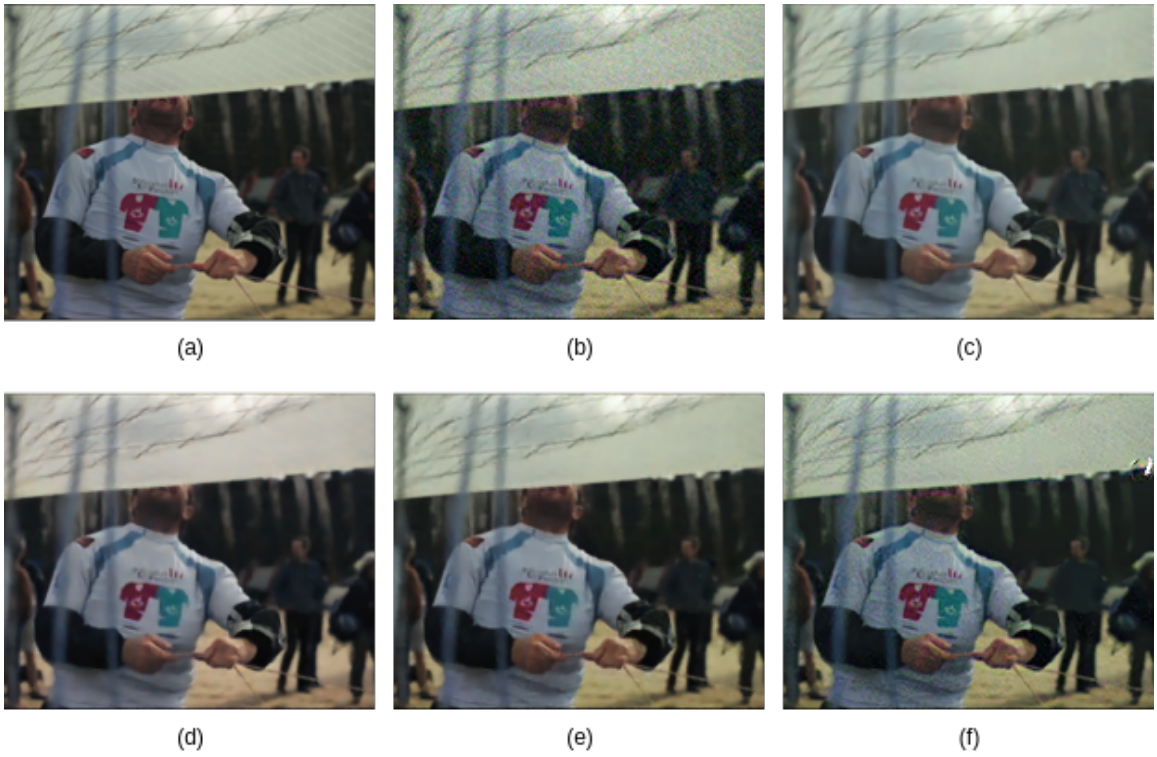


Figure 6.10: Comparison of the denoising video sequence from the test set. (a) Clean image. (b) Noisy image. (c) UNet Baseline. (d) ResUNet. (e) ResUNet Temporal. (f) ViDeNN.

This chapter explores the training, validation, and performance of the proposed models for video denoising. The experiments compared these models against each other and against the ViDeNN pre-trained model, which is known for its effectiveness in video denoising.

The UNet Baseline model showed good results in terms of PSNR and SSIM as a starting point of the research. However, during the evaluation process, the model showed that it struggles with maintaining consistency in video sequence, indicating a need for further improvement in handling temporal information. Despite effectively reducing noise in individual frames, it often resulted in slight blurring and could not prevent video flickering.

The UNet Baseline model showed good results in terms of PSNR and SSIM. However, it struggled with maintaining consistency in video sequences, indicating a need for further improvement in handling temporal information.

The ResUNet and ResUNet Temporal models introduced improvements, particularly in managing video dynamics. ResUNet provided better temporal consistency than the baseline, even though it still had issues with proper noise reduction in specific complex scenes. ResUNet Temporal further enhanced consistency between frames and reduced the amount of artifacts, showing potential as a strong candidate for practical applications.

The final comparison of the proposed models against the pre-trained ViDeNN model highlighted the significant role that the dataset plays in model performance. Since the ViDeNN model was primarily trained on a dataset with Gaussian noise, it performed worse than the proposed models on the test set. Notably, the ResUNet Temporal model outperformed in terms of both PSNR and SSIM metrics, producing better results without any noticeable artifacts or anomalies.

