VANDERBILT
UNIVERSITY

Master of Science in Computer Science

Non-Thesis Portfolio

Haoli Yin

Commodore ID: 000768997

VUNetID: yinh4

Masters Start: Fall 2023

Expected Graduation: Spring 2025

Academic Advisor

Dr. Tyler Derr

Advisor Signature

Date

Department of Computer Science
School of Engineering
Vanderbilt University

April 12, 2025

Contents

Summary Data

Chapter 1

Summary Data

Personal Information

- **Name:** Haoli Yin
- **Date Entered Program:** Fall 2023
- **Status:** Full Time
- **Principal Source of Support:** Cornelius Vanderbilt Scholarship

Academic Advisor

- **Name:** Dr. Tyler Derr

Courses and Grades

UNOFFICIAL DOCUMENT ISSUED TO STUDENT – NOT OFFICIAL

Name : Haoli Yin
Student # : 000768997
Birth Date : 09/24

Academic Program(s)

School of Engineering BS
Computer Science Major

Test Credit
Applied Toward Undergraduate Program

CS	1101	AP: Programming & Prob Solv	3.00
ENGL	1230W	AP: Lit/Analytical Thinking	3.00
ENGL	1220W	Drama: Forms/Tech	3.00
ENGL	1300W	AP: Intermediate Composition	3.00
HIST	2061	AP: No Eq (HIST-United States)	3.00
MATH	1300	AP: Accel Single-Var Calc I	4.00
MATH	1301	Accel Single-Var Calc II	4.00
PSCI	1100	AP: Intro Amer Govt/Politics	3.00
PSY	1200	AP: General Psychology	3.00
		TOTAL:	29.00

Undergraduate Academic Record (4.0 Grade System)

2021 Fall

BSCI	1510	**Intro to BioSci: Lecture**	3.00	**A**	12.00
BSCI	1510L	**Intro to BioSci: Lab**	1.00	**A**	4.00
CHEM	1601	**General Chemistry**	3.00	**A**	12.00
CHEM	1601L	**General Chemistry Lab**	1.00	**A-**	3.70
ES	1115	**Engr Freshman Seminar**	1.00	**A**	4.00
Course Topic:		Moore's Law & Engr Econ			
ES	1401	**Intro to Engr, Module 1**	1.00	**A**	4.00
ES	1402	**Intro to Engr, Module 2**	1.00	**A**	4.00
ES	1403	**Intro to Engr, Module 3**	1.00	**A**	4.00
SPAN	2201	**Intermediate Spanish I**	3.00	**A**	12.00

	EHRS	QHRS	QPTS	GPA
SEMESTER:	15.00	15.00	59.70	3.980
CUMULATIVE:	44.00	15.00	59.70	3.980

2022 Spring

BSCI	1511	**Intro to BioSci: Lecture**	3.00	**A**	12.00
BSCI	1511L	**Intro to BioSci: Lab**	1.00	**A**	4.00
CHEM	1602	**General Chemistry**	3.00	**A**	12.00
CHEM	1602L	**General Chemistry Lab**	1.00	**A**	4.00
CS	2201	**Prog Design Data Struct**	3.00	**A**	12.00
EECE	3860	**Undergraduate Research**	3.00	**A**	12.00
MATH	2300	**Multivariable Calculus**	3.00	**A**	12.00

	EHRS	QHRS	QPTS	GPA
SEMESTER:	17.00	17.00	68.00	4.000
CUMULATIVE:	61.00	32.00	127.70	3.990

2022 Fall

CS	2212	**Discrete Structures**	3.00	**A**	12.00
CS	3251	**Intermed Software Design**	3.00	**A**	12.00
EECE	2123	**Digital Systems**	3.00	**A**	12.00
EECE	2123L	**Digital Systems Laboratory**	1.00	**A**	4.00
MATH	2410	**Methods Linear Algebra**	3.00	**A**	12.00
NSC	2201	**Neuroscience**	3.00	**A**	12.00

	EHRS	QHRS	QPTS	GPA
SEMESTER:	16.00	16.00	64.00	4.000
CUMULATIVE:	77.00	48.00	191.70	3.993

2023 Spring

CS	1151	**Computers & Ethics**	3.00	**A**	12.00
CS	3270	**Programming Languages**	3.00	**A**	12.00
CS	3281	**Principles Operating Systems I**	3.00	**A**	12.00
CS	3860	**Undergrad Research**	3.00	**A**	12.00
MATH	2820	**Intro Prob/Math Stats**	3.00	**A-**	11.10

	EHRS	QHRS	QPTS	GPA
SEMESTER:	15.00	15.00	59.10	3.940
CUMULATIVE:	92.00	63.00	250.80	3.980

2023 Fall

CS	3250	**Algorithms**	3.00	**A**	12.00
CS	3861	**Undergraduate Research**	3.00	**A**	12.00
		Quantitative Contribution of Individual Samples in Diffusion Models: A Leave-One-Out Approach			
MATH	2420	**Methods of Ord Diff Eqs**	3.00	**A**	12.00

	EHRS	QHRS	QPTS	GPA
SEMESTER:	9.00	9.00	36.00	4.000
CUMULATIVE:	101.00	72.00	286.80	3.983

Date: 04/11/2025

UNOFFICIAL DOCUMENT ISSUED TO STUDENT – NOT OFFICIAL

Name : Haoli Yin
Student # : 000768997
Birth Date : 09/24

2024 Spring

CS	4260	**Artificial Intelligence**	3.00	**A**	12.00
MATH	3890	**Selected Topics**	3.00	**A**	12.00
Course Topic:		Computing with Splines			

	EHRS	QHRS	QPTS	GPA
SEMESTER:	6.00	6.00	24.00	4.000
CUMULATIVE:	107.00	78.00	310.80	3.984

2024 Fall

CS	3891	**Special Topics**	3.00	**A-**	11.10
Course Topic:		Reinforcement Learning			
CS	4287	**Principles of Cloud Computing**	3.00	**A**	12.00
CS	4959	**Computer Science Seminar**	1.00	**A**	4.00

	EHRS	QHRS	QPTS	GPA
SEMESTER:	7.00	7.00	27.10	3.871
CUMULATIVE:	114.00	85.00	337.90	3.975

2025 Spring

CS	4240	**Internet of Medical Things**	(3.00)	0.00
CS	4267	**Deep Learning**	(3.00)	0.00

	EHRS	QHRS	QPTS	GPA
SEMESTER:	0.00	0.00	0.00	0.000
CUMULATIVE:	114.00	85.00	337.90	3.975

IMMERSION VANDERBILT

Experiential Learning: **Research**

Project Title: **AI-powered Accessibility Tool**

---------- NO ENTRIES BELOW THIS LINE ----------

Date: 04/11/2025

UNOFFICIAL DOCUMENT ISSUED TO STUDENT – NOT OFFICIAL

Name : Haoli Yin
Student # : 000768997
Birth Date : 09/24

Academic Program(s)

Grad Schl Master of Science
Computer Science Major

Graduate Academic Record (4.0 Grade System)

2023 Fall

CS	5283	**Computer Networks**	3.00	**A**	12.00
CS	8395	**Special Topics**	3.00	**A**	12.00
Course Topic:		Deep Learning: Representation			
ECE	8395	**Special Topics**	3.00	**B**	9.00
Course Topic:		Engineering for Surgery			

	EHRS	QHRS	QPTS	GPA
SEMESTER:	9.00	9.00	33.00	3.666
CUMULATIVE:	9.00	9.00	33.00	3.666

2024 Spring

CS	5262	**Found of Machine Learnin**	3.00	**A**	12.00
CS	5891	**Special Topics**	3.00	**A**	12.00
Course Topic:		Mach Lrn / Nat Lang Proc Hlthc			
CS	8395	**Special Topics**	3.00	**A**	12.00
Course Topic:		Selected Topics in Deep Learni			
CS	8395	**Special Topics**	3.00	**A**	12.00
Course Topic:		AI for Cyber-Physical Systems			
RCRG	6303	**Responsible Conduct Research**	0.00	**P**	0.00

	EHRS	QHRS	QPTS	GPA
SEMESTER:	12.00	12.00	48.00	4.000
CUMULATIVE:	21.00	21.00	81.00	3.857

2024 Fall

CS	6362	**Advanced Machine Learning**	3.00	**A**	12.00
CS	7999	**Master's Thesis Research**	3.00	**S**	0.00

	EHRS	QHRS	QPTS	GPA
SEMESTER:	6.00	3.00	12.00	4.000
CUMULATIVE:	27.00	24.00	93.00	3.875

2025 Spring

CS	8390	**Individual Studies**	(3.00)	0.00
		Understanding Systems for ML		
CS	8395	**Special Topics**	(3.00)	0.00
Course Topic:		Security & Privacy in Pervasiv		

	EHRS	QHRS	QPTS	GPA
SEMESTER:	0.00	0.00	0.00	0.000
CUMULATIVE:	27.00	24.00	93.00	3.875

---------- NO ENTRIES BELOW THIS LINE ----------

Date: 04/11/2025

Statement of Professional Goals and Achievements

Chapter 2

Statement of Professional Goals and Achievements

Professional Goals

My journey in computer science has shaped ambitious yet focused professional aspirations that blend technical innovation with real-world impact. I aim to advance large-scale machine learning infrastructure, with particular emphasis on vision-language systems that can transform how humans interact with technology. Specializing in distributed computing solutions for extremely large-scale deployments (100TB+), I seek to address fundamental challenges in data processing that arise at the intersection of different modalities.

At the heart of my professional goals is a commitment to making multimodal systems more efficient and accessible. My master's degree has equipped me with both the theoretical foundation and practical experience to tackle complex problems in this domain, preparing me to contribute meaningfully to research teams and industry projects that push the boundaries of what's possible in machine learning and distributed systems.

Achievements During Graduate Studies

During my graduate studies, I've been fortunate to engage in research and professional experiences that have significantly shaped my development as a computer scientist. I've authored several impactful publications, including "UniCat: Crafting a Stronger Fusion Baseline for Multimodal Re-Identification," which was accepted at the NeurIPS 2023 UniReps Workshop and achieved state-of-the-art performance on several benchmark datasets. Additionally, my work on "GraFT: Gradual Fusion Transformer for Multimodal Re-Identification" received borderline acceptance at WACV 2024, representing the culmination of my research internship at Modern Intelligence.

My research in biomedical applications has been particularly rewarding, with projects like "Digital Staining of Unpaired White and Blue Light Cystoscopy Videos for Bladder Cancer Detection" and "SpecReFlow: A Specular Reflection Restoration Framework." The latter allowed me to deliver an oral presentation at the SPIE Photonics West 2023 Conference.

Professionally, I've made significant contributions at organizations like DatologyAI, where I orchestrated data pipelines at multi-billion sample scale and improved CLIP pretraining efficiency by over 10x. At Modern Intelligence, I led the development of novel multimodal fusion techniques while enhancing the team's infrastructure for model training, achieving a 400% improvement in training speed.

My academic achievements have been recognized through prestigious honors, including being named a Neo Scholar Finalist (2024), receiving the Goldwater Scholarship (2023), participating in the Google CS Research Mentorship Program, and maintaining my Cornelius Vanderbilt Scholarship throughout my studies.

Interests in Computer Science

I'm drawn to the fascinating intersection of systems engineering and machine learning, where theoretical advancements meet practical implementation challenges. This space offers abundant opportunities to solve meaningful problems that impact how we process, understand, and derive value from data at scale.

My passion centers on addressing scale challenges in multimodal processing—finding elegant solutions to the computational and architectural problems that arise when working with diverse data types simultaneously. The complexity of these systems, with their intricate performance trade-offs and optimization requirements, presents intellectually stimulating puzzles that I find deeply engaging.

My academic journey has followed a deliberate path from cloud computing foundations to advanced machine learning techniques. This progression has allowed me to build a comprehensive understanding of both the infrastructure that supports AI systems and the algorithms that drive them forward. Through coursework, research projects, and industry experience, I've cultivated expertise in distributed systems, optimization techniques, and multimodal fusion approaches that form the foundation of next-generation AI applications.

Curriculum Vitae

Chapter 3

Curriculum Vitae

Haoli Yin

Updated April 11, 2025

Email: haoli.yin@vanderbilt.edu
Phone: (346) 307 - 4568
Website: https://haoliyin.me
GitHub: https://github.com/Nano1337
LinkedIn: https://linkedin.com/in/haoliyin
Twitter: https://twitter.com/HaoliYin

Research Interests

My research interest mainly focuses on incorporating rich information from multiple modalities, times, and views to create a better structured understanding of the world and the objects contained within it. In particular, I've been tackling practical and real-world challenges in various application domains such as multi-sensor fusion and video completion. I currently focus on the following problems:

- Multi-Sensor Fusion: Multimodal Classification, Retrieval, & Re-Identification
- Video Completion: Optical Flow, Domain Transfer, Style Transfer

Education

Vanderbilt University — Nashville, Tennessee
M.Sc. in Computer Science — May 2025

Vanderbilt University — Nashville, Tennessee
B.S. in Computer Science and Mathematics — May 2024
GPA: 3.98
Relevant Coursework: Algorithms, Computer Networking, Data Structures, Systems, Linear Algebra, Machine Learning, Multivariable Calculus, Operating Systems, Probability and Statistics, Programming Languages, Representation Learning

Experience

DatologyAI — June. 2024 - Present
Member of Technical Staff

- Orchestrated data pipeline at ***multi-billion sample scale***, curating image-text multimodal datasets to ***speed up CLIP pretraining by >10x*** to reach the same performance as the uncurated raw data baseline and ***>2x*** vs CLIPScore filtering
- Ported the OpenCLIP repository for internal use, enabling ***multi-node, multi-gpu distributed training*** with SLURM and implemented comprehensive eval suite. Monitored CLIP pretraining in WandB and managed artifact storage with AWS S3.
- Fine-tuned Multimodal Data Filtering Networks for improved scoring and curating ***multimodal pretraining datasets***
- Led a ***synthetic data generation*** project to leverage vLLM, improving pretraining time by 40% and avg eval performance by up to 50%.

Modern Intelligence — Jan. 2023 – Dec 2023
AI Research Scientist Intern

- Authored two manuscripts and developed the Gradual Fusion Transformer, which achieved ***state-of-the-art performance*** in multimodal vehicle re-identification benchmarks, reduced transformer model size by ***62%***, and introduced a novel ***multimodal contrastive learning*** objective, thereby establishing a unique market differentiator for the company.
- Led ***3-day sprints***, swiftly evolving ideas into fully-realized experiments with intern team, catalyzing project momentum.
- Engineered a robust, modular ***PyTorch infrastructure***, harnessing Lightning Fabric for ***multi-GPU training*** to supercharge model training speed by ***400%***, advancing overall project timeline.
- Authored a ***custom job scheduler*** with a user-centric interface, optimizing load balancing and training run management, improving resource utilization by ***40%***

Bowden Biomedical Optics Lab @ Vanderbilt University Nov 2021 – Present
Research Assistant, Computer Science

- Spearheaded deep learning research on specular reflection restoration in white-light endoscopy videos as ***first author*** in Bowden Biomedical Optics Lab, achieving state-of-the-art results with ASPP model for segmentation (92.8% Dice Score, 52.3% sensitivity increase over U-net models) and flow-guided video completion pipeline leveraging ***optical flow*** estimation and ***vision transformer*** models (16.8% PSNR and 10.1% SSIM improvements over spatial inpainting methods).
- Project areas include 3D hollow organ model reconstruction, ***video artifact restoration***, and a current project of using GAN models for semantically aware ***modality transfer*** to enhance sensitivity of bladder cancer detection.

Yoomi Health Sept. 2022 – Feb. 2023
Machine Learning Engineer

- Spearheaded research initiatives as ***third hire*** in pre-seed physical therapy startup, delivering a state-of-the-art ***pose estimation model*** with ***98% mAP*** using the EfficientFormerV2 transformer backbone for ***mobile optimization***.
- Pioneered efficient in-browser ***edge-deployment*** of the core 2D pose estimation model using ***TF.js*** and ***int8 quantization***, achieving real-time inference optimization and driving significant improvements in ***speed and performance***.
- Leveraged networking to secure startup demo with ***Mark Cuban***, resulting in $46k pre-seed funding win.

Lynntech Inc May 2022 – August 2022
Data Science Research Intern

- Designed and implemented a ***GPU-accelerated*** state estimation engine using C++ and MAGMA (CUDA wrapper), ***reducing runtime by 21.8%*** compared to the MATLAB baseline.
- Developed over 20 GPU-accelerated ***linear algebra utility functions*** with unit tests using C++.
- Validated the effectiveness of 20+ ***adversarial ML patterns*** on 30+ state-of-the-art ***PyTorch classification models*** using Anaconda, Jupyter Notebook, NumPy, OpenCV, and Pandas

Arion Blue LLC December 2020 – May 2021
Software Engineering Intern

- Coordinated metadata collection on a Software-Defined Network with information from ***20+ literature reviews***
- Trained a ***Proximal Policy Optimization*** reinforcement learning model on a custom environment to predict the best ***Azure Files*** storage tier to place data based on daily size and read/write frequencies
- Optimized file archival process and minimized data storage costs by ***74%*** during backtesting

Summer STEM Institute June 2020 – August 2020
Data Science Researcher

- Built an ***explainable AI*** pipeline to assess prostate tumor detection from mpMRI scans with Logistic Regression, SVM, Random Forest, XGBoost, and PyTorch CNN with saliency mapping for each of three prostate zones.

- ***Published paper*** "Prostate Lesion Detection and Salient Feature Assessment Using Zone-Based Classifiers" in the Summer STEM Institute Journal, selected as 1 of 10 papers out of 80+ submissions

Conference Publications/Talks

[C3] ***UniCat: Crafting a Stronger Fusion Baseline for Multimodal Re-Identification***
Jennifer Crawford, **Haoli Yin**, Luke McDermott, Daniel Cummings
Unifying Representations in Neural Models **(UniReps) Workshop NeurIPS 2023**
arXiv: 2310.18812, 2023

[C2] ***Digital Staining of Unpaired White and Blue Light Cystoscopy Videos for Bladder Cancer Detection in the Clinic***
Shuang Chang, **Haoli Yin**, Kristen Scarpato, Amy Luckenbaugh, Sam Chang, Soheil Kolouri, Audrey Bowden.
Medical Imaging in Deep Learning **(MIDL) 2023**, Nashville, TN

[C1] ***SpecReFlow: A Specular Reflection Restoration Framework using Flow-Guided Video Completion***
Haoli Yin, Rachel Eimen, Daniel Moyer, Audrey Bowden.
Under Review for SPIE Journal of Medical Imaging.
SPIE Photonics West 2023, San Francisco, CA
Oral Presentation for Advanced Biomedical and Clinical Diagnostic and Surgical Guidance Systems XXI

Preprints

[P2] ***GraFT: Gradual Fusion Transformer for Multimodal Re-Identification***
Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, Daniel Cummings.
Under Review. Previously Borderline reviews at WACV 2024 Round 2.
arXiv:2310.16856, 2023

[P1] ***Prostate Lesion Detection and Salient Feature Assessment Using Zone-Based Classifiers***
Haoli Yin and Nithin Buduma. arXiv:2208.11522, 2020

Reviewer Services

International Conferences:
2023 *Neural Information Processing Systems* (NeurIPS)

Honors and Scholarships

Neo Scholar	*2024*
Goldwater Scholar	2023
Google CS Research Mentorship Program	2023
Cornelius Vanderbilt Scholar	2021
Equitable Excellence $10k Scholarship	2021
National Merit Scholar	2021
Coca-Cola Scholarship Semifinalist	2020
Science Olympiad National Medalist	2019-2021
USA Biology Olympiad Semifinalist	2018-2020

Leadership

VandyHacks February 2022 – Present
Sponsorship Assistant Director
- Spearhead the coordination of cold calls and lead the management of existing sponsor relations to exceed our goal of raising ***$80,000 in funding*** for Vanderbilt's fall hackathon event

- Represent VandyHacks organization at conferences and networking events, ***increasing lead generation by 40%*** while building strong partnerships with potential sponsors for future events.

Vanderbilt Commodore Orchestra August 2021 – Present
Viola Section Leader

- ***Scheduled and led*** viola section practices, fostering a collaborative environment that improved the overall quality of performances.
- Provide motivation and ***coaching*** to slower members, ensuring that all members are able to play to their full potential.
- Collaborated with fellow section leaders to enhance the overall ***cohesion and excellence*** of the orchestra.

References

Matthew Leavitt, Chief Scientist, DatologyAI, US
Email: matthew@datologyai.com

Prof. Audrey Bowden, Associate Professor, Vanderbilt University, US
Email: a.bowden@vanderbilt.edu

Dr. Daniel Cummings, Staff AI Research Scientist, Modern Intelligence, US
Email: daniel@modernintelligence.ai

Knowledge and Mastery of Computer Science Concepts

Chapter 4

Knowledge and Mastery of Computer Science Concepts

Application of Computer Science Concepts

Paper: "Unimodal Ensemble: Addressing Modality Laziness in Multimodal Fake News Detection" from the class AI for Cyberphysical Systems with Dr. Meiyi Ma.

Problem Statement

The paper addresses the challenge of "modality laziness" in multimodal fake news detection systems. This phenomenon occurs when multimodal models primarily learn from a dominant modality while underutilizing other modalities, leading to suboptimal performance. In multimodal fake news detection, where both textual and visual information are important for accurate classification, this bias significantly limits effectiveness, sometimes causing multimodal models to perform worse than unimodal alternatives.

Methodology

The authors propose a novel Unimodal Ensemble (UME) architecture that independently trains unimodal image and text models in parallel. Key aspects of the methodology include:

1. Separate backbones for text and image processing, using transfer learning with pretrained CLIP variants (SigLIP for Fakeddit dataset and Chinese CLIP for Weibo dataset)
2. Independent training with modality-specific loss functions to ensure each modality learns its task-relevant features without cross-modal influence
3. Simple averaging of unimodal logits during inference to create joint predictions
4. Evaluation on two datasets: Fakeddit (English) and Weibo (Chinese)

This approach intentionally avoids complex fusion mechanisms, focusing instead on maximizing the potential of each unimodal model before combination.

Results and Discussion

The UME architecture outperformed state-of-the-art multimodal fake news detection models (EANN, SpotFake, HMCAN, and CAFE) on both datasets:

- Achieved highest accuracy on both Fakeddit (91.9
- Demonstrated superior performance for fake news detection metrics, with significant improvements in precision and recall
- Ablation studies confirmed UME's effectiveness compared to other approaches addressing modality laziness (late fusion, OGM-GE, and QMF)

The results suggest that allowing each modality to independently learn its optimal features before combination is more effective than traditional fusion-based approaches for multimodal fake news detection. The authors note this finding implies strong intra-modal signals for fake news detection exist within each modality separately, and that combining these signals through simple ensembling creates a more robust multimodal system.

Personal Contribution

As the primary technical contributor to this project, I led the implementation efforts across multiple phases. I developed the data pipeline for both the Fakeddit and Weibo datasets, including automated download scripts and preprocessing modules that standardized the multimodal inputs. For the model architecture, I implemented the parallel unimodal training approach with separate backbones for text and image processing, adapting pre-trained CLIP variants (SigLIP and Chinese CLIP) for our specific task. I also designed and executed the training framework with modality-specific loss functions, implemented the ensemble prediction mechanism, and created comprehensive evaluation protocols that measured performance across multiple metrics. Throughout the project, I maintained the codebase, ensuring reproducibility and documentation quality while collaborating with team members on experiment design and results analysis.

Software Artifact

Project Overview

The code linked below contains the full reproducible study for the paper.

Design

The repository follows a modular architecture centered around reproducible machine learning workflows. The codebase is organized into distinct components for data processing, model training, and evaluation, with configuration files enabling experiment customization. This design facilitates independent training of unimodal models while maintaining a clean separation between preprocessing, feature extraction, and ensemble prediction mechanisms.
GitHub Repository: `https://github.com/Nano1337/ume-fakenews`

Unimodal Ensemble: Addressing Modality Laziness in Multimodal Fake News Detection

David Gao
Department of Computer Science
Vanderbilt University
Nashville, United States
david.gao@vanderbilt.edu

Lincoln Murr
Department of Computer Science
Vanderbilt University
Nashville, United States
lincoln.d.murr@vanderbilt.edu

Haoli Yin
Department of Computer Science
Vanderbilt University
Nashville, United States
haoli.yin@vanderbilt.edu

***Abstract*—The growth of fake news and misinformation on social media platforms has become a significant challenge in the digital age. Although there are existing multimodal fake news detection approaches, they often suffer from modality laziness, where the model primarily learns from a dominant modality, leading to suboptimal performance. We propose a novel Unimodal Ensemble (UME) architecture that independently trains unimodal image and text models in parallel to address this issue. By allowing each modality to learn task-relevant features without cross-modal influence, our method ensures that textual and visual information are effectively used in the fake news detection task. We benchmark our approach on two widely used datasets, Fakeddit and Weibo, and compare UME's performance against state-of-the-art models. Our evaluation demonstrates that our model outperforms these benchmarks, exhibiting competitive or state-of-the-art performance in multimodal fake news detection for various metrics. Our ablation study further validates the effectiveness of the UME approach compared to past proposed solutions for modality laziness. The code can be found at https://github.com/Nano1337/ume-fakenews.**

***Index Terms*—Fake News Detection, Multimodal, Modality Laziness, Ensemble Learning**

I. Introduction

A. Motivation

In the era of rapid information growth and the proliferation of generative AI tools, spreading misinformation has become a significant challenge [22]. The internet, while serving as an invaluable resource for learning, has also become a breeding ground for fake news [8]. This issue is particularly concerning in educational settings, where students often explore new subjects and may lack the experience to discern between credible sources and fabricated content. It is also rampant across social media platforms, a medium that can potentially influence many real-world events, such as elections.

The impact of misinformation on the real world cannot be overstated [12]. When students are exposed to fake news, it can lead to misconceptions, hinder critical thinking skills, and ultimately undermine the educational process [21]. In the context of political campaigns, where technology is increasingly integrated into the voter experience, addressing the threat of fake news becomes even more crucial [2]. As such, there is a pressing need for effective fake news detection systems to support people in their quest for knowledge and help foster a culture of information literacy.

B. Challenges

Developing a fake news detection model presents several challenges. One major hurdle is overcoming bias in the data and ensuring the model can generalize well to various topics and information sources encountered in educational contexts. Information comes in many shapes and forms, so we must find datasets that encapsulate various scenarios.

Another significant challenge lies in contextualizing the information being analyzed. Fake news often relies on subtle manipulations of language and imagery that can be difficult to detect without considering the broader context. Models must understand the relationships between different information modalities and reason about their veracity holistically. This requires sophisticated techniques that can capture the semantic and contextual cues present in the data.

Furthermore, the accessibility and interpretability of fake news detection models are critical factors in their successful adoption in educational settings. Students and educators must be able to interact with the system and understand its outputs easily. This necessitates the development of intuitive interfaces and clear explanations of the model's decision-making process. Striking the right balance between model complexity and user-friendliness is a key challenge that must be addressed.

On a technical level, multimodal learning suffers from the "modality laziness" problem [29], which prevents the fusion-based multimodal model from effectively learning from multimodal data, leading to possible underperformance, even when compared to its unimodal counterpart. We aim to more carefully explore how this phenomenon affects modeling multimodal fake news detection and if a solution exists to bypass this issue.

C. Novelty

Our proposed solution aims to tackle the challenges of fake news detection through a novel application of multimodal learning techniques. By leveraging textual and visual information, we seek a more comprehensive understanding of the analyzed content. Integrating multiple data modalities allows our model to identify subtle inconsistencies and manipulations that single-modality approaches may miss for greater robustness.

Central to our solution is a specific data fusion process tailored to the unique requirements of fake news detection. We explore techniques such as late fusion, on-the-fly gradient modulation with generalization enhancement (OGM-GE) [15], quality-aware multimodal fusion (QMF) [28], and unimodal model ensembling (UME), to determine the most effective approach for combining the outputs of the text and image models (see Section VI.B for benchmarking). This comparative study of fusion techniques sets our work apart from previous efforts and contributes to the broader understanding of multimodal learning in the context of fake news detection.

D. Contributions

By addressing the challenges of fake news detection in a real-world context through a novel multimodal approach, our work aims to empower people with the tools and knowledge necessary to navigate the complexities of the modern information landscape. We believe that our contributions will enhance the learning experience and foster a culture of critical thinking and information literacy that extends to more applications in the future.

II. Related Work

Fake news and misinformation detection have garnered significant attention recently due to the increasing prevalence of visual and textual disinformation on social media platforms. Researchers have proposed various approaches for tackling the problem, focusing on textual and visual modalities.

Several works emphasize improving the detection of out-of-context captions in multimodal fake news. [1] builds upon the COSMOS framework [3] and proposes four methods to improve its detection accuracy, including fake-or-fact checking, differential sensing, object-caption matching, and threshold adjustment. Similarly, [9] incorporates textual semantics understanding from large corpora and combines it with various combinations of text-image matching and image captioning methods to classify triples of (image, caption1, caption2) into out-of-context and no out-of-context labels.

Another research gap focuses on the data-intensive nature of deep neural networks and the need to preserve structural and semantic information in multimodal fake news detection. [30] introduces a self-supervised model grounded in contrastive learning to extract features from text and image simultaneously, achieving strong visual feature extraction with minimized data training requirements. Liu et al. [11] propose a semantic distillation and structural alignment (SDSA) network to reduce redundant information and effectively preserve structural and semantic information.

"Synthetic Misinformers: Generating and Combating Multimodal Misinformation" [14] explores manners by which we can automatically generate multimodal misinformation synthetically to tackle the labor-intensive process of manual annotation. The authors comparatively study existing and new Synthetic Misinformers and demonstrate that CLIP-based Named Entity Swapping can lead to better multimodal detection models.

[10] introduces a consistency-learning fine-grained fusion network (CFFN) that explores the consistency and inconsistency from high and low relevant word-region pairs. The authors employ a cross-modal attention mechanism to evaluate consistency and calculate inconsistency scores, demonstrating CFFN's superior performance compared to baselines.

There are four papers representing the current state of the art.

Wang et al. [23] introduce an end-to-end framework that learns event-invariant features enabling fake news detection on newly-emerged events. Their Event Adversarial Neural Network (EANN) includes a multimodal feature extractor, a fake news detector, and an event discriminator. The feature extractor and fake news detector together learn discriminable representations for fake news detection, while the event discriminator removes event-specific features to retain shared features among events.

SpotFake [20] introduces a multimodal approach for fake news detection that addresses the issue of relying on subtasks like event discriminators or cross-modality correlations. The authors exploit textual and visual features of an article without considering other subtasks. The BERT language model is used to learn text features, and VGG-19, pre-trained on ImageNet, learns image features. It performs highly on the Twitter and Weibo datasets. In a follow-up work, SpotFake+ [19] extends SpotFake by utilizing transfer learning to capture semantic and contextual information from news articles and associated images. It is the first work to perform a multimodal approach for fake news detection on a dataset with full-length articles.

In [16], authors propose a hierarchical multimodal contextual attention network (HMCAN) that can jointly model multimodal context information and the hierarchical semantics of text in a unified deep model. HMCAN employs BERT and ResNet to learn better representations for both text and images, which is then used as data in a multimodal contextual attention network to combine inter-modality and intra-modality relationships. Fake news detection is achieved using a hierarchical encoding network that captures hierarchical semantics.

Lastly, "Cross-Modal Ambiguity Learning for Multimodal Fake News Detection" [5] addresses ambiguity across different modalities that leads to inferior detection. CAFE, an ambiguity-aware multimodal fake news detection method, consists of a cross-modal alignment module, a cross-modal ambiguity learning module, and a cross-modal fusion model. CAFE aggregates unimodal features and cross-modal correlations adaptively based on the strength of the cross-modal ambiguity, which leads to more accurate fake news detection.

Our proposed solution differs from the aforementioned works in several key aspects. Despite the advancements made by these papers, a key weakness they share is the problem of modality laziness. This phenomenon occurs in multimodal fusion models (especially late-fusion models) where the multimodal model underperforms its unimodal counterpart by primarily learning from the dominant modality that emerges from gradient-based training dynamics that's faster/easier to learn [29]. This in turn saturates the joint multimodal loss (e.g.

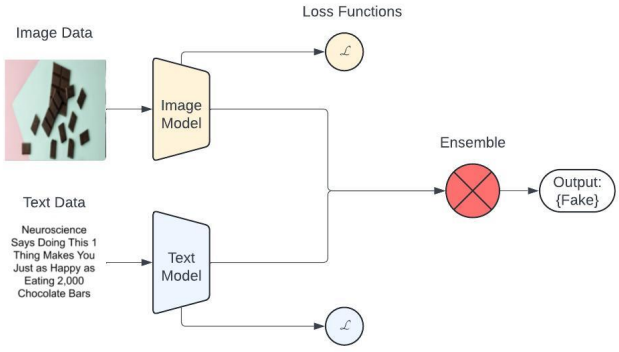


Fig. 1. Unimodal Ensemble (UME) Architecture with unimodal loss functions. The ensemble operation averages the unimodal logits to attain a joint prediction.

cross entropy loss given a joint prediction in classification) but may not be the most task-relevant, which means that the weaker modality is not pushed to learn more as the gradient norm for the joint embedding is near zero. This biases the model to mostly depend on this dominant modality during inference. Our method takes an ensemble approach, leveraging a unimodal loss per unimodal backbone to encourage each model to learn its respective unique features as best as it can without having to depend on other modalities, which leads to better overall multimodal predictions when fused (e.g. weighted average of unimodal logits) during inference time.

III. Overview

A. Problem Formulation

First, we restricted the definition of multimodal data in this study to mean textual and image data which are presented together to form information which is consumed by a person. We would like to define the concept of fake news in this study, utilizing the descriptions of the datasets.

The Fakeddit dataset [13] requires that fake news fall under one of five categories: Satire, Misleading Content, Imposter Content, False Connection, or Manipulated Content.

- Satire: True content is phrased in a satirical way that makes it false.
- Misleading Content: Manipulated information with intent to trick people.
- Imposter Content: Bot-generated content.
- False Connection: Images do not connect to the text which it is paired with.
- Manipulated Content: Images are photoshopped.

The Weibo dataset [7] classified fake news based on tweets which were reported and investigated by a committee.

We want to detect fake news, turning this into a binary classification task where we utilize machine learning to classify a sample as either fake news or real news. This is the problem we try to solve in this study.

B. Solution Overview

Our proposed solution utilizes a Unimodal Ensemble (UME) architecture to incorporate both textual and visual information. The solution architecture consists of separate image and text processing pipelines. Each input modality is passed through its corresponding backbone and classifier to obtain unimodal output logits. These logits represent the individual predictions of the image and text models regarding the verifiability of the news content, which are then evaluated with their respective binary cross entropy loss. We then directly average the logits to combine the unimodal predictions, creating joint logits used in the final inference.

A key aspect of our approach is the separate but parallel training of each unimodal model rather than a joint training strategy. By independently training image and text models, we allow the gradient descent optimization to bring each unimodal model to its maximum potential in learning to predict fake or real news. Each modality focuses on its task-relevant features and patterns without being influenced by the other modality during the learning process.

Compared to fusion-based models that attempt to learn cross-modal interactions and correlations, our Unimodal Ensemble approach offers several advantages. First, it mitigates the problem of modality laziness, as each modality is equally important and trained to make predictions independently. This intentional design decision ensures that both textual and visual information are both effectively used in the fake news detection task. Second, by avoiding the need for complex fusion mechanisms and deviations from the standard training pipeline, our solution remains computationally efficient and straightforward, making it easy to adopt to various datasets.

IV. Methods

To introduce formal notation, in this supervised learning problem, we are given a multimodal dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i = (x_{i1}, x_{i2})$ represents the input data with two modalities, x_{i1} and x_{i2}, and y_i is the corresponding output or target variable. We have two modalities, x_1 and x_2, for text and image data, respectively, and unimodal models for each will extract relevant features for downstream fusion and inference. We will denote $f(x_1)$ as the unimodal text encoder and $g(x_2)$ as the unimodal image encoder. For simplicity, we will assume that a linear classifier is included in each unimodal encoder function to output binary prediction logits.

A. Architecture Choices

Since we were computationally limited in scope, we chose to use transfer learning by initializing our unimodal backbones for each of image and text using strongly pretrained weights, particularly those acquired from Contrastive Language-Image Pretraining (CLIP) [18].

We employ CLIP-variants as the image and text unimodal backbones in our ensemble, specifically using SigLIP [26] for the Fakeddit dataset and Chinese CLIP [25] for the Weibo dataset. This choice was made to capture the semantic and contextual information from the news articles and associated

images in a domain-specific manner with proper language support. For instance, the text data in the Weibo dataset (see Section V.A for more information) consisted of chinese characters, which is not in the typical training corpora of English-based CLIP models. This called for a special chinese tokenizer and pre-trained model that handled this domain, which was the justification in choosing Chinese CLIP.

B. Training Pipeline

Once the pre-trained weights were loaded, the models were then fully unfrozen and fine-tuned on each of fake news datasets. The unimodal forward pass can be represented as:

$$z_{i1} = f(x_{i1}) \tag{1}$$

$$z_{i2} = g(x_{i2}) \tag{2}$$

$$z_i = \frac{z_{i1} + z_{i2}}{2} \tag{3}$$

$$p_i = \sigma(z_i) \tag{4}$$

where $z_i \in \mathbb{R}^{nx2}$ is the joint logit after directly averaging each unimodal logit. These joint logits are then passed through a softmax function $\sigma(\cdot)$ to attain the joint prediction p_i. Although we use this joint prediction during inference time, we do not directly use it to generate gradients for gradient descent, as this would cause the modality laziness problem in the late-fusion setting [29]. To enforce each modality to completely learn its respective predictive features, we take a unimodal ensemble approach by which each unimodal encoder would have its own loss function. Let L be the standard (binary) cross entropy loss for classification. We can attain unimodal predictions from the unimodal logits shown in equations 1 and 2, which would then be separately evaluated:

$$L_1 = \sum_{i=1}^{n} L(\sigma(z_{i1}), y_i) \tag{5}$$

$$L_2 = \sum_{i=1}^{n} L(\sigma(z_{i2}), y_i) \tag{6}$$

where n is the batch size. From multi-task learning, the total loss function would be additive combination of the unimodal losses:

$$L_{total} = L_1 + L_2 \tag{7}$$

C. Training Details

All experiments were run with Ubuntu 22.04 on a machine with an NVIDIA RTX 4090 GPU with 24GB VRAM. For modular code and efficient training, we leveraged the PyTorch Lightning [6] training framework with fp16 mixed precision and 12 CPU workers. To monitor training and assess evaluation, we used Weights and Biases for logging. For all runs, we used the SGD optimizer with a momentum of 0.9 and weight decay of 1.0e-4.

For the Fakeddit dataset, we used a batch size of 144 samples with a learning rate of 0.01. This learning rate was modulated by a StepLR scheduler that halved the learning rate (i.e. gamma=0.5) every 500 steps. Only one epoch was used for training (similar to what is done in large language model training) since there were sufficiently many training samples necessary for convergence.

For the Weibo dataset, we used a batch size of 50 samples with a learning rate of 0.01. We also used a StepLR scheduler here that halved the learning rate every 250 steps. The model was then trained for 20 epochs with validation metrics logged between every epoch.

Since there is some class imbalance as seen in Table I, we accounted for this during training by balancing out the sampler using the provided PyTorch WeightedRandomSampler class. The robust metrics utilized during evaluation in Section V.C also accounted for this fact.

V. Evaluation

A. Data

We use the Fakeddit dataset [13] and Weibo dataset [7] to evaluate our model. Both are public datasets which contain labeled samples of textual data (Fakeddit is English, Weibo is Chinese) paired with an image. Such a sample is given a binary classification of Fake News (label 0) or Real News (label 1). Both datasets were pre-processed to centralize all the metadata and text data into a single comma-separated values (CSV) file, of which each row contained a sample's text data, supervised label, and a corresponding ID for the respective image file in the image folder directory. Because the CLIP model was pretrained with a maximum of 77 text token length, we first tokenized the text accordingly by either padding or truncating to that length to effectively leverage parallelism in training.

The multimodal subsection of the Fakeddit dataset is a very large dataset (around 110GB data) consisting of around 680,000 samples of text and images scraped from various subreddits off the Reddit social media platform.

This data is split into training, validation, and testing splits. The training split is around 82% of the dataset and consists of 61% Fake News and 39% Real News. This adds up to more than 560,000 samples. The validation and testing sets are approximately the same size and follow the same distribution as the training split. Each are around 60,000 samples in total.

The Weibo dataset is much smaller and is scraped from the Chinese social media platform, Weibo. The labels are given by the official "rumor debunking system" of the platform. This multimodal dataset consists of more than 9,000 samples with Chinese characters paired with images.

This data is only split into two splits, so we use the smaller split as both validation and testing. The training split consists of around 80% of the dataset (7,481 samples) with fake news making up around 80% of samples. The other split (Validation and Testing) consists of 1,930 samples with fake news making up around 77% of samples.

B. Baselines

We compare our solution against the following models which have also benchmarked their algorithms on the same multimodal fake news detection task. EANN [24], SpotFake

TABLE I
DATASET STATISTICS

Dataset	Split	Total Samples	Fake News	Real News
Weibo	Training	7,481	6,044 (˜80%)	1,437(˜20%)
	Validation	1,930	1,480 (˜77%)	450 (˜23%)
	Testing	1,930	1,480 (˜77%)	450 (˜23%)
Fakeddit	Training	563,613	341,532 (˜61%)	222,081 (˜39%)
	Validation	59,299	35,979 (˜61%)	23,320 (˜39%)
	Testing	59,299	35,781 (˜61%)	23,518 (˜39%)

[20], HMCAN [17], and CAFE [5]. These represent the state-of-the-art models in fake news detection and utilize various approaches as described in Section II. All of these models utilize various multimodal approaches to detect fake news, and are the most heavily cited for this area of work.

Our comparative analysis validates our approach and gives relevant comparisons for readers to see our relative performance against the top models out there.

C. Metrics

We evaluate our model on the typical metrics for a classification task: accuracy, precision, recall, and F1 score.

Accuracy measures the overall correctness of the model across all predictions, which is the most natural measurement of performance in this classification task. Precision measures the proportion of true positives within positive predictions given statistics from the confusion matrix. This measures the ability of a solution to identify samples as positive. Recall measures the proportion of predicted positives within actual positive samples, measuring the ability of a solution to catch as many true positives as possible. Finally, the F1 score is a combination of precision and recall which is calculated by:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

.

Precision, recall, and F1 score have separate scores for Fake News and Real News since we are measuring the performance of the model in identifying these and both must perform well for a comprehensive evaluation.

VI. RESULTS

A. Comparison Against State-of-the-Art

We compare our model against the four state-of-the-art models mentioned in previous sections. Table II summarizes the performance of the models on the two datasets.

Our model, unimodal ensemble (UME), has the highest accuracy for both datasets which is a positive indication that our model outperforms the other models. We can see that the margin is slimmer on Fakeddit dataset, but the improvement in accuracy is more than 3% on the Weibo dataset.

As we dive deeper into our evaluation metrics, we see that we outperform all benchmarks on precision and recall for fake news on the Fakeddit dataset and are competitive for the F1 score for fake news. We see a small increase of around 1% for precision and a larger increase in recall (more than 2%). For real news, we score very closely to the other models, but we do not outperform the state-of-the-art models.

For the Weibo dataset, we outperform Fake News metrics by a significant margin. Precision is increased by more than 3% from the best model, and F1 score is 2% higher than the best model. We also quite competitive with the results of SpotFake in the recall metric. For the real news performance, we do not outperform the other models, but still have comparable performance.

B. Ablation Study

As a part of this study, we experiment with various fusion techniques that attempt to counteract the problem of modality laziness that our solution primarily addresses.

We first baseline with using late fusion, which we know suffers from modality laziness and serves as a lower bound on multimodal performance. We then take a closer look by benchmarking with two proposed solutions: on-the-fly gradient modulation with generalization enhancement (OGM-GE) [15] and quality-aware multimodal fusion (QMF) [28]. As a brief summary, OGM-GE is able to identify the dominant modality using a gradient-based method and artificially slows down the learning rate of the dominant modality so it doesn't immediately saturate the loss and the weaker modality has a chance to learn. QMF does almost the opposite by identifying "hard" samples through loss trajectories and encouraging the multimodal model to learn more from those samples to prevent loss saturation. Finally, we show the results of our method of unimodal ensembling (UME). We compare the results of these methods in Table III.

As you can see, UME performed the best across almost all metrics across both datasets. The next best was QMF with the best fake news precision and real news recall on the Fakeddit dataset; however, it is important to note that the QMF training collapsed on the Weibo dataset, so we lack true results in that portion of our results. We hypothesize that this is due to the smaller size of the dataset leading to representational collapse.

VII. DISCUSSION

A. Limitations

One limitation of our Unimodal Ensemble architecture is the lack of synergistic interactions between modalities since there it no further depth-wise processing beyond the point of fusion. While this does not seem to hold back performance based on our benchmarking efforts, there is definite room for improvement to better model the case of modality disagreement (i.e. synergy) in the real world, which commonly seen in the task of fake news detection.

This leads us to another limitation of our work. Since we only evaluated the model on two datasets, the robustness of the model in the real world (integration into cyber-physical systems) is unknown. The model may simply learn the features of the datasets very well during training (data dependent), resulting in the excellent results that we present in this paper. Overall, the term "fake news" is quite subjective, so the task

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON TWO DATASETS.

Model	Fakeddit							Weibo						
	Acc	Fake News			Real News			Acc	Fake News			Real News		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1
EANN	0.724	0.727	0.719	0.723	0.722	0.729	0.726	0.782	0.827	0.697	0.756	0.752	0.863	0.804
SpotFake	0.819	0.801	0.848	0.824	0.839	0.790	0.813	0.892	0.902	**0.964**	0.932	0.847	0.656	0.739
HMCAN	0.881	0.880	0.882	0.881	**0.882**	0.880	0.881	0.885	0.920	0.845	0.881	**0.856**	**0.926**	**0.890**
CAFE	0.912	0.946	0.886	**0.959**	0.878	**0.942**	**0.909**	0.840	0.855	0.830	0.842	0.825	0.851	0.837
UME	**0.919**	**0.953**	**0.910**	0.931	0.872	0.932	0.901	**0.927**	**0.957**	0.948	**0.952**	0.833	0.859	0.846

TABLE III
ABLATION STUDY RESULTS.

Method	Fakeddit							Weibo						
	Acc	Fake News			Real News			Acc	Fake News			Real News		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1
Late Fusion	0.891	0.927	0.889	0.908	0.841	0.893	0.867	0.921	0.950	0.947	0.948	0.826	0.836	0.826
OGM-GE	0.865	0.885	0.891	0.888	0.832	0.825	0.828	0.915	0.950	0.939	0.944	0.805	0.836	0.820
QMF	0.917	**0.955**	0.904	0.929	0.865	**0.935**	0.899	0.767	0.767	-	0.868	0	0	0
UME	**0.919**	0.953	**0.910**	**0.931**	**0.872**	0.932	**0.901**	**0.927**	**0.957**	**0.948**	**0.952**	**0.833**	**0.859**	**0.846**

itself is not well-defined. However, we make a best effort to capture nuances by using multi-lingual datasets from different cultural perspectives in this work.

B. Implications

Since our model outperforms state-of-the-art models, we can infer that there is a gap in the understanding of multimodal model training which necessitates correction. If separating the modalities can increase performance, it means that there is some unimodal information lost in the fusion-based training approaches, and the field is not fully leveraging all modalities to the degree that they can be learned.

On the contrary, this finding also implies that there are task-relevant, intra-modal signals for fake news detection, meaning that within each modality, there are patterns which discern fake news from real news. These signals are present within each modality separately, and allows our unimodal ensemble to accurately detect fake news without any cross-modal interactions. Following this line of logic, fusing these unimodal signals (e.g. averaging) allows the multimodal model to be more robust due to the effect of ensembling.

Another interesting finding is that our model performs better on detecting fake news than detecting real news, implying that media inherently might be a biased data source, which could impact the understanding of fake news detection.

While our work provides promising results in detecting fake news, we also want to point out that any deployment of this system in a real world context would require much more testing and human-in-the-loop capabilities to ensure that First Amendment rights are not infringed upon and that bias is minimized.

C. Future Work

In the future, our technique can be applied to other modalities to see if it can consistently outperform other models (or at least achieve similar performance). In the realm of fake news detection, modalities of interest include video and audio, since those two have become deeply integrated into modern social media and the internet, in general. We did not include video as a modality in this initial study, as we believe that most media intake in an academic setting comes from text and images. This would have added additional complexity and computational burden that we leave for future work.

We can also apply the model architecture to other multimodal problems such as the sign language translation problem [27] and smart classrooms [4]. Since sensor data is even less studied than textual and image data, we could see interesting results in overcoming modality laziness.

Lastly, we must evaluate our approach on more datasets to ensure robustness of the unimodal ensemble approach. This will ensure that the performance translates across many settings and is not specific to these two datasets.

VIII. CONCLUSION

In this work, we have proposed a novel Unimodal Ensemble (UME) architecture for multimodal fake news detection, aiming to tackle the issue of modality laziness in existing fusion-based approaches. By training the unimodal image and text models in independently and in parallel, each modality learns task-relevant features without being influenced by the other during the learning process. This approach ensures that both textual and visual information are effectively utilized in the fake news detection task.

We evaluated our UME model on two widely used datasets, Fakeddit and Weibo, and compared its performance against state-of-the-art models like EANN, SpotFake, HMCAN, and CAFE. The results demonstrate that our model outperforms these baselines regarding overall accuracy across both datasets. Furthermore, our model performs strongly in detecting fake news, as evidenced by its high precision, recall, and F1 scores.

Our ablation study, benchmarking existing solutions for the modality laziness problem, further validates the effectiveness of the UME approach. UME consistently outperforms late fusion, OGM-GE, and QMF across most metrics on both

datasets, highlighting the benefits of ensembling separately-trained unimodal models.

As the field of multimodal learning continues to evolve, we believe that our work contributes to the ongoing effort to combat the spread of fake news and misinformation in the digital age.

ACKNOWLEDGMENT

Thank you to Dr. Meiyi Ma for her mentorship and guidance throughout this semester. This project would not have been possible without her.

REFERENCES

[1] Tankut Akgul, Tugce Erkilic Civelek, Deniz Ugur, and Ali C. Begen. Cosmos on steroids: a cheap detector for cheapfakes. In *Proceedings of the 12th ACM Multimedia Systems Conference*, MMSys '21, page 327–331, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[3] Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: Catching out-of-context misinformation with self-supervised learning, 2021.

[4] Wilson Chango, Juan A Lara, Rebeca Cerezo, and Cristóbal Romero. A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(4):e1458, 2022.

[5] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905, 2022.

[6] William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019.

[7] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816, 2017.

[8] Srijan Kumar and Neil Shah. False information on web and social media: A survey, 2018.

[9] Tuan-Vinh La, Minh-Son Dao, Duy-Dong Le, Kim-Phung Thai, Quoc-Hung Nguyen, and Thuy-Kieu Phan-Thi. Leverage boosting and transformer on text-image matching for cheap fakes detection. *Algorithms*, 15(11), 2022.

[10] Jun Li, Yi Bin, Jie Zou, Jiwei Wei, Guoqing Wang, and Yang Yang. Cross-modal consistency learning with fine-grained fusion network for multimodal fake news detection. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, MMAsia '23, New York, NY, USA, 2024. Association for Computing Machinery.

[11] Shangdong Liu, Xiaofan Yue, Fei Wu, Jing Sun, Yujian Feng, and Yimu Ji. Semantic distillation and structural alignment network for fake news detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6620–6624, 2024.

[12] Michael Luca and Georgios Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Manage. Sci.*, 62(12):3412–3427, dec 2016.

[13] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019.

[14] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, MAD '23, page 36–44, New York, NY, USA, 2023. Association for Computing Machinery.

[15] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[16] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 153–162, New York, NY, USA, 2021. Association for Computing Machinery.

[17] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162, 2021.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[19] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13915–13916, Apr. 2020.

[20] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE, 2019.

[21] Borhan Uddin, Nahid Reza, Md Saiful Islam, Hasib Ahsan, and Mohammad Ruhul Amin. Fighting against fake news during pandemic era: Does providing related news help student internet users to detect covid-19 misinformation? In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, page 178–186, New York, NY, USA, 2021. Association for Computing Machinery.

[22] Xiangyu Wang, Min Zhang, Weiguo Fan, and Kang Zhao. Understanding the spread of covid-19 misinformation on social media: The effects of topics and a political leader's nudge. *Journal of the Association for Information Science and Technology*, 73(5):726–737, 2022.

[23] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA, 2018. Association for Computing Machinery.

[24] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.

[25] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.

[26] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[27] Qian Zhang, JiaZhen Jing, Dong Wang, and Run Zhao. Wearsign: Pushing the limit of sign language translation using inertial and emg wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–27, 2022.

[28] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023.

[29] Yedi Zhang, Peter E. Latham, and Andrew Saxe. A theory of unimodal bias in multimodal learning, 2023.

[30] Peng Zheng, Hao Chen, Shu Hu, Bin Zhu, Jinrong Hu, Ching-Sheng Lin, Xi Wu, Siwei Lyu, Guo Huang, and Xin Wang. Few-shot learning for misinformation detection based on contrastive models. *Electronics*, 13(4), 2024.

Communication Skills in Computer Science

Chapter 5

Communication Skills in Computer Science

Artifact Demonstrating Communication Skills

Context

This paper presentation was prepared for the graduate special topics class called "Security and Privacy in Pervasive Environments". The paper itself is called "Privacy Leakage via Unrestricted Motion-Position Sensors in VR: Snooping Typed Input on Virtual Keyboards" and was a study on privacy leakage in virtual reality environments.

Slides Presentation

CS8395 Security & Privacy in Pervasive Environments

VANDERBILT UNIVERSITY

Privacy Leakage via Unrestricted Motion-Position Sensors in VR: Snooping Typed Input on Virtual Keyboards

by Yi Wu et al. (IEEE S&P 2023)

Haoli Yin

Roadmap

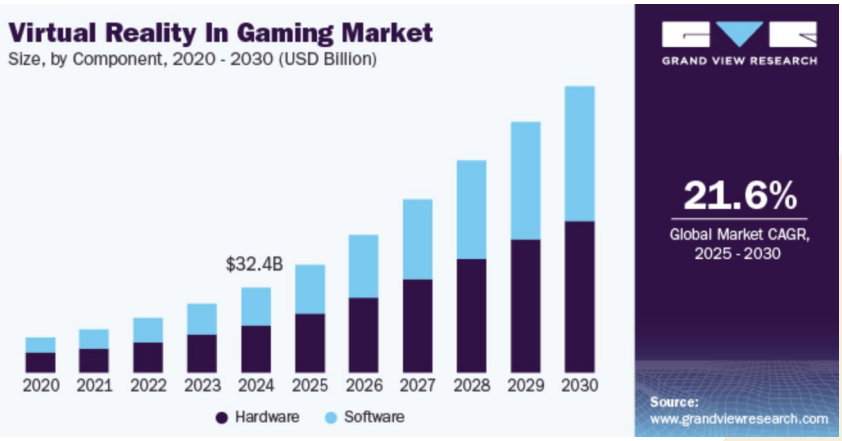
- Introduction
- Key Ideas
- Methods
- Experimentation
- Discussion
- Limitations
- Countermeasures

Introduction

- Background
- Motivation
- Related Work
- Approach Overview

Background

- **VR is Growing** [1]
- VR Devices have:
  - motion, position, orientation sensors
  - Camera, mic, headset, controllers
- Sensitive Inputs in VR
  - Typing passwords
- Privacy Risk
  - Sensors can encode private info



[1] https://www.grandviewresearch.com/industry-analysis/virtual-reality-in-gaming-market

Motivation - Privacy Vulnerabilities in VR

- **Unrestricted Sensors**
  - Most VR sensor data requires no user permission on current platforms (OpenVR, Oculus, WebXR)
- **Smartphones Case Study**
  - Similar “zero-permission” sensor attacks were known on phones (e.g. accelerometer used to infer keystrokes) [2]
- **Attack Scenario**
  - Imagine you log into a VR app and type a password on a virtual keyboard
  - How would you feel if a malicious process was reading your every movement?
  - Could it decipher what you typed?

[2] Emmanuel Owusu, Jun Han, Sauvik Das, Adrian Perrig, and Joy Ying Zhang. Accessory: password inference using accelerometers on smartphones. In HotMobile ’12, 2012

Video Demo

https://youtu.be/xaXDmjhTTTc?si=mn9d2BjNpmjr7ikz

Related Works (and their drawbacks)

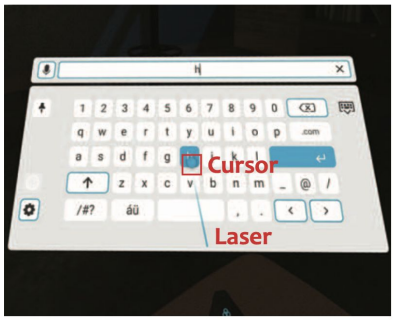
- User Authentication [3] → Narrow focus, leaves sensor data insecure
- Smartphone VR keyboard detection (Samsung Gear) [4] → Strong assumption of fixed controller rotation between keys

[3] Markus Funk, Karola Marky, Iori Mizutani, Mareike Kritzler, Simon Mayer, and Florian Michahelles. LookUnlock: Using Spatial-Targets for User-Authentication on HMDs. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA'19, pages 1–6. Association for Computing Machinery, 2019.
[4] Zhen Ling, Zupei Li, Chen Chen, Junzhou Luo, Wei Yu, and Xinwen Fu. I know what you enter on gear vr. In 2019 IEEE Conference on Communications and Network Security (CNS), pages 241–249. IEEE, 2019

Approach

- Assumptions:
  - Keyboard Layout (QWERTY)
  - Controller Typing Mechanism
- Deciphers keyboard inputs from sensor data

(a) Drum-based typing



(b) Laser-based typing

Key Ideas

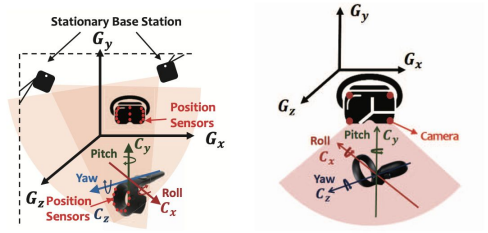
- Threat Model
- Attack Overview

Threat Model

- **Attacker's Method**: Deploy malware or malicious webpage to continuously capture sensor streams (motion, orientation, controller button states)
- **Limited Knowledge Assumption**: The attacker does **NOT** know the user's VR setup or environment (e.g. unknown keyboard app, unknown room layout)
- **Goal**: Infer the sensitive text the user types (passwords or messages) purely from sensor data
- **Realism**: Tested on two popular VR systems (HTC Vive Pro and Oculus Quest)

(a) Sensors on the headset (b) Sensors on the controller

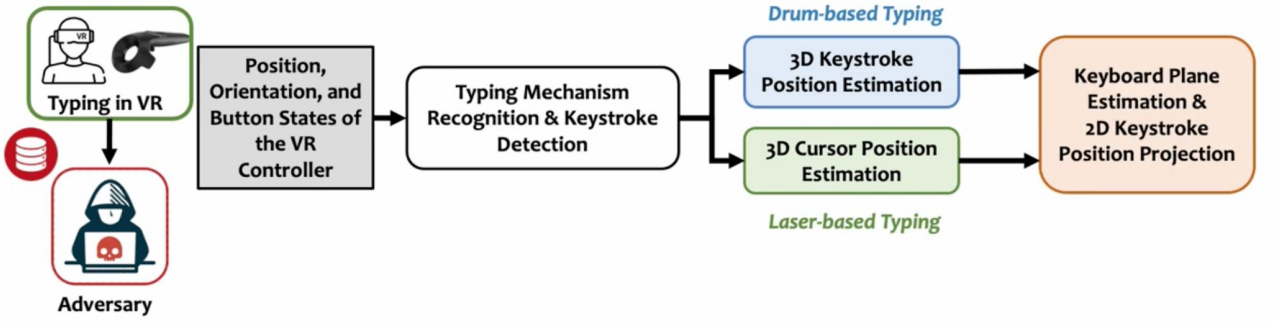
Fig. 1. Sensors in a VR system (i.e., HTC Vive Pro).



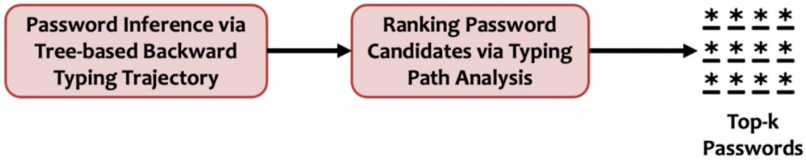
(a) Outside-in tracking (HTC Vive Pro) (b) Inside-out tracking (Oculus Quest)

Fig. 4. Position tracking & coordinate systems in VR.

Attack Overview



Inferring Passwords



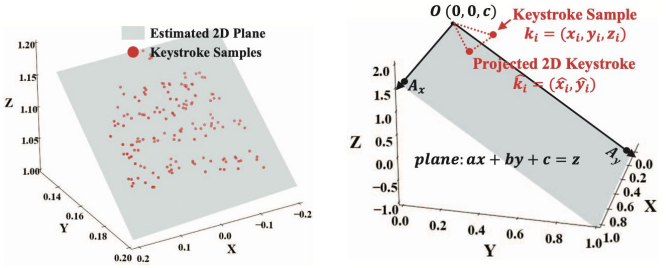
Inferring Sentences

Methods

- Keystroke Position Estimation
- Inferring Passwords
- Inferring Sentences

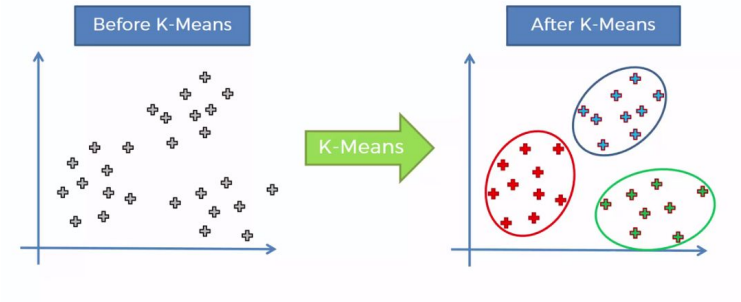
Keystroke Position Estimation

- **Goal**: Determine *which key* was hit by figuring out where the controller was at each keystroke
- **3D to 2D Projection**:
  - Each detected keystroke has a 3D position (controller coordinates)
  - We know what a QWERTY keyboard layout → virtual keyboard
  - Using a least-squares plane fit, project points onto that virtual keyboard plane
  - ✅ 2D coordinates relative to the keyboard surface.
- K-means cluster → use centroids as key mappings for calibrated virtual keyboard
- We know sequence of key presses based off timestamps



(a) Keyboard plane estimation (b) 2D keystroke position projection

Fig. 8. Illustration of Keyboard Plane Estimation & 3D-to-2D Projection.



Inferring Passwords

- **Goal**: Decipher password Input (random characters). Passwords lack context (no dictionary words), so approach is brute-force guided by geometry.
- **Tree-Based Backwards Typing Trajectory → Predict multiple password options**
  - "Enter" key will always be at the end of sequence (and serves as root of tree)
  - Recursively calculate which key could precede it based on the distance to other keys
  - Yields a set of *likely password candidates* best matching hand motions
- **Ranking Password Candidate**
  - Only using distance can be ambiguous since there are multiple likely candidates
  - Leverage directional info in trajectory analysis of similarity to other candidates to rank
  - The top-ranked candidates (e.g., top 3) are used as guesses

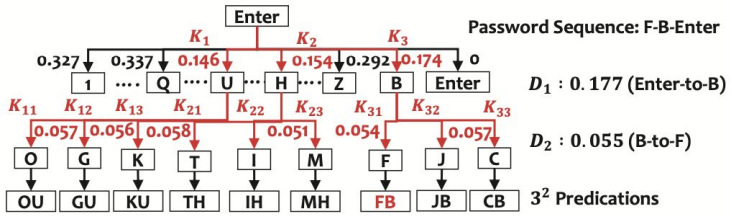


Fig. 9. Tree-based Backward Typing Trajectory Estimation for Password Recovery.

Inferring Sentences

- **Goal**: Reconstruct and label natural language keystrokes
- **Cluster Keystrokes:**
  - Use DBSCAN (density-based clustering) to group 2D keystroke points based on spatial proximity. Use minimum 2 instances/cluster & distance threshold 0.03 units
- **Align Keyboards:**
  - Apply Least Squares Estimation (LSE) to align victim's keyboard with virtual one
  - Randomly select n keys (matching DBSCAN clusters) and solve transformation matrix
- **Label Keystrokes and Correct Errors:**
  - Use K-Nearest Neighbors (K=1) to classify each keystroke by mapping to reconstructed keyboard
    - Example: if a victim's keystroke is closest to the "T" centroid on the attacker's layout, that keystroke is labeled as "T."
  - Refine natural language output grammar correct (e.g. Google Docs Spell Check)

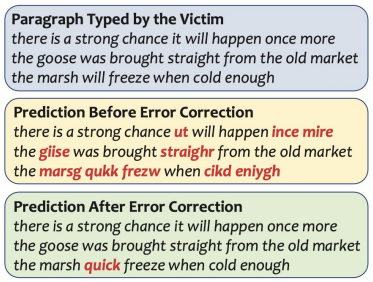


Fig. 16. Examples of Recovered Paragraph.

Experimentation

- Experiment Setup
- Metrics
- Results

Experiment Setup

Study settings:

- 2 systems: HTC Vive Pro (outside-in system) and Oculus Quest (inside-out system)
- 7 participants for each system (14 total)
- 38 keys (26 alpha, 10 num, space key, enter key)
- Used both drum-based and laser-based systems

Simulation Setup:

- Randomly generate passwords of {4, 6, 8} characters
- Randomly selected 10 sentences from Harvard sentences dataset [5]

[5] EH Rothauser. Ieee recommended practice for speech quality measurements. IEEE Trans. on Audio and Electroacoustics, 17:225–246, 1969.

Metrics

- For single keystroke/**character classification**: Accuracy, Precision, Recall
- For Password Inference Metrics:
  - **Top-k Success Rate**: fraction of trials that victim's password was successfully recovered among the top-k candidate predictions
- For Paragraph Inference Metrics:
  - **Word Recognition Rate** (WRR) = correct words / total words

Results (with brevity)

- **Keystroke Recognition**:
  - The attack can recognize over 89.7% of keystrokes correctly overall
- **Password Recovery**:
  - For random passwords (length 4–8 characters), the attacker's top-3 guesses contain the correct password about 84.9% of the time
  - Even with just a single guess (top-1), success rates were significant (around 50–75% depending on length)
- **Sentence Recovery**:
  - For natural language input, the attacker achieves an average 87.1% Word Recognition Rate
  - Most words in a sentence are reconstructed correctly, often with minor spelling errors. After language-model corrections, many sentences are almost fully readable.
- **Drum vs. Laser**:
  - Drum-based typing had slightly higher accuracy than laser-based in experiments
  - Per-key recognition averaged ~91.7% on drum vs ~81.1% on laser in one test
  - Drum keystrokes produce bigger motion signals, making them a bit easier to classify, but both methods were vulnerable.

Discussion

- Limitations
- Privacy Implications
- Countermeasures

Limitations

1. Have to assume QWERTY layout
2. Environmental variability (controller noise in real environments)
3. Not many participants (k=14), conclusive power is low

Privacy Implications

- **New Side-Channel Threat**:
  - immersive VR isn't purely visual – it's leaking data
  - An attacker doesn't need to "see" your VR screen; motion sensors suffice to know what you type.
- **User Trust & Awareness**:
  - VR users today likely assume typing in a virtual environment is secure (no one looking over your shoulder).
  - But a background app or website could be "shoulder surfing" via sensors. This is an unseen risk – literally invisible to the user.
- **Urgency for Solutions**:
  - These findings put pressure on VR platform providers (like Meta/Oculus, HTC, Valve) to re-evaluate sensor policies.

Countermeasures - Protecting VR Users

- **Restrict Sensor Access:**
  - Implement permissions for motion/position sensors similar to camera or mic
- **UI Indicators**:
  - Hardware or software indicators (lights or on-screen icons) on VR devices to notify the user when sensors are being recorded
  - Similar to webcam LED on computers
- **Anomaly Detection**:
  - VR anti-malware tools could try to detect suspicious sensor logging
  - If an app that shouldn't need your motion data is constantly polling it, that might be flagged

Questions?

Conduct Independent Inquiry in Computer Science

Chapter 6

Conduct Independent Inquiry in Computer Science

Paper Title: SpecReFlow: An Algorithm for Specular Reflection Restoration Using Flow-Guided Video Completion

Publication Details: Journal of Medical Imaging, Volume 11, Issue 2, Pages 024012-024012, Society of Photo-Optical Instrumentation Engineers (SPIE), March 1, 2024.

Problem Statement

Specular reflections (SRs) are highlight artifacts commonly found in endoscopy videos that can severely disrupt a surgeon's observation and judgment. Despite numerous attempts to restore SR, existing methods are inefficient and time consuming and can lead to false clinical interpretations. Therefore, we propose the first complete deep-learning solution, SpecReFlow, to detect and restore SR regions from endoscopy video with spatial and temporal coherence.

Research Methodology

SpecReFlow consists of three stages: (1) an image preprocessing stage to enhance contrast, (2) a detection stage to indicate where the SR region is present, and (3) a restoration stage in which we replace SR pixels with an accurate underlying tissue structure. Our restoration approach uses optical flow to seamlessly propagate color and structure from other frames of the endoscopy video.

Results and Findings

Comprehensive quantitative and qualitative tests for each stage reveal that our SpecReFlow solution performs better than previous detection and restoration methods. Our detection stage achieves a Dice score of 82.8% and a sensitivity of 94.6%, and our restoration stage successfully incorporates temporal information with spatial information for more accurate restorations than existing techniques.

Conclusions

SpecReFlow is a first-of-its-kind solution that combines temporal and spatial information for effective detection and restoration of SR regions, surpassing previous methods relying on single-frame spatial information. Future work will look to optimizing SpecReFlow for real-time applications. SpecReFlow is a software-only solution for restoring image content lost due to SR, making it readily deployable in existing clinical settings to improve endoscopy video quality for accurate diagnosis and treatment.

GitHub Repository

GitHub Repository: `https://github.com/Nano1337/SpecReFlow`

JMI Journal of Medical Imaging

RESEARCH PAPER

SpecReFlow: an algorithm for specular reflection restoration using flow-guided video completion

Haoli Yin,[a] **Rachel Eimen**,[b,c] **Daniel Moyer**,[a] **and Audrey K. Bowden**[b,c,d,*]

[a]Vanderbilt University, Department of Computer Science, Nashville, Tennessee, United States
[b]Vanderbilt University, Vanderbilt Biophotonics Center, Nashville, Tennessee, United States
[c]Vanderbilt University, Department of Biomedical Engineering, Nashville, Tennessee, United States
[d]Vanderbilt University, Department of Electrical and Computer Engineering, Nashville, Tennessee, United States

ABSTRACT. **Purpose:** Specular reflections (SRs) are highlight artifacts commonly found in endoscopy videos that can severely disrupt a surgeon's observation and judgment. Despite numerous attempts to restore SR, existing methods are inefficient and time consuming and can lead to false clinical interpretations. Therefore, we propose the first complete deep-learning solution, SpecReFlow, to detect and restore SR regions from endoscopy video with spatial and temporal coherence.

Approach: SpecReFlow consists of three stages: (1) an image preprocessing stage to enhance contrast, (2) a detection stage to indicate where the SR region is present, and (3) a restoration stage in which we replace SR pixels with an accurate underlying tissue structure. Our restoration approach uses optical flow to seamlessly propagate color and structure from other frames of the endoscopy video.

Results: Comprehensive quantitative and qualitative tests for each stage reveal that our SpecReFlow solution performs better than previous detection and restoration methods. Our detection stage achieves a Dice score of 82.8% and a sensitivity of 94.6%, and our restoration stage successfully incorporates temporal information with spatial information for more accurate restorations than existing techniques.

Conclusions: SpecReFlow is a first-of-its-kind solution that combines temporal and spatial information for effective detection and restoration of SR regions, surpassing previous methods relying on single-frame spatial information. Future work will look to optimizing SpecReFlow for real-time applications. SpecReFlow is a software-only solution for restoring image content lost due to SR, making it readily deployable in existing clinical settings to improve endoscopy video quality for accurate diagnosis and treatment.

Keywords: specular reflection; image artifacts; image restoration; optical flow; multiview restoration; endoscopy

Paper 23324GR received Oct. 29, 2023; revised Feb. 4, 2024; accepted Apr. 3, 2024; published Apr. 24, 2024.

1 Introduction

Specular reflection (SR) is an optical phenomenon wherein light is strongly reflected from a surface in a unified direction based on Snell's law. On a camera, SR often leads to saturation of the pixels (causing them to appear white) and is well known for obscuring the underlying information in the image—rendering it a functional blind spot. In clinical settings, that blind

*Address all correspondence to Audrey K. Bowden, a.bowden@vanderbilt.edu

spot can critically impair all three stages of surgical operations: diagnosis, planning, and post-operative observation. Moreover, SR artifacts impact the interpretation of clinical endoscopy data and can hinder the utility of augmented/virtual reality (AR/VR) tools for surgical navigation and the creation of accurate three-dimensional (3D) organ models that have been found to improve surgical outcomes.[1–3] Thus, there is a significant positive impact associated with improving video quality by restoring SR regions. Several studies have directly linked enhanced image restoration to improved performance in downstream medical tasks. For instance, Ali et al.[4] found that the restoration of SR and saturated regions significantly improved image feature matching, which could be used for 3D organ reconstruction for further analysis. This and other studies[5,6] collectively suggest that advancements in image restoration quality not only are beneficial in their own right but also play a crucial role in enhancing the accuracy and effectiveness of various medical applications.

We propose a novel method for accurate, multiview restoration of SR regions of endoscopy data based on the propagation of optical flow[7] from adjacent frames of a video sequence. Optical flow is defined as the pattern of apparent movement of objects in a visual scene caused by relative motion between an observer and a scene. The optical flow concept has been commonly applied to tasks such as object detection and tracking, movement detection, robot navigation, and visual odometry;[8] however, we are the first to apply it to SR restoration of medical endoscopy data. The optical flow method that we employ utilizes pixel movement data, enabling a transformer deep learning model to effectively integrate pixel information into the SR region. Pretrained on various video data, this model is proficient at adapting to different lighting conditions in SR regions. Its adaptability is crucial for maintaining spatiotemporal consistency in SR regions informed by optical flow. The flow-guided transformer (FGT) algorithm[9] that we use implicitly accounts for depth information via optical flow, which aids in lighting adjustments during the restoration process. As a result, the restored pixels in the SR regions are not merely copied from another frame; their colors are also appropriately transformed. Intuitively, this method preserves the texture of the restored image content while allowing color to adapt seamlessly to its new background. We evaluate the effectiveness of our method in the context of other state-of-the-art restoration strategies and propose new assessment metrics appropriate for this task, termed mPSNR and mSSIM, to evaluate free-form SR and masked regions.

The ability to remove and restore SR regions in an accurate and color-consistent manner is useful for improving the potential for both accurate analysis and diagnosis of endoscopy images. Our algorithm represents a major contribution to the field by being the first full system to combine both temporal and spatial information from endoscopy videos for effective detection and restoration of SR regions, surpassing previous methods that rely solely on spatial information. Importantly, unlike other work for optical flow applied to SR restoration,[10] our implementation is a software-only solution that does not require additional equipment beyond the clinical standard of care for endoscopy, making it readily deployable for clinical use.

2 Related Works

One strategy to eliminate SR artifacts is to replace them with new information that infers the original content. Efforts to restore the appearance of the clinical data underlying SR artifacts can be broadly categorized into three groups: single-frame conventional methods, single-frame learning methods, and multi-frame matching methods.

Single-frame conventional methods include traditional image processing methods such as thresholding or filtering. One limitation of using traditional techniques for detecting SR is that it may result in some SR regions—especially smaller regions—going undetected, which prevents restoration.[11–13] In addition, conventional methods for SR restorations largely depend on empirical parameter setting, which can lead to slow runtimes and inaccurate feature appearance such as blurriness, structural inaccuracy, and poor color blending with the surrounding non-SR region.[6,11–17]

Many single-frame methods use deep learning approaches, such as convolutional neural networks (CNNs)[18,19] or generative adversarial networks,[4,5,20–22] to directly inpaint the region for restoration. Although learned methods may lead to fast and high-quality restorations, the

generated regions may not be truly representative of the ground-truth tissue, as they only inpaint the missing SR region using local spatial information obtained from a single perspective.

Multi-frame matching methods involve the use of multiple frames (e.g., video datasets). SR regions can be restored by directly replacing SR pixels that are present in one frame with the corresponding pixels of the same region that are not covered by SR in another frame. This is possible because SR can be thought of as a dynamic object with no apparent pattern of movement. SR regions vary among frames due to the varying angle of incidence of the light source, which may be caused by changes in the camera angle relative to the tissue, changes in the smoothness of the tissue, or other variations in the optical density of the intervening medium. These factors affect how light is reflected off the surface, leading to changes in the appearance and location of SRs in each frame. A significant drawback of simple replacement approaches (e.g., homography transform), however, is that the restored SR regions do not have the same brightness of color as the surrounding area due to lighting inconsistencies caused by camera or light source movement from one frame to another. The result of these lighting aberrations is seen as temporal inconsistency and color flickering.[23,24] Recent efforts, such as those by Daher et al.,[25] have tried to harness temporal data using traditional nonlearning methods to detect SR regions. However, their limited accuracy in SR detection causes temporal-aware generative adversarial models to unintentionally include undetected SR from other regions, resulting in noisy reconstructions.

Finally, a general limitation of all previous specular restoration studies performed to date is that the standard metrics used to evaluate the effectiveness of the restoration techniques [e.g., peak signal-to-noise ratio (PSNR) and structural similarity metric (SSIM)] have settings (e.g., kernel size) that are often not fully explained or standardized, which can make it difficult to accurately compare the results of different methods. Furthermore, as we will show, these standard metrics are not sensitive enough to capture important details relevant to the scope of the restoration task, or they may not be applied to the most relevant regions of the images being restored. As a result, it is difficult to accurately assess the true improvement in image quality achieved by existing methods.

3 Methodology

3.1 Overview

SpecReFlow is an SR detection and restoration algorithm with three stages: (1) image preprocessing, (2) SR detection, and (3) SR restoration. Figure 1 presents an overview of the framework. The preprocessing stage increases the contrast of SR in the original image to simplify the detection task. The SR detection stage combines a modified U-net model that detects small SR regions with a thresholding method to detect the larger SR regions on which deep learning methods commonly fail. The SR restoration stage leverages recent progress in flow-based video completion using optical flow and vision transformers to distinguish the dynamic foreground image (SR) from the dynamic background image (ground truth tissue)[26] and inpaint the detected SR regions. All work was implemented in Python 3.10 using PyTorch version 1.13.1+cu117 with Ubuntu 22.04 on an NVIDIA GeForce RTX 3090 GPU with 24 GB VRAM, 32 GB RAM, and an AMD Ryzen 9 5950X with 16 logical processors.

3.2 Dataset

For model training and pipeline development, we used the CVC-EndoSceneStill from Vázquez et al.,[27] a compilation of 912 frames obtained from 44 endoscopy videos from 36 patients. We chose this dataset because it is one of the only known datasets with manually annotated SR region masks.

To increase the amount of training data for the deep learning detection model, we employed data augmentation techniques: each image was cropped into four equal-sized, smaller images of 288×384 pixels and subjected to horizontal/vertical flips at random. The order of the entire dataset was then randomized to ensure sample independence. The train/validation/test split was 80/10/10 based on common splits in other papers.[4,19]

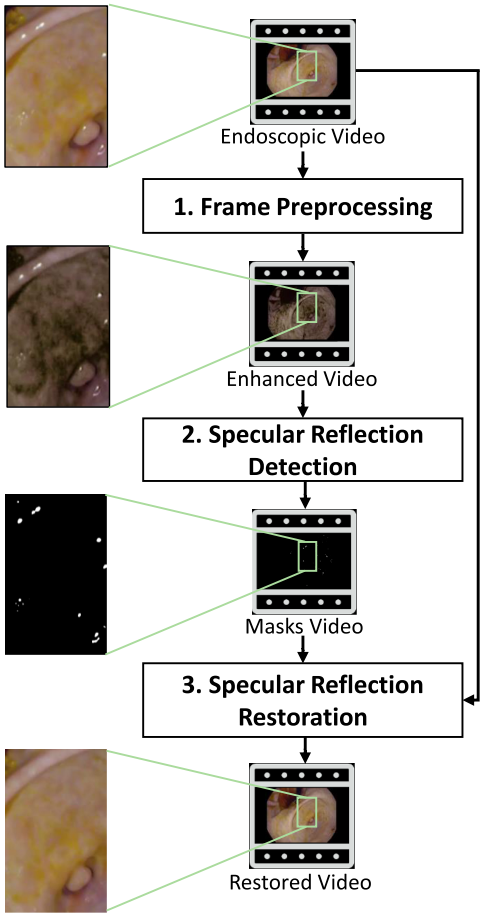


Fig. 1 Simplified overview of the SpecReFlow algorithm. A white-light endoscopy video is used as input; it is then preprocessed to create an enhanced video for input to the detection stage for better results. After the detection stage, SR binary masks are generated for each video frame, and both the original endoscopic video and the mask video are used during the restoration stage to create the restored video.

3.3 Algorithm Design

3.3.1 *Stage 1: image preprocessing*

To improve the contrast between SR regions and surrounding regions, we used an image enhancement algorithm proposed by Saint-Pierre et al.[11] that modifies the pixel intensity histogram distribution by multiplying the red, green, and blue (RGB) planes with the 1-S plane of the hue, saturation, and value (HSV) color model, as shown in Fig. 2. When observing the distribution of pixel intensities, we found that this step leads to an increased gap between the SR values and the rest of the histogram, which makes it easier to separate SR and non-SR regions in the detection stage.

3.3.2 *Stage 2: SR detection*

Our SR detection algorithm combines deep learning with a traditional thresholding algorithm. The deep learning detection model, which we call "Light U-net," is built upon the popular and proven U-net architecture for biomedical segmentation tasks.[28] The original U-net model[28] included 64 convolutional layers and went five layers deep (i.e., it had four downsampling operations on the encoder side). We use only four layers (with three downsampling operations on the encoder side) and eight convolutional layers, yielding a total of 121,641 trainable parameters. The resulting model, shown in Fig. 3, is lighter and more efficient for SR detection. We used fewer layers because a smaller model allowed for less complexity to detect relatively simple objects such as SR. In

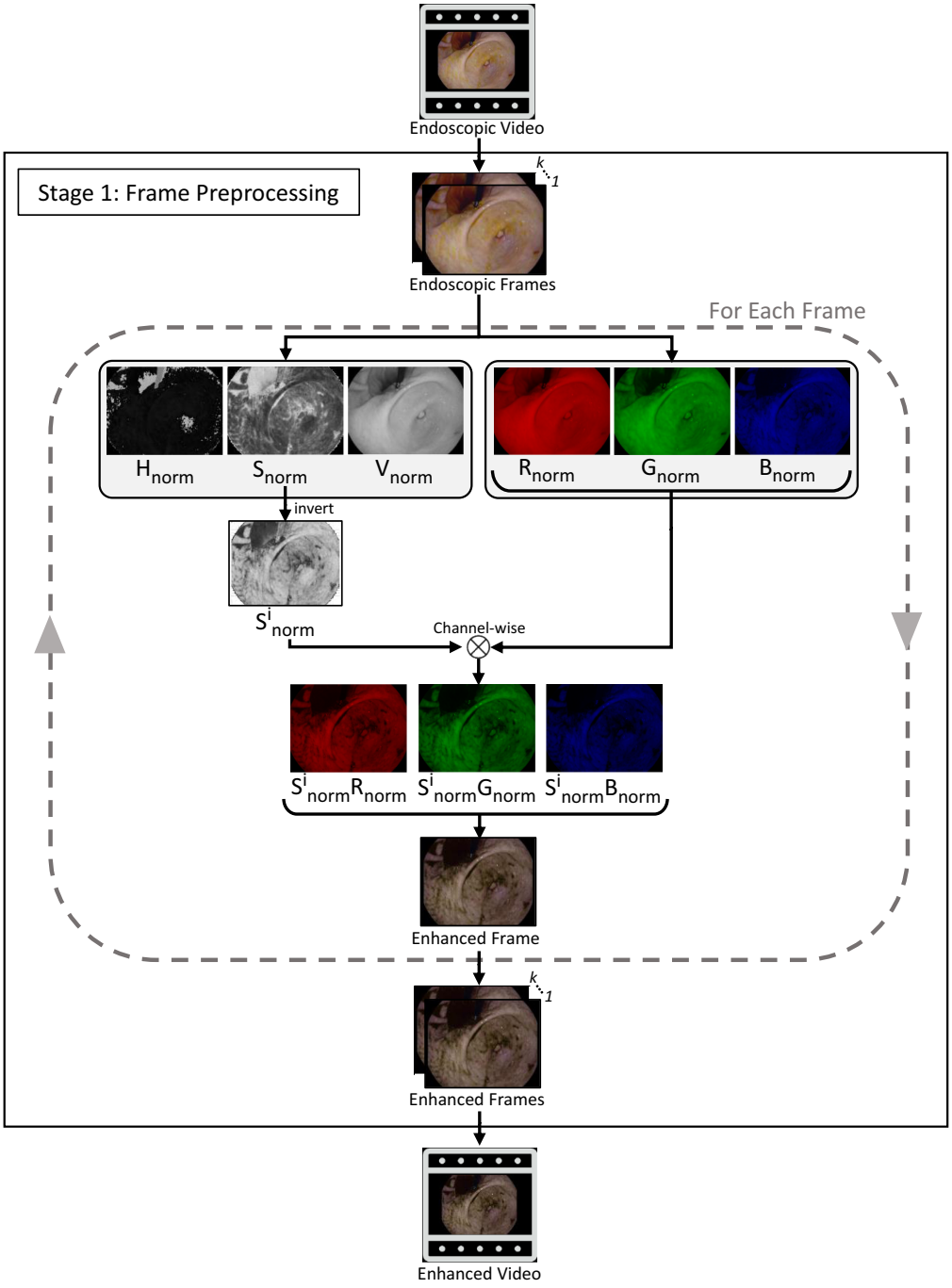


Fig. 2 Image preprocessing stage visualization. A white-light endoscopy video is used as input, and each video frame is processed to create an enhanced video. For each frame, HSV and RGB images are created. Then, the saturation (S) channel is inverted through a 1-S operation and multiplied with each channel of the RGB image to ultimately create the enhanced frame.

addition, the reduced complexity of the model allows for real-time SR detection, which can be combined with an appropriate SR restoration algorithm to fix SRs in real time.

To train our Light U-net model, we used a combination of Dice loss and binary cross entropy loss, a U-net four levels deep (i.e., three pooling operations), an Adam optimizer (initial learning rate of 0.01, $\beta_1 = 0.9$, $\beta_1 = 0.999$), and a learning rate scheduler of ReduceLrOnPlateau (patience = 10, factor = 0.1). These settings were then used to train the model on 120 epochs with a batch size of eight. We found that training the model on more epochs led to overfitting and a decreased performance on the validation and test sets.

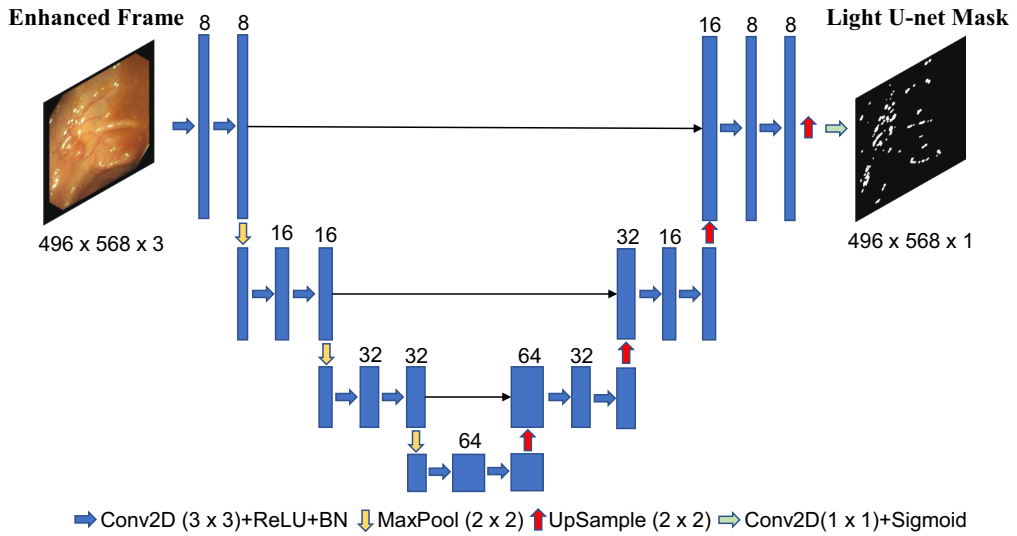


Fig. 3 Architecture of our Light U-net model. We use four layers and eight initial convolutional filters to create an efficient and optimized architecture as determined by a neural architecture search.

On its own, the Light U-net model learns to detect smaller SR regions significantly better than larger regions, which can be explained in two ways. First, the training dataset was relatively unbalanced as there were significantly more samples of smaller, scattered SRs; thus, the model often misidentified larger SR regions as non-SR objects. Second, the fixed receptive field of the convolutional kernels caused the model to learn more of the smaller SR regions. Because the training frames were of a lower resolution than true endoscopy frames, the model was primed to detect smaller SR regions in higher-resolution frames.

Hence, to enable the detection of large SR regions, we utilized a simple thresholding method. It is well known that SR regions have higher pixel intensity values than surrounding regions based on the thresholding algorithm proposed by Yu et al.[29] and the analysis of a histogram of pixel intensities within a frame with SR. We chose the threshold $I(x, y) > 194$, where $I(x, y)$ is the pixel intensity in the value channel of the HSV image and (x, y) is the location of the pixel, by optimizing the detection metrics [i.e., Dice score and intersection over union (IoU)] through a greedy search on the same training set used for U-net training. Although this basic method worked well on large SR regions, our testing (as shown in Sec. 4.1) revealed that thresholding alone is insufficient for detecting all SR as it misses relatively low-intensity SR commonly found in smaller SR regions. Inspired by Xu et al.,[30] to harness the strength of both algorithms, our final strategy was a combination of deep learning mask prediction (good at small, scattered SR) and thresholding mask prediction (good at large SR) through an AND operation of the masks generated by each prediction method. In summary, the final, optimal SR detection stage algorithm that we adopted is the Light U-net + thresholding algorithm, with an additional dilation step that we describe and justify in Sec. 4.1.

3.3.3 *Stage 3: SR restoration*

In this work, we chose to use an existing flow-based algorithm for the SR restoration stage. We tested three state-of-the-art algorithms for flow-based video completion to determine which performs best for our task: (1) the Flow edge-Guided Video Completion (FGVC) algorithm from 2020,[26] modified by us to use the state-of-the-art single-frame inpainting model, LaMa,[31] instead of DeepFillv1 to test more recent methodologies; (2) the End-to-End Flow-Guided Video Inpainting (E2FGVI) algorithm,[32] which comprises three main modules—flow completion, feature propagation, and content hallucination—all tied together in a streamlined manner to allow for fast, efficient, and accurate restoration of SR regions; (3) the FGT,[9] which improves upon FGVC by replacing the edge completion part of flow estimation with an edge loss function in a novel flow estimate network, mitigating additional computational complexity. After color propagation along

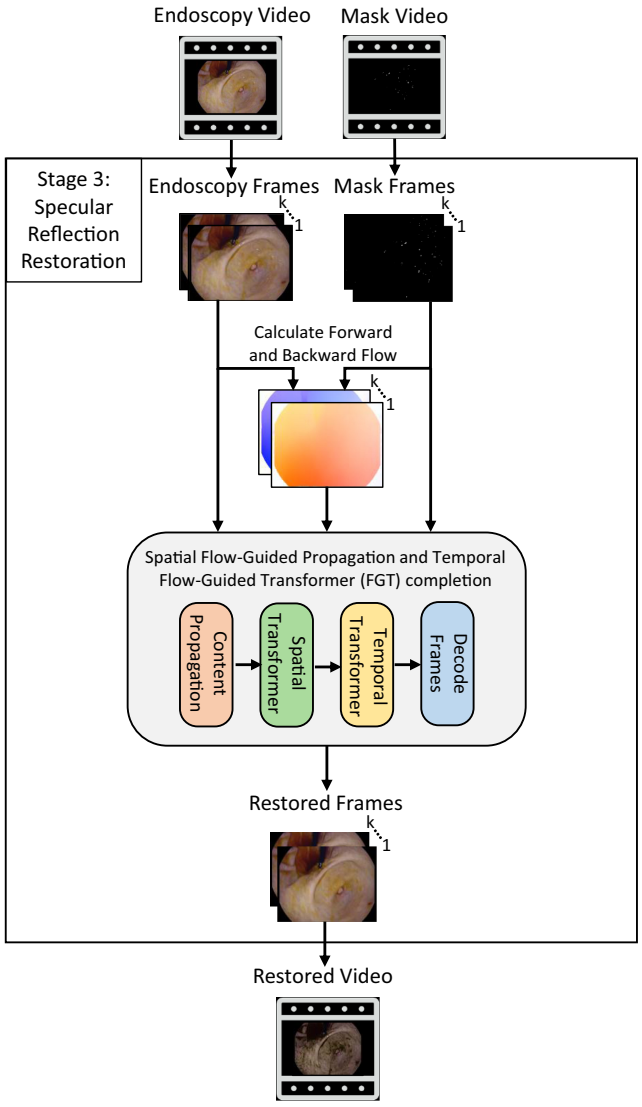


Fig. 4 SR restoration stage visualization. The original endoscopy video and SR binary masks generated from the SR detection stage are used as inputs. Once frames are sampled from the video, forward and backward flows are calculated for adjacent frames, and all three datasets are used as inputs to the flow propagation and the FGT, which performs frame restoration. The restored frames are then combined to synthesize the restored video.

the estimated flow, FGT completes the missing pixels of multiple frames at once using a spatial and temporal transformer model, incorporating more information to complete missing pixels more efficiently than the single-frame and iterative inpainting version of FGVC. To compare these flow-based models against a baseline, we also analyzed two single-frame inpainting methods representing the current standard in SR restoration: LaMa[31] and DeepFillv2.[33]

Based on the experimentation and discussion in the following sections, we determined that FGT accomplished the optical flow-based restoration task better than the other two. Hence, SpecReFlow uses FGT as its third stage; an overview of the restoration algorithm is given in Fig. 4.

3.4 Evaluation Strategy and Metrics

3.4.1 *SR detection*

To test the effectiveness of the SR detection stage, we reserved ~10% of the CVC-EndoSceneStill dataset detailed in Sec. 3.2 (separate from the other 90% utilized for training), which was 184 images. The following quantitative metrics were evaluated: sensitivity, Jaccard/

IoU score, and Dice score. Here, we define sensitivity to be the number of true-positive SR pixels detected divided by the total number of SR pixels in an image. Sensitivity, rather than specificity, is relevant because our end goal is not detection but restoration; hence, we recognize that the detected SR mask does not have to be exact to the pixel (i.e., high specificity), but rather needs to cover as many true-positive regions as possible. Note that using sensitivity alone will give a false impression of the performance as over-detection (of non-SR pixels) can also lead to high sensitivity; thus, we first optimized the detection algorithm for high IoU and Dice scores to limit the detected regions to locations of true SR pixels. Our qualitative analysis sought to validate whether our method masks all SR regions and any surrounding regions affected by increased saturation. In both cases, we compared our results to baselines of existing, open-source SR region-detection methods: a conventional histogram thresholding method by Tchoulack et al.,[34] a more recent non-deep learning adaptive robust principal component analysis (RPCA) method,[16] and a denoising convolutional neural network (DnCNN) modified to detect SR regions by Zhang et al.[35] We chose these methods because they represent the broad range of approaches taken to solve the detection problem in the past and are the current state-of-the-art methods for open-source implementations.

3.4.2 *SR restoration*

To identify which SR restoration method performs best for endoscopy data, we generated a synthetic dataset. We overlaid the GLENDA dataset (360×640-pixel resolution) from Leibetseder et al.,[36] which contains over 13,000 unannotated (no SR labels/masks), nonpathological images from 20+ laparoscopy endometriosis video surgeries, with synthetic SR masks pulled from the CVC-EndoSceneStill dataset and resized to the appropriate dimensions. The latter dataset contained 200 masks, from which 50 independent random masks were chosen to be applied to frames as described in the following experimental procedure. Note that SR masks were not applied to regions of GLENDA frames that had pre-existing SR.

We chose to test three flow-based algorithms, FGVC, FGT, and E2FGVI, along with two single-frame inpainting methods, DeepFillv2 and LaMa, as these were the state-of-the-art algorithms for their respective categories. Each experimental trial comprised an independent set of 100 sequential frames, defined as a sequence of video frames that do not overlap temporally with another set of 100 sequential frames. To provide the flow-based algorithms with sufficient temporal data (not relevant for the single-frame inpainting algorithms as they only require spatial data), the first 50 frames were considered "warm-up" data, and the following 50 frames were used for "assessment" for both classes of algorithms. No masks were applied to the "warm-up" dataset. To account for the potential spatial overlap (i.e., footage circles back to the same region) between temporally independent sets of frames, the process of selecting random masks also allowed for different regions to be restored despite the overlap, preserving the spatial independence of each frame sequence. In total, 13 independent restoration trials were conducted for each algorithm.

Our metrics of assessment were modified versions of the established PSNR and the SSIM, which we term mPSNR and mSSIM, respectively. The need for modifications stems from the specifics of our use case. In general, PSNR is used as a quality measurement to indicate how well the restored image compares to the fidelity of the original image. When applied to evaluate the effectiveness of various restoration algorithms, we observed that the difference in PSNR outcomes [Fig. 5(a)] is statistically insignificant ($p > 0.05$ in a two-sample t-test); we believe that this result is due to the introduction of excess noise from non-SR regions included in the PSNR calculation. Similarly, SSIM measures perceived changes in structural information. Because SSIM uses a sliding window/kernel across the entire image, the resulting structural similarities are nearly indistinguishable for the various algorithms and are close to ideal across all restoration methods, as shown in Fig. 5(b). We believe that this result is due to the relatively small size of SR regions compared with the image size. Hence, we observed that the original formulas for PSNR and SSIM do not fit our use case well to distinguish restoration algorithms.

We thus modified the metrics as follows. Our proposed mPSNR differs from PSNR in that we zero out the error terms for non-SR pixels and replace the standard mean square error (MSE) term with a modified MSE in which we divide the MSE by the number of SR pixels instead of the total number of pixels in the frame. The overall result is that only SR pixels from the image

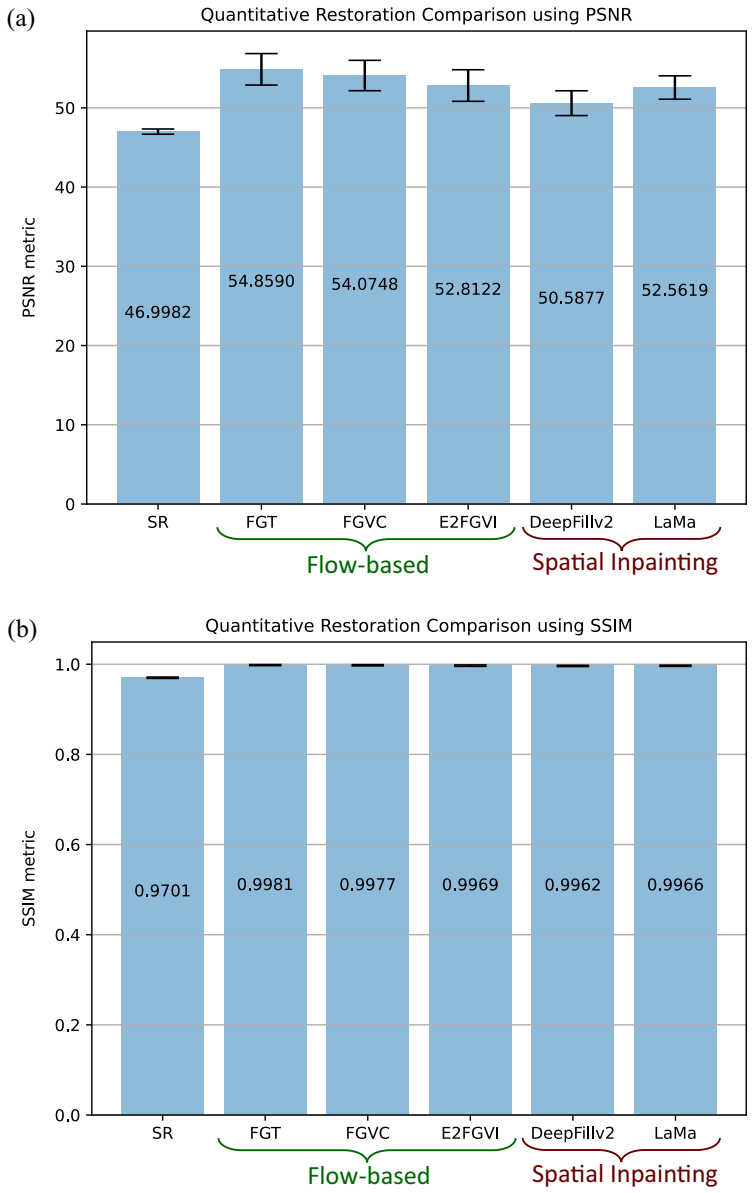


Fig. 5 Original restoration metrics for quantitative restoration analysis. (a) PSNR and (b) SSIM performance of various restoration algorithms show insignificant differences. The following restoration algorithms and baselines are compared: SR unrestored image as the baseline; optical flow-based methods such as FGT, FGVC, and E2FGVI; and single-frame inpainting methods such as DeepFillv2 and LaMa.

contribute to the metric assessment. The formula to describe the mPSNR between the ground truth and a restored image is

$$\mathrm{mPSNR} = 20 * \log\left(\frac{\mathrm{MAX}_I}{\sqrt{\mathrm{mMSE}}}\right), \tag{1}$$

$$\mathrm{mMSE} = \frac{1}{C}\frac{1}{S}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 * w, \tag{2}$$

where C is the number of color channels, S is the number of SR region pixels, I is the original image of size $m \times n$, K is the restored image, i and j are pixel indexes, and w is a binary weight, which is 1 for SR pixels and 0 for non-SR pixels. For Eq. (1), MAX is the maximum possible pixel value of the image, which is 255 for an 8-bit image and 1.0 for a normalized image.

Algorithm 1 mSSIM calculation

Require: Ground Truth Image (origImg), Restored Image (restoredImg) and SR Mask

Ensure: mSSIM Metric

```
Initialize ssims, weights = [ ], [ ]
boundingBoxes = Get bounding boxes for each separate SR region in SR Mask
for bBox in boundingBoxes do
    Initialize boxHeight, boxWidth
    if boxHeight <8 then
        add to each side until minimum height of sliding window dimension
    end if
    if boxWidth <8 then
        add to each side until minimum width of sliding window dimension
    end if
    ssim = Calculate ssim with origImg[bBox], restoredImg[bBox], windowSize=7
    ssims.append(ssim)
    weights.append(sum(mask[bBox]))
end for
totalSpecularReflections = sum(weights)
weights = weights/totalSpecularReflections
return weights × ssims
```

Our mSSIM also provides focus on the SR regions rather than the entire frame as this change better captures structural inconsistencies of the restored regions. Because SSIM requires rectangular regions, we first extracted bounding boxes from the ground truth SR masks. We chose a minimum bounding box size of eight to facilitate the use of sliding windows (of size seven) during the SSIM operation. Once the SSIM of each bounding box was found, the frame-level mSSIM was computed as a weighted average of the number of SR pixels in a particular bounding box divided by the total number of SR pixels in the frame. Note that the possible ranges of values for these metrics are still the same as that of the unmodified versions, i.e. [0, unbounded) for mPSNR and [0, 1] for mSSIM. The algorithm is detailed in Algorithm 1.

4 Results

4.1 SR Detection

The results reported in Table 1 reveal that our initial detection algorithm (SpecReFlow Det(-)) achieved Dice score and IoU values that were more than 1.17× and 1.22× higher, respectively, than the next best established method (modified DnCNN), and it was nearly 4.25× faster. As observed, SpecReFlow detection performs favorably against existing SR detection algorithms of all types in both measures of effectiveness and efficiency.

Although our initial SpecReflow detection algorithm already shows the best sensitivity of all methods, we sought to increase its sensitivity to ensure that all SR regions are detected. The bottom row of Table 1 shows the performance that results when it is updated to include a step that dilates the generated segmentation mask with a 5×5 elliptical kernel. Although this step leads to a noticeable decrease in Dice score and IoU, the behavior is expected as the enlarged masks intentionally overstep the bounds of the ground truth SR regions. Importantly, this step

Table 1 Quantitative comparison of SR detection metrics.

Detection method	Dice ↑	IoU ↑	Sensitivity ↑	Time (ms) ↓
Tchoulack algorithm[34]	0.5012	0.3500	0.5122	1212
Adaptive RPCA[16]	0.6094	0.4590	0.5960	**3.70**
Modified DnCNN[35]	0.7096	0.5853	0.6210	53.17
SpecReFlow Det. (-)	**0.8285**	**0.7158**	0.8158	11.06
SpecFeFlow Det.	0.6517	0.4947	**0.9457**	11.32

Note: bold values highlight the best performing algorithm for each metric.

leads to a notable sensitivity increase. We believe that mask dilation is warranted for our application because it helps capture SR-adjacent pixel regions that have a manipulated appearance due to their proximity to SR, which can affect the restoration quality due to disruptions in the true color and feature propagation. We explore this topic more in depth in Sec. 4.2. Thus, our final implementation of SpecReFlow detection (SpecReFlow Det.) includes dilation. We thus term the original detection algorithm that does not include dilation as SpecReFlow Det.(-).

To assess our results qualitatively, we used a representative subset of the same test dataset from Sec. 3.4.1. The algorithms were evaluated on representative types of SR: (a) a large SR region and SR present in both light and dark regions, (b) medium-sized SR regions scattered across a relatively homogeneous region, and (c) small, speckled SR regions scattered in localized regions. As seen in Fig. 6, the green annotations indicate true positives, and the blue annotations indicate false positives. The red asterisks indicate regions of increased saturation around the central SR region that are not strictly SR but require correction; thus, our algorithm is sensitive to detecting such regions of information loss. SpecReFlow Det. Successfully detects large SR regions [Fig. 6(a)(6)], whereas the modified DnCNN model is unable to [Fig. 6(a)(5)]. In addition, the blue arrow in Fig. 6(a)(2) indicates a location where SR was not labeled in the ground truth image but was still detected with SpecReFlow Det., suggesting that our method is robust to mislabeled annotations.

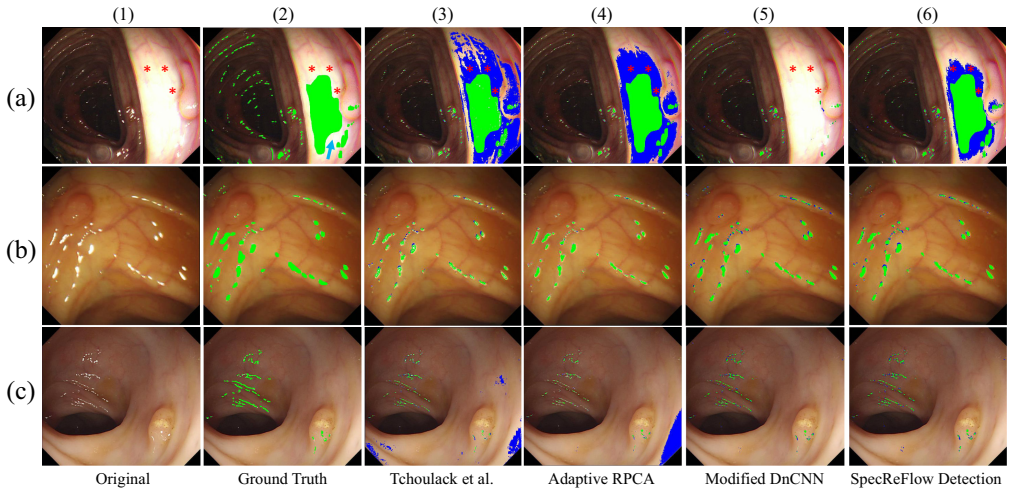


Fig. 6 Qualitative comparison of SR detection performance. Performance of algorithms on representative types of SR: (a) a large SR region and SR present in both light and dark regions, (b) medium-sized SR regions scattered across a relatively homogenous region, and (c) small, speckled SR regions scattered in localized regions. The green annotations indicate true positives, and the blue annotations indicate false positives. The red asterisks indicate regions of increased saturation around the central SR region. The blue arrow in the ground truth image (a)(2) indicates a region of high saturation that our detection method picked up but was not indicated in the ground truth.

Considering Figs. 6(c)(3) and 6(c)(4), we see that SpecReFlow Detection [Fig. 6(c)(6)] also led to fewer false positives than the conventional methods (adaptive RPCA and Tchoulack's algorithm). Furthermore, our method generated a smoother, more continuous mask that captured the entire SR region rather than just the most intense SR regions (as seen in Tchoulack's algorithm), which can lead to more accurate restorations without random spots of SR. Notably, we observed that the ground truth was manually labeled with a thick marker size, which makes the labeled SR regions appear larger than they are. Thus, some of the false negative regions in Figs. 6(b) and 6(c) (unlabeled) are not actually wrong but appear so due to the overly extensive marker labels in the ground truth. Qualitatively, the "thickness" of the algorithm-marked labels in Fig. 6(c) matches what is expected from the original image.

As mentioned, the final SpecReFlow detection stage includes a marginal dilation step to increase sensitivity and capture adjacent regions that are not strictly SR but that require inpainting. Hence, in some locations, our detection mask covers a larger area than the "ground truth" SR. However, some of these additional regions are warranted: SpecReFlow Det. can accurately identify highly saturated regions (red asterisks) that are adjacent to the main SR region and contribute to information loss, as demonstrated by their coverage indicated with blue regions in Fig. 6(a)(6). Furthermore, our method does not contribute false positives that still contain relevant information, unlike the blue areas in Figs. 6(a)(3) and 6(a)(4) that mask regions with useful spatial information. This exemplifies the effectiveness of our method in identifying only highly saturated regions of interest with information loss.

We also performed an ablation study to assess the effectiveness of each of the SR detection pipeline components: preprocessing, deep learning (Light U-net), and thresholding. The input to each run comprised a raw or preprocessed RGB image, and the output was an SR mask. We analyzed the time and improvement in quantitative detection metrics as each component was systematically added or removed from the pipeline. The results can be observed in Table 2. Notably, the impact of preprocessing on quality performance metrics (Dice and IoU) could not be assessed as preprocessing alone is not a form of detection and yields no detected mask output.

Using only raw images as input, simple thresholding takes less than a millisecond (0.51 ms) of runtime on average. We found that using the Light U-net model alone takes significantly longer—9.39 ms of runtime on average—but improves the Dice score and IoU metrics by nearly 2.03× compared with only using thresholding. Thresholding, however, yields better sensitivity than only using Light U-net (0.8367 versus 0.7614). Furthermore, when comparing the results of using Light U-net + thresholding versus using only thresholding, we surprisingly find the Dice and IoU metrics to be similarly low, which can be explained by the high amount of detection error contributed by the simple thresholding component.

Image preprocessing adds 1.20 ms of runtime on average but makes significant improvements across all detection metrics: Light U-net metrics improved by 1.13× on average, thresholding metrics improved by 1.44× on average, and Light U-net + thresholding metrics improved

Table 2 Quantitative ablation study of SR detection performance.

Detection component(s)	Dice ↑	IoU ↑	Sensitivity ↑	Time (ms) ↓
Raw image				
Light U-net	0.6892	0.5560	0.7614	9.39
Thresholding	0.2945	0.2087	0.8367	**0.51**
Light U-net + thresholding	0.3020	0.2117	**0.9462**	9.83
Preprocessed image				
Light U-net	0.7881	0.6837	0.7636	10.76
Thresholding	0.5540	0.4047	0.4247	1.64
Light U-net + thresholding	**0.8285**	**0.7158**	0.8158	10.92

Note: bold values highlight the best performing algorithm for each metric.

by 2.37× on average. When comparing the results of using preprocessed images with the Light U-net + thresholding method versus using only thresholding, we find a significant 1.63× improvement in Dice score and IoU metrics—a major difference from when only raw images were used as input.

Thresholding on preprocessed images is significantly worse than Light U-net alone in terms of sensitivity (0.4247 versus 0.7636). Although thresholding does seem to have superior sensitivity to Light U-net when using raw images, it is important to note the significantly lower Dice and IoU metrics produced by thresholding for both raw and preprocessed images. This can be explained by the fact that, although thresholding detects more of the "true" SR region, it does so by also detecting many more non-SR regions (low specificity), which will cause the restoration to be more difficult.

Although mask dilation is the last step of the SpecReFlow detection stage, we chose not to include it in our ablation study for the following reasons. (1) We wanted to measure detection performance using Dice and IoU metrics before optimizing for sensitivity, so the predicted SR mask is as close to the ground truth as possible. (2) We wanted the restored SR region to seamlessly integrate with the context of its surroundings, which makes the mask dilation step more relevant for the SpecReFlow restoration stage rather than for the detection stage on which the ablation study focuses.

Overall, the results of the ablation study of our detection pipeline components show that each component adds minimal runtime to SR detection while contributing significant improvements in all detection metrics.

Figure 7 corroborates the results of our ablation study. Figure 7(a)(1) shows that the deep learning model works poorly in detecting larger SR regions (indicated by the gold arrow), but it performs well in covering all of the relatively smaller SR regions. Switching to using preprocessed images, Fig. 7(a)(2) shows how the model can recognize a hotspot in the large SR region, indicating an improvement in detection.

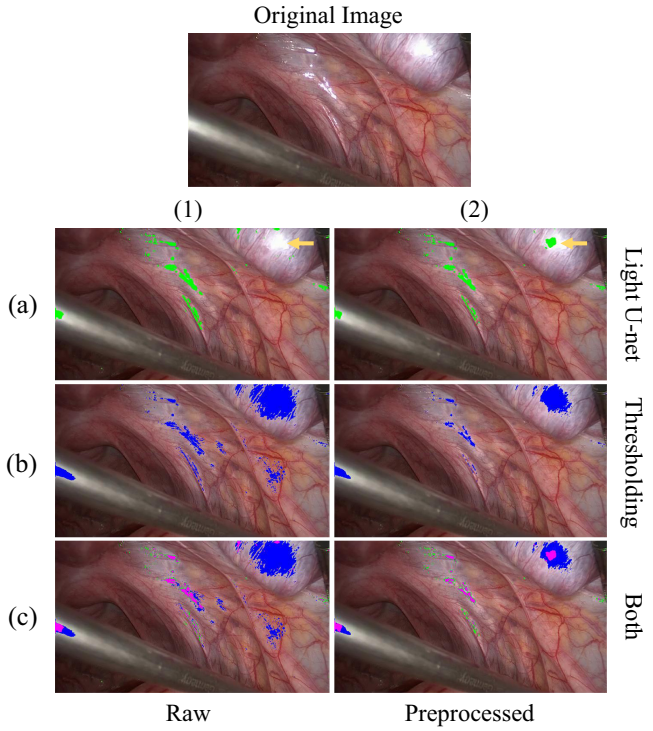


Fig. 7 Qualitative ablation study of our detection pipeline components. To visually explain the contributions of each component of our detection pipeline, masks generated from combinations of components are shown. The green mask indicates predictions from only using Light U-net, the blue mask indicates predictions from only using thresholding, and the magenta mask indicates intersections between the green and blue masks. As expected, thresholding detects large SR regions well, and the Light U-net detects smaller regions well.

In Fig. 7(b)(1), we see that thresholding alone of raw images includes significant amounts of unsaturated pixels, which makes the restoration task more difficult. By applying preprocessing [Fig. 7(b)(2)], thresholding works much better to only detect saturated regions with information loss; however, the smaller SR regions are weakly detected. By combining preprocessing, Light U-net, and thresholding [Fig. 7(c)(2)], it can be seen that the Light U-net detects the smaller SR regions where thresholding fails (seen by the green mask enveloping the magenta mask), and thresholding detects the larger SR regions where the Light U-net fails (seen by the blue mask

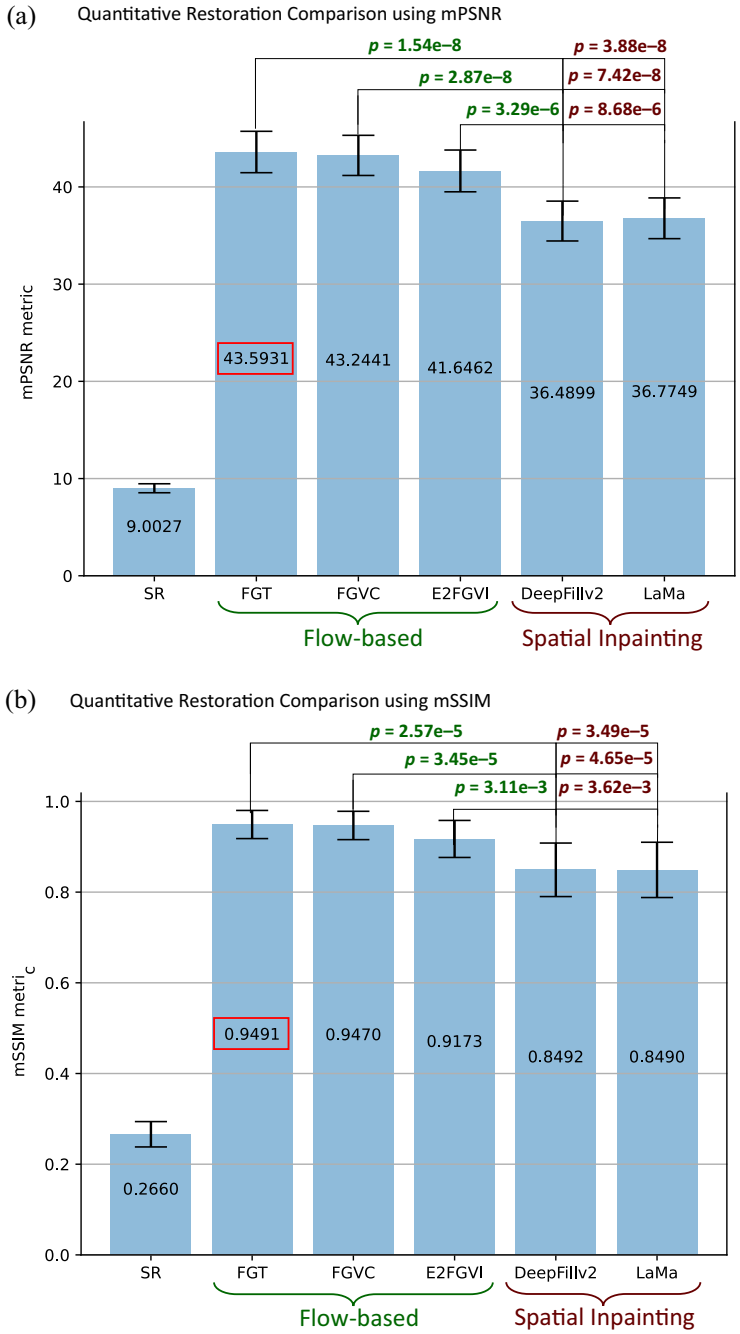


Fig. 8 Visual quantitative comparison of SR restoration methods. Restoration (a) mPSNR and (b) mSSIM are measured for 13 independent SR restoration trials. The following restoration algorithms and baselines are compared: an SR unrestored image as the baseline, optical flow-based methods such as FGT, FGVC, and E2FGVI, and single-frame inpainting methods such as DeepFillv2 and LaMa. The *p*-values associated with two-tailed, two-sample *T*-tests of means comparing each flow-based algorithm with the inpainting options are displayed above the bars.

enveloping the magenta mask), as the magenta mask represents the intersection between the blue and green mask. Thus, both small and large SR regions are fully detected, which is reflected well in the quantitative detection metrics.

4.2 SR Restoration

Figure 8 compares the mPSNR and mSSIM performance of several restoration algorithms. In general, the flow-based algorithms (FGT, FGVC, and E2FGVI) performed better than single-frame inpainting methods (DeepFillv2 and LaMa) across both metrics. As expected, FGT outperforms FGVC because it incorporates optimizations by design. All algorithms outperform the baseline of not restoring the region at all. Importantly, the differences are statistically significant.

From visual observation, FGT and FGVC did not have overlapping error bars with those of the single-frame inpainting methods, suggesting that those two algorithms performed the restoration task significantly better than that of the single-frame inpainting methods. The end-to-end method (i.e., E2FGVI) had overlapping bars with all other methods, showing weak evidence to support a significant difference.

In both bar charts, comparing the restoration methods within their classes results in overlapped bars, showing strong evidence that there was not a significant difference. To analyze the relationship quantitatively, for each restoration metric (i.e., mPSNR and mSSIM), we conducted six combinations of two-tailed two-sample t-tests for means with the baseline being a spatial inpainting algorithm and the "after" being a flow-based algorithm. The resulting p-values are shown in Fig. 8. Because the null hypothesis is that there is no significant difference between a selected spatial inpainting algorithm and a flow-based algorithm, with all resulting p-values much less than 0.01, we find strong evidence that there is a significant difference and visually observe that flow-based algorithms have significantly higher restoration metrics than do spatial inpainting algorithms. Indeed, FGT has the best performance of all (highest mPSR, mSSIM, and

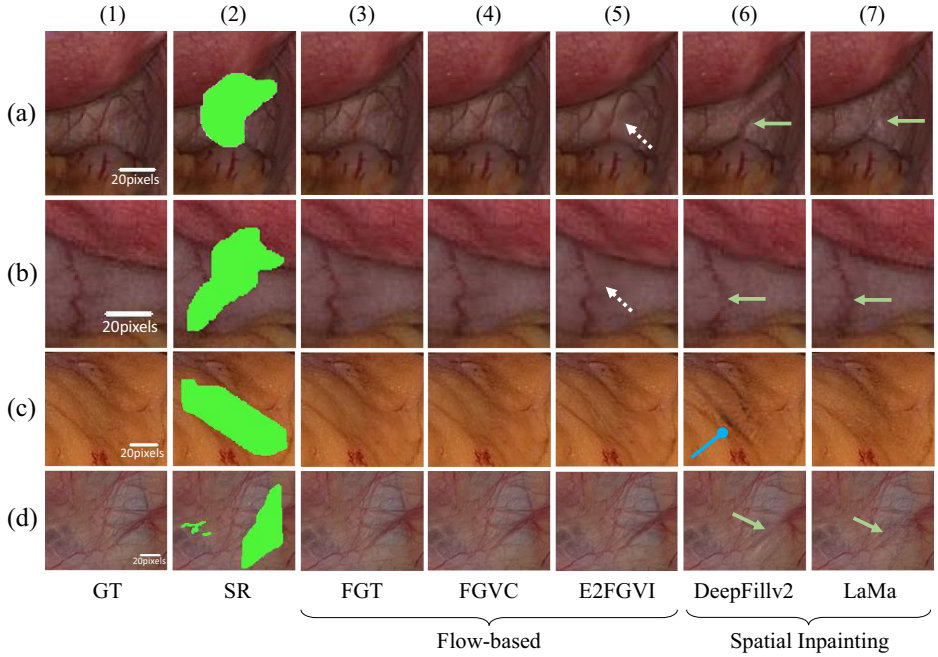


Fig. 9 Qualitative comparison of SR restoration performance. Qualitative comparison of flow-based and single-frame restoration algorithms. Four representative SR masks (green) are applied to qualitatively test restoration in (a)–(d). DeepFillv2 and LaMa fail to restore key structural features such as blood vessels (green arrows), with distortions or omissions evident. DeepFillv2 also struggles with larger areas, creating darker pixels and artifacts (blue rounded arrow). In contrast, optical flow-based algorithms like FGT, FGVC, and E2FGVI preserve structural integrity and color accuracy. However, E2FGVI shows undesirable smoothing (white dashed arrows), affecting its quantitative scores. Pixel-wise scale bars illustrate relative resolution between samples.

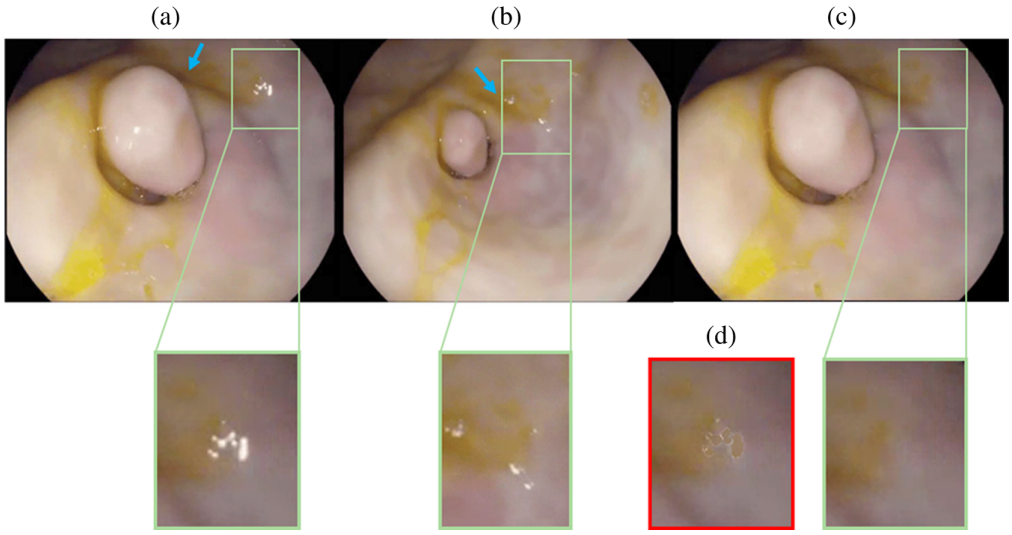


Fig. 10 Restoration under different lighting conditions. Green insets show enlarged versions of each picture covering the region of intended restoration. The blue arrows emphasize regions that have significantly different lighting across frames. Panel (a) shows a frame with an SR region. Panel (b) is a later frame with target pixels that are not obscured by SR; however, it has different lighting so direct replacement is not reasonable. Panel (c) shows the restored version of panel (a) that is free from SR artifacts with appropriate consideration of lighting changes. Panel (d) represents the baseline algorithm for restoration using homography transform for direct pixel replacement, which leads to color and lighting inconsistencies (Video 1, MP4, 1.02 MB [URL: https://doi.org/10.1117/1.JMI.11.2.024012.s1]).

most statistically significant difference from conventional methods) and was therefore selected as the restoration algorithm of choice for SpecReFlow.

Our qualitative analysis also supports the superior performance of flow-based restoration algorithms compared with single-frame counterparts. For example, in Fig. 9, both DeepFillv2 and LaMa fail to accurately restore structurally important features such as blood vessels (indicated by the green arrows), which appear distorted in some cases and not at all in others. Furthermore, we observed that DeepFillv2 struggled to correctly inpaint larger areas such as in Fig. 9(c)(6), which resulted in darker pixels and more artifacts (indicated by the rounded blue arrow). Compared with the single-frame inpainting restoration algorithms, optical flow-based algorithms (FGT, FGVC, and E2FGVI) performed favorably: not only do structural features appear in their correct likeness, but the colors are highly matched as well. E2FGVI, however, demonstrated undesirable smoothing (indicated by dashed white arrows), which may be responsible for its lower performance on the quantitative metrics. Overall, the qualitative results reflect the conclusions drawn from the quantitative analysis. Note that Fig. 9 contains pixel-wise scale bars that compare the relative resolution among samples, as physical measurements were not available. In Fig. 10, we demonstrate that our algorithm works in different lighting conditions, in which direct replacement is not as accurate as a restoration option. A demonstration video of the full SpecReFlow algorithm in action is provided as Video 1.

5 Conclusion

The need to improve the visualization of internal structures using endoscopy video by eliminating SR image artifacts motivates a complete and systematic approach to SR detection and restoration. To this end, we have created a complete solution for removing SR artifacts from endoscopy videos, called SpecReFlow. Our approach combines image preprocessing for contrast enhancement, deep learning, and conventional methods for complete SR detection, and it uses transformer and optical flow-based methods to leverage both temporal and spatial data for SR restoration.

SpecReFlow addresses the limitations of conventional detection algorithms by being able to detect all types of SR regions while doing so without the time-consuming, inefficient, and

ungeneralizable empirical parameter setting needed in conventional methods. Using flow-based restoration algorithms, our approach not only removes the need to predict features and colors directly, as done in other spatial inpainting methods, but also fixes the issue of lighting inconsistencies from direct pixel replacement in multiview restoration methods by operating in the color gradient domain. Ultimately, we showed how incorporating temporal information from adjacent slides leads to superior restorations compared with directly inpainting the SR region. In addition, we proposed two new metrics, mPSNR and mSSIM, to evaluate the quality of SR restoration more accurately than existing metrics.

Although the algorithm itself works well, a particularly challenging aspect of this work is the memory requirement. Although the vision transformer in the restoration algorithm can process relatively many frames at once, the virtual memory overhead is cost inefficient and not scalable to longer-duration endoscopy videos. Future investigation of memory-efficient transformer architectures, efficient loading of data from CPU to GPU, or a chunking method to encode new temporal information can drastically increase the restoration performance on longer duration frames. Another edge case is that, when both the camera/lighting and tissue are truly stationary across many frames, the calculated optical flow (which would be minimal because there is no movement) could be exploited to remove redundant frames and allow for more efficient processing. Once the set of frames is de-duplicated and the remaining frame is restored using our proposed algorithm, the result can be broadcasted to the initial set of duplicate frames. We leave the implementation of this memory-saving method to future work.

By providing a reliable and efficient solution for SR artifact removal in endoscopy videos, SpecReFlow can greatly improve diagnostic accuracy and efficiency, facilitating clearer visualization for early detection and management of diseases. Furthermore, we hope that our approach can be extended to other imaging modalities and environments to remove image artifacts in an efficient and reliable manner.

One promising future use case for SpecReFlow Det. is to generate accurate labels for other endoscopy datasets and build a database of annotated SRs that can be used to train a better detection network. Future work can be done with this new dataset to encode the predictions of a multicomponent detection algorithm into a singular more efficient and well-trained model architecture that performs just as well as, if not better than, our current detection algorithm.

Disclosures

There are no financial conflicts of interest with any of the authors regarding the material in this paper.

Code and Data Availability

The code, supporting documentation, and publicly available datasets for this paper can be found at https://osf.io/brmg9/

Acknowledgments

This work was supported in part by the National Institutes of Health (Grant No. DK117236). H.Y. acknowledges members of Bowden Biomedical Optics Lab for their editorial critiques. Special thanks go to Grace (Shuang) Chang for assistance in reviewing the data analysis.

References

1. C. Boedecker, "Using virtual 3D-models in surgical planning: workflow of an immersive virtual reality application in liver surgery," *Langenbecks Arch. Surg.* **406**, 911–915 (2021)
2. F. Sánchez-Margallo et al., "Comparative study of the influence of three-dimensional versus two-dimensional urological laparoscopy on surgeons' surgical performance and ergonomics: a systematic review and meta-analysis," *J. Endourol.* **35**(2), 123–137 (2021).
3. J. Shirk, "Effect of 3-dimensional virtual reality models for surgical planning of robotic-assisted partial nephrectomy on surgical outcomes: a randomized clinical trial," *JAMA Network Open* **2**(9), e1911598 (2019).
4. S. Ali et al., "A deep learning framework for quality assessment and restoration in video endoscopy," *Med. Image Anal.* **68**, 101900 (2021).

5. C. Nie et al., "Specular reflections detection and removal for endoscopic images based on brightness classification," *Sensors* **23**(2), 974 (2023).
6. F. Queiroz and T. Ren, "Automatic segmentation of specular reflections for endoscopic images based on sparse and low-rank decomposition," in *Braz. Symp. of Comput. Graphic and Image Process.*, pp. 282–289 (2014).
7. D. Fleet and Y. Weiss, "Optical flow estimation," in *Handbook of Mathematical Models in Computer Vision*, pp. 237–257, Springer, Boston, Massachusetts (2006).
8. K. Aires, A. Santana, and A. Medeiros, "Optical flow using color information: preliminary results," in *Proc. ACM Symp. Appl. Comput.*, pp. 1607–1611 (2008).
9. K. Zhang, J. Fu, and D. Liu, "Flow-guided transformer for video inpainting," *Lect. Notes Comput. Sci.* **13678**, 74–90 (2022).
10. S. Iwata et al., "Specular reflection removal with high-speed camera for video imaging," in *IECON 2015 - 41st Annu. Conf. IEEE Ind. Electron. Soc.*, pp. 1735–1740 (2015).
11. C. Saint-Pierre et al., "Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images," *Mach. Vision Appl.* **22**(1), 171–180 (2011).
12. B. Chwyl et al., "Specular reflectance suppression in endoscopic imagery via stochastic Bayesian estimation," *Lect. Notes Comput. Sci.* **9164**, 385–393 (2015).
13. S. Alsaleh et al., "Adaptive segmentation and mask-specific Sobolev inpainting of specular highlights for endoscopic images," in *Proc. Annu. Int. Conf. of the IEEE Eng. in Med. and Biol. Soc., EMBS*, pp. 1196–1199 (2016).
14. X. Wang et al., "Detection and inpainting of specular reflection in colposcopic images with exemplar-based method," in *Proc. Int. Conf. Anti-Counterfeit., Secur. and Identification, ASID*, pp. 90–94 (2019).
15. M. Asif et al., "Intrinsic layer based automatic specular reflection detection in endoscopic images," *Comput. Biol. Med.* **128**, 104106 (2021).
16. L. Ranyang et al., "Specular reflections removal for endoscopic image sequences with adaptive-RPCA decomposition," *IEEE Trans. Med. Imaging* **39**, 328–340 (2019).
17. S. Alsaleh, A. Aviles-Rivero, and J. Hahn, "ReTouchImg: fusioning from-local-to-global context detection and graph data structures for fully-automatic specular reflection removal for endoscopic images," *Comput. Med. Imaging Graphics* **73**, 39–48 (2019).
18. T. Liu, J. Wang, and L. Yang, "Specular reflections detection and removal based on deep neural network for endoscope images," Institute of Electrical and Electronics Engineers (IEEE) (2021).
19. P. Monkam et al., "EasySpec: automatic specular reflection detection and suppression from endoscopic images," *IEEE Trans. Comput. Imaging* **7**, 1031–1043 (2021).
20. S. Funke et al., "Generative adversarial networks for specular highlight removal in endoscopic images," *Proc. SPIE* **10576**, 1057604 (2018).
21. R. Sizyakin, "Automated visual inspection algorithm for the reflection detection and removing in image sequences," *Proc. SPIE* **11433**, 114332B (2020).
22. S. Ali et al., "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy," *Sci. Rep.* **10**, 2748 (2020).
23. J. Jiao et al., "Highlight removal for camera captured documents based on image stitching," in *Int. Conf. Signal Process. Proc., ICSP*, pp. 849–853 (2016).
24. S. Shah, S. Marshall, and P. Murray, "Removal of specular reflections from image sequences using feature correspondences," *Mach. Vision Appl.* **28**, 409–420 (2017).
25. R. Daher, F. Vasconcelos, and D. Stoyanov, "A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence," *Med. Image Anal.* **90**, 102994 (2022).
26. C. Gao et al., "Flow-edge guided video completion," *Lect. Notes Comput. Sci.* **12357**, 713–729 (2023).
27. D. Vázquez et al., "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthc. Eng.* **2017**, 4037190 (2017).
28. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
29. B. Yu et al., "Specular highlight detection based on color distribution for endoscopic images," *Front. Phys.* **8**, 575 (2021).
30. Y. Xu et al., "U-net with optimal thresholding for small blob detection in medical images," in *IEEE 15th Int. Conf. Autom. Sci. and Eng. (CASE)*, pp. 1761–1767 (2019).
31. R. Suvorov, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vision (WACV)*, pp. 2149–2159 (2022).
32. Z. Li et al., "Towards an end-to-end framework for flow-guided video inpainting," in *IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)* (2022).
33. J. Yu et al., "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)* (2019).

34. S. Tchoulack, J. Langlois, and F. Cheriet, "A video stream processor for real-time detection and correction of specular reflections in endoscopic images," in *Joint IEEE North-East Workshop Circuits and Syst. and TAISA Conf., NEWCAS-TAISA*, pp. 49–52 (2008).
35. K. Zhang et al., "Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.* **26**(7), 3142–3155 (2017).
36. A. Leibetseder et al., "GLENDA: gynecologic laparoscopy endometriosis dataset," *Lect. Notes Comput. Sci.* **11962**, 439–450 (2019).

Haoli Yin is an accelerated MS/BS in computer science student at Vanderbilt University and a research assistant at the Bowden Biomedical Optics Laboratory. He is also a Cornelius Vanderbilt and Goldwater scholar, researching multimodal representation learning and data-centric artificial intelligence with an applied focus on biomedical imaging.

Rachel Eimen is a PhD candidate in the Department of Biomedical Engineering at Vanderbilt University. She is a member of the Bowden Biomedical Optics Laboratory and the Vanderbilt Biophotonics Center, where her research focuses on creating software to enable the creation of three-dimensional virtual reconstructions of hollow organs. She earned her BS degree in computer engineering at Clemson University, where her research focused on neuronal circuitry and bioinformatics.

Daniel Moyer: Biography is not available.

Audrey K. Bowden is a Dorothy J. Wingfield Phillips Chancellor Faculty Fellow and an associate professor of biomedical engineering (BME) and electrical and computer engineering (ECE) at Vanderbilt University. She received her BSE degree from Princeton University and her PhD from Duke University and completed her postdoctoral training in chemistry and chemical biology at Harvard University. She is a fellow of SPIE, a fellow of AIMBE, a fellow of Optica, and a recipient of numerous awards.