BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Thesis

IMAGE-BASED CLASSIFICATION OF HOARDING CLUTTER USING DEEP LEARNING

by

ZHENGHAO SUN

B.Eng., Huaqiao University, 2022

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2024

Approved by

First Reader

Janusz Konrad, PhD
Professor of Electrical and Computer Engineering

Second Reader

Jordana Muroff, PhD
Associate Professor & Chair of Clinical Practice
Boston University, School of Social Work

Third Reader

Prakash Ishwar, PhD
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

Acknowledgments

First of all, I would like to thank my advisor Professor Konrad. Over the past year, the careful guidance and assistance from him have been greatly beneficial to me. In our weekly meetings, he always provided thoughtful suggestions and feedback, and pointed out the shortcomings in my experiments. He also spent considerable amount of time reading and revising my thesis. From the structure and logic of individual chapters to the grammar and orthography in individual sentences, he provided detailed feedback on my thesis. I have learned a lot from him in the process of writing my thesis and I am very grateful for his help.

Secondly, I would also like to thank Professor Muroff for sharing her expertise in the field of hoarding disorder and for her help with the evaluation of new hoarding-clutter images collected in the course of this project - both were instrumental in advancing my research.

In addition, I would like to thank Professor Ishwar for providing extremely valuable feedback on my data-processing and machine-learning efforts, and especially for providing numerous optimization references, which turned out to be critical for my work.

Finally, my appreciation goes to my friends and family. Their help, care, and, of course, financial support have enabled me to continue on the journey of pursuing my Master's degree. Being able to successfully complete my thesis is not solely my achievement but also belongs to everyone who has supported and helped me along this journey. Thank you all !

Zhenghao Sun

IMAGE-BASED CLASSIFICATION OF HOARDING CLUTTER USING DEEP LEARNING

ZHENGHAO SUN

ABSTRACT

Hoarding disorder (HD) is characterized by difficulty letting go of items in living space resulting in excessive clutter, which can lead to significant health and safety risks. Traditionally, HD is assessed in an interview with a practitioner, but can be complemented by evaluating room clutter from pictures. To formalize the assessment of room clutter, a numerical scale was developed, called the Clutter Image Rating (CIR) scale: CIR = 1 corresponds to an uncluttered room, while CIR = 9 corresponds to a fully-cluttered room. CIR assessment is performed by social workers or other trained health or human-service professionals, which is time-consuming (and, therefore, costly), subjective, and can lack consistency in its repeatability. To address these challenges, deep-learning methods have been developed to automatically assess CIR from pictures, achieving up to 81% accuracy in estimating CIR on a dataset of 1,233 images of room clutter. However, this is a relatively small dataset for training large deep-learning models, and its CIR-class composition is imbalanced.

This thesis focuses on issues associated with the dataset size and imbalance, and also adopts a new deep-learning architecture for CIR scoring. First, data augmentation is applied to enlarge the training dataset and a novel weighted loss function is introduced to combat the dataset imbalance. Jointly, these two techniques improve the CIR scoring accuracy by 1% point compared to a ResNet-18-based method previously developed by Tezcan et al. Secondly, a Vision Transformer (ViT) architecture is adopted for CIR scoring, resulting in additional 5% points improvement in accuracy

over ResNet-18. Thirdly, in order to further address the dataset imbalance, DALL.E, a generative AI tool, is employed to synthesize new images with room clutter based on existing natural images in the dataset. This can be considered a novel type of data augmentation - AI-driven. New images are generated for underrepresented CIR classes in order to minimize the dataset imbalance. This also increases the overall dataset size which is beneficial for training the ViT model. Extensive experiments conducted using ResNet-18 and ViT models demonstrate that augmenting the original training dataset by AI-generated images enhances the performance for most under-represented classes but that the overall CIR-estimation accuracy is not improved. A detailed analysis of AI-generated clutter images against natural images from the dataset performed using t-SNE visualization suggests that for some CIR classes the new images exhibit outlying properties compared to the natural images, which likely affects the trained model's performance. While the novel idea of AI-driven data augmentation is beneficial for improving performance for some CIR classes, more research is needed to extend these gains across all classes.

