UNIVERSITY OF CALIFORNIA,
IRVINE

Deep Learning Applications in Biomedical Sciences and Bioinformatics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Siwei Chen

Dissertation Committee:
Professor Pierre Baldi, Chair
Professor Roy Fox
Professor Amal Alachkar

2022

DEDICATION

To my parents, grandparents, and the rest of my family for their love and support.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor Pierre Baldi for his guidance and support.

I would like to thank my labmates for their support and friendship.

I wish to acknowledge publishers of my previous publications for permission to incorporate the material for the sole purpose of writing this thesis. Portions of Chapter 2 were previously published as "Weakly Supervised Polyp Segmentation in Colonoscopy Images Using Deep Neural Networks, Journal of Imaging, April 2022. Siwei Chen, Gregor Urban, Pierre Baldi". Permissions to reuse the text were granted by MDPI. Portions of Chapter 3 were previously published as "Deep learning to enable color vision in the dark, PLOS ONE, April 2022. Andrew W. Browne, Ekaterina Deyneka, Francesco Ceccarelli, Josiah K. To, Siwei Chen, Jianing Tang, Anderson N. Vu, Pierre F. Baldi". Permissions to reuse the text were granted by PLOS.

I wish to thank all co-authors of those publications in enabling the research which forms the basis for this dissertation.

I wish to acknowledge financial support from grant NIH GM123558, NSF NRT 1633631, and NIH R01-EB029751 issued to Pierre Baldi.

VITA

Siwei Chen

EDUCATION

Doctor of Philosophy in Computer Science **2022**
University of California, Irvine *Irvine, California*

Master of Science in Statistics **2016**
University of California, Irvine *Irvine, California*

RESEARCH EXPERIENCE

Graduate Research Assistant **2017–2022**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2017–2018**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Weakly Supervised Polyp Segmentation in Colonoscopy Images Using Deep Neural Networks **2022**
Journal of Imaging

Deep learning to enable color vision in the dark **2022**
Plos one

Atlas of exercise metabolism reveals time-dependent signatures of metabolic homeostasis **2022**
Cell Metabolism

An atlas of posttranslational modifications on RNA binding proteins **2022**
Nucleic Acids Research

Interrogating Metabolic Interactions Between Skeletal Muscle and Liver Circadian Clocks In Vivo **2022**
bioRxiv

The central clock suffices to drive the majority of circulatory metabolic rhythms **2022**
bioRxiv

Dynamic Changes of Brain Cilia Transcriptomes across the Human Lifespan **2021**
International journal of molecular sciences

Age-Related Neurometabolomic Signature of Mouse Brain **2021**
ACS Chemical Neuroscience

Brain histone beta-hydroxy-butyrylation couples metabolism with gene expression **2021**
bioRxiv

Metabolomic and transcriptomic signatures of prenatal excessive methionine support nature rather than nurture in schizophrenia pathogenesis **2020**
Communications Biology

MiR-29 coordinates age-dependent plasticity brakes in the adult visual cortex **2020**
EMBO reports

haracterizing the Skin and Gut Microbiome of Alopecia Areata Patients **2020**
SKIN The Journal of Cutaneous Medicine

Defining the independence of the liver circadian clock **2019**
Cell

Time of exercise specifies the impact on muscle metabolic pathways and systemic energy homeostasis **2019**
Cell Metabolism

Epigenetic regulation of the circadian gene Per1 contributes to age-related changes in hippocampal memory **2018**
Nature Communications

CircadiOmics: circadian omic web portal **2018**
Nucleic acids research

ABSTRACT OF THE DISSERTATION

Deep Learning Applications in Biomedical Sciences and Bioinformatics

By

Siwei Chen

Doctor of Philosophy in Computer Science

University of California, Irvine, 2022

Professor Pierre Baldi, Chair

Deep learning has been applied to solve complex problems in a variety of scientific domains including biological and medical sciences. In particular, convolutional neural networks (CNNs) have displayed state-of-the-art results in computer vision tasks, outperforming humans in numerous cases. In this work, we explore several deep learning applications in medicine and bioinformatics. We show that deep learning can be applied to real-world medical procedures such as colon screening, and assist physicians with polyp detection and segmentation, and that deep learning can be applied to optical surgery, and assist physicians with image quality enhancement. In the field of bioinformatics, advances in high-throughput sequencing technologies have greatly lowered the cost of genome-wide sequencing, allowing researchers to include a larger number of experimental replicates than they previous could. This allows an easier application of deep learning as well as other statistical models to solve the problems in bioinformatics, such as detecting circadian rhythm in high-throughput omic data. To access and mine circadian datasets in a comprehensive and integrated way, we curate over 227 high-thoughput circadian datasets across different species and tissues, apply a deep learning model named BIO_CYCLE together with other statistical methods to detect circadian pattern in the omic data, and build a web portal `http://circadiomics.ics.uci.edu` for search and visualization.

Chapter 1

Introduction

1.1 Deep Learning in Medicine

Deep learning has been applied to solve complex problems in a variety of scientific domains including the biological and medical sciences [4]. Deep learning models are applicable to model for a wide range of data types in those fields including text data, scanning data, high throughput gene expression data, and so on. For instance, convolutional neural networks (CNNs) have displayed impressive results in image analysis tasks such as image recognition [34, 91], object detection [26], and segmentation [63], outperforming humans in several cases, and they have been applied extensively to biomedical imaging problems with many hundreds of articles published in the scientific literature (e.g. [104], [13], [103]. [107], [108] from our group alone). Recurrent neural network (RNN) with long short-term memory units (LSTM) have showed great capability in modeling and predicting healthcare outcomes using sequential or time-series data [78].

Some challenges associated with modeling for medical datasets include: (1) the size of the dataset is usually small. Models trained on small-sized dataset are harder to converge, and

easier to overfit. To solve this problem in image analysis tasks, a lot of data augmentation techniques are applied. These techniques add variations to the original data in colors, shape, noises, etc. and thus increase the amount of training data. In addition, transfer learning techniques are often used so that the knowledge learned on a much larger dataset can be transferred or "borrowed" to solve a similar task in a much smaller dataset. Also, public data or synthetic data can be a helpful addition if the current dataset is too small. (2) The second problem associated with the medical datasets is the poor quality of data. Medical datasets often suffer from missing data, such as missing records of patients or missing time points in a record, which require imputation using mean, median, or even predicted values to fill in for the missing ones. Image data such as CT and MRI scans or screening videos, often have problems of missing labels or having low image resolutions. These images need to be preprocessed or augmented before feeding into the models for training.

In the next two chapters, I will discuss two deep learning applications in medicine in detail, on their architectures, performances, and significance: (1) a real-time polyp segmentation system in the colonoscopy screening procedure; and (2) a deep learning aid to enable colored vision under infrared light for optical surgeries.

1.2 Deep Learning and Software Systems in Bioinformatics

With recent advances in high-throughput sequencing technologies, the costs of sequencing applications such as RNA-Seq, ChIP-Seq, and Ribo-Seq, have been greatly reduced. Therefore, researchers can afford to include more experiment samples (mice, rats, or even human subjects) to measure the biological properties of the interest. This improvement mitigate one of the challenges associated with the field of bioinformatics, which is the small number

of sample points for data analysis. In this work, we curate the largest repository of circadian datasets across 8 species and over 277 experiments, and apply a deep learning model named BIO_CYCLE [2] to detect circadian oscillations in transcriptomic and metabolic time-series data. We create a web server named Circadiomics [12] for search and visualization.

Chapter 2

Weakly Supervised Polyp Segmentation in Colonoscopy Images using Deep Neural Networks

2.1 Introduction: Deep Learning Applications in Colonoscopy Screening

Colorectal cancer (CRC) is the third most lethal and third most commonly diagnosed cancer in the United States [1]. Most CRC cases arise from growths of adenomatous polyps, most of which are benign but can become malignant over time, leading to death unless detected and treated [90]. Colonoscopy is the established method of choice and gold standard for detection and removal of colorectal polyps. However, the effectiveness of this procedure is jeopardized by the number of polyps missed depending upon the colonoscopist's skill level, polyp size and morphology, location in the colon, bowel preparation, and other operator- and procedure-dependent factors [3]. Studies have reported that between 6% and 28% of

present colorectal polyps are missed during colonoscopy [55], and as a result, they could develop into post-screening CRC or interval CRC, which amount to 5% to 8% of all CRCs [57]. In patients with inflammatory bowel disease (IBD), interval CRCs can even account for up to 50% of CRCs [101]. A study by le Clercq et al. reported that 86.4% of interval CRCs could be explained by procedural factors, including 57.8% of interval CRCs caused by missed polyps during colonoscopy, 19.8% caused by inadequate examination, and 8.8% caused by incomplete removal of polyps [54]. A study on interval CRCs in Sweden reported that individuals with interval CRCs had shorter survival times than individuals with CRCs detected in colonoscopy, and that the hazard ratios for men with interval CRCs were 2.75 with 2.00 for women [24]. Macken et al. also report worse survival for patients with interval CRCs than those without: 80% of the patients with interval CRCs survived for 1.6 years compared to 2.8 years for patients without; 60% of the patients with interval CRCs survived for 4.7 years compared to 6.7 years for patients without interval CRCs [64]. Many technologies have been developed to reduce interval CRCs and to improve the quality of colonoscopy screening, including high-definition imaging, narrow-band imaging (NBI), magnification endoscopes, and more recently, deep-learning-based diagnostic aid.

2.2 Overview and Challenges in Applying Deep Learning Models to Aid Colonoscopy Screening

Deep learning has been applied to solve complex problems in a variety of scientific domains including the biological and medical sciences [4]. In particular, convolutional neural networks (CNNs) have displayed impressive results in image analysis tasks such as image recognition [34, 91], object detection [26], and segmentation [63], outperforming humans in several cases, and they have been applied extensively to biomedical imaging problems with many hundreds of articles published in the scientific literature (e.g., [104, 13, 103, 107, 108]

from our group alone). Image segmentation is especially important for analyzing medical images, as the segmented image can provide insights into the structure's size, volume, and morphology and can help physicians with detection and diagnosis [32]. For instance, machine learning models have been developed to segment and classify pulmonary nodules in chest scan datasets, assisting physicians in early-stage lung cancer diagnosis [59]. Machine learning and deep learning approaches have also been proposed to help improve colonoscopists' skill level in detecting polyps during screening colonoscopy [105, 48, 75, 81], and several clinical trials have investigated their effectiveness [97, 109, 95].

The use of deep learning for diagnostic aid in colonoscopy falls into one of three categories: (1) predicting polyp pathology, such as Li et al.'s work [60] classifying images of polyps into sessile serrated adenomas (SSA) and non-SSA, or Byrne et al. [11], who propose a CNN model to classify neoplastic vs. non-neoplastic polyps. (2) The second category is polyp detection with bounding boxes, such as Shin et al. [90], who use region-based CNN (R-CNN) [27] for polyp detection with a processing time of 39ms per frame, which is borderline too slow for processing colonoscopy videos in real time. Urban et al. [105] implement a YOLO-derived [80] architecture for polyp detection in videos and achieve an accuracy of 96.4% with a processing time of 10ms per frame, making it possible to detect polyps with high accuracy in real time. (3) The third category is polyp segmentation at the pixel level of video frames, which is computationally more demanding. Examples are Brandao et al. [9], who propose a fully convolutional neural network (FCN) with 51ms prediction time per image, which is too slow for use during procedures, and other attempts such as Vazquez et al. [106] using Mask R-CNN [33] or the work by Jie et al. [69] – but none of these meet real-time processing speed constraints. The work by Guo et al. summarizes the performances of several polyp segmentation models on the widely used public CVC-ClinicDB [7] data set [30], where a baseline ResUNet model obtains a dice coefficient of 79.55% [47], and an ensemble of multiple models, referred to as MED-Net, achieves a dice coefficient of 91.3% [73]. In a study by Mahmud et al., a modified encoder-decoder segmentation model referred

to as PolypSegNet achieves a dice coefficient of 91.5% on CVC-ClinicDB and 88.7% on another commonly benchmarked public data set, Kvasir-SEG [46]; however, the model only has a near-real time inference speed of 39ms per image (25 frames per second) [65]. Tomar et al. implement a feedback attention network FANet, and achieve a dice coefficient of 93.55% on CVC-ClinicDB and 88.03% on Kvasir-SEG [100]. However, the above-mentioned segmentation models are all fully supervised and must train on pixel-level, hand-annotated masks, which are extremely time-consuming to obtain. To leverage unlabeled data in model training, Wu et al. propose a semi-supervised polyp segmentation model with an adversarial learning method. When trained on 30% mask-labeled images and 70% unlabeled images, this model achieves a dice coefficient of 89.29% on CVC-ClinicDB and 80.95% on Kvasir-SEG, and when trained on 15% mask-labeled images and 75% unlabeled images, the dice coefficient score drops to 82.18% on CVC-ClinicDB and 76.76% on Kvasir-SEG [112].

A real-time polyp segmentation model can be a major step toward automation of non-invasive procedures using capsule endoscopy [71], and in the future, full robotic automation of the entire screening procedure. However, most published segmentation approaches suffer from the drawback that prediction throughput is too low to be real-time-capable even with a GPU (graphics processing unit); furthermore, most segmentation models require pixel-level annotated training data, which is very labor-intensive to obtain. In this study, we propose a weakly supervised [18] U-Net model [83] that leverages easy-to-obtain bounding box annotations to predict pixel-level segmentation masks of polyps in real time. The main contributions of this work are (1) a model that can predict pixel-level segmentation masks despite using only bounding box annotation as supervision source during training and, thus, avoids the need for expensive data annotation and (2) a model can generate polyp segmentations in real time at 45 FPS and, thus, assist with real-world colonoscopy screening procedures. The weakly supervised model is trained through an iterative process, where the model iteratively refines its own training targets (imperfect predictions of polyp segmentation masks), while weak external feedback is used to guide the training process by ensuring that all intermediary

predictions on training images are consistent with the corresponding bounding box annotations. This process demonstrably recovers the correct segmentation mask in almost all cases as the network is trained. The model is trained on the publicly available CVC-ClinicDB dataset and a novel Polyp-Box-Seg dataset of 4070 colonoscopy images obtained from over 2000 patients. We show that the iterative training process achieves the same performance level (dice coefficient and pixel-wise accuracy) as would have been obtained by supervised training on human-annotated segmentation masks. In addition, we find that pooling together all labeled data (bounding box annotations and segmentation masks) during training further improves the model's accuracy. Various CNN models are evaluated on the CVC-ClinicDB dataset; the Kvasir-SEG dataset; and the Polyp-Box-Seg dataset, in particular, on the subset of sessile serrated adenomas (SSA) in Polyp-Box-Seg, which are the most challenging polyps to correctly identify and segment due to their morphology and appearance [60]. The Polyp-Box-Seg dataset, together with a real-time video demonstration of the segmentation system, are available at `www.igb.uci.edu/colonoscopy/AI_for_GI2.html`.

2.3 Methods used in Real-time Polyp Segmentation

2.3.1 Deep Learning Architecture

The architecture used in this study is a U-Net [83] architecture, which is a fully convolutional network with a contracting path (encoder) and an expanding path (decoder), producing a segmentation prediction that is the same size as the input image. We chose to use five down-sampling blocks and five up-sampling blocks in the model, as shown in Figure 2.1. We used VGG16 weights trained on ImageNet as initial weights for the encoder component and random initial weights for the rest of the network [91]. A detailed description of the U-Net architecture is as below:

The first down-sampling blocks consisted of two $3 \times 3 \times 64$ padded convolutions followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2. At each down-sampling step, the number of feature channels was doubled. The second down-sampling blocks consisted of two $3 \times 3 \times 128$ padded convolutions followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2. The following three down-sampling blocks consisted of three 3×3 padded convolutions with numbers of feature channels of 256, 512, and 512, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2. Each of the up-sampling block consisted of a 2×2 up-convolution, a concatenation with the feature map from the down-sampling path, and two 3×3 convolution, each followed by a ReLU. At each up-sampling step, the number of feature channels was halved and concatenated with an equal number of feature channels from the corresponding block from the down-sampling path. The final layer used a 1×1 convolution followed by a sigmoid activation to map the 64 feature maps to a probability between 0 and 1. The closer to 1, the more likely this pixel contained polyp, and the closer to 0, the more likely this pixel contained background.

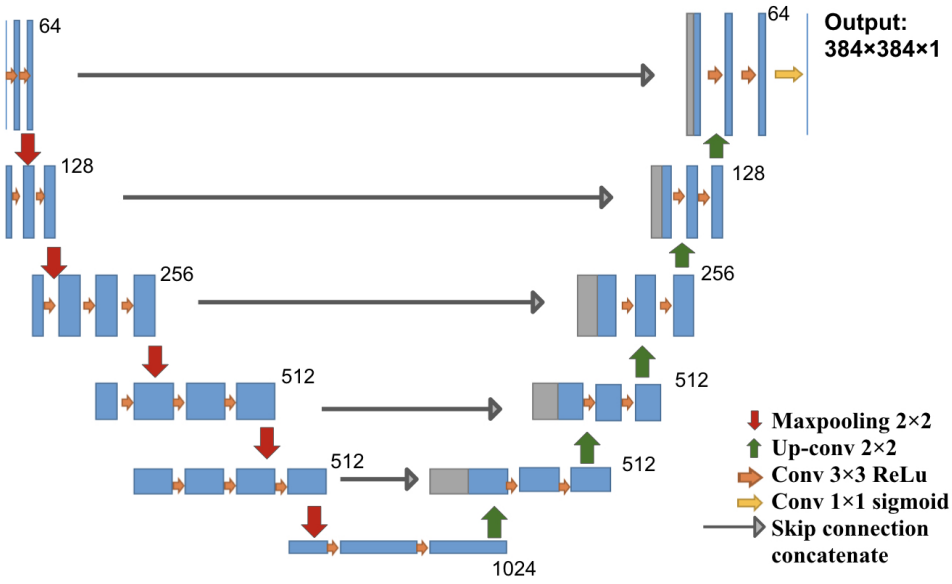


Figure 2.1: U-Net architecture: each block in the down-sampling path consists of convolution and max pooling operations, and each block in the up-sampling path consists of up-convolution and convolution operations. Each blue box corresponds to a multi-channel feature map. The number of features at the end of each block is denoted on top of the box. Arrows of different colors denote the different operations.

2.3.2 Iterative Weakly Supervised Training

To start the weakly supervised segmentation training process, initial segmentation masks for polyps in the training images must be generated, even if they are inaccurate. Three different approaches to obtain these masks were tested: (1) the initial mask for each polyp is approximated by a solid circle located in the center of the bounding box and sized to fill the box but not exceed any of its edges; (2) a segmentation model is pre-trained on the public polyp datasets, and its predictions serve as initial training targets after removing all parts of the predictions that exceed the annotated bounding boxes; and (3) a combination of approaches (1) and (2): initial masks are generated as in (2) using a pre-trained network, but if the predicted area within one bounding box occupies less than 30% of the box, then this initial prediction is rejected and replaced by a solid circle, as in (1). The third approach resulted in initial masks, of which approximately half were solid circles, as in (1), and half

were predicted masks, as in (2).

These masks were then used as training targets for the corresponding input images in the iterative training scheme. After each iteration, the masks were updated using the network's predictions. The bounding box annotations were used to define the weighted training loss function (as discussed in Section 2.3.3) and to remove all positive polyp predictions that fell outside of the bounding box by setting those mask values to 0. Then, the neural network was re-trained on the updated masks, and the process of updating masks and network was repeated for several iterations.

This iterative training process is illustrated in Figure 2.2. The model was trained with a batch size of 12 for 3 epochs per iteration for the first 2 iterations and then for 6 epochs for another 6 iterations. The initial learning rate was 0.0001 for the first two iterations and decay with a factor of 0.0005 at each epoch starting from the third iteration. The Adam optimizer was used for training. Training was terminated after the validation loss stopped improving for two consecutive iterations.

The dice coefficient (F1 score), accuracy and confusion matrix are used as metrics to evaluate model performances [5].

2.3.3 Masked Loss Function

Using this weakly supervised training procedure entails that the training targets are sometimes not a perfect representation of the polyps' outlines, especially in the early iterations. To reduce the impact of these inaccuracies on the neural network performance, we used a pixel-wise weighted loss function with the pixel weights accounting for our confidence on the most likely location of the polyps within a bounding box. The loss mask is implemented using oval shaped rings, each extending to the four edges of its corresponding bounding box.

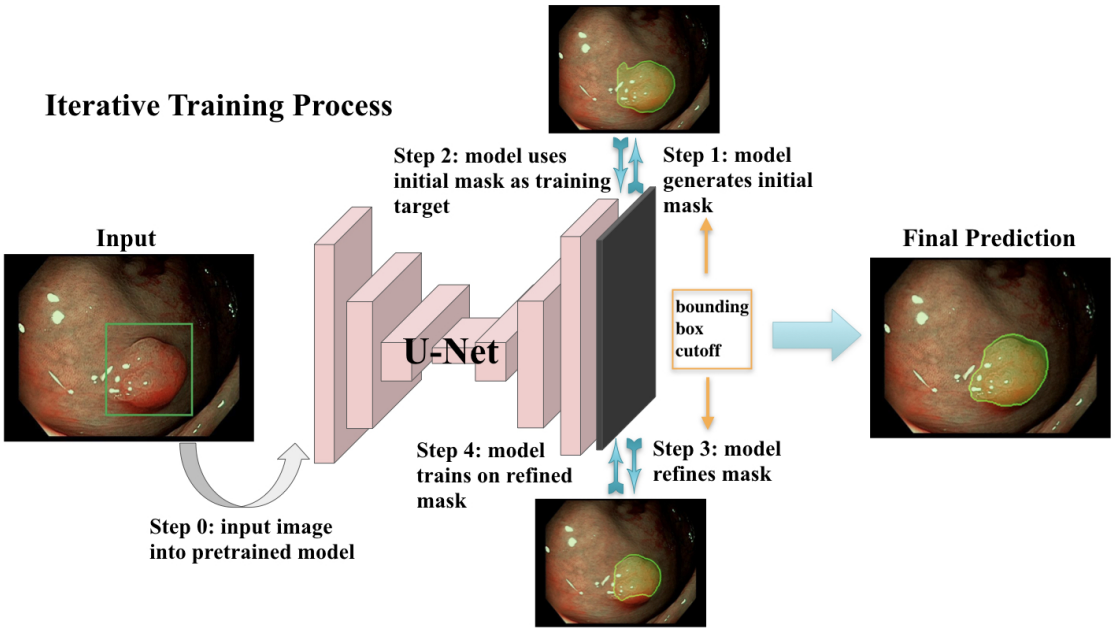


Figure 2.2: Evolution of the segmentations starting from their initialization state over the course of training.

The ring has a value of zero along its length, while its center and surrounding area have a mask value of one, with a gradual transition between the two regimes. The effect of this construction is that the neural network's predictions are heavily scrutinized only in areas outside of the bounding boxes (which are guaranteed to be free of polyps) and areas in the center of the bounding boxes (which are guaranteed to belong to a polyp), while the model's predictions in the areas in between (close to the bounding box's edges) are left mostly untouched. Please note that the bounding boxes and loss were only used during training, and neither are present or used when the system is in live operation on videos or test images.

2.3.4 Dataset and Preprocessing

Four datasets were used for training and/or evaluation: (1) the ImageNet competition dataset, on which the VGG16 weights were trained [86]; (2) the public CVC-ClinicDB dataset of 612 polyp images with ground truth segmentation masks; (3) the public Kvasir-

SEG dataset of 1000 polyp images with ground truth segmentation masks and bounding box coordinates of the polyp regions; and (4) a novel dataset, Polyp-Box-Seg, of 4070 images containing unique polyps collected from over 2000 patients. In the Polyp-Box-Seg dataset, all images with a polyp were annotated with bounding boxes over the polyp region, and a randomly chosen subset of 1300 polyp images was further annotated with segmentation masks by human experts. The set of 4070 polyp images was hand-selected from colonoscopy screening videos so that each image contains an unique polyp. This is to prevent correlation between images and thus to avoid a possible intra-patient polyp similarity bias. The Polyp-Box-Seg dataset contains polyps of all sizes and morphologies and covers all portions of the colorectum. Locations and dimensions of bounding boxes over the polyp regions were recorded by a team of colonoscopists. The original resolution of these images is 480 $\times$ 640 pixels, which was resized and padded to dimensions of 384 $\times$ 384. The data were normalized by subtracting its mean pixel value and by dividing by its standard deviation before training. While modern scopes generally operate at a higher resolution than 384 $\times$ 384, a prior study by Urban et al. [105] found that changes in input resolution have an almost negligible effect on automated polyp detection accuracy (comparing a resolution of 480 $\times$ 480 to 224 $\times$ 224 pixels). The dataset contains polyps of all histologies shown in Table 2.1, with a total of 349 sessile serrated adenomas (SSA), which is the polyp type that is the most difficult to delineate and annotate. Out of those 349 images, 90 have human labeled segmentation masks (chosen at random). Data augmentation techniques that were applied during training include (1) random translations of the image in any direction by up to 60 pixels but limited to never move parts of a polyp outside of the image; (2) zooming in or out of the images by up to +/− 10%; (3) random mirroring of the image (horizontal and vertical); and (4) color augmentations performed by shifting each of the three color channels separately using a Gaussian distribution (mean 0 standard deviation 0.1) and thus slightly changing the images' overall hue.

Table 2.1: Histology information for 4070 polyp images from the Polyp-Box-Seg dataset.

Histology	Count
Tubular adenoma	2102
Hyperplastic	909
Sessile serrated adenoma	349
Non-serrated sessile	446
Tubulovillous adenoma	64
Inflammatory	42
Traditional serrated adenoma	33
Lymphoid nodule	19
CA adenocarcinoma	14
Hamartomatous	11
Juvenile polyp	6
CA lymphoma	5
Sessile serrated adenoma w dysplasia	3
Mucosal prolapse	3
CA squamous/epidermoid	1
Other	63

2.3.5 Model Training

An U-Net model was constructed as depicted in Figure 2.1 and initialized as described in Section 2.3.6. The following experiments were conducted:

1. A fully supervised model was trained on the 612 polyp images and segmentation masks of the public CVC-ClinicDB dataset. The weights of this model were used as a starting point for further experiments. This model is denoted as Full-Sup-1-VGG.

2. A fully supervised model was trained and evaluated via 10-fold cross-validation on the 1300 polyp images with segmentation masks from the Polyp-Box-Seg dataset. While the entire dataset contains 1300 images with segmentation masks, 100 of those were randomly selected and reserved for hyperparameter selection, with the 10-fold cross-validation being performed on the remaining 1200 images. A detailed breakdown of the data subsets used can be found in Figure 4.4. Weights pre-trained on the CVC-ClinicDB dataset were used as initial weights for this model. This model is denoted

as Full-Sup-2. The cross-validation procedure trains 10 independent models and tests each on a different set of 120 images. This approach both significantly reduces the variance of the estimated (average) test score compared with a naïve single train-test data split and allows for an estimation of the variance of test accuracy estimates.

3. A weakly supervised model was trained using the bounding box annotations on the 4070 polyp images from the Polyp-Box-Seg dataset in a 10-fold cross-validation, as described in Section 2.3.2. Three approaches to generating the initial segmentation targets were evaluated:

   (a) All initial training targets are set to be a solid circle in the center of the bounding boxes, with a diameter equal to 4/5 of the box's shorter side (width or height). The area outside of the bounding boxes is thereby assumed to not contain any polyps and serves as "background" category. This model is referred to as Weak-Sup-Box-CI in the following (Circular Initialization).

   (b) Predictions of the model trained on the public CVC-ClinicDB dataset are used as initial targets. All pixels with a predicted probability over 0.5 that lie within a bounding box are assumed to show part of a polyp, while all other pixels are considered to be background pixels. Models trained on this initialization method will be denoted as Weak-Sup-Box-PI (Prediction Initialization).

   (c) The same initialization scheme as in 3b is used, but bad predictions are replaced with a solid circle as in 3a. Polyp predictions are considered bad if they occupy less than 30% of the bounding box area. Models trained using this initialization method are denoted as Weak-Sup-Box-HI (Hybrid Initialization). Weak-Sup-Box-HI is evaluated on the Kvasir-SEG dataset.

4. A segmentation model was trained using the 1300 images with human segmentation labels together with the 2770 images with bounding boxes (a total of 4070 images) from the Polyp-Box-Seg dataset. Initial masks for the 2770 images with bounding boxes were

generated in the hybrid manner, as described in 3c, and these masks were updated at the end of each training iteration, as described in Section 2.3.2. Segmentation mask labels were used for the 1300 images throughout training without being iteratively updated, as these labels are already accurate. In each iteration, the model trains on the mixture of 4070 images from both supervision sources, and at the end of the iteration, the model updates the masks for the 2770 images using model prediction. Models trained with a combination of weakly supervised bounding box targets and fully supervised segmentation mask targets are denoted as Weak-Sup-Mix.

5. A Weak-Sup-Mix model was trained as described in 4 and evaluated on sessile serrated adenomas using 10-fold cross-validation.

6. A Weak-Sup-Mix model was trained as described in 4 and evaluated on CVC-ClinicDB using 10-fold cross-validation. In each fold, the model was trained on all images from the Polyp-Box-Seg dataset plus 90% of the CVC-ClinicDB data and validated on the remaining 10% of the CVC-ClinicDB dataset. In the first two iterations, one extra training epoch on CVC-ClinicDB data was added at the end of three training epochs on all data, and in the remaining six iterations, two extra training epochs on CVC-ClinicDB data were added at the end of six training epochs on all data.

7. A fully supervised segmentation model was trained and evaluated using 10-fold cross-validation on the CVC-ClinicDB dataset using initial weights pre-trained on the Weak-Sup-Mix model, as described in 4. This model is named Full-Sup-3. This pre-trained model essentially transfers knowledge of polyp shapes learned on the Polyp-Box-Seg dataset to the CVC-ClinicDB dataset, where it is fine-tuned to adjust to the differences between the datasets (such as different cameras and lighting conditions). The model was trained for 30 epochs with an initial learning rate of 1e-4, learning rate decay of 5e-4 after each epoch, and a batch size of 1.

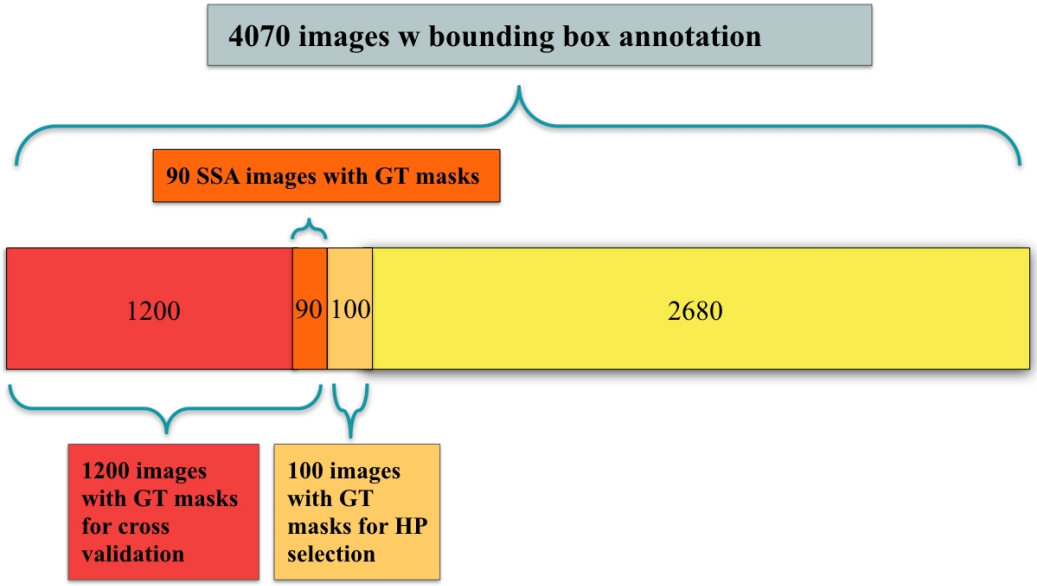


Figure 2.3: A breakdown of the 4070 polyp-containing images in training and testing: all 4070 images contain bounding box annotations of polyps, and a subset of 1300 images additionally contains human labeled ground truth (GT) segmentation masks. Among the 1300 images, 100 are used for hyperparameter (HP) selection and 1200 are used for model evaluation using cross-validation; 90 out of the 1200 images are sessile serrated adenomas (SSA).

2.3.6 Network Initialization

Neural networks are commonly initialized with small random weights before training, but it has been common practice to instead use weights from networks trained on broadly related tasks, as this is extremely beneficial. In this work, all models with suffix "-VGG" were initialized with weights from a VGG16 model trained on the ImageNet dataset of 1.2 million images. Although that dataset does not contain medical polyp images, it has been shown repeatedly in the past that this transfer learning process greatly aids subsequently trained polyp detector models. Other models were initialized with weights trained from Full-Sup-1-VGG. Different initializations allow us to assess the importance of the public polyp dataset on which Full-Sup-1-VGG was trained.

2.4 Results

2.4.1 Pre-Trained Weakly Supervised Models on Polyp-Box-Seg

A series of models pre-trained on the CVC-ClinicDB dataset are trained and evaluated on the Polyp-Box-Seg dataset. First, as a baseline, a fully supervised model called Full-Sup-2 is trained on the 1300 segmentation masks annotated images from Polyp-Box-Seg alone. Full-Sup-2 reaches an average dice coefficient of 81.52% and a pixel-level accuracy of 98.76% using 10-fold cross-validation (Full-Sup-2 in Table 2.2).

Second, a series of weakly supervised models are trained using bounding box annotations alone (4070 images from Polyp-Box-Seg), with three different segmentation mask initialization methods (circle, prediction, and hybrid). With the same 10-fold cross-validation, these models achieve average dice scores of 77.58 $\pm$ 0.66% (Weak-Sup-Box-CI in Table 2.2), 77.78 $\pm$ 0.87% (Weak-Sup-Box-PI in Table 2.2), and 81.36 $\pm$ 0.43% (Weak-Sup-Box-HI in Table 2.2). The Weak-Sup-Box-HI model performs significantly better than the two others, which highlights that the quality of the initial training targets is of significant importance and that the iterative training process cannot compensate for this entirely. The Weak-Sup-Box-HI model, trained iteratively on 4070 bounding box annotations, achieves a comparable pixel-level segmentation performance with the fully supervised Full-Sup-2 model (dice coefficient 81.52 $\pm$ 0.41%), which is trained on 1300 manually labeled segmentation masks. There is no significant difference between the performances of those two models according to a t-test (p-value = 0.87). This shows that, even though bounding box annotations lack pixel-level information on the shape of polyps, we can achieve competitive results on the pixel-wise segmentation task using the iterative weakly supervised training procedure.

Third, the weakly supervised model Weak-Sup-Mix in Table 2.2 is trained using both annotations combined, i.e., 1300 images with segmentation masks and 2770 images with bounding

Table 2.2: Average test-set scores and standard deviation using 10-fold cross-validation on the Polyp-Box-Seg images. The neural networks in this table are pre-trained on the public CVC-ClinicDB dataset and then further trained as described in the text.

	Full-Sup-2	**Weak-Sup-Box-CI**	**Weak-Sup-Box-PI**	**Weak-Sup-Box-HI**	**Weak-Sup-Mix**
Dice Coefficient	81.52 ± 0.41%	77.58 ± 0.66%	77.78 ± 0.87%	81.36 ± 0.43%	**85.53 ± 0.33%**
Accuracy	98.76 ± 0.09%	98.42 ± 0.10%	98.50 ± 0.11%	98.67 ± 0.09%	**98.96 ± 0.07%**

box annotations. The iterative weakly supervised procedure is applied to the 2770 images (the annotations for the 1300 images are already correct and are excluded from being changed during this procedure). This model reaches a dice score of 85.53 ± 0.33%, significantly surpassing the performance of all other models, including Full-Sup-2, which was trained on segmented masks and reaches a dice coefficient of only 81.52 ± 0.41% (t-test p-value of 0.00096 ¡ 0.05). This suggests that training with the maximum amount of data and heterogeneous annotations works better than utilizing either single supervision source on its own.

An evolution of the segmentation masks predicted by the network is shown in Figure 4.5 and discussed in the following. At the beginning of the training process, the initial masks are far from perfect (see Figure 4.5 b,f,j,n) but improve significantly over the course of training the network. In Figure 4.5, (b) and (f) depict masks initialized with predictions made from Full-Sup-1-VGG, a model pre-trained on the public CVC-ClinicDB polyp dataset (see 3b). Differences between CVC-ClinicDB and our dataset reduce this model's segmentation accuracy on our dataset, causing the initial segmentation masks in (b) and (f) to not to overlap with the true polyps very well. Nevertheless, several iterations of the weakly supervised training scheme allow the model to segment polyps accurately (see (d) and (h) in Figure 4.5). The last two rows of Figure 4.5 show masks that are initialized with circles (images (j) and (n); see 3a for details), and despite these poor initial labels, the model still

gradually recovers the contour of the polyps close to ground truth during training ((l) and (p) in Figure 4.5). Notice that image (i) contains a surgical instrument in the bounding box area that is initially part of the circular labels and falsely predicted in an early iteration (j) but correctly excluded from the predicted mask after six iterations of training. Compared with the bounding boxes in (a) and (e), which are tightly placed around the polyps, bounding boxes in (i) and (m) contain a large non-polyp area. Nevertheless, even with loosely placed bounding boxes where the polyp center, size, and shape are harder to determine, the model eventually finds the polyp contour.

To compare the effectiveness of the four weakly supervised models from Table 2.2, we use a fixed validation set and compute the average number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in their predictions (Table 2.3). The performances between Weak-Sup-Box-PI and Weak-Sup-Box-CI are similar, with Weak-Sup-Box-PI having a higher TP and higher FP. Weak-Sup-Box-CI yields the highest number of FN among the four models. This suggests that simply assuming polyps to be circles is not enough for initial mask generation. Weak-Sup-Box-HI is more effective than the first two models because it generates better initial masks by using the hybrid method. When trained on the combination of bounding-box-labeled data together with mask-labeled data, the Weak-Sup-Mix model is the most effective among these four weakly supervised models and reaches the highest TP and lowest FP in these predictions. This shows the benefit of effectively pooling all images together in training, even when their supervision sources are different.

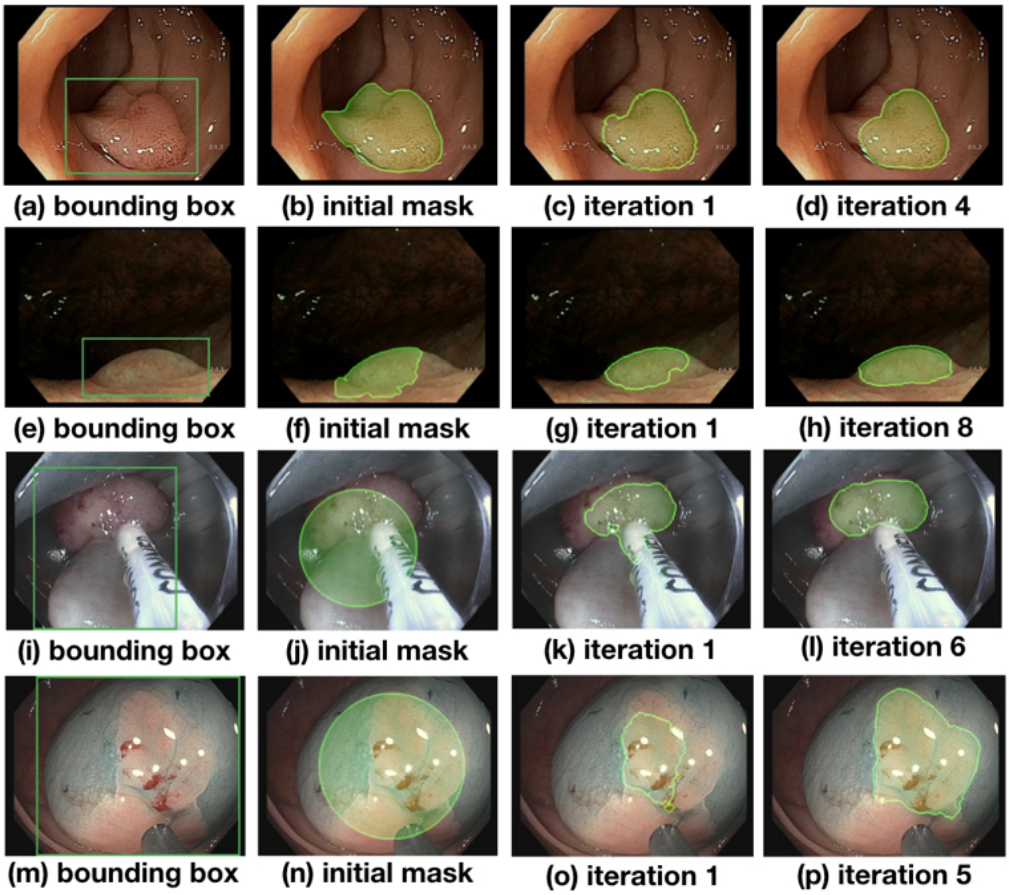


Figure 2.4: Evolution of the segmentations starting from their initialization state over the course of training.

Table 2.3: Confusion matrices from weakly supervised models' predictions on one split of the validation set.

	Weak-Sup-Box-CI	**Weak-Sup-Box-PI**	**Weak-Sup-Box-HI**	**Weak-Sup-Mix**
TP	5039.01	5947.08	6180.07	**6333.31**
TN	139604.96	139414.03	139341.57	**139755.01**
FP	512.68	703.62	776.07	**362.63**
FN	2299.32	1391.26	1158.26	**1005.02**

2.4.2 Weakly Supervised Models Initialized with VGG16 Weights on Polyp-Box-Seg

Models initialized with VGG16 weights in Table 2.4 performed slightly worse than those that benefited from pre-training on the CVC-ClinicDB dataset. However, even under these circumstances, the model with the addition of weakly supervised training (Weak-Sup-Mix-VGG in Table 2.4) performed significantly better than the model trained on only the human segmentation annotations (Full-Sup-2 in Table 2.2), with a dice coefficient of 85.08 ± 0.60% compared with only 81.52 ± 0.41% (t-test p-value of 0.0018).

Table 2.4: Average test-set scores and standard deviation using 10-fold cross-validation on the Polyp-Box-Seg images. Models are initialized with VGG16 weights and then trained as described in the text.

	Full-Sup-2-VGG	**Weak-Sup-Box-CI-VGG**	**Weak-Sup-Mix-VGG**
Dice Coefficient	79.11 ± 0.93%	76.14 ± 0.67%	**85.08 ± 0.60%**
Accuracy	98.56 ± 0.14%	98.31 ± 0.08%	**98.93 ± 0.09%**

2.4.3 Weakly Supervised Models Tested on Sessile Serrate Adenomas Alone

The best-performing weakly supervised model, Weak-Sup-Mix, which utilizes both bounding boxes and human annotations, is evaluated on sessile serrated adenomas (SSA), the most difficult category of polyps to segment correctly. Weak-Sup-Mix reaches a dice score of 84.56 ± 2.38% on SSA segmentation, as in Table 2.5. The same model, but initialized with VGG16 weights instead of pre-training on the CVC-ClinicDB data (Weak-Sup-Mix-VGG), reaches a dice score of 83.66 ± 1.95%. As one would expect, the segmentation accuracy on SSA is worse than that on an average polyp, but it turns out that the difference in dice scores is only 1.0% between those two groups (t-test p-value 0.77).

Table 2.5: Average test-set scores and standard deviation using 10-fold cross-validation on the 90 sessile serrated adenoma images with human-labeled segmentation masks.

	Weak-Sup-Mix on SSA	**Weak-Sup-Mix-VGG on SSA**
Dice Coefficient	**84.56 ± 2.38%**	83.66 ± 1.95%
Accuracy	**98.23 ± 0.34%**	98.05 ± 0.39%

2.4.4 Weakly Supervised Models on CVC-ClinicDB

The model Weak-Sup-Mix-VGG is initialized with VGG16 weights and trained on bounding boxes annotation and human-annotated images and achieves an average dice score of 90.43 ± 0.43% using 10-fold cross-validation on the CVC-ClinicDB dataset (Weak-Sup-Mix-VGG on CVC in Table 2.6). Full-Sup-3 in Table 2.6 is the fully supervised model trained and evaluated on the CVC-ClinicDB dataset with 10-fold cross-validation. This model uses the weights of a fully trained Weak-Sup-Mix-VGG model as initial weights (see details in 4). This

transfer learning procedure allows Full-Sup-3 to utilize features learned from the Polyp-Box-Seg dataset and to obtain an average dice score of 91.79 ± 0.43% on the CVC-ClinicDB data, surpassing the best test score of 91.3% reported for MED-NET [73] as well as 91.5% reported for PolypSegNet [65].

Indeed, the current best-performing models on the CVC-ClinicDB dataset include MED-NET (dice coefficient 91.3%), PolypSegNet (dice coefficient 91.5%), and FANet (dice coefficient 93.55%) [73, 65, 100]. However, several issues with these three approaches must be considered. First, the images in the CVC-ClinicDB dataset are highly correlated with each other, as they correspond to consecutive video frames. In contrast, the Polyp-Box-Seg dataset contains 4070 unique polyp images, and thus, these images are highly uncorrelated with each other. Additionally, CVC-ClinicDB contains ordinary and clean polyps with few artifacts, while the Polyp-Box-Seg data includes challenging polyps with additional features such as forceps, snares, debris, and fluid, to name a few. Second, the results reported for MED-NET and FANet are obtained on a fixed test set containing 20% of the CVC-ClinicDB dataset, without varying this test set and without doing any systematic cross-validation experiments. In all our experiments, we see differences in test scores of up to 3% using different, randomly selected, test sets. Thus, the reported results come with no error bars and could provide overly optimistic performance estimates. And finally, these models have relatively larger inference times due to their complex architectures. For instance, MED-NET consists of an ensemble of several deep encoder–decoder models, which significantly increases its inference time, preventing deployment on video data and real-time applications. While PolypSegNet is faster and can process up to 25 frames per second, this is barely enough for video processing, and slower than the models presented here.

Table 2.6: Average test-set scores and standard deviation using 10-fold cross-validation on CVC-ClinicDB from the Weak-Sup-Mix-VGG model and from the Full-Sup-3 model as described in text.

	Weak-Sup-Mix-VGG on CVC	**Full-Sup-3 on CVC**
Dice Coefficient	90.43 ± 0.43%	**91.79 ± 0.43%**
Accuracy	98.87 ± 0.08%	**99.06 ± 0.07%**

2.4.5 Weakly Supervised Models on Kvasir-SEG

The model Weak-Sup-Box-HI uses weights were pre-trained on CVC-ClinincDB data as initial training weights (as described in 3c) and therefore cannot be evaluated on the CVC-ClinicDB dataset. This model is instead evaluated on the public Kvasir-SEG dataset using a random 90/10 training/testing split. Weak-Sup-Box-HI uses bounding box annotations as the only source of supervision and trains in the iterative manner. This weakly supervised model achieves a dice coefficient of 82.81% on the testing set, as shown in Table 2.7. The fully supervised ResUnet model used in the Kvasir-SEG paper achieved a dice coefficient of 78.77% on the testing set [46]. A semi-supervised model proposed by Wu et al. achieved a dice coefficient of 80.95% when training with 30% mask-labeled images and 70% unlabeled images [112]. Compared with one of the current best-performing models, PolypSegNet, which has a 88.7% dice coefficient testing on Kvasir-SEG, Weak-Sup-Box-HI reaches a competitive score with a faster inference time and only requires bounding box labels for training.

Table 2.7: Test-set scores on Kvasir-SEG from the Weak-Sup-Box-HI model.

	Weak-Sup-Box-HI on Kvasir-SEG
Dice Coefficient	82.81%
Accuracy	95.41%

All of the trained models above use the same underlying U-Net architecture, which is capable of processing video data at 45 frames per second (fps) using a Titan RTX GPU. Thus, these models can be easily deployed in real-time colonoscopy screenings, even when using a cheaper consumer-grade GPU. A video demonstrating real-time polyp segmentation in colonoscopy can be found at `www.igb.uci.edu/colonoscopy/AI_for_GI2.html`. Training and validation losses for each model, and the confusion matrix for each model predictions can be found in supplementary table S1 and S2.

2.5 Discussion

Several convolutional neural networks were applied to the task of real-time polyp segmentation using either only bounding box annotations as supervision source and/or human annotations of the polyp contours. Bounding box annotations are relatively easy to obtain but contain much less information compared to human-labeled segmentation masks. Nevertheless, our proposed iterative weakly supervised training procedure enables training models on bounding boxes that reach a competitive performance level (81.36% dice coefficient) compared with fully supervised models trained on human-labeled segmentation masks (81.52% dice coefficient). Moreover, our combined training approach on human-labeled segmentation masks and bounding box annotations further improves the model's polyp segmentation accuracy (85.53% dice coefficient). The proposed model performs well in the most challenging

cases corresponding to the segmentation of sessile serrated adenomas (84.56% dice coefficient), and when tested on the public CVC-ClinicDB, our model is comparable with other state-of-the-art approaches. Furthermore, our models can process video data at 45 fps and thus can be easily deployed in real-world colonoscopy videos.

The proposed weakly supervised training algorithm can greatly reduce the cost of annotating large volumes of data by human experts, thus resolving a major limitation in obtaining large amounts of training data for semantic segmentation. Instead of annotating segmentation masks for all images, researchers can simply use bounding box annotations as supervision source or only annotate segmentation masks for a small fraction of images and can train models on a mixture of ground truth masks and bounding box annotations.

Computer vision for object detection entails a hierarchy of at least four different tasks: (1) the detection of an object (presence/absence), (2) its localization (bounding box), (3) its segmentation (contour), and (4) its representation in three dimensions (3D model). In a previous study, we tackled the problem of using deep learning for real-time polyp detection and localization with bounding boxes in colonoscopy videos [105]. The newly proposed system extends this approach to the prediction of detailed pixel-level segmentations of polyps in real time, while also avoiding the problem of increased human annotation effort. Thus, the first three problems in the hierarchy are largely solved. A possible future extension is to tackle the fourth problem and to develop a multi-view stereo (MVS) model to reconstruct the 3-dimensional structure of any polyp: for instance, by combining the current system's predicted 2D segmentations with depth estimation derived from multiple viewpoints of the same polyp [31, 89]. Whether the problem can be solved without using multiple viewpoint polyp images or using the limited and constrained multiple viewpoints provided by existing colonoscopy video frames, and whether it requires new carefully acquired data are open questions that need to be investigated.

2.6 Conclusions

To date, several AI-based tools have been designed to improve the quality of colonoscopy and to reduce the rate of post-screening CRCs by reducing the amount of polyp missed during screening. The proposed polyp segmentation approach can assist colonoscopists by drawing contours around polyps in real time. It also brings the technology one step closer toward partially or even fully automated colonoscopy, especially in the context of capsule colonoscopy, where it could be used to detect polyps and to generate automated reports. These reports could include size and volume information to guide clinical interventions, such as automated surgical polyp excision.

Chapter 3

Deep Learning to Enable Color Vision in the Dark

3.1 Introduction: Deep Learning Applications in Vision

The human eye perceives light in the visible spectrum using opsin proteins bound to a light-sensitive chromophore in retinal photoreceptors. The peak spectral opsin absorbances with surrounding Gaussian distributions for rods, S-cones, M-cones, and L-cones are: 498 nm, 420 nm, 534 nm, and 564 nm respectively [8]. Cone photoreceptors perceive color and are responsible for high acuity vision with rapid phototransduction kinetics. Rod photoreceptors are sensitive to low light conditions offering night vision, but with phototransduction kinetics an order of magnitude slower than cones [87]. Therefore, cones enable the daytime vision of quickly or slowly changing visual scenes, and rods provide night vision but with lower temporal resolution. The peak spectral absorbance for each opsin class defines what is termed the visible spectrum because humans vision occurs in the range of 400-700 nm. Humans do

not perceive scenes outside the visible spectrum. In this work, we sought to evaluate the ability of deep learning (DL) coupled with infrared spectroscopy to render visible spectrum images using infrared light illumination and no visible spectrum light. Night vision systems seek to illuminate the environment with infrared (IR) light that is detected and visualized by conventional camera sensors and render a scene in the visible spectrum on a digital display. Historically, night vision systems render scenes as a monochromatic green display. Newer night vision systems use ultrasensitive cameras to detect and amplify visible light. Computer vision tasks with low illuminance imaging have employed image enhancement and deep learning to aid in object detection and characterization from IR spectrum, but not with accurate interpretation of the same scene in the visible spectrum [114]. Conventional cameras acquire blue (B), green (G), or red (R) pixels of data to produce a color image perceptible to the human eye (3.1 top row). We investigated if a combination of infrared illuminants in the red and near-infrared (NIR) spectrum could be processed using deep learning to recompose an image with the same appearance as if it were visualized with visible spectrum light (3.1 bottom row). We established a controlled visual context with limited pigments to test our hypothesis that DL can render human-visible scenes using NIR illumination that is, otherwise, invisible to the human eye.

Modern color printers produce color images using 4 ink colors: cyan (C), magenta (M), yellow (Y), and black (K). Under broad visible light and human perception, each of these ink colors appears cyan, magenta, yellow or black based on the incident light that they reflect. Spectral reflectance is the percent of light energy that a compound reflects for each specific wavelength of illuminating light. Ink dye's spectral reflectance is determined by systematically illuminating the dyes across a range of the electromagnetic spectrum and measuring the reflected light. Therefore, printed images from a CMYK printer provide a simplified context to test our hypothesis that NIR illuminated images can be processed to render visible spectrum images that match the images illuminated with visible light.

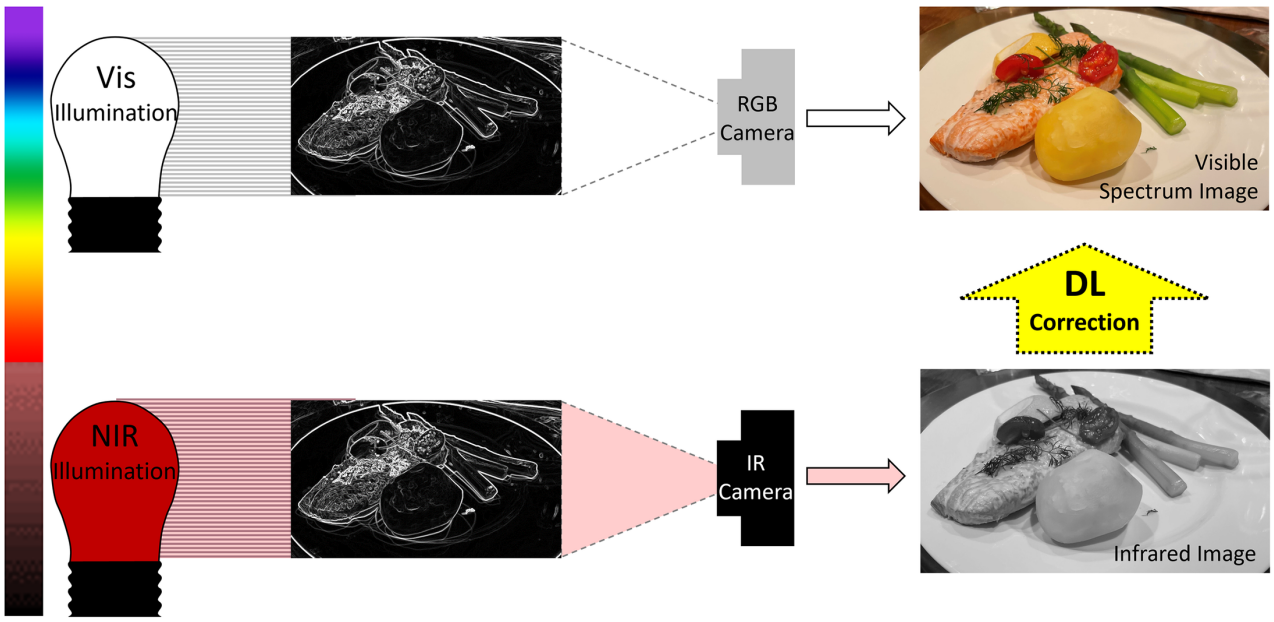


Figure 3.1: Deep Learning enable vision from infrared illuminants.

3.2 Overview and Challenges in Deep Learning Application in Vision

The NIR image colorization problem shares some similarities with the general image colorization and color transfer problems [16, 117, 43] with several significant distinctions. In grayscale-to-RGB reconstruction, luminance, i.e. the intensity of light emitted from a surface, can be directly obtained from the grayscale images so only the chrominance needs to be estimated. This greatly simplifies the reconstruction task. Moreover, in color transfer problems, the input and output are both three-channel RGB images, whereas NIR images have only a single channel making the color reconstruction more challenging. Early image colorization techniques required human input to correctly identify similar neighborhood intensity values and assign them a similar color, i.e Scribble [58] or Similar images [110, 42]. Other approaches [14, 77, 21] automatically matched patches of the input image to patches of reference image via feature extraction and matching. Despite being fully automatic, these methods require a fitting reference image database and retrieval mechanism to be imple-

mented. More recent approaches [53, 41] leveraged deep neural networks to propose a single framework where image colorization can be learned end-to-end.

Other groups have explored colorizing IR images by several approaches. Toet employed a look-up-table approach where first order statistics were used to develop a color transfer method to evaluate luminance distributions in monochromatic images (such as from Infrared cameras) and assign chromaticity values to each pixel [98, 99]. Liu et al. applied two stages of DL to first apply textural details to a thermal image acquired using a single IR camera. They subsequently added color information to the texturalized thermal image [61]. Zheng et al. studied qualitative and quantitative metrics for night vision colorization and identified computationally rapid colorization by a combination of histogram matching and statistical matching between source IR image and a target color image [118]. They identified trade-offs in contrast and saturation at the expense of computational time. Suarez et al. employed Deep Convolutional Generative Adversarial Network to colorize single-channel near IR images from city scenery and succeeded in colorization at the expense of image clarity [96]. Mehri et al. proposed a Cycle-Consistent adversarial network to predict color channels from a single NIR input on unpaired dataset [68]. Each of these infrared to color image techniques rely on single IR image input for images acquired in the real world. Brown and Susstrunk demonstrated improved scene category recognition when color images of scenes were supplemented by a NIR image channel [10]. We, therefore, sought to systematically study infrared colorization for multiple infrared wavelengths in a controlled imaging context.

Deep neural networks, particularly Convolutional Neural Networks (CNNs), have achieved state-of-the-art performances in a variety of computer vision problems. CNNs, originally inspired by the seminal work of Hubel and Wiesel [38], were developed by Fukushima [25], using however the wrong learning algorithm. They were perfected in the 1980s-1990s using the backpropagation learning algorithm and are used today across most computer vision tasks [52, 22], including biomedical imaging tasks, with thousands of references in the liter-

ature. Recognition CNNs convolve the input image through a hierarchical series of adaptive filters that are learned from the data and produce increasingly more abstract representations of the data supporting translation invariant recognition. Generative CNNs operate in the reverse direction and both recognition and generation networks can be combined in various architectures. In particular, in this work we use architectures inspired by the U-Net architecture [84] in which a contracting path is combined with an expansive path (yielding a "U" shape architecture). The U-Net architecture has been used effectively in several tasks, from image segmentation to photoacoustic imaging reconstruction [84, 23, 28].

3.3 Methods

3.3.1 Multispectral image acquisition

To learn the spectral reflectance for cyan, magenta and yellow ink, we printed a rainboy color pallete using a using a Canon office printer with CMYK ink. Examples of the color palette under multispectral illumination (center wavelengths: 397, 408, 427, 447, 466, 480, 495, 507, 519, 529, 530, 562, 594, 604, 618, 630, 636, 660, 665, 692, 697, 718, 734, 748, 760, 777, 807, 854, 910, 950, 971 nm) is shown in 3.2. Photographs of each printed image under multispectral illumination were acquired using a monochromatic camera (Zeiss Axiocam 305, CarlZeiss) mounted on a dissection microscope focused on the image. To account for diminishing camera quantum efficiency in the red and infrared spectrum, each image channel was normalized to the white background of the paper upon which the images were printed. For both panels in 3.2, the right inset is an image combined by assigning the 447 nm, 529 nm, and 604 nm images to the red, green, and blue channels respectively. Channel assignment and merging were performed in imageJ [85]. Spectral reflectance was determined for each C, M, Y printer dye using a normalized plot profile function (imageJ) for each of the 31 LED

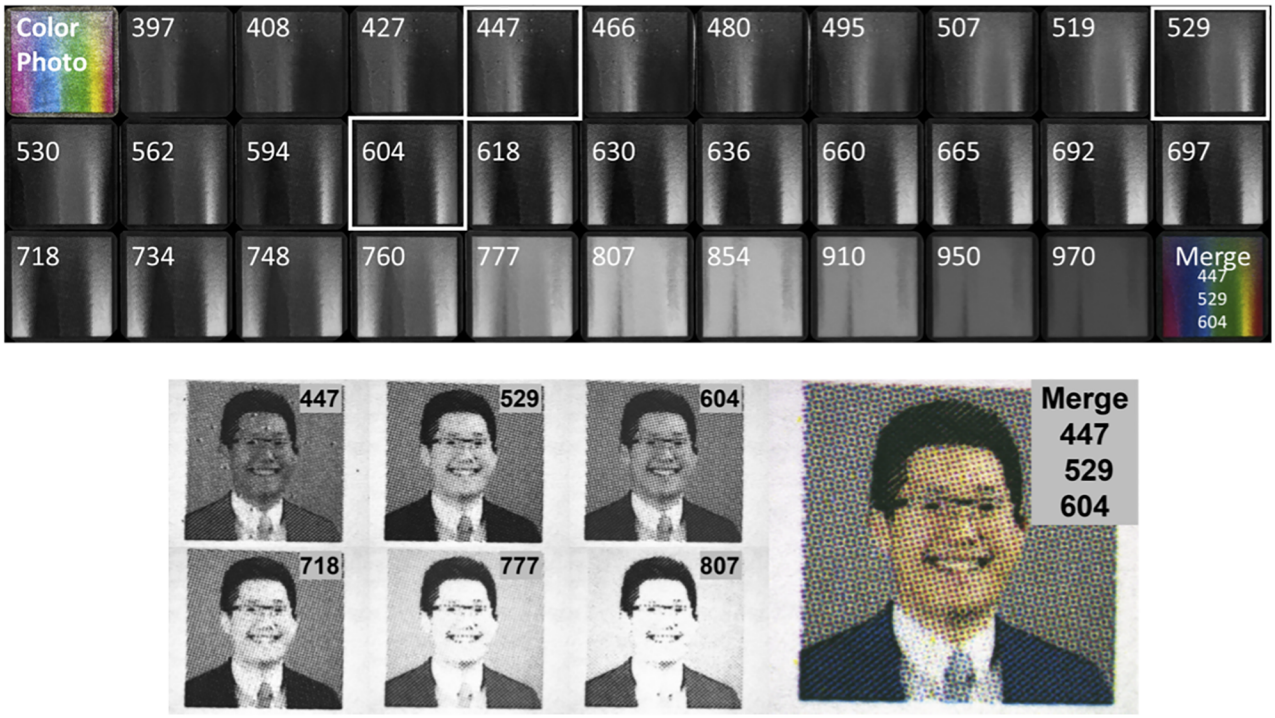


Figure 3.2: Sample images from the human portraits library.

illuminants.

We printed a library of over 200 human faces available from the publicly available "Labeled Faces in the Wild" repository (`http://vis-www.cs.umass.edu/lfw/`) [37]. The images were printed using a using a Canon office printer with CMYK ink. We acquired photographs of each image under different wavelengths of illumination to then be used to train machine learning models to predict RGB color images from individual or combinations of single wavelength illuminated images. Each image in the dataset was illuminated with diffuse uniform illumination by 6 different LED wavelengths spanning the visible and near-infrared spectrum (center wavelengths: 447, 529, 604, 718, 777, 807 nm). The datasets are available for downloading (`https://dx.doi.org/10.6084/m9.figshare.19291166`). The individual pictured in manuscript Figures 2,3 and 9 provided written informed consent (as outlined in PLOS consent form) to publish their image alongside the manuscript.

3.3.2 Baseline and network architectures

To predict RGB color images from individual or combinations of wavelength illuminations, we evaluated the performance of the following architectures: a baseline linear regression, a U-Net inspired CNN (UNet), and a U-Net augmented with adversarial loss (UNet-GAN).

Linear regression baseline

We implemented a simple linear regression model as a baseline to compare the results of the deep neural architectures and assess their performances. As input to the linear model, we evaluated patches of several sizes and predicted target patches separately for each color channel (R, G, and B).

Conventional U-Net

A traditional U-Net is a fully convolutional network that consists of a down-sampling path (encoder) that helps to capture the context and an up-sampling path (decoder) that is responsible for precise localization. The down-sampling path (left side of 3.3) repeatedly applies a series of operations which consists of three 3×3 convolutions followed by rectified linear unit (ReLU) activation and a 2×2 max pooling operation. At the end of one series of operations, the number of feature maps is doubled and the size of feature maps is halved. The up-sampling path (right side of 3.3) repeatedly performs another series of operations consisting of an up-convolution operation of the feature maps by a 2×2 filter, a concatenation with the corresponding feature map from the down-sampling path, and two 3×3 convolutions, each followed by a ReLU activation. In the output layer, a 3×3 convolution is performed with a logistic activation function to produce the output. Padding is used to ensure that the size of output matches the size of the input.

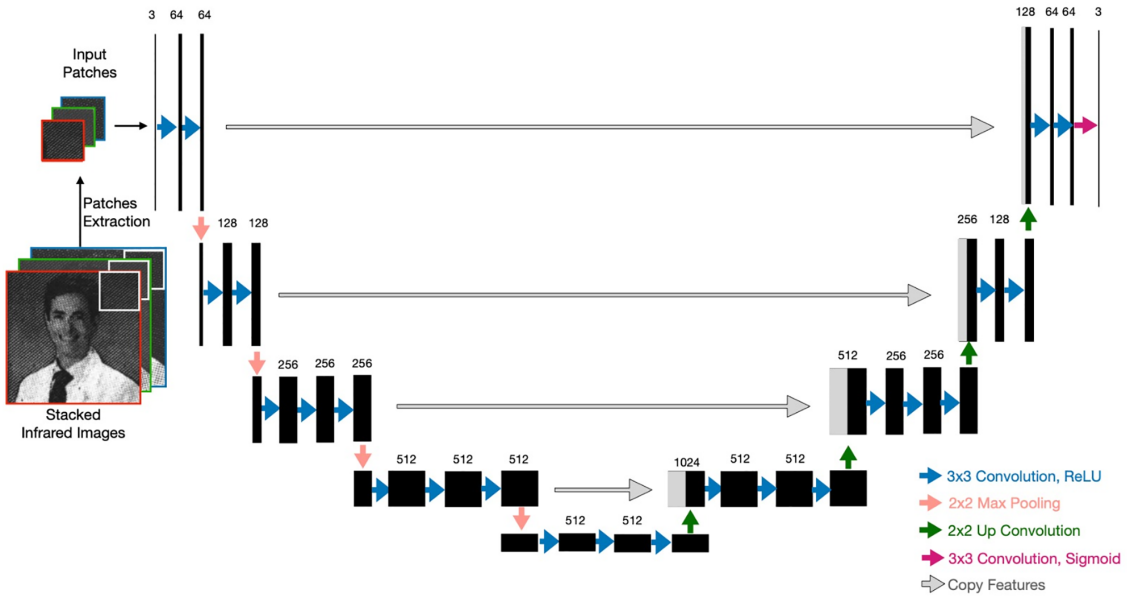


Figure 3.3: U-Net architecture for infrared colorization.

VGG network as a U-Net encoder (UNet)

VGG neural networks [92], named after the Visual Geometry Group that first developed and trained them, were initially introduced to investigate the effect of the convolutional network depth on its accuracy in large-scale image recognition settings. There are several variants of VGG networks like VGG11, VGG16, and VGG19, consisting of 11, 16, and 19 layers of depth, respectively. VGG networks process input images through a stack of convolutional layers with a fixed filter size of 3×3 and a stride of 1. Convolutional layers are alternated with max-pooling filters to down-sample the input representation. In [40], using VGG11 with weights pre-trained on ImageNet as U-Net encoder [91] was reported to improve the performance in binary segmentation. In this study, we tried a U-Net architecture with a VGG16 encoder [93] using both randomly initialized weights and weights pre-trained on ImageNet. We refer to this architecture simply as U-Net.

U-Net with adversarial objective (U-Net-GAN)

To enhance the performance of the U-Net architecture, we also explored a Conditional Generative Adversarial Network approach (CGAN). A conventional CGAN consists of two main components: a generator and a discriminator. The task of the generator is to produce an image indistinguishable from a real image and "fool" the discriminator. The task of the discriminator is to distinguish between real and fake images produced by the generator, given the reference input images. In this work, we examined the performances of two U-Net-based architectures: 1) a stand-alone U-Net implementation as described in the previous section and 2) a CGAN architecture with a U-Net generator along with an adversarial discriminator. For the second approach, we used a PatchGAN discriminator as described in [43], which learns to determine whether an image is fake or real by looking at local patches of 70×70 pixels, rather than the entire image. This is advantageous because a smaller PatchGAN has fewer parameters, runs faster, and can be applied on arbitrarily large images. We refer to this latter architecture (U-Net generator and PatchGAN discriminator) as U-Net-GAN.

3.4 Experimental settings and training

For all the experiments, following standard machine learning practice, we divided the dataset into 3 parts and reserved 140 images for training, 40 for validation and 20 for testing. To compare performances between different models, we evaluated several common metrics for image reconstruction including Mean Square Error (MSE), Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), Angular Error (AE), DeltaE and Frechet Inception Distance (FID). FID is a metric that determines how distant real and generated images are in terms of feature vectors calculated using the Inception v3 classification model [35]. Lower FID scores usually indicate higher image quality. Since FID has been used in many image generation tasks including NIR colorization study [68]. However, Mehri et

al. used a single-channel NIR input whereas we feed three stacked images of different NIR wavelengths. Therefore, we cannot entirely compare FID values between our studies as the settings are different. We report FID in the main manuscript and comprehensively present additional metrics in the Supplementary section.

3.4.1 Linear regression baseline

For the linear regression model, the input images were divided into patches. We explored several patch sizes of $6 \times 6, 12 \times 12, 24 \times 24$ and 64×64 pixels. Every patch from the selected input infrared images was used to predict the corresponding R, G, and B patch of the target image. A different linear model was used for each color channel and the results for one image were computed as the average over the three channels.

3.4.2 U-Net and U-Net-GAN

To train the proposed deep architectures, we divided the input images into random patches of size 256×256. Random cropping of the patches was used as a data augmentation technique. Both inputs and outputs were normalized to $[-1, 1]$. The deep architectures were trained for 100000 iterations with a learning rate starting at 1×10^{-4} and cosine learning decay. Given the fully convolutional nature of the proposed architectures, the entire images of size 2048×2048 were fed for prediction at inference time. As a loss function for neural networks, i.e. U-Net and U-Net-GAN, we used mean absolute error (MAE).

3.4.3 Human grader evaluation of model performance

We performed a human evaluation of model performance. We randomly sampled 10 images (ground truth and the 3 prediction output by UNet-GAN, UNet and Linear regression). We provided a blinded multiple choice test to 5 graders to select which of the 3 outputs subjectively appeared most similar to the ground truth image. We subsequently randomly selected 10 patches (portion of images) and asked graders to compare similarity of predicted patches from UNet-GAN and UNet with the corresponding ground truth patch.

3.5 Results

In this section, we first describe our spectral reflectance findings for CMY printed inks imaged under different illuminant wavelengths and then report and discuss the performances for the different architectures to accurately predict RGB images from input images acquired by single and combinations of different wavelength illumination.

3.5.1 CMY spectral reflectance

To identify the range of wavelength with the most distinct reflectance profile, we conducted CMY spectral reflectance analysis. Inluminant wavelengths with spectral reflectances that are unique for each ink color were assumed to contain the most valuable information for color reconstruction.

Spectral reflectance of cyan, magenta, and yellow inks was determined by multispectral illumination and plotted according to the detected grayscale value (3.4). The top inset demonstrates the grayscale reflectance values for each ink with ascending wavelength of illumination (total of 31 wavelengths of illumination). The plot in 3.4 is the normalized

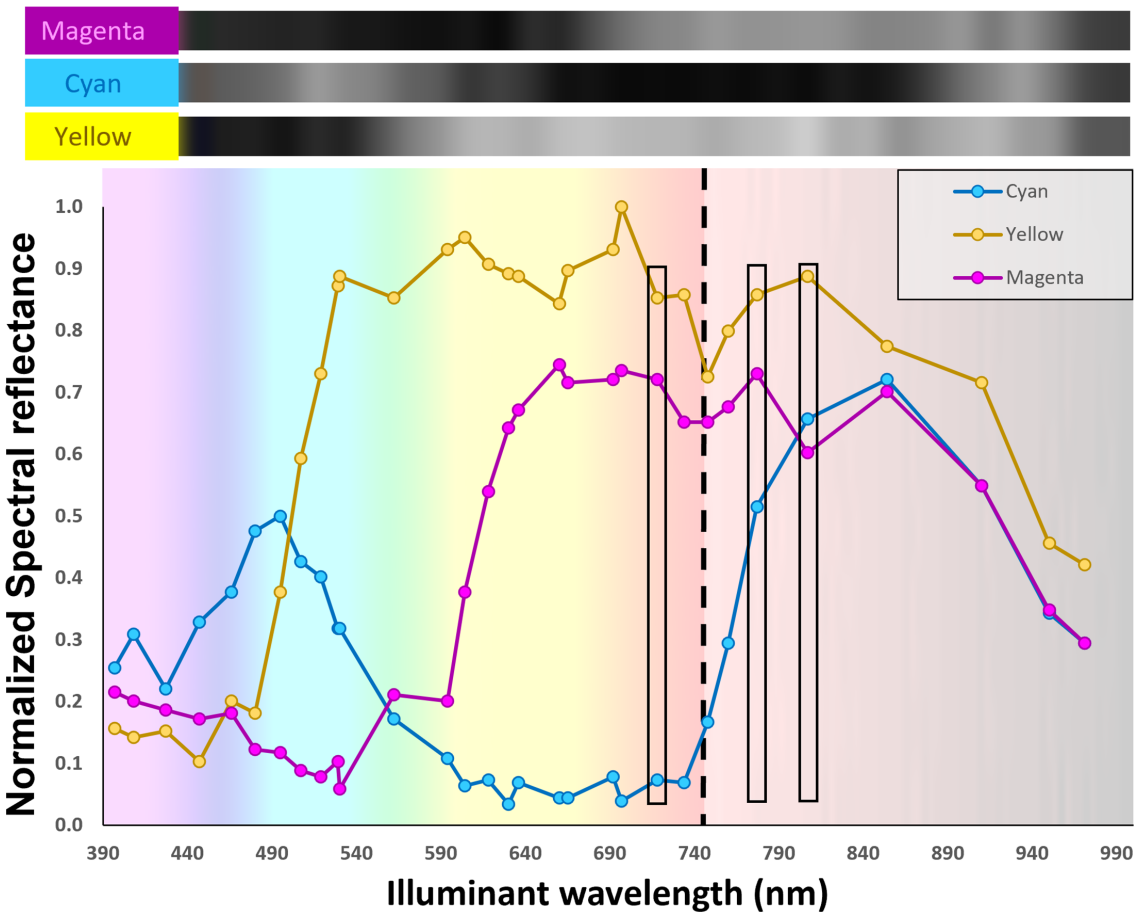


Figure 3.4: Spectral reflection analysis of cyan, magenta, and yellow inks illuminated with 31 LED channels spanning visible to near-infrared illumination.

quantified reflectance across visible and NIR spectra. Cyan and magenta inks have a peak reflectance with the 495 and 660 nm LED illuminants respectively. Yellow ink had a broader reflectance with peak reflectance for 604 and 697 nm illuminants. Cyan ink demonstrated an isolated reflectance maximum in the NIR spectrum around 854 nm, while yellow and magenta inks demonstrated broad reflectance in the NIR spectrum. Spectral reflectance profiles were distinct for each ink color, but cyan and magenta demonstrated overlapping reflectance profiles above 854 nm. Spectral reflectance demonstrated no overlap for 718, 777 and 807nm (black box outlines in 3.4).

3.5.2 Infrared to RGB reconstruction

RGB reconstruction

Since transfer learning has been widely used to improve deep model performances and reduce the risk of overfitting, we explored four slight variations of the main learning architectures. That is, using the three selected wavelengths (718, 777, and 807), we evaluated a UNet-GAN and a UNet architecture, both with or without ImageNet pretrained weights. Quantitative results are shown in 3.5, where the x-axis denotes the number of iterations and the y-axis the FID scores. From 3.5, it is observed that the adverserial objective enables predictions to reach lower FID scores. Furthermore, the use of ImageNet pretraining (3.5 does not boost the performances, leading in some cases to higher FID scores. A possible explanation could be the difference in the domain of the ImageNet dataset, which does not contain a lot of human images, while human portraits used in the current study. The experimentation that followed did not include ImageNet pretrained weights. To further qualitatively assess the difference in performances and to validate the use of the FID as a metric reflective of human perception, 3.6 shows the reconstruction of a ground truth RGB patch (*right*) by UNet (*left*) and UNet-GAN (*middle*). From visual inspection of 3.6, it is evident that adopting adverserial loss results in significantly sharper reconstructions which are much closer to the ground truth both in terms of colors and patterns.

To determine if three input wavelengths (718, 777, and 807) provided benefit over single or doublet wavelength inputs, we evaluated the UNet-GAN architecture using all the possible combinations of two wavelengths, three wavelengths, and single infrared illuminations as inputs. The quantitative results of these experiments are shown in 3.7. From 3.7, it appears that using single infrared illuminations generally leads to poor reconstructions, while combining two infrared inputs, like 718 and 777 or 718 and 807, is useful to improve performances. The use of the three selected wavelengths results in the lowest FID score, which is

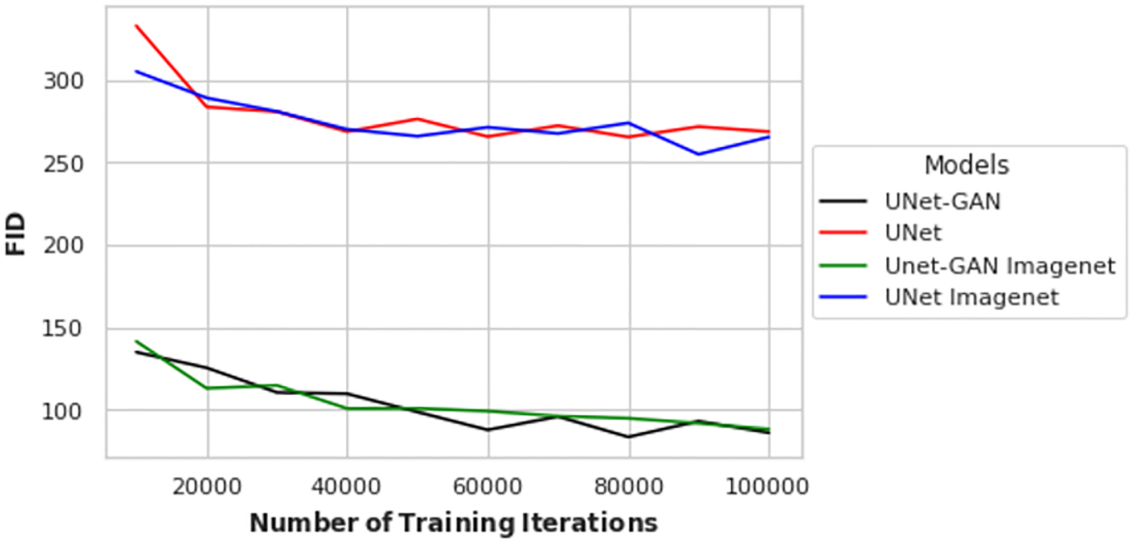


Figure 3.5: Comparison of reconstruction performances.

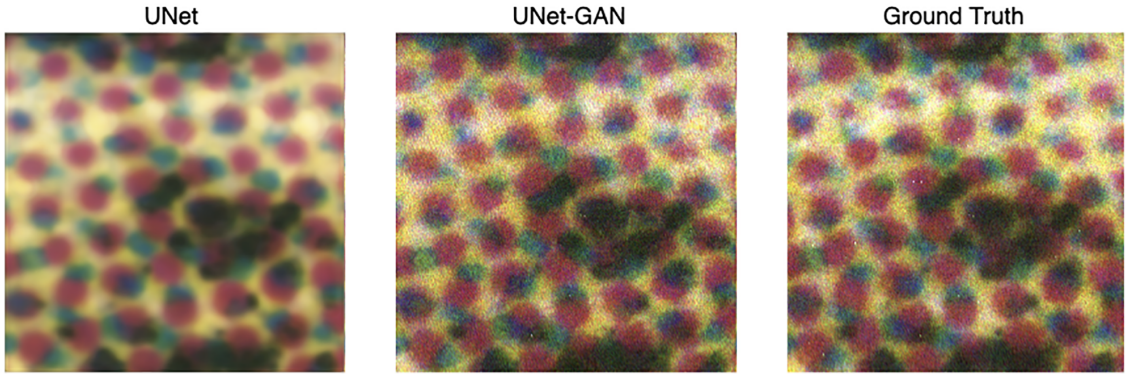


Figure 3.6: Visualization of reconstructed RGB patch.

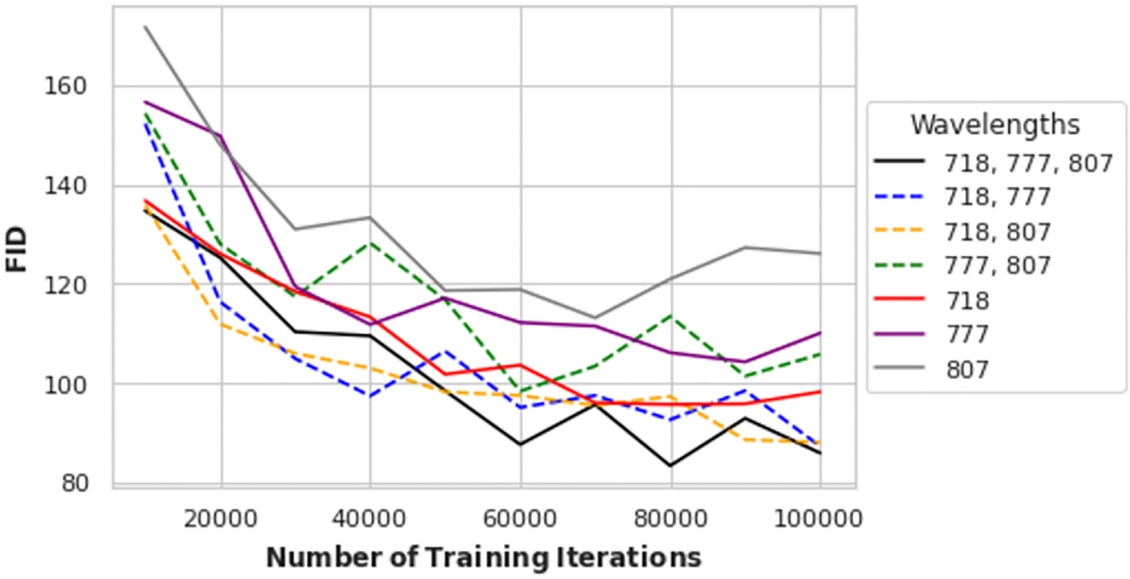


Figure 3.7: Reconstruction performances with different input illuminations.

reached at 80000 iterations. For further evaluations we used the models at 80000 iterations.

To compare the deep architectures with a baseline, we assessed the performances of UNet, UNet-GAN at 80000^{th} iterations and a linear regression model on the same test set composed of 20 images. The quantitative results are shown in 3.8. The simple linear regression baseline failed to accurately reconstruct the ground truth RGB images and results in the highest FID scores. Furthermore, as qualitatively presented in 3.6, the use of adversarial loss allows for sharper reconstructions and results in a much lower FID score (140.8 vs 87.2). Qualitative comparison of RGB reconstructions are presented in 3.9. Simple linear regression model produces images lacking color information seen in ground truth images and images produced from UNet and UNet-GAN dee architectures. While the UNet and UNet-GAN reconstructions appear similar upon gross inspection of recontructed images, the patch analysis shown in 3.6 demonstrates the qualitatively closer match between the adversarial network and ground truth than the non-adversarial network and ground truth. The qualitative observations are mirrored by the lower FID score of 87.2 seen with the adversarial network (3.8).

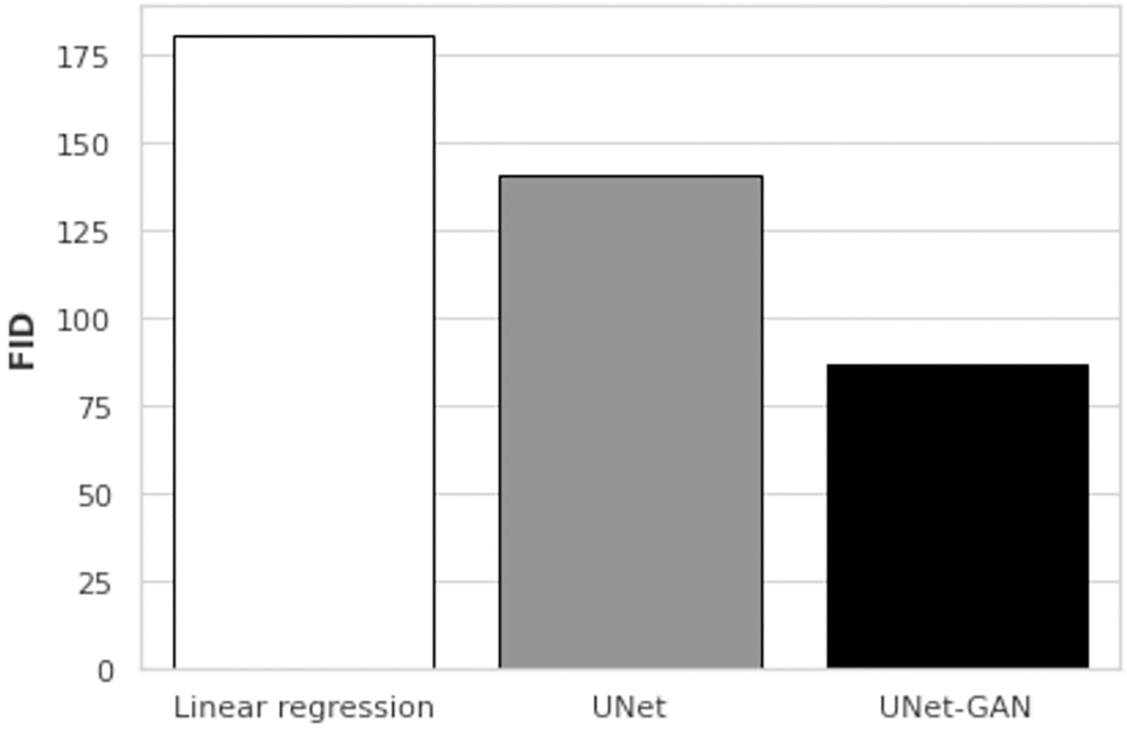


Figure 3.8: Models comparison with a regression baseline.

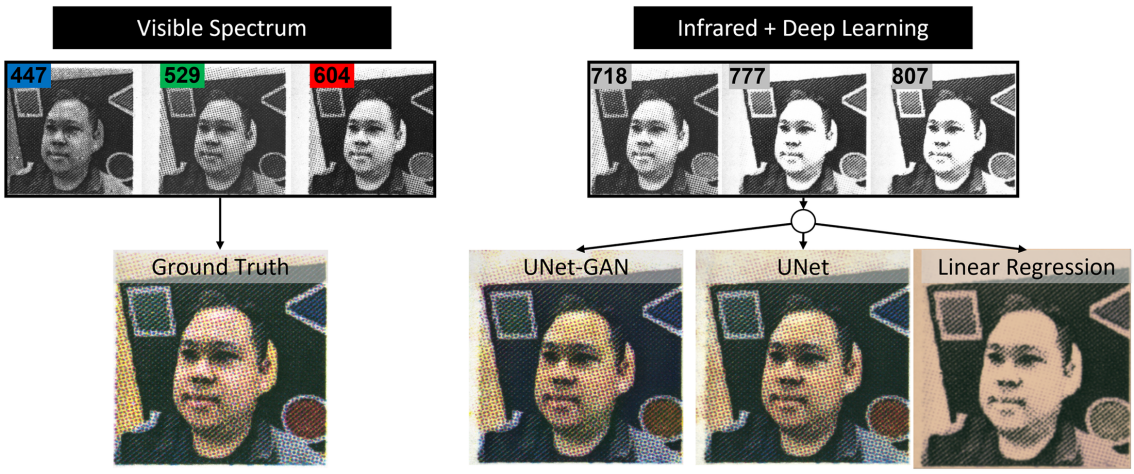


Figure 3.9: Visualization of the reconstructed RGB images by the deep architectures and baseline regression model.

Finally, human subject comparison of model outputs with ground truth inputs was used to determine perceived model performance. In all cases, the graders determined that linear regression had worse performance than UNet-GAN or UNet. In 52% of cases, human graders determined that UNet outperformed UNet-GAN. However, when zooming into predicted images at the patch level, UNet-GAN was perceived to out-perform UNet 100% of the time. UNet patch predictions, while similar in color content, were more blurry than UNet-GAN patches.

3.6 Discussion and conclusion

We have investigated the reconstruction of RGB images, given their representation under infrared illuminations, using deep neural networks. We have evaluated two slight variations of deep architectures along with a linear regression baseline model and different weights initialization techniques using an in-house generated dataset. The deep U-Net-based architectures proved able to generate visually consistent RGB reconstructions using only three input infrared images. These reconstructions were significantly better than the ones produced by the linear model, both quantitatively and qualitatively through visual inspection. The use of adversarial training was useful to produce significant improvements, often resulting in lower FID scores. Human subject evaluation consistently determined that UNet and UNet-GAN models outperformed linear regression, and UNet moderately outperformed Unet-GAN at colorizing infrared images. This demonstrates the limitations of objective computational metrics, like FID, to fully describe human perception.

This work suggests that predicting high-resolution images is more dependent upon training context than on spectroscopic signatures for each ink. In this work, four compounds (C, M, Y, K inks) were characterized spectroscopically using a monochromatic camera sensor in the contexts of rainbow input images and naturally occurring images. One might hypoth-

esize that larger separation of in spectral reflectance curves for each ink would aid CNNs to predict a color image with lower error because curve overlap could introduce represent spectroscopically indistinguishable information. For example, spectral reflectance curves for CMY inks at 718 and 777 nm illumination demonstrate significantly different curve separations (3.4). For 718 nm illumination, the reflectance of cyan, magenta, and yellow are 0.1, 0.7, and 0.9 respectively. For the 777 nm illuminant, the spectral reflectances are much closer in value with cyan, magenta, and yellow inks having normalized reflectances of 0.5, 0.7, and 0.9 respectively.

Another important aspect we aimed to address was the speed of the predictions. To enable real-time color vision in complete darkness for video data, the inference time should be small enough so that a sufficient number of frames can be processed each second. The proposed U-Net architectures are capable of producing three images per second, without using any particular accelerations. We tried several generator backbones and reported the inference time in SFig. 1. This provides a starting point for further speed improvements that could use, for instance, smaller architectures, or multi-threading, or better hardware.

This proof-of-principle study using printed images with a limited optical pigment context supports the notion that the deep learning approach could be refined and deployed in more practical scenarios. These scenarios include vision applications where little visible light is present either by necessity or by goal. Night vision, for example, has applications in security, military operations, and animal observation. Similarly, handling, processing, and studying samples sensitive to visible light may require a technology that uses no visible light. Studying the light-sensitive retinal tissue, for example, may require processing the sample in darkness to avoid altering its biochemistry and function [67, 76]. Likewise, performing eye surgery benefits from low light exposure to avoid retinal damage [116]. Similarly, handling restoration of light-sensitive artifacts may benefit from minimizing exposure to visible light.

In short, this study suggests that CNNs are capable of producing color reconstructions

starting from infrared-illuminated images, taken at different infrared wavelengths invisible to humans. Thus, it supports the impetus to develop infrared visualization systems to aid in a variety of applications where visible light is absent or not suitable.

Chapter 4

Transcriptomic Organization of Circadian Rhythms

4.1 Introduction: Detecting Circadian Rhythm in High-throughput Omic Data

The introduction of high-throughput sequencing technologies has revolutionized the field of bioinformatics. The high throughput gives biologists the ability to measure biological features such as gene expression on a genome-wide scale in a short time frame. Common high-throughput sequencing applications include transcriptome sequencing (RNA-Seq), chromatin immunoprecipitation sequencing (ChIP-seq), methylation sequencing (Methyl-seq), and ribosomal sequencing (Ribo-seq) [17]. RNA-seq is used to assess the gene expressions on the transcriptome in different experiment conditions. ChIP-seq is used to assess epigenetic modifications in the DNA regions, by performing chromatin immunoprecipitations with antibodies towards various histone modification states and sequencing the resulting immunopreciptated DNA fragments. After the sequencing is completed, the level of gene

or metabolite expression is calculated, and further analyses such as differential analysis, linear regression, and analysis using deep learning models, are carried out to determine the biological features detected in each experimental setting.

Even though the costs of carrying out biological experiments have been reduced thanks to the advances in high-throughput sequencing technologies, it is still expensive for researchers to include a large number of samples (mice, rats, human subjects) in an experiment. For instance, it is typical to find 3-4 replicates per experiment group in a RNA-seq an ChIP-seq experiment, with about 22000 genes across the mouse genome measured. Therefore, the limited number of samples and replicates in each experiment group still poses a big challenge to data analysis in bioinfromatics.

4.2 Overview and Challenges in Detecting Circadian Rhythm

Detecting circadian rhythm has long been a topic of interest in research [51, 88]. Disrupted circadian rhythm are associated with health problems including cancer, diabetes, premature aging, and so on [79, 113]. In a circadian experiment, it is typical to find 3-4 high-throughput omic measurements taken at multiple time points (TP) along the circadian cycle under both control and treated conditions. This setting poses several challenges to detecting circadian rhythm in a transcriptomics or metabolic dataset: (1) the time points are sparse - measurements are taken at 4 hours, 6 hours, or an even longer duration apart; (2) the variability in each biological sample can be big; (3) the problem of a small sample size; and (4) the multi-testing problem of the large number of measurements (about 22000 genes) [2].

To study the circadian rhythms in all tissues and all species as a whole, we have processed and complied a large number of omic datasets from project collaborations, conducted circadian

rhythm detection with BIO_CYCLE and a commercially available software JTK_CYCLE [2, 39], and created a web portal named CircadiOmics for search and visualization. The collected datasets cover over 6400 individual time points spanning 227 separate circadian experiments. Eight species are currently available on CircadiOmics. The majority are collected from Mus musculus and Papio anibus [12]. In aggregate, these datasets form the largest single repository of circadian data available. CircadiOmics is available as a pubic domain website at `http://circadiomics.ics.uci.edu`. Examples of circadian oscillating genes from CircadioOmics can be found in Figure 4.1.

4.3 Results

To identify a model of organization for transcriptomic circadian oscillations, we perform a series of analyses of increasing complexity using novel computational metrics.

To achieve robustness and overcome noise in the data and incomplete knowledge, we present results obtained consistently at different statistical threshold as well as results that are supported by multiple lines of evidence. In total, 87 datasets from mouse were used to generate each set of results. The most represented tissues are liver (37 datasets), skin (14 datasets) and brain (13 datasets). In all cases, aggregated results, as well as tissue specific results, were generated.

4.3.1 Frequency Analysis

The frequency at which a TF or RBP is found to oscillate in a collection of datasets provides a simple metric for estimating its consistent role in circadian oscillation. Figure 4.2 illustrates this frequency distribution for mouse at a BIO_CYCLE p-value < 0.01. Additionally, 64 datasets from Papio anubis (baboon) were used for comparison to validate the

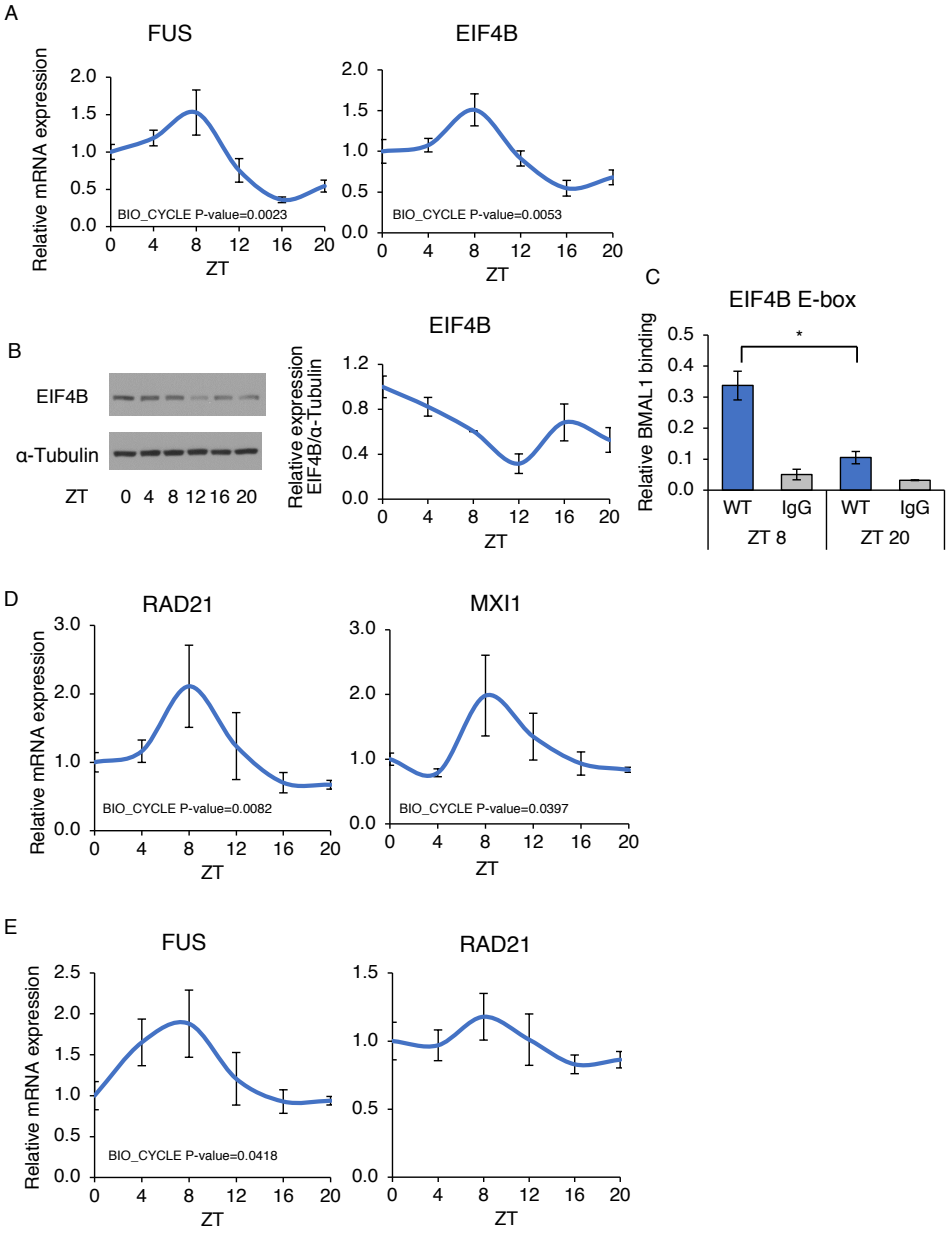
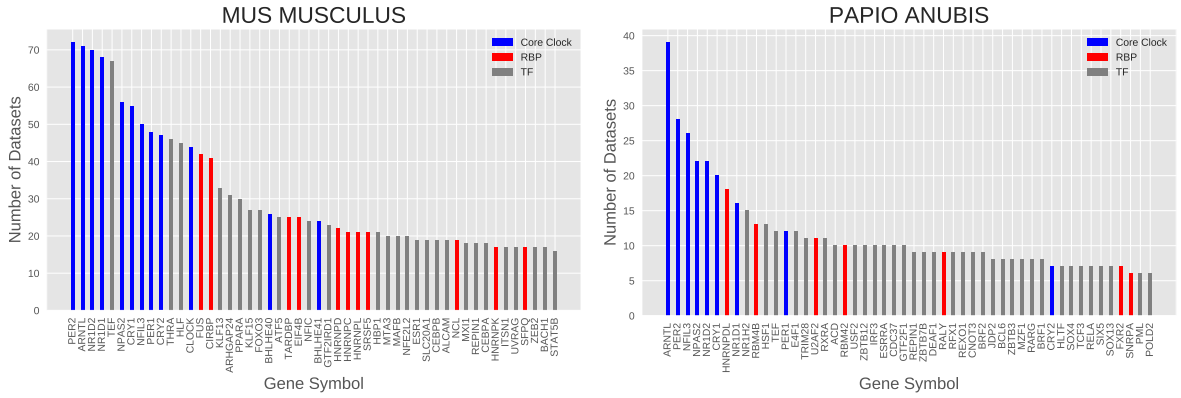


Figure 4.1: Examples showing circadian genetic expressions from CircadiOmics.

methods. Both analyses show that TFs involved in the circadian core clock are found to be the most frequently oscillating. This purely data-driven approach automatically discovers the circadian core clock. Furthermore, it identifies additional TFs and RBPs that must play an important role in circadian oscillation.



(a) 87 Transcriptome Datasets (b) 64 Transcriptome Datasets

Figure 4.2: Most Frequent Oscillating TFs and RBPs

4.3.2 The Circadian Regulatory Control Graphs

Measuring the circadian regulatory influence of the TFs and RBPs identified in the previous analysis requires further investigation using more sophisticated computational methods. To this end, a novel computational method was used to identify and score directed regulatory edges in oscillating loops. The Circadian Regulatory Control (CRC) method can be understood as a proxy for circadian regulation between a TF or RBP source and a transcript target. There are three major components of the CRC method applied to each experiment. First, as a prerequisite, the source and target must be oscillating, as assessed by BIO_CYCLE. Second, the source must have at least one high quality binding site on the target for transcriptional or post-transcriptional regulation, as assessed by MotifMap and MotifMap-RNA [115, 19, 62]. For a TF, binding sites were assessed at the promoter region of the target transcript. For an RBP, binding sites were assessed at the introns or UTRs of

the target transcript. Third, there must be a correlative relationship between the phases of the source and the target. Recent studies have shown a significant lag between the transcript expression and the concentration of the corresponding protein [82]. We addressed this issue by computing and modeling the distribution of this lag, using transcriptomic and proteomic datasets produced from the same study on CircadiOmics.

After filtering on p-value for the first criteria, the remaining two criteria were combined into two different CRC scores. The *B-score* is a binary indicator of circadian regulation at various filtering thresholds for the number of high quality binding sites and the likelihood of phase correlation. The *E-score* is an exponentially weighted combination of these two criteria. In general, results generated using both scores tend to agree. However, as a binary indicator, the *B-score* is more convenient for large scale analysis of graph structures. In contrast, as a real valued metric, the *E-score* has more sensitivity and is used for ranking nodes and edges.

In either case, the CRC method yields a CRC graph, which is a directed weighted graph, where the vertices correspond to circadian protein-coding transcripts in a given experiment, and the directed edges correspond to regulatory relationships with *E-score* > 0, and the weights correspond to the *E-score* or the *B-score*. The CRC graphs obtained for each experiment can be superimposed, for instance, across tissues or across all experiments, by aggregating the scores of the corresponding edges.

4.3.3 Node Influence Analysis

For each source associated with a TF or RBP, an influence score was computed by aggregating all the CRC E-scores from all its outgoing edges either in all experiments or in tissue-specific experiments. The highest scoring TFs and RBPs are shown in Table 4.3. When looking at aggregated results, core clock TFs such as CLOCK and BMAL1 were found to have the largest scores, a finding consistent with both the frequency of oscillation and previous

Mouse All (n=81)		
TF/RBP	All Score	All Ranking
FUS	8.83	9
THRA	8.44	10
BHLHE40	8.30	11
NFIC	8.09	12
HNRPDL	7.56	15
CIRBP	7.42	16
MXI1	7.35	17
EIF4B	7.34	18
CEBPB	6.65	20
HNRNPK	5.61	21
TARDBP	5.40	22
KLF13	4.75	23
PPARA	4.73	24
FOXO3	4.64	25
RAD21	4.56	26
KHDRBS1	4.38	27
ALCAM	4.30	28
MTA3	4.25	29
ARHGAP24	4.23	30
YY1	4.22	31
HNRNPL	4.15	32
NFYA	4.14	33
MAFK	4.12	34
KLF15	4.12	35
PCBP4	4.06	36
HNRNPC	4.06	37
ESR1	4.00	38
SREBF2	3.92	39
BACH1	3.77	40

Mouse Brain (n=13)		
TF/RBP	Brain Score	Brain Rank
CIRBP	1.54	2
SFPQ	1.20	6
KLF15	0.96	8
FUS	0.94	9
ZC3H11A	0.88	10
MXI1	0.86	11
RBM28	0.81	14
EGR1	0.75	15
CHD1	0.71	16
CREB1	0.71	17
HIF1A	0.67	18
HNRPDL	0.66	19
CEBPG	0.65	20
BHLHE40	0.64	21
HNRNPK	0.63	22
SP2	0.63	23
RAD21	0.62	24
NFE2L2	0.60	25
EGR2	0.59	26
GTF2I	0.58	28
KLF12	0.58	29
HCFC1	0.58	30
CHD2	0.57	31
GTF2F1	0.56	32
SRPR	0.54	33
CEBPD	0.53	34
ETV1	0.49	35
GABPA	0.49	36
A1CF	0.49	37

Mouse Skin (n=14)		
TF/RBP	Skin Score	Skin Rank
NFIC	3.44	6
E2F1	2.68	7
MXI1	2.34	10
RUNX1	1.59	13
BRCA1	1.55	14
TCF4	1.46	15
HCFC1	1.41	16
MEF2A	1.32	17
ETV5	1.27	18
THRA	1.17	19
CHD1	1.16	20
FOXM1	1.12	21
NFATC1	1.12	22
FUS	1.12	23
ALCAM	1.11	24
HNRPDL	1.08	25
NFYA	1.03	26
ZFP161	1.00	27
CHD2	0.98	28
RBM5	0.97	29
TCF12	0.95	30
SREBF2	0.94	31
KHDRBS1	0.91	32
EIF4B	0.87	33
KLF13	0.85	34
ELK4	0.85	35
RFX5	0.83	36
SF1	0.81	37
ELAVL1	0.79	38

Mouse Liver (n=31)		
TF/RBP	Liver Score	Liver Rank
CEBPB	4.25	4
BHLHE40	4.22	5
FUS	3.92	7
HNRNPK	3.58	10
EIF4B	3.51	12
PCBP4	3.45	14
THRA	3.30	15
MXI1	2.83	18
YY1	2.80	19
MAFK	2.79	20
ATF5	2.75	21
MTA3	2.75	22
PPARA	2.63	23
BACH1	2.59	24
RXRA	2.58	25
ESR1	2.57	26
RFX4	2.54	27
HNRPDL	2.53	28
FOXO3	2.52	29
CIRBP	2.52	30
STAT5B	2.49	31
CRP	2.43	32
TARDBP	2.37	33
HNRNPC	2.35	34
ARHGAP24	2.33	35
DCTN2	2.28	36
NFIC	2.27	37
KLF1	2.25	38
USF2	2.25	39

Figure 4.3: Tables showing the ranking of circadian TFs and RBPs by CRC E-score in different tissue types. The leftmost table shows ranking in mouse transcriptome across all datasets. RBPs are labeled in red while TFs are labeled in black. Core clock TFs have been removed from the listing.

literature [102]. Extended members of the core clock were also identified in the ranking including THRA and BHLHE40.

In the results across all datasets, additional TFs and RBPs were identified that seem to have a much broader regulatory role than reported in the literature. For instance, FUS and CIRBP have been reported to affect the core circadian factor PER2 via alternative splicing, but only in the mouse liver [50, 70, 74]. In contrast, we find that FUS and CIRBP are found to be high scoring also in both brain and skin. EIF4B has been identified in the circadian regulation of translation in mouse liver [49]. We find that EIF4B is also top scoring in skin. HNRPDL is listed as a potential target of circadian regulation via microRNA in the brain [15]. Strikingly, these RBPs and TFs are found to have very high CRC scores across all mouse datasets. This suggest that they play a broader, previously uncharacterized, role in circadian regulation.

When looking at tissue specific results, many additional TFs and RBPs with high CRC scores are discovered. Although literature evidence has shown that these factors interact with circadian pathways, they are not known to be regulators of oscillation. These TFs may explain tissue specific circadian reprogramming. Within brain tissue, SFPQ is functionally involved in the cell cycle pathway, which also includes NONO and PER2 [29]. EGR1 has been found to oscillate and under direct regulation by the core clock [56]. Within our results, EGR1 potentially regulates a large number of downstream transcripts in the brain. CHD1 is known to be involved in circadian chromatin remodeling in brain [6]. KLF15 is well known to be regulated by the peripheral clock in relation to circadian nitrogen homeostasis in liver and muscle [45]. Within skin tissue, RUNX is a top TF and is known to be regulated in a circadian fashion in epidermal cells [44]. E2F1 is regulated by circadian factors SIRT1 and CLOCK [72]. BRCA1 is known to interact with core clock TFs such as PER2 [111]. Within liver tissue, CEBPB is top ranking excluding core clock TFs. This agrees with the literature finding that it interacts with the core clock through REV-ERB [36]. PCBP4 is known to be involved in circadian alternative splicing in the liver [66].

Additionally, there are many other novel findings that have been linked to very few circadian studies. These findings include: NFIC, RAD21, MXI1, and TARDBP across all tissues; ZC3H11A, RBM28, and CEBPG in brain; HCFC1 and ETV5 in skin; and HNRNPK, ATF5, and BACH1/MAFK in liver (Supplementary Table 2), and may provide leads for investigations of previously unknown circadian regulatory mechanisms.

4.3.4 Edge Influence Analysis

While the previous analysis was performed on the nodes of the CRC graph (aggregated or tissue-specific), to further understand the organization of circadian transcripts, it is necessary to study the regulatory edges (aggregated or tissue-specific). Because we are focused on the

A

Regulation	Mouse Score	Ranking	Type
FUS->PER2	29.49	19	RBP->Core Clock TF
ARNTL->NFIC	24.13	28	Core Clock TF->TF
BHLHE40->PER2	23.64	29	TF->Core Clock TF
ARNTL->FOXO3	22.41	31	Core Clock TF->TF
NFIX->PER2	21.03	36	TF->Core Clock TF
MXI1->PER2	20.54	37	TF->Core Clock TF
CEBPB->PER2	17.59	48	TF->Core Clock TF
NFYA->TEF	17.42	50	TF->Core Clock TF
ARNTL->KLF13	16.61	52	Core Clock TF->TF
ARNTL->BHLHE41	15.91	53	Core Clock TF->TF
KLF15->ARNTL	15.59	55	TF->Core Clock TF
RXRA->TEF	15.37	56	TF->Core Clock TF
ARNTL->KLF15	14.84	60	Core Clock TF->TF
PER2->KLF15	14.76	61	Core Clock TF->TF
ARNTL->NCL	14.67	62	Core Clock TF->RBP
NFYA->ARNTL	14.63	63	TF->Core Clock TF
PER1->FOXO3	14.52	64	Core Clock TF->TF
ARNTL->BHLHE40	14.52	65	Core Clock TF->TF
REPIN1->TEF	14.09	66	TF->Core Clock TF
ARNTL->SLC20A1	13.94	68	Core Clock TF->TF
USF2->TEF	11.65	72	TF->Core Clock TF
ARNTL->XBP1	11.29	73	Core Clock TF->TF
CEBPB->NFIL3	10.63	74	TF->Core Clock TF
PER1->KLF15	9.40	75	Core Clock TF->TF
ARNTL->UVRAG	9.39	76	Core Clock TF->TF
CRY1->REPIN1	9.34	77	Core Clock TF->TF
NFYA->PER2	9.30	78	TF->Core Clock TF
ARNTL->MAFK	8.58	79	Core Clock TF->TF
BHLHE41->CRY1	8.55	80	TF->Core Clock TF

B

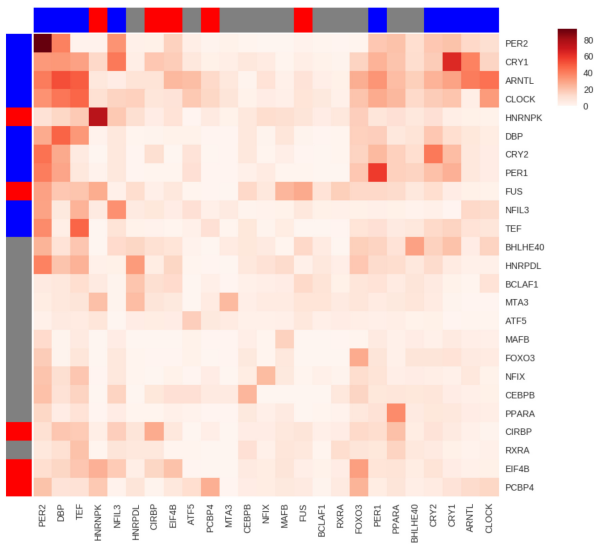


Figure 4.4: Edge influence analysis. **A**: Ranking of top edges between TFs and RBPs. Edges between core clock TFs have been omitted. **B**: Heatmap of inter-regulator (TF/RBP) circadian CRC score (*E-score* aggregates) in mouse. The score is calculated by aggregating CRC *E-scores* from the directed edges starting from row TF/RBP to the column TF/RBP across all datasets. Stronger colors in the heatmap indicate higher total scores (normalized for visualization). Color on row and column indicates the type of regulators: blue indicates core clock TF, red indicates RBP, and gray indicates regular TF. The heatmap visualizes the 100 strongest edges by aggregate scores.

regulatory architecture, we restrict the results to edges where the source is either a TF or RBP and the target is also either a TF or RBP. This is a subgraph of the whole CRC graph. An edge analysis was performed using CRC E-score weighted edges in this restricted graph. We rank these edges by their aggregated *E-scores* as shown in Figure 4.4.

The results compiled from all 87 mouse datasets are shown in Figure 4.4. Additional tissue specific results can be found in Supplementary Material 2. The heatmap demonstrates strong correlation within core clock TFs, as well as between the core clock and other top CRC scoring TFs and RBPs such as FUS and CIRBP. Figure 4.4 A lists some of the top interactions. Some of the findings are consistent with literature, such as the regulation of PER2 by FUS [74]. However, some top interactions seem to be underestimated in the circadian literature,

such as the potential regulation of PER2 by NFIX, or TEF by NFYA and RXRA. These novel interactions may be important in circadian regulation. Furthermore, we visualized the weighted edge scores between top TFs and RBPs, including the core clock, in Figure 2 B. It can be clearly seen that while there exists strong interaction between core clock TFs, the core clock does not strongly regulate a large number of additional TFs and RBPs. In fact, we find that using the CRC graph, the core clock can directly regulate only 324 TFs and RBPs. Given the limited direct regulatory influence of the core clock, it is necessary to examine indirect regulatory influence beyond the immediate neighborhood of the core clock within the CRC graphs.

4.3.5 Graph Analysis

Ranking CRC scores by node or edge provides a measure of direct influence for each TF and RBP. To understand the role of these TFs and RBPS in the architecture of circadian transcriptional oscillations, we examined regulatory paths within the CRC graphs to examine indirect influences. We performed this analysis on both the individual experiment specific CRC graphs, as well as the aggregated CRC graph.

Regulatory distance was computed as the length of the shortest directed path in a CRC graph from a source node to a target node. The mean percentage of oscillating transcripts found at increasing regulatory distances across all experiment specific CRC graphs is shown in Figure 4. The set of oscillating transcripts that are found to have a regulatory distance-one from the core clock were considered to be directly regulated by the core clock. The mean percentage of distance-one transcripts across all CRC graphs (individual, tissue-specific, or aggregated) is roughly 35% of all oscillating transcripts. While the majority of transcripts are not found to be directly regulated by the core clock, almost any transcript can be connected through a regulatory path in the CRC graph to the core clock. On average, in an individual

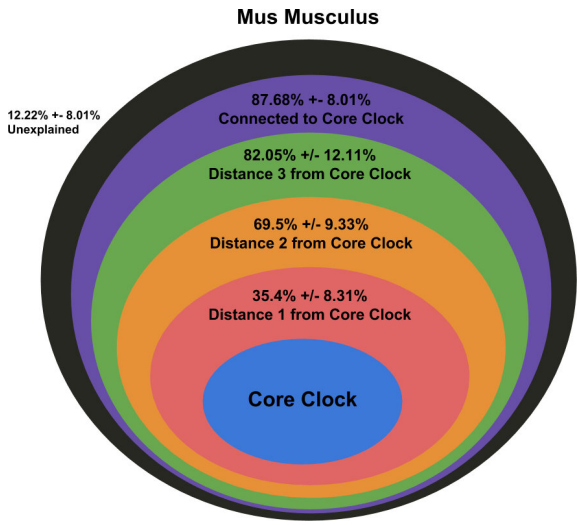


Figure 4.5: Mean percentages of transcriptome explained by TF/RBP at fixed regulatory distances from the core clock across mouse datasets.

experiment, 80% of all oscillating transcripts are within distance-three from the core clock (Figure 4.5).

Given that over 80% of oscillating transcripts can be connected to the core clock within a distance of three in experiment specific CRC graphs, we performed analysis of the aggregated CRC graph to quantify the total number of oscillating transcripts over the entire mouse transcriptome that can be connected to the core clock. We find that not only do 95% of protein-coding transcripts oscillate in some condition, but that this 95% can be connected to the core clock at some regulatory distance. The remaining 5% of non-oscillatory transcripts are found to be significantly enriched (p-value < 0.001, q-value < 0.01) for molecular function of GPCRs involved in olfactory reception.

Within the aggregated CRC graph, the number of out-going edges for TFs and RBPs at increasing regulatory distances decreases. Core clock TFs have an average of 21 out-going edges. At distance-one, this number decreases to nine. For TFs and RBPs at a distance of three or greater, the average number of outgoing edges is less than one. The decreasing number of out-going edges at each regulatory distances illustrates the decreasing regulatory

influence of these TFs and RBPs.

Functional enrichment by Gene Ontology (GO) term was performed on the subset of transcripts found at each regulatory distance. Transcripts found at regulatory distance-1 from the core clock exhibited significant enrichment for terms related to Circadian Rhythms, Cell Cycle, Metabolic Processes, and Neuronal Processes. The sets of transcription factors found at this distance were grouped into these functional categories including the set of known RBPs.

By combining the results from both the individual CRC graphs and the aggregated CRC graph, we identified a cascading hierarchical organization of transcriptomic regulation. This hierarchical organization begins at the core clock and is composed of several layers corresponding to increasing regulatory distance from the core clock.

In the aggregated CRC graph, the core clock can directly regulate 324 circadian TFs and RBPs along with 6300 transcripts (regulatory distance-one). Strikingly, only 17.1 ± 14.7 in any experiment specific CRC graph (Supplementary Table 4) are directly controlled by the core clock. The experiment specific distance-one TFs and RBPs are mostly distinct across experiments. These TFs and RBPs form the first layer of the hierarchical organization. At regulatory distance-two, the core clock can indirectly regulate 519 total TFs and RBPs through the 324 TFs and RBPs found at distance-one. As a result, it can also indirectly regulate 9080 total circadian transcripts. The hierarchy extends similarly to regulatory distance-three and greater. In addition, there are feedback connections that extend predominantly from one layer to the previous one.

Throughout this hierarchy, we observe that the fanning of the regulatory nodes decreases along with its total influence. As a consequence, to observe the significant shifts in the set of oscillating transcripts found in reprogramming events, a perturbation must occur in the expression of TFs and RBPs with high fan-out and high CRC scores, and thus predomi-

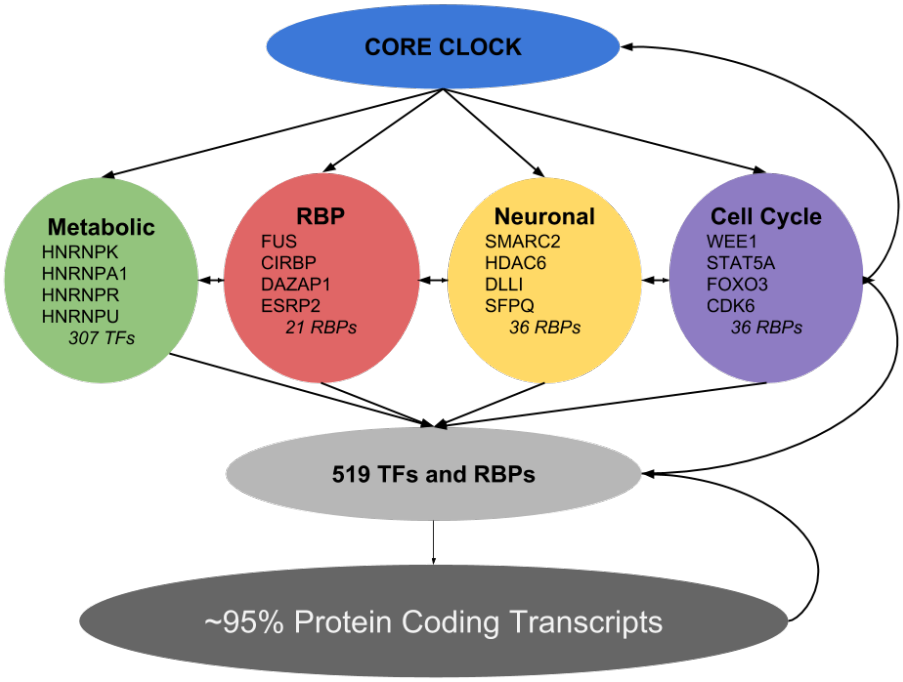


Figure 4.6: Network view of TFs and RBPs that are found at regulatory distance = 1. These TFs predominantly fall into three broad categories labeled from GO annotations that includes *Cell Cycle*, *Neuronal Function*, and *Metabolic Process*.

nantly at distance-one to the core clock. Perturbations affecting low CRC scoring TFs and RBPs at large regulatory distances can only induce changes in proportionally smaller sets of oscillating transcripts. Therefore, we conclude that experimental conditions that relate to changes in metabolism, neuronal function, cell cycle, or RBP modification likely instigate large reprogramming events through perturbations found in distance-one TFs or RBPs.

4.4 Discussion

The question of how specific subsets of oscillators are selected under specific genetic, epigenetic, and environmental conditions has remained an open question in the study of circadian rhythms. An organizational model that allows this selection within the network of coupled circadian oscillators must exist. This model must lie within a spectrum extending from a

purely centralized model, entirely orchestrated by the core circadian clock, to a completely decentralized model. In order to determine where in this spectrum the network of coupled oscillators operates, we obtained results using the CRC method and CRC graph with the large collection of mouse transcriptomic data available on CircadiOmics.

Using 87 mouse transcriptomic datasets, we analyzed the frequency with which TFs and RBPs oscillate. With no prior information, this purely data driven results rediscovered the core circadian clock. Remarkably, these results were confirmed using the 64 Baboon datasets in CircadiOmics.

We developed the Circadian Regulatory Control (CRC) method for identification of regulatory edges in circadian feedback loops. Two CRC scores, *B-Score* and *E-score*, incorporate three sources of evidence. These sources include the statistical significance of oscillation as assessed by BIO_CYCLE and high quality predicted binding sites assessed from MotifMap and MotifMap-RNA. Additionally, the scores take into account the delay between transcript and protein expression using available proteomic datasets included in CircadiOmics. CRC scores provide a measure circadian influence from a TF or RBP to a target (promoter or transcript).

Using the CRC method, we introduced the concept of the CRC graph as a directed weighted graph where nodes are oscillating transcripts and regulatory edges are measured by CRC scores. We constructed CRC graphs from individual experiments and superimposed these graphs to generate tissue-specific and aggregated CRC graphs.

We performed a node influence analysis on the CRC graphs by aggregating all of the outgoing edges of a source node across experiments. The node influence analysis highlighted more than 20 novel TFs and RBPs that are either previously underestimated or completely uncharacterized as important circadian regulators. Our analysis of directed edges within the CRC graph demonstrated the limited direct influence of the core clock. We found that core

clock TFs can directly regulate only 324 TFs and RBPs in the aggregated CRC graph.

We performed further graph analysis on both experiment specific CRC graphs and the aggregated CRC graph. We computed the average number of oscillating protein coding transcripts at increasing regulatory distances centering on the core clock. We found that 35% ($\pm$8%) of oscillating protein coding transcripts are directly regulated by the core clock. Additionally, we showed that on average 80% of oscillating transcripts can be connected within a regulatory distance of three to the core clock.

Analysis of the aggregated CRC graph revealed that not only do 95% of protein coding transcripts oscillate in some condition, but that any oscillating transcript can be connected at some regulatory distance to the core clock. Functional analysis of the remaining 5% non-oscillatory transcripts showed significant enrichment for molecular function of GPCRs involved in olfactory reception. At increasing regulatory distances, we observed that TFs and RBPs have decreasing aggregated CRC edge scores and a decreasing number of outgoing edges. This observation highlights the decreasing regulatory influence at increasing regulatory distances. Finally, functional annotation by GO terms revealed that at a regulatory distance of one, TFs can be grouped together into a few categories including Cell Cycle, Metabolic Process, and Neuronal Function. These sets, along with the set of RBPs found at distance-one, are functionally related and highly influential circadian regulators. RBPs play an important and previously underestimated role in circadian regulation. Reprogramming events, where the core circadian clock continues to oscillate, must originate from perturbations of expression in these TFs and RBPs. Furthermore, we show that the core clock can directly control 324 TFs and RBPs but only selects 17.1 ± 14.7 of them in any given condition. We conclude from these analyses that the organization of circadian transcriptomic regulation is structured as a hierarchy of regulatory steps originating at the core circadian clock.

4.5 Online Methods

The methods presented provide tools for identifying the organization of circadian rhythms within the context of the transcriptome. This type of analysis is made possible by the large repository of circadian omic data in CircadiOmics. CircadiOmics currently provides access to 231 omic datasets from eight species in a broad collection of tissue and conditions. To address bias in the repository of data related to tissue and species, separate results were computed from only mouse liver, skin, and brain for each method. All datasets that support the analyses can be found on the Circadiomics site

4.5.1 Frequency Histograms

Transcript frequency is defined as the total number of datasets where a given protein coding transcript is found to be oscillating at a BIO_CYCLE p-value < 0.01. Protein-coding transcripts were identified from the BioMart ENSEMBL gene database [94].

4.5.2 Circadian Regulation Control Scoring

The Circadian Regulation Control method is a measure of circadian regulation from a source TF/RBP to a target (promoter or transcript), consisting of three components. First, both the source and the target must be oscillatory according to BIO_CYCLE [2] with p$\leq$ 0.01. Second, if the source is a TF, there must exist at least one high quality binding site ($BBLS \geq 1, NLOD \geq 0.9$ and $FDR \leq 0.25$) in the promoter region of the target (defined as -5,000 to +500 bp of the transcription start site) according to MotifMap [115, 19]. For RBP sources, there must be high quality binding sites ($BBLS > 0, NLOD \geq 0.9$) predicted by MotifMap-RNA [62] in the introns or UTRs of its target transcripts.

Third, a measure of the correlative relationship between the phases of the source and the target is computed from the lag between peak expressions. The distribution of lags between the transcript and the protein product of 2,400 genes [20] was modeled as a beta distribution with $\mu = 0.35, SD = 0.25$. The empirical distribution and the fitted distribution are shown in Supplementary Material 3. The lag between a circadian transcript and the protein product is roughly 4 $\pm$2 hours corresponding to the mode of the fitted distribution. CRC scores are computed from the value of the highest probability density interval (HPD) associated with a given lag between a source and its target.

The values of these two components are summarized in two ways. A binary edge indicator, the *B-score*, uses a threshold number of binding sites (≥ 1) and phase correlation as measured by HPD value (≥ 0.8). An exponentially weighted edge score, *E-score*, is defined as $E = 0.5 * HPD - 0.5 * exp(-NumberOfBindingSites)$.

4.5.3 Circadian Regulatory Control Graphs

CRC Graphs are constructed for each of the 87 mouse datasets. Nodes are defined as oscillating transcripts using a threshold BIO_CYCLE p-value ≤ 0.01. Edges with a CRC *B-score* $= 1$ are included in the aggregated CRC graph. Edges with an *E-Score* ≥ 0 are included in experiment and tissue specific CRC graphs. The shortest directed path length is used to define regulatory distance between two nodes. The number of out-going edges for each TF and RBP was computed by NetworkX (https://networkx.github.io/). Results for mouse liver, skin, and brain tissue are provided in Supplementary Table 4. GO enrichment analysis and functional annotation for transcripts found at each distance was completed using the Python library *goatools* (https://github.com/tanghaibao/goatools) using NCBI gene to GO term associations.

4.5.4 Code availability

BIO_CYCLE was used for determining p-values for circadian oscillation and is available for download on the CircadiOmics website at circadiomics.ics.uci.edu/biocycle. MotifMap was used for generating TF binding sites and is available for download and querying at http://motifmap.ics.uci.edu/. MotifMap-RNA was used for generating RNA binding sites and is available for download and querying at http://motifmap-rna.ics.uci.edu/. Other analyses were performed using custom software.

Chapter 5

Summary and Conclusion

We have successfully applied deep learning models to a variety of tasks in medicine and bioinformatics. One of the most common problem we encounter in modeling for the biological datasets is the problem of small sample sizes. We mitigate or solve this problem by using data augmentation, data synthesis, public data, transfer learning, and so on. When it comes to applying well-known deep learning models to a real-world biological dataset, we are often required to customize certain components of existing network so that the network is better suited for the particular features from the dataset. This often means we have to customize the loss function by adding weights to different regions (weighted loss function), or choose a less commonly used loss function (SSIM), etc.

We have curated over 227 circadian omic dataset from various tissues from more than 8 species, detected the circadian oscillation with deep learning software BIO_CYCLE, and created a web hosting server `http://circadiomics.ics.uci.edu` for search and visualization. This is the largest circadian data repository available. CircadiOmics, with its advanced search features and unprecedented amount of high quality circadian data, is an enabling tool for accessing and comparing circadian studies. To date, the web server has contributed to

multiple studies that have been published in high impact journals.

Bibliography

[1] Colorectal cancer statistics — center of disease control and prevention.

[2] F. Agostinelli, N. Ceglia, B. Shahbaba, P. Sassone-Corsi, and P. Baldi. What time is it? deep learning approaches for circadian rhythms. *Bioinformatics*, 2016.

[3] J. C. Anderson and L. F. Butterly. Colonoscopy: Quality indicators. *Clinical and translational gastroenterology*, 6(2):e77, 2015.

[4] P. Baldi. *Deep Learning in Science*. Cambridge University Press, 2021.

[5] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:3412–424, 2000.

[6] W. J. Belden, Z. A. Lewis, E. U. Selker, J. J. Loros, and J. C. Dunlap. Chd1 remodels chromatin and influences transient dna methylation at the clock gene frequency. *PLoS genetics*, 7(7):e1002166, 2011.

[7] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput Med Imaging Graph*, pages 99–111, 2015.

[8] J. K. Bowmaker and H. Dartnall. Visual pigments of rods and cones in a human retina. *The Journal of physiology*, 298(1):501–511, 1980.

[9] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov. Fully convolutional neural networks for polyp segmentation in colonoscopy. In S. G. A. III and N. A. Petrick, editors, *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, pages 101 – 107. International Society for Optics and Photonics, SPIE, 2017.

[10] M. Brown and S. Süsstrunk. Multi-spectral sift for scene category recognition. pages 177–184, 06 2011.

[11] M. F. Byrne, N. Chapados, F. Soudan, C. Oertel, M. L. Linares Pérez, R. Kelly, N. Iqbal, F. Chandelier, and D. K. Rex. Su1614 artificial intelligence (ai) in endoscopy–deep learning for optical biopsy of colorectal polyps in real-time on unaltered endoscopic videos. *Gastrointestinal Endoscopy*, 85(5, Supplement):AB364–AB365, 2017.

[12] N. Ceglia, Y. Liu, S. Chen, F. Agostinelli, K. Eckel-Mahan, P. Sassone-Corsi, and P. Baldi. Circadiomics: circadian omic web portal. *Nucleic Acids Res.*, 2018.

[13] P. Chang, J. Grinband, B. Weinberg, M. Bardis, M. Khy, G. Cadena, M.-Y. Su, S. Cha, C. Filippi, D. Bota, P. Baldi, L. Poisson, R. Jain, and D. Chow. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *AJNR Am J Neuroradiol.*, 39(7):1201–1207, 2018.

[14] G. Charpiat, M. Hofmann, and B. Schölkopf. Automatic image colorization via multimodal predictions. In *European conference on computer vision*, pages 126–139. Springer, 2008.

[15] H.-Y. M. Cheng, J. W. Papp, O. Varlamova, H. Dziema, B. Russell, J. P. Curfman, T. Nakazawa, K. Shimizu, H. Okamura, S. Impey, et al. microrna modulation of circadian-clock period and entrainment. *Neuron*, 54(5):813–829, 2007.

[16] Z. Cheng, Q. Yang, and B. Sheng. Deep colorization, 2016.

[17] J. Churko, G. Mantalas, M. Snyder, and J. C. Wu. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circ Res.*, 112, 2013.

[18] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *CoRR*, abs/1503.01640, 2015.

[19] K. Daily, V. R. Patel, P. Rigor, X. Xie, and P. Baldi. MotifMap: integrative genome-wide maps of regulatory motif sites for model species, 2011.

[20] R. Dallmann, A. U. Viola, L. Tarokh, C. Cajochen, and S. A. Brown. The human circadian metabolome. *Proceedings of the National Academy of Sciences*, 109(7):2625–2629, 2012. [PubMed:22308371] [PubMed Central:22308371] [doi:10.1073/pnas.1114410109].

[21] A. Deshpande, J. Rock, and D. Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 567–575, 2015.

[22] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27:766–774, 2014.

[23] J. Feng, J. Deng, Z. Li, Z. Sun, H. Dou, and K. Jia. End-to-end res-unet based reconstruction algorithm for photoacoustic imaging. *Biomedical optics express*, 11(9):5321–5340, 2020.

[24] A. Forsberg, L. Widman, M. Bottai, A. Ekbom, and R. Hultcrantz. Postcolonoscopy colorectal cancer in sweden from 2003 to 2012: Survival, tumor characteristics, and risk factors. *Clinical Gastroenterology and Hepatology*, 18(12):2724–2733.e3, 2020.

[25] K. Fukushima. Neocognitron: A self-organizing neural network for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[26] R. B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[27] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[28] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis. Fully dense unet for 2-d sparse photoacoustic tomography artifact removal. *IEEE journal of biomedical and health informatics*, 24(2):568–576, 2019.

[29] F. Guillaumond, B. Boyer, D. Becquet, S. Guillen, L. Kuhn, J. Garin, M. Belghazi, O. Bosler, J.-L. Franc, and A.-M. François-Bellan. Chromatin remodeling as a mechanism for circadian prolactin transcription: rhythmic nono and sfpq recruitment to hltf. *The FASEB Journal*, 25(8):2740–2756, 2011.

[30] Y. Guo, J. Bernal, and B. J. Matuszewski. Polyp segmentation with fully convolutional deep neural networks—extended evaluation study. *Journal of Imaging*, 6(7), 2020.

[31] X. Han, H. Laga, and M. Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *CoRR*, abs/1906.06543, 2019.

[32] H. Hassan, Z. Ren, H. Zhao, S. Huang, D. Li, S. Xiang, Y. Kang, S. Chen, and B. Huang. Review and classification of ai-enabled covid-19 ct imaging models based on computer vision tasks. *Computers in Biology and Medicine*, 141:105123, 2022.

[33] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

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

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017.

[36] G. Huang, F. Zhang, Q. Ye, and H. Wang. The circadian clock regulates autophagy directly through the nuclear hormone receptor nr1d1/rev-erbα and indirectly via cebpb/(c/ebpβ) in zebrafish. *Autophagy*, 12(8):1292–1309, 2016.

[37] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[38] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[39] M. Hughes, J. Hogenesch, and K. Kornacker. Jtk_{cycle} : $an efficient nonparametric algorithm for detecting rhythmic components in genome-scale datasets. J Biol Rhythms.$, 2010.

[40] V. Iglovikov and A. Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. 01 2018.

[41] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016.

[42] R. Ironi, D. Cohen-Or, and D. Lischinski. Colorization by example. *Rendering techniques*, 29:201–210, 2005.

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[44] P. Janich, K. Toufighi, G. Solanas, N. M. Luis, S. Minkwitz, L. Serrano, B. Lehner, and S. A. Benitah. Human epidermal stem cell function is regulated by circadian oscillations. *Cell stem cell*, 13(6):745–753, 2013.

[45] D. Jeyaraj, F. A. Scheer, J. A. Ripperger, S. M. Haldar, Y. Lu, D. A. Prosdocimo, S. J. Eapen, B. L. Eapen, Y. Cui, G. H. Mahabeleshwar, et al. Klf15 orchestrates circadian nitrogen homeostasis. *Cell metabolism*, 15(3):311–323, 2012.

[46] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen. Kvasir-seg: A segmented polyp dataset, 2019.

[47] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H. D. Johansen. Resunet++: An advanced architecture for medical image segmentation, 2019.

[48] E. H. Jin, D. Lee, J. H. Bae, H. Y. Kang, M.-S. Kwak, J. Y. Seo, J. I. Yang, S. Y. Yang, S. H. Lim, J. Y. Yim, J. H. Lim, G. E. Chung, S. J. Chung, J. M. Choi, Y. M. Han, S. J. Kang, J. Lee, H. Chan Kim, and J. S. Kim. Improved accuracy in optical diagnosis of colorectal polyps using convolutional neural networks with visual explanations. *Gastroenterology*, 158(8):2169 – 2179.e8, 2020.

[49] C. Jouffe, G. Cretenet, L. Symul, E. Martin, F. Atger, F. Naef, and F. Gachon. The circadian clock coordinates ribosome biogenesis. *PLoS biology*, 11(1):e1001455, 2013.

[50] B. Kornmann, O. Schaad, H. Bujard, J. S. Takahashi, and U. Schibler. System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock. *PLoS biology*, 5(2):e34, 2007.

[51] K. B. Koronowski, K. Kinouchi, P.-S. Welz, J. G. Smith, V. M. Zinna, J. Shi, M. Samad, S. Chen, C. N. Magnan, J. M. Kinchen, W. Li, P. Baldi, S. A. Benitah, and P. Sassone-Corsi. Defining the independence of the liver circadian clock. *Cell*, 177(6):1448–1462.e14, 2019.

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[53] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016.

[54] C. M. C. le Clercq, M. W. E. Bouwens, E. J. A. Rondagh, C. M. Bakker, E. T. P. Keulen, R. J. de Ridder, B. Winkens, A. A. M. Masclee, and S. Sanduleanu. Postcolonoscopy colorectal cancers are preventable: a population-based study. *Gut.*, 63(6):957–63, 2014.

[55] J. Lee, S. W. Park, Y. S. Kim, K. J. Lee, H. Sung, P. H. Song, W. J. Yoon, and J. S. Moon. Risk factors of missed colorectal lesions after colonoscopy. *Medicine(Baltimore)*, 96(27):e7468, 2017.

[56] J. H. Lee and A. Sancar. Circadian clock disruption improves the efficacy of chemotherapy through p73-mediated apoptosis. *Proceedings of the National Academy of Sciences*, 108(26):10668–10672, 2011.

[57] Y. M. Lee and K. C. Huh. Clinical and biological features of interval colorectal cancer. *Clin Endosc.*, 50(3):254–260, 2017.

[58] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*, pages 689–694. 2004.

[59] R. Li, C. Xiao, Y. Huang, H. Hassan, and B. Huang. Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: A review. *Diagnostics*, 12(2), 2022.

[60] T. Li, J. Cohen, M. Craig, K. Tsourides, N. Mahmud, and T. Berzin. Mo1979 the next endoscopic frontier: A novel computer vision program accurately identifies colonoscopic colorectal adenomas. *Gastrointestinal Endoscopy*, 83:AB482, 05 2016.

[61] S. Liu, M. Gao, V. John, Z. Liu, and E. Blasch. Deep learning thermal image translation for night vision perception. *ACM Trans. Intell. Syst. Technol.*, 12(1), dec 2020.

[62] Y. Liu, S. Sun, T. Bredy, M. Wood, R. C. Spitale, and P. Baldi. MotifMap-RNA: a genome-wide map of RBP binding sites. *Bioinformatics*, 33(13):2029–2031, jul 2017.

[63] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

[64] E. Macken, S. V. Dongen, I. D. Brabander, S. Francque, A. Driessen, and G. V. Hal. Post-colonoscopy colorectal cancer in belgium: characteristics and influencing factors. *Endosc Int Open*, 2019.

[65] T. Mahmud, B. Paul, and S. A. Fattah. Polypsegnet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. *Computers in Biology and Medicine*, 128:104119, 2021.

[66] N. J. McGlincy, A. Valomon, J. E. Chesham, E. S. Maywood, M. H. Hastings, and J. Ule. Regulation of alternative splicing by the circadian clock and food related cues. *Genome biology*, 13(6):R54, 2012.

[67] B. T. McLelland, B. Lin, A. Mathur, R. B. Aramant, B. B. Thomas, G. Nistor, H. S. Keirstead, and M. J. Seiler. Transplanted hesc-derived retina organoid sheets differentiate, integrate, and improve visual function in retinal degenerate rats. *Investigative ophthalmology & visual science*, 59(6):2586–2603, 2018.

[68] A. Mehri and A. D. Sappa. Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 971–979, 2019.

[69] J. Meng, L. Xue, Y. Chang, J. Zhang, S. Chang, K. Liu, S. Liu, B. Wang, and K. Yang. Automatic detection and segmentation of adenomatous colorectal polyps during colonoscopy using mask r-cnn. *Open Life Sciences*, 15(1):588–596, 2020.

[70] J. Morf, G. Rey, K. Schneider, M. Stratmann, J. Fujita, F. Naef, and U. Schibler. Cold-inducible rna-binding protein modulates circadian gene expression posttranscriptionally. *Science*, 338(6105):379–383, 2012.

[71] M. Mylonaki, A. Fritscher-Ravens, and P. Swain. Wireless capsule endoscopy: a comparison with push enteroscopy in patients with gastroscopy and colonoscopy negative gastrointestinal bleeding. *Gut*, 52(8):1122–1126, 2003.

[72] Y. Nakahata, M. Kaluzova, B. Grimaldi, S. Sahar, J. Hirayama, D. Chen, L. P. Guarente, and P. Sassone-Corsi. The nad+-dependent deacetylase sirt1 modulates clock-mediated chromatin remodeling and circadian control. *Cell*, 134(2):329–340, 2008.

[73] N.-Q. Nguyen, D. M. Vo, and S.-W. Lee. Contour-aware polyp segmentation in colonoscopy images using detailed upsampling encoder-decoder networks. *IEEE Access*, 8:99495–99508, 2020.

[74] H. Oster, S. Damerow, R. A. Hut, and G. Eichele. Transcriptional profiling in the adrenal gland reveals circadian regulation of hormone biosynthesis genes and nucleosome assembly genes. *Journal of biological rhythms*, 21(5):350–361, 2006.

[75] T. Ozawa, S. Ishihara, M. Fujishiro, Y. Kumagai, S. Shichijo, and T. Tada. Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. *Therap Adv Gastroenterol.*, 2020.

[76] G. Palczewska, P. Stremplewski, S. Suh, N. Alexander, D. Salom, Z. Dong, D. Ruminski, E. H. Choi, A. E. Sears, T. S. Kern, et al. Two-photon imaging of the mammalian retina with ultrafast pulsing laser. *JCI insight*, 3(17), 2018.

[77] J. Pang, O. C. Au, K. Tang, and Y. Guo. Image colorization using sparse representation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1578–1582. IEEE, 2013.

[78] T. Pham, T. Tran, D. Phung, and S. Venkatesh. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics*, 69:218–229, 2017.

[79] S. qun Shi, T. S. Ansari, O. McGuinness, D. Wasserman, and C. Johnson. Circadian disruption leads to insulin resistance and obesity. *Current Biology*, 23(5):372–381, 2013.

[80] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[81] M. Riegler, K. Pogorelov, S. L. Eskeland, P. T. Schmidt, Z. Albisser, D. Johansen, C. Griwodz, P. Halvorsen, and T. D. Lange. From annotation to computer-aided diagnosis: Detailed evaluation of a medical multimedia system. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 13(3), 2017.

[82] M. S. Robles, J. Cox, and M. Mann. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. *PLoS genetics*, 10(1):e1004047, 2014.

[83] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[84] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[85] C. T. Rueden, J. Schindelin, M. C. Hiner, B. E. DeZonia, A. E. Walter, E. T. Arena, and K. W. Eliceiri. Imagej2: Imagej for the next generation of scientific image data. *BMC bioinformatics*, 18(1):1–26, 2017.

[86] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[87] S. Ryan. *Retina*. Elsevier, 2005.

[88] S. Sato, A. L. Basse, M. Schönke, S. Chen, M. Samad, A. Altıntaş, R. C. Laker, E. Dalbram, R. Barrès, P. Baldi, J. T. Treebak, J. R. Zierath, and P. Sassone-Corsi. Time of exercise specifies the impact on muscle metabolic pathways and systemic energy homeostasis. *Cell Metabolism*, 30(1):92–110.e4, 2019.

[89] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528, 2006.

[90] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham. Automatic colon polyp detection using region based deep cnn and post learning approaches. *IEEE Access*, 6:40950–40962, 2018.

[91] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[92] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[93] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[94] D. Smedley, S. Haider, S. Durinck, L. Pandini, P. Provero, J. Allen, O. Arnaiz, M. H. Awedh, R. Baldock, G. Barbiera, P. Bardou, T. Beck, A. Blake, M. Bonierbale, A. J. Brookes, G. Bucci, I. Buetti, S. Burge, C. Cabau, J. W. Carlson, C. Chelala, C. Chrysostomou, D. Cittaro, O. Collin, R. Cordova, R. J. Cutts, E. Dassi, A. D. Genova, A. Djari, A. Esposito, H. Estrella, E. Eyras, J. Fernandez-Banet, S. Forbes, R. C. Free, T. Fujisawa, E. Gadaleta, J. M. Garcia-Manteiga, D. Goodstein, K. Gray, J. A. Guerra-Assunção, B. Haggarty, D.-J. Han, B. W. Han, T. Harris, J. Harshbarger, R. K. Hastings, R. D. Hayes, C. Hoede, S. Hu, Z.-L. Hu, L. Hutchins, Z. Kan, H. Kawaji, A. Keliet, A. Kerhornou, S. Kim, R. Kinsella, C. Klopp, L. Kong, D. Lawson, D. Lazarevic, J.-H. Lee, T. Letellier, C.-Y. Li, P. Lio, C.-J. Liu, J. Luo, A. Maass, J. Mariette, T. Maurel, S. Merella, A. M. Mohamed, F. Moreews, I. Nabihoudine, N. Ndegwa, C. Noirot, C. Perez-Llamas, M. Primig, A. Quattrone, H. Quesneville, D. Rambaldi, J. Reecy, M. Riba, S. Rosanoff, A. A. Saddiq, E. Salas, O. Sallou, R. Shepherd, R. Simon, L. Sperling, W. Spooner, D. M. Staines, D. Steinbach, K. Stone, E. Stupka, J. W. Teague, A. Z. DayemUllah, J. Wang, D. Ware, M. Wong-Erasmus, K. Youens-Clark, A. Zadissa, S.-J. Zhang, and A. Kasprzyk. The biomart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–W598, 2015.

[95] J.-R. Su, Z. Li, X.-J. Shao, C.-R. Ji, R. Ji, R.-C. Zhou, G.-C. Li, G.-Q. Liu, Y.-S. He, X.-L. Zuo, and Y.-Q. Li. Impact of real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with video). *Gastrointestinal Endoscopy*, 91, 08 2019.

[96] P. Suarez, A. Sappa, and B. Vintimilla. Infrared image colorization based on a triplet dcgan architecture. pages 212–217, 07 2017.

[97] W. Tavanapong, J. Oh, G. Kijkul, J. Pratt, J. Wong, and P. deGroen. Real-time feedback for colonoscopy in a multicenter clinical trial. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 13–18, 2020.

[98] A. Toet. Colorizing single band intensified nightvision images. *Displays*, 26:15–21, 01 2005.

[99] A. Toet and M. Hogervorst. Progress in color night vision. *Optical Engineering*, 51, 02 2012.

[100] N. K. Tomar, D. Jha, M. A. Riegler, H. D. Johansen, D. Johansen, J. Rittscher, P. Halvorsen, and S. Ali. Fanet: A feedback attention network for improved biomedical image segmentation. *CoRR*, abs/2103.17235, 2021.

[101] F. S. Troelsen, H. T. Sørensen, S. D. Crockett, L. Pedersen, and R. Erichsen. Characteristics and survival of patients with inflammatory bowel disease and postcolonoscopy colorectal cancers. *Clinical Gastroenterology and Hepatology*, 2021.

[102] H. R. Ueda, S. Hayashi, W. Chen, M. Sano, M. Machida, Y. Shigeyoshi, M. Iino, and S. Hashimoto. System-level identification of transcriptional circuits underlying mammalian circadian clocks. *Nature genetics*, 37(2):187, 2005.

[103] G. Urban, K. Bache, D. T. Phan, A. Sobrino, A. K. Shmakov, S. J. Hachey, C. C. Hughes, and P. Baldi. Deep learning for drug discovery and cancer research: Automated analysis of vascularization images. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3):1029–1035, 2018.

[104] G. Urban, N. Feil, E. Csuka, K. Hashemi, C. Ekelem, F. Choi, N. A. Mesinkovska, and P. Baldi. Combining deep learning with optical coherence tomography imaging to determine scalp hair and follicle counts. *Lasers in Surgery and Medicine*, 53:171–178, 2021.

[105] G. Urban, P. Tripathi, T. Alkayali, M. Mittal, F. Jalali, W. Karnes, and B. Pierre. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4):P1069–1078.E8, 2018.

[106] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. C. Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *CoRR*, abs/1612.00799, 2016.

[107] J. Wang, H. Ding, F. A. Bidgoli, B. Zhou, C. Iribarren, S. Molloi, and P. Baldi. Detecting cardiovascular disease from mammograms with deep learning. *EEE Trans Med Imaging.*, 36(5):1172–1181, 2017.

[108] J. Wang, Z. Fang, N. Lang, H. Yuan, M.-Y. Su, and P. Baldi. A multi-resolution approach for spinal metastasis detection using deep siamese neural networks. *Computers in Biology and Medicine*, 84:137–146, 2017.

[109] P. Wang, T. Berzin, J. Brown, S. Bharadwaj, A. Becq, X. Xiao, P. Liu, L. Li, Y. Song, D. Zhang, Y. Li, G. Xu, M. Tu, and X. Liu. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. *Gut*, 68:gutjnl–2018, 02 2019.

[110] T. Welsh, M. Ashikhmin, and K. Mueller. Transferring color to greyscale images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 277–280, 2002.

[111] S. L. Winter, L. Bosnoyan-Collins, D. Pinnaduwage, and I. L. Andrulis. Expression of the circadian clock genes pert, per2 in sporadic, familial breast tumors. *Neoplasia*, 9(10):797–800, 2007.

[112] H. Wu, G. Chen, Z. Wen, and J. Qin. Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3469–3478, 2021.

[113] Y. Wu, B. Tao, T. Zhang, Y. Fan, and R. Mao. Pan-cancer analysis reveals disrupted circadian clock associates with t cell exhaustion. *Frontiers in Immunology*, 10, 2019.

[114] Y. Xiao, A. Jiang, J. Ye, and M.-W. Wang. Making of night vision: Object detection under low-illumination. *IEEE Access*, 8:123075–123086, 2020.

[115] X. Xie, P. Rigor, and P. Baldi. MotifMap: A human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, 25(2):167–174, 2009.

[116] P. Youssef, N. Sheibani, and D. Albert. Retinal light toxicity. *Eye*, 25(1):1–14, 2011.

[117] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization, 2016.

[118] Y. Zheng, W. Dong, and E. Blasch. Qualitative and quantitative comparisons of multispectral night vision colorization techniques. *Optical Engineering*, 51:087004–1–087004–16, 2012.