Overall, the models developed in this study demonstrate potential for real-world applications, yet there remains a space for improvement. Future efforts could focus on enhancing the dataset by integrating different types of noise into the video sequences. This would likely improve the model's generalization capabilities, potentially stabilizing its performance. Another suggestion for exploration is experimenting with different training setups, such as utilizing enhanced optimizers and custom training loops, which would allow for more effective monitoring of model performance. Additionally, integrating the model into a user-friendly application could provide valuable feedback through subjective metrics of model performance.

Chapter 7

Conclusion

This thesis explores the challenge of noise in digital video, which significantly impacts video quality across various applications, from entertainment to video processing and analysis. Traditional noise reduction techniques often struggle to effectively balance noise reduction with maintaining details and consistency, prompting a shift towards more innovative deep learning approaches.

The study began with an investigation of the fundamental concepts of noise in videos and the techniques used for video denoising. Once the core challenges were understood, the research delved into the neural networks themselves, which form a main component of this study.

As a starting point for tackling the video denoising challenges using neural networks, two start-of-the-art models, FastDVDNet and ViDeNN, were selected and analyzed. Key concepts of the UNet architecture were adopted based on the architectures of these models.

An essential first step in the experimental phase involved creating a suitable dataset for the video denoising task. The Vimeo-90k dataset was selected as the baseline and was enhanced by integrating the C2N noise generation model. This process produced a unique dataset that closely mimics real-world synthetic noise.

The initial experiment involved training a baseline model to establish a starting point for video denoising and to familiarize with the neural network training process. The UNet Baseline model was trained on the proposed dataset. Validation of the neural network proposed PSNR and SSIM scores of 33.4340 and 0.9440, respectively, which are excellent for a baseline model. However, visual inspection revealed that the model struggles with maintaining consistency between frames. Additionally, it produced visual artifacts such as blinking in some video sequences, leading to video degradation towards the end of the sequences.

The second experiment focused on enhancing the UNet Baseline model, leading to the creation of the ResUNet model. This iteration utilizes residual blocks, allowing for a deeper model architecture. Additionally, modifications were made to the loss function, combining SSIM and MAE for improved effectiveness. Validation of the model revealed lower performance compared to the UNet Baseline, with PSNR and SSIM scores of 31.5674 and 0.9344, respectively. However, visual inspection showed significant improvements in maintaining consistency between frames.

The last experiment built upon the outcomes of the two previous models to develop the ResUNet Temporal model. This model emphasizes temporal consistency in video sequences by altering the way it processes input frames and what it outputs. Unlike previous models that output a single denoised frame, the ResUNet Temporal model outputs three denoised

frames. This change allowed for the integration of a temporal loss function, enhancing the model's ability to maintain consistency between frames. Validation results showed that the model successfully denoised the video sequences, achieving PSNR and SSIM scores of 34.6109 and 0.9344, respectively. Furthermore, visual inspection confirmed that this model produced the best-denoised videos among all tested models, maintaining smooth motion without any highly noticeable artifacts.

The final comparison against the pre-trained ViDeNN model showed that the proposed ResUNet Temporal model outperformed this state-of-the-art model on the proposed dataset. It achieved better validation metrics scores and produced higher-quality videos, demonstrating the success of the experiments.

In conclusion, the application of deep learning techniques in the video-denoising sphere has made a significant impact, providing tools that can outperform traditional methods. The models proposed in this study showed excellent results on the selected dataset, underscoring the potential of deep learning to enhance video quality effectively.

Bibliography

[1] Agarap, A. F. Deep Learning using Rectified Linear Units (ReLU). *CoRR*, 2018, abs/1803.08375. Available at: http://arxiv.org/abs/1803.08375.

[2] Ahamed, B.; Yuvaraj, D. and Priya, S. S. Image Denoising With Linear and Non-linear Filters. In: *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. 2019, p. 806–810.

[3] Bai, Y.; Yang, E.; Han, B.; Yang, Y.; Li, J. et al. Understanding and Improving Early Stopping for Learning with Noisy Labels. *CoRR*, 2021, abs/2106.15853. Available at: https://arxiv.org/abs/2106.15853.

[4] Bao, W.; Lai, W.; Zhang, X.; Gao, Z. and Yang, M. MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement. *CoRR*, 2018, abs/1810.08768. Available at: http://arxiv.org/abs/1810.08768.

[5] Boyat, A. K. and Joshi, B. K. *A Review Paper: Noise Models in Digital Image Processing*. 2015.

[6] Brauwers, G. and Frasincar, F. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*. Institute of Electrical and Electronics Engineers (IEEE), april 2023, vol. 35, no. 4, p. 3279–3298. ISSN 2326-3865. Available at: http://dx.doi.org/10.1109/TKDE.2021.3126456.

[7] Claus, M. and Gemert, J. van. ViDeNN: Deep Blind Video Denoising. *CoRR*, 2019, abs/1904.10898. Available at: http://arxiv.org/abs/1904.10898.

[8] Fan, L.; Zhang, F.; Fan, H. et al. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art*. Springer, 2019, vol. 2, no. 7. Available at: https://doi.org/10.1186/s42492-019-0016-7.

[9] Hadj Fredj, A. and Malek, J. GPU-based anisotropic diffusion algorithm for video image denoising. *Microprocessors and Microsystems*, 2017, vol. 53, p. 190–201. ISSN 0141-9331. Available at: https://www.sciencedirect.com/science/article/pii/S0141933117300807.

[10] He, K.; Zhang, X.; Ren, S. and Sun, J. Deep Residual Learning for Image Recognition. *CoRR*, 2015, abs/1512.03385. Available at: http://arxiv.org/abs/1512.03385.

[11] Hu, W.; Li, H.; Pan, L.; Li, W.; Tao, R. et al. Feature Extraction and Classification Based on Spatial-Spectral ConvLSTM Neural Network for Hyperspectral Images. *CoRR*, 2019, abs/1905.03577. Available at: http://arxiv.org/abs/1905.03577.

[12] Ioffe, S. and Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR*, 2015, abs/1502.03167. Available at: http://arxiv.org/abs/1502.03167.

[13] Jang, G.; Lee, W.; Son, S. and Lee, K. M. C2N: Practical Generative Noise Modeling for Real-World Denoising. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. October 2021, p. 2350–2359.

[14] Kingma, D. P. and Ba, J. *Adam: A Method for Stochastic Optimization*. 2017.

[15] Lim, S. Characterization of noise in digital photographs for image processing. *Proc. SPIE Digital Photography II, vol. 6069,*, february 2006, vol. 6069.

[16] Mohammadi, P.; Ebrahimi Moghadam, A. and Shirani, S. *Subjective and Objective Quality Assessment of Image: A Survey*. 2014.

[17] O'Shea, K. and Nash, R. An Introduction to Convolutional Neural Networks. *CoRR*, 2015, abs/1511.08458. Available at: http://arxiv.org/abs/1511.08458.

[18] Perazzi, F.; Pont Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M. et al. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, p. 724–732.

[19] Pizurica, A.; Zlokolica, V. and Philips, W. Noise Reduction in Video Sequences Using Wavelet-Domain and Temporal Filtering. *Proceedings of SPIE - The International Society for Optical Engineering*, february 2004, vol. 5266.

[20] Rajni, R. and Anutam, A. Image Denoising Techniques - An Overview. *International Journal of Computer Applications*, december 2013, vol. 86.

[21] Ronneberger, O.; Fischer, P. and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR*, 2015, abs/1505.04597. Available at: http://arxiv.org/abs/1505.04597.

[22] Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P. et al. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *CoRR*, 2016, abs/1609.05158. Available at: http://arxiv.org/abs/1609.05158.

[23] Tang, J.; Sun, Q.; Liu, J. and Cao, Y. An Adaptive Anisotropic Diffusion Filter for Noise Reduction in MR Images. In: *2007 International Conference on Mechatronics and Automation*. 2007, p. 1299–1304.

[24] Tassano, M.; Delon, J. and Veit, T. DVDNET: A Fast Network for Deep Video Denoising. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, September 2019. Available at: http://dx.doi.org/10.1109/ICIP.2019.8803136.

[25] Tassano, M.; Delon, J. and Veit, T. FastDVDnet: Towards Real-Time Video Denoising Without Explicit Motion Estimation. *CoRR*, 2019, abs/1907.01361. Available at: http://arxiv.org/abs/1907.01361.

[26] Thiruvenkadam, K.; Ravindran, V. and Padmanaban, S. A Study on Validation Metrics of Digital Image Processing. In:. January 2017.

[27] Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L. and Paluri, M. C3D: Generic Features for Video Analysis. *CoRR*, 2014, abs/1412.0767. Available at: http://arxiv.org/abs/1412.0767.

[28] Wang, C.; Zhou, S. K. and Cheng, Z. First image then video: A two-stage network for spatiotemporal video denoising. *CoRR*, 2020, abs/2001.00346. Available at: http://arxiv.org/abs/2001.00346.

[29] Xue, T.; Chen, B.; Wu, J.; Wei, D. and Freeman, W. T. Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision (IJCV)*. Springer, 2019, vol. 127, no. 8, p. 1106–1125.

Appendix A

Contents of the included storage media

- **xnaume01.pdf** Thesis report file.
- **video.mp4** Demonstration video.
- **src/** Folder with source files.
- **examples/** Folder with image and video examples of the models' results.
- **dataset/** Folder with example samples of used dataset.
- **pretrained_models/** Folder with pre-trained models.
- **requirements.txt** List of Python libraries dependencies.
- **README.md** README manual for the project.
- **latex/** Folder with LaTeXsource files.