Contents

List of Tables

List of Figures

List of Abbreviations

BERT	.....	Bidirectional Encoder Representations from Transformer
CCR	.....	Correct Classification Rate
CIR	.....	Clutter Image Rating
CNN	.....	Convolutional Neural Network
DL	.....	Deep Learning
GPT	.....	Generative Pre-trained Transformer
HD	.....	Hoarding Disorder
HOG	.....	Histogram of Oriented Gradients
HRS-I	.....	Hoarding Rating Scale-Interview
LLM	.....	Large Language Model
MAE	.....	Mean Absolute Error
ML	.....	Machine Learning
MLP	.....	Multi-Layer Perceptron
NLP	.....	Natural Language Processing
RNN	.....	Recurrent Neural Network
SGD	.....	Stochastic Gradient Descent
SI-R	.....	Saving Inventory - Revised
SVC	.....	Support Vector Classification
SVM	.....	Support Vector Machine
SVR	.....	Support Vector Regression
UHSS	.....	UCLA Hoarding Severity Scale
VIT	.....	Vision Transformer

Chapter 1

Introduction

Hoarding disorder (HD) is a complex and impairing mental health and public health problem. Its characterized by persistent difficulty and distress associated with discarding ordinary items regardless of their value, resulting in clutter in the living space (Tezcan et al., 2018; American Psychiatric Association, 2013). In some cases, the clutter extends beyond the active living areas and interferes with the use of other spaces, such as vehicles, front and back yards, the workplace, and relatives' homes. In severe cases, hoarding can pose a range of health risks, including fire, falling, and poor sanitation (Frost et al., 2000). It can also increase the risk of death from a house fire, or from being trapped under a "clutter avalanche." In general, the quality of life of a person with HD is substantially negatively affected (Saxena et al., 2011), and family relationships are often strained (Tolin et al., 2008). In the United States, HD affects about 5% of adult population (Iervolino et al., 2009; Samuels et al., 2008) and is a serious social issue.(Tolin et al., 2008)

HD is usually identified through a detailed psychological interview with the individual involved, preferably carried out in their home to properly evaluate the clutter and how it affects their life (Mataix-Cols, 2014). In 2008, a novel method, called "Clutter Image Rating" (CIR), was introduced (Frost et al., 2008). It utilizes a set of reference images with different levels of clutter to help assess the severity of hoarding and clutter. The CIR method allows individuals with hoarding challenges, their family members, trained experts, or independent evaluators to measure the clutter in

a patient's living space, such as living room, bedroom, or kitchen, utilizing a series of nine standardized images that showcase various levels of clutter (CIR = 1 corresponds to an uncluttered space, whereas CIR = 9 corresponds to a fully-cluttered space). However, this approach is time-consuming (and, therefore, costly), subjective, and can lack consistency in its repeatability.

Over the last decade, the surge in available computational power has dramatically advanced the application of Machine Learning (ML) and Deep Learning (DL) in tackling complex regression and classification tasks across various domains. Since the CIR method is basically a classification task (CIR = 1, 2, ..., 9) performed by humans, it is only natural to leverage ML and DL techniques to estimate the CIR value automatically. A computer-based CIR assessment would be instantaneous, objective (not dependent on assessor's mood, subjectivity, etc.) and repeatable (the same image would always result in the same CIR value).

Prior research at Boston University has resulted in two automated CIR assessment algorithms. Tooke *et al.* (Tooke et al., 2016) introduced a method combining Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) feature extractor with Support Vector Machine (SVM) classifier to assess CIR value from an image. On a set of 620 images of hoarding clutter collected on-line and rated by trained professionals specializing in hoarding disorder, they achieved 72% accuracy in assessing CIR value within ± 1 off the ground truth using 4-fold cross-validation. Subsequently, Tezcan *et al.* (Tezcan et al., 2018) proposed to use ResNet-18 deep-learning model in combination with a novel loss function combining traditional single-label loss with a 3-label loss to afford ± 1 departure from the ground truth. Since ResNet-18 requires much more training data than the HOG+SVM approach, the study team collected additional clutter images bringing the dataset size to 1,233 images. Using a ResNet-18 model pre-trained on ImageNet dataset (He et al., 2016), they fine-tuned and

tested the model in 4-fold cross-validation on the new, expanded dataset achieving 81% accuracy within ±1 off the ground-truth CIR. This was a significant result since the HOG+SVM approach of Tooke *al.* re-tested on the new dataset achieved only 60%. These findings underscore the feasibility of employing ML and DL methods to accurately assess the degree of room clutter from images. Although some progress has been made, the accuracy of assessing the severity of clutter in real-world scenarios is still insufficient; there exists considerable room for improvement.

As discovered by Tezcan *et al.*, the primary challenge in enhancing accuracy of CIR classifiers is data scarcity. The most recent clutter-image dataset comprises merely 1,233 images for training and 90 images for testing. This size is inadequate for training recent large DL models effectively. A sufficient volume of data is crucial for such models to avoid simply memorizing specific training examples, but instead to learn the underlying patterns. Models struggle with generalizing to new, unseen data if they don't have access to enough diverse training examples, resulting in poor performance.

Additionally, the distribution of the current 1,233 images across CIR classes is highly imbalanced. For instance, the number of images in CIR class 7 is more than double that found in classes 4 and 9. Such imbalance can lead to model bias towards classes with more samples, consequently resulting in diminished performance on classes with fewer samples. The limited data available for these less-represented classes can make it challenging for the model to learn effectively and predict accurately.

To address the challenge of a limited dataset, this thesis first proposes employing diverse data augmentation techniques to enhance training. Next, to mitigate the adverse effects of dataset imbalance, a novel class-weighted loss function is introduced. A combination of these two techniques incorporated into the original ResNet-

18 framework yields measurable but modest improvements in accuracy. Given the rising prominence of Large Language Models (LLMs) and the demonstrated effectiveness of multi-head architectures in handling image-related tasks, this thesis subsequently explores the Vision Transformer (ViT) (Dosovitskiy et al., 2020) whose classification performance surpasses that of ResNet-18. Although data augmentation and the new class-weighted loss function help mitigate some deficiencies of the dataset, neither approach adds new content to the dataset. In order to truly expand our dataset, this thesis proposes to leverage the power of generative AI methods, that have recently gained wide recognition, and adopts a new generative AI tool DALL.E for creating clutter images. To ensure that the generated images are consistent with the CIR ratings of their reference samples and that their content is realistic, we seek assistance from professionals specializing in hoarding disorder. We task them with verification whether the AI-generated images maintain the CIR rating of the reference sample, and revising the rating if need be. We also ask them to remove image samples that in their judgement are unrealistic (too cartoonish, overly sterile, etc.) We incorporate the retained AI-generated images into CIR-matching classes of the dataset, thereby expanding our training data.

This thesis is structured as follows. Chapter 2 reviews relevant literature and prior work in the field. Chapter 3 details data-augmentation methods proposed in this study and formulates a novel class-weighted loss function to address class imbalance in the current dataset. Chapter 4 outlines our implementation of the Vision Transformer for CIR classification of clutter images. Chapter 5 explores the application of generative AI tools to expand our dataset and details a methodology for evaluating the newly-generated images. Chapter 6 summarizes main contributions of the thesis, draws conclusions and proposes some ideas for future work.

Chapter 2

Relevant Work

Research into developing hoarding-specific assessments and diagnosing the severity of HD has increased over the past 20 years. This chapter begins by discussing an auxiliary method for HD assessment called "Clutter Image Rating" (CIR), that complements traditional environmental observation and clinical interview with a patient. Subsequently, evaluation metrics, needed to assess performance of automated CIR estimation algorithms, are introduced. This is followed by a description of two recent, automated methods for CIR estimation, one using machine-learning methods and one using deep-learning models. Finally, the last section discusses Large Language Models (LLMs) that have recently gained widespread prominence, and, more specifically, transformer models that have been adopted for image classification and we adopt here for CIR estimation.

2.1 Clutter Image Rating (CIR)

In 1993, Frost and Gross explicitly proposed the definition of hoarding for the first time (Frost and Gross, 1993), primarily encompassing three symptoms or behaviors: (1) excessive acquisition and inability to discard a large number of items that appear useless or of no value; (2) living spaces so cluttered that they can't be used for their intended purposes; (3) significant distress caused by hoarding, leading to impairment in psychological and behavioral functioning (Frost and Hartl, 1996). Traditionally, the symptoms of hoarding were mainly identified through interviews, self reports,

and home visits by a practitioner (e.g., social worker) or human-service personnel. In order to diagnose the severity of a patient's HD, researchers have dedicated significant efforts to developing various criteria to assist the health professional in assessing the patient's condition, for example: the "Saving Inventory - Revised" or SI-R (Frost et al., 2004; Kellman-McFarlane et al., 2019), the "Hoarding Rating Scale-Interview" or HRS-I (Tolin et al., 2010; Tolin et al., 2018), and the "UCLA Hoarding Severity Scale" or UHSS (Saxena et al., 2015).

A novel pictorial assessment method using images, called "Clutter Image Rating" (CIR), was introduced as an auxiliary approach (Frost et al., 2008). It provides a visual scale for evaluating the severity of clutter within homes, allowing for ratings to be given by the individuals with clutter, their family members, healthcare professionals, or external evaluators. The CIR image set consists of nine carefully-generated photographs that show varying levels of clutter across three primary living spaces: the living room, kitchen, and bedroom (Figure 2·1). Each photo set is designed to illustrate a continuum of clutter for each respective room: CIR = 1 corresponds to no clutter while CIR = 9 corresponds to fully-cluttered space. The development of the CIR aimed to mitigate the inaccuracies often found in self-reported assessments of clutter, offering a more objective measure.

2.2 Evaluation Metrics

In order to evaluate performance of a CIR-estimation algorithm, quality metrics are needed. The first automated CIR assessment method proposed by Tooke et al. (Tooke et al., 2016) approached the problem in two ways: as an estimation problem and as a classification problem. In the first case, the authors proposed to use the Mean-Absolute Error (MAE) metric between estimated CIR values, denoted $\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_N$, and the corresponding ground-truth values, denoted $y_1, y_2, .., y_N$, where N is the num-

(1) CIR=1 **(2)** CIR=2 **(3)** CIR=3

(4) CIR=4 **(5)** CIR=5 **(6)** CIR=6

(7) CIR=7 **(8)** CIR=8 **(9)** CIR=9

Figure 2·1: Reference bedroom images proposed by Frost et al. for image-based assessment of hoarding clutter according to CIR scale (Frost et al., 2008). Numbers shown below images are the assigned CIR values.

ber of computed estimates (i.e., the number of images for which CIR is assessed). In the second case, similarly to Tezcan et al. (Tezcan et al., 2018), they treated CIR assessment as a classification problem, where classification accuracy is a more appropriate metric. Typically, this can be measured by computing the Correct Classification Rate (CCR) as follows:

$$CCR = \frac{1}{N}\sum_{k=1}^{N}\mathbf{1}(|y_k - \widehat{y}_k| = 0), \tag{2.1}$$

where $\mathbf{1}(x)$ is an indicator function that equals 1 if x is true and 0 if x is false. We adopt the CCR as one of evaluation metrics used in this thesis.

However, in the case of CIR assessment, even trained professionals have difficulty when deciding between close CIR values. Therefore, CIR variations within ± 1 are deemed acceptable, leading to a modified version of CCR, called CCR-1 (Tooke et al., 2016), and defined as follows:

$$CCR_1 = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}(|y_k - \widehat{y}_k| \leq 1). \tag{2.2}$$

2.3 CIR assessment using traditional machine learning

In traditional machine learning, feature extraction is typically followed by estimation or classification of the target based on these features. Since rooms of patients with HD are filled with physical clutter (e.g., piles of boxes, newspapers, clothing), which corresponds to "busy" image areas with a high density of edges, Tooke et al. (Tooke et al., 2016) proposed to use the Histogram of Gradients (HOG) (Dalal and Triggs, 2005) as the feature extractor in conjunction with either Support Vector Regression (SVR), treating CIR assessment as an estimation problem, or Support Vector Classification (SVC), treating the assessment as a classification problem.

They used their own dataset of 620 hoarding images collected on-line and CIR-rated by trained professionals specializing in hoarding disorder. After re-sizing each image to 320×240-pixel resolution, they used Prewitt operator for gradient calculations, 20×20-pixel cells, 2×2-cell blocks and 4-bin histograms $(0°, 45°, 90°, 135°)$ to capture the angle of gradients. After suitable normalization, a 16×12×4=768-long feature vector was associated with each image. Subsequently, they applied either estimation, using SVR, or classification, using SVC, to predict CIR values.

Since the dataset was relatively small, they implemented data augmentation as

follows. From each training image, they extracted 16 sub-images of size 300×225 by randomly shifting them by 0, 5, 10 or 15 pixels in both the horizontal and vertical directions, and performing a horizontal flip.

In a 4-fold cross-validation on this 620-image dataset, they achieved CCR-1 of 67% by means of estimation (SVR approach) and 72% by means of classification (SVC approach).

2.4 CIR assessment using Convolutional Neural Networks

While neural networks have been researched for several decades, only in the last decade have they gained widespread adoption owing to the rapid growth of computing power and availability of large datasets. Among such networks, Convolutional Neural Networks (CNNs) have demonstrated huge gains over state-of-the-art model-based methods in various image classification tasks (Krizhevsky et al., 2012). The objective of CIR assessment aligns well with the type of classification challenges that CNNs excel at solving. In this context, building upon Tooke et al.'s foundational research, Tezcan et al. (Tezcan et al., 2018) adopted ResNet-18 CNN (He et al., 2016) for CIR assessment, treating it as a classification problem.

The adoption of ResNet-18 for CIR assessment was not accidental. ResNet-18 is a deep CNN composed of 18 layers. Its core characteristic is the concept of residual learning, which directly adds the input to subsequent layers through the so-called short-circuit (or skip) connections, thereby solving the gradient disappearance and gradient explosion problems in deep-network training. This allows the network to learn more efficiently (He et al., 2016). Such a design endows ResNet-18 with a relatively small model size and high efficiency, ensuring its effective performance under real-time constraints or when resources are limited.

A unique contribution of Tezcan et al.'s study was the introduction of a weighted

combination of two loss functions: one aiming to maximize CCR (single-label classification) and the other striving to maximize CCR-1 (3-label classification allowing ± 1 departure from the ground truth). This allows to tune the focus of ResNet-18 towards either goal.

As the single-label loss function aiming to maximize CCR, they used traditional cross-entropy as follows:

$$\mathcal{L}_k^{SL} = -\sum_{i=1}^{C} \mathbf{y}_k^1[i] \log \frac{\exp(\widehat{\mathbf{y}}_k[i])}{\sum_{j=1}^{C} \exp(\widehat{\mathbf{y}}_k[j])}, \tag{2.3}$$

where C is the number of classes ($C = 9$ in CIR assessment), $\mathbf{y}_k^1$ is a one-hot encoded vector of the ground-truth CIR value for image number k (i.e., $\mathbf{y}_k^1[i] = 1$ and otherwise 0, if the CIR value for image number k equals i), and $\widehat{\mathbf{y}}_k$ is the output of the last layer of the network (before *softmax*). This loss function is designed to train a CNN with the goal of achieving high accuracy in exactly matching the ground-truth value.

They also used a multi-label, binary cross-entropy loss function between the *sigmoid* output of ResNet-18's last layer and a three-hot encoded ground truth. This loss aims to maximize CCR-1 and is formulated as follows:

$$\mathcal{L}_k^{ML} = -\sum_{i=1}^{C} \Big(\mathbf{y}_k^3[i] \log \frac{1}{1+\exp(-\widehat{\mathbf{y}}_k[i])} + (1 - \mathbf{y}_k^3[i]) \log \frac{\exp(-\widehat{\mathbf{y}}_k[i])}{1+\exp(-\widehat{\mathbf{y}}_k[i])} \Big), \tag{2.4}$$

where $\mathbf{y}_k^3[i]$ is a three-hot encoded vector of the ground truth, i.e., $\mathbf{y}_k^3[i]$ equals 1 for i corresponding to the ground truth or within ± 1 off the ground truth.

While targeting a high CCR-1 aligns well with the objective of CIR assessment, achieving a high CCR (exact match) is equally important. Designing a CNN solely for CCR-1 might lead to the network largely ignoring the boundary labels (i.e., class 1 and class 9). To mitigate this scenario, Tezcan et al. combine both loss functions

by assigning different weights to $\mathcal{L}_k^{SL}$ (2.3) and $\mathcal{L}_k^{ML}$ (2.4) as follows:

$$\mathcal{L}_k^{CIR} = (1-\lambda)\mathcal{L}_k^{SL} + \lambda\mathcal{L}_k^{ML}, \tag{2.5}$$

where λ is a weight parameter that can be used to control performance. This approach aims to meet the requirements for improved CCR-1 outcomes while preventing the scenario where the exact match rates for boundary classes are extremely low. The overall loss function for the whole training set of N images is the sum of the individual loss functions, i.e., $\mathcal{L}^{CIR} = \sum_{k=1}^{N} \mathcal{L}_k^{CIR}$.

Tezcan et al. have also expanded the clutter-image dataset. Unlike in Tooke et al.'s study, which utilized 620 images, the new expanded dataset comprised 1,233 images for training and 90 images for testing. However, even this double-size dataset is still considered insufficient for training ResNet-18. Consequently, Tezcan et al. employed the same data augmentation as proposed by Tooke et al. (Tooke et al., 2016).

The use of ResNet-18 with a weighted loss function resulted in a significant performance boost. The HOG+SVM approach of Tooke et al. (Tooke et al., 2016) re-tested on the new 1,233-image dataset via 4-fold cross-validation produced CCR-1 of only 60% while ResNet-18 with $\lambda = 0.9$ resulted in CCR-1 of 81%.

2.5 Self-attention mechanism in deep-learning models

In 2017, Google's machine-translation team introduced seminal work on the use of attention mechanism (Vaswani et al., 2017). This revolutionized machine translation by replacing Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) with an attention mechanism. This approach excels in modeling dependencies within sequences, irrespective of the elements' separation, capturing intricate and long-range patterns crucial for complex data like text or time series. As a result,

its effectiveness in Natural Language Processing (NLP) led to the rapid adoption of attention in leading models such as BERT, GPT, and Transformer.

The introduction of attention mechanism has also impacted other fields of application, beyond NLP. In 2020, Google adapted the Transformer architecture and self-attention mechanism to computer vision by introducing the Vision Transformer (ViT). ViT divides an image into a series of fixed-size patches, treats each patch as an input token similar to words in a sentence, and then processes these patches with the Transformer's self-attention mechanism (Dosovitskiy et al., 2020). This approach allows ViT to consider the global context of the entire image, making it highly effective for image classification tasks. The success of ViT has been validated by its impressive performance on standard image classification benchmarks, demonstrating the versatility and power of self-attention schemes beyond text-based applications.

In this thesis, we adopt the ViT model as a replacement for ResNet-18 in order to seek further performance gains in automated CIR assessment.

Chapter 3

Enhanced Data Augmentation and Class-Weighted Loss Function

In this thesis, we consider Tezcan et al.'s study (Section 2.4) as the baseline algorithm upon which to improve. In this chapter, we start by proposing more extensive data augmentation during training and a novel weighted loss function to address class imbalance in our clutter-image dataset.

3.1 Extension of data augmentation

Tezcan et al. (Tezcan et al., 2018) used a dataset that consists of 1,233 training images and 90 test images (Table 3.1), all with CIR ratings assigned by trained professionals specializing in hoarding disorder. This dataset was an expanded version of the dataset used by Tooke et al. (Tooke et al., 2016).

Table 3.1: Number of training and test images in each CIR class used by Tezcan et al. (Tezcan et al., 2018).

CIR	1	2	3	4	5	6	7	8	9	All
Training	118	153	117	97	146	181	215	119	87	1,233
Testing	10	10	10	10	10	10	10	10	10	90
Total	128	163	127	107	156	191	225	129	97	1,323

Due to a relatively small dataset size, both studies applied data augmentation by means of horizontal and vertical shifts of 5, 10, or 15 pixels, and a horizontal "flip".

However, the maximum shift of 15 pixels is rather small for 224×224 images accepted by ResNet-18, so very little visual information is changed. To allow more significant visual "jitter", we increased the maximum range of shifts to 30 pixels while keeping 5-pixel increments, and also applying a horizontal "flip". Furthermore, since the camera angle, when taking a picture, does not need be aligned with room's features (doors, window frames, etc.), we added an additional geometric augmentation by means of clockwise or anti-clockwise image rotation by 0, 3, 6, or 9 degrees. Finally, because of the diversity of cameras that can be used as well as wide range of possible illumination conditions, we also applied color-jitter augmentation. This method increases data diversity by randomly altering the visual attributes of images, such as brightness and contrast, as well as color saturation and hue, thereby aiding the model in better generalizing to unseen data. Figure 3·1 shows one example for each type of data augmentation.

(1) Original **(2)** Shift down **(3)** Shift left

(4) Rotation **(5)** Flip **(6)** Color jitter

Figure 3·1: Examples of data augmentation applied during training.

3.2 Novel class-weighted loss function

Tezcan et al. developed a weighted loss function to strike balance between maximizing CCR and CCR-1 performance metrics (Section 2.4). However, this does not address the dataset imbalance clearly visible in Table 3.1. The number of samples for CIR = 4 and CIR = 9 is less than 100, while there are 215 samples for CIR = 7. In this case, if we perform 4-fold cross-validation with non-overlapping splits, the number of samples from a majority class in each fold will be significantly higher than that from a minority class. This will lead to the model's predictive ability being biased towards majority classes, with poor recognition capability for minority classes. This is because the model finds it easier to improve overall accuracy by focusing on the more frequently-occurring data samples. However, our goal is to accurately assess CIR across all CIR classes, and avoid poor performance for the minority classes.

To address this issue, we propose to employ a cost-sensitive learning approach, assigning higher mis-classification cost to minority classes. By adjusting the loss function to include a weighted mis-classification cost for each class, we aim to more severely penalize mis-classifications of the minority classes. The equation for calculating weight w_i for class i is as follows:

$$w_i = C \times \frac{\frac{1}{c_i}}{\sum_{j=1}^{C} \frac{1}{c_j}}, \quad i = 1, ..., C, \tag{3.1}$$

where C is the number of classes and c_i is the number of samples in class i. Thus, weight w_i is inversely proportional to the frequency of occurrence of samples from class i, and the sum of all weights equals the total number of classes. This ensures that different attention is given to classes of different sizes within the dataset.

We apply weights w_i (3.1) to the multi-class, single-label loss function (2.3), that

maximizes CCR, as follows:

$$\mathcal{L}_k^{WSL} = -\sum_{i=1}^{C} w_i \cdot \mathbf{y}_k^1[i] \log \frac{\exp(\widehat{\mathbf{y}}_k[i])}{\sum_{j=1}^{C} \exp(\widehat{\mathbf{y}}_k[j])}. \tag{3.2}$$

It is also essential to incorporate weights into the loss function for multi-label CIR classification (2.4). Considering the ±1 error margin in CCR-1, each prediction necessitates the application of three weights. However, in the context of cluttered-image classification, boundary categories can only utilize two weights; for instance, the first category involves CIR classes 1 and 2, while the last category involves CIR classes 8 and 9. Consequently, this study adopts an average-weight approach for multi-label scenarios. Based on the previously-calculated weights for each class (3.1), we also take into account the weights of adjacent classes within ±1 range. An arithmetic mean of three weights (or two weights for edge categories) is computed as follows:

$$\bar{w}_i = \frac{w_{\max(i-1,1)} \cdot 1_{\{i>1\}} + w_i + w_{\min(i+1,C)} \cdot 1_{\{i<C\}}}{1_{\{i>1\}} + 1 + 1_{\{i<C\}}} \tag{3.3}$$

where $1_{\{condition\}}$ is an indicator function that equals 1 when *condition* is met, and 0 otherwise.

We apply the average weights $\bar{w}_i$ (3.3) to the multi-label loss function (2.4) that maximizes CCR-1, as follows:

$$\mathcal{L}_k^{WML} = -\sum_{i=1}^{C} \bar{w}_i \Big(\mathbf{y}_k^3[i] \log \frac{1}{1+\exp(-\widehat{\mathbf{y}}_k[i])} + (1-\mathbf{y}_k^3[i]). \log \frac{\exp(-\widehat{\mathbf{y}}_k[i])}{1+\exp(-\widehat{\mathbf{y}}_k[i])} \Big) \tag{3.4}$$

Combining loss functions (3.2) and (3.4), as proposed by Tezcan et al. (2.5), the final loss function for an image-CIR pair number k is as follows:

$$\mathcal{L}_k^{CIR} = (1-\lambda)\mathcal{L}_k^{WSL} + \lambda \mathcal{L}_k^{WML}. \tag{3.5}$$

3.3 Experimental results

In order to gauge improvements offered by the proposed enhanced data augmentation and new class-weighting mechanism against the baseline method, we performed an ablation study using all 1,323 images available in the original dataset (Table 3.1), that is we combined all training and testing images into one set.

We applied 4-fold cross-validation with non-overlapping splits to ensure robust performance evaluation. To generate precise performance metrics, this procedure was repeated 10 times, with each iteration running over 50 epochs for every fold. Within the last 10 epochs of each run, the highest and lowest CCR-1 value was recorded, along with the corresponding CCR value (not necessarily highest or lowest). This helps us understand the range of variability in our model's performance due to random initialization and other factors influencing the outcomes of each run. In Table 3.2 and other tables in this thesis, we report the maximum and minimum CCR-1 value recorded over the last 10 epochs in all 10 runs (last column: "[Max/Min]"). We also report the average of 10 maximum CCR-1 values recorded in 10 runs (last column: "Average Max"). Finally, we report the average of 10 CCR values recorded along with the maximum CCR-1 values (CCR column: "Average"), but we caution that CCR Average is not necessarily the average of 10 maximum CCR values recorded in the last 10 epochs in each of the 10 runs.

Table 3.2 shows the performance of ResNet-18 using various combinations of data augmentation and class-weighted loss function. The baseline method of Tezcan et al. (Tezcan et al., 2018) reported in the second data row, slightly improves the Average Max CCR-1 value compared to no augmentation at all (first data row). Although CCR value slightly suffers, as already mentioned this value may not correspond to the best average CCR value. Adding the proposed class weighting, slightly boosts performance in all metrics. The proposed enhanced data augmentation further improves

Table 3.2: ResNet-18 performance in CIR classification for various combinations of data augmentation and class-weighted loss function.

Class-Weighted Loss	Data Augmentation	CCR Average	CCR-1 Average Max [Max/Min]
No	No	0.4376	0.8121 [0.8347/0.7994]
No	Baseline	0.4341	0.8176 [0.8293/0.7867]
Yes	Baseline	0.4446	0.8183 [0.8383/0.7964]
No	Enhanced	0.4551	0.8216 [0.8383/0.7754]
Yes	**Enhanced**	**0.4723**	**0.8299 [0.8503/0.8144]**

performance against the baseline method. Finally, combining the enhanced augmentation with class-weighted loss achieves the most significant performance gain; CCR value increased by almost 0.04 to 0.4723 compared to the baseline method and the Average Max CCR-1 values improved by 0.0123 to 0.8299. For CCR-1 expressed in percent, this would be a 1.23% point increase to almost 83% compared to the baseline method.

Our proposed enhanced data augmentation and class-weighted loss improve ResNet-18 performance compared to the Tezcan et al.'s baseline. However, the improvements brought about by these two methods are quite limited. In the next chapter, we study another deep-learning model hoping to further improve CIR classification accuracy.

Chapter 4

CIR Classification using the Vision Transformer

4.1 Vision-transformer architecture

In the Vision Transformer (ViT) (Dosovitskiy et al., 2020), each input image is divided into fixed-size patches (typically, 16×16). Subsequently, each patch (block) is projected into a fixed-length vector to form the input to the encoder, along with block-position information and a learnable embedding with the class label of the image. Subsequent encoding operations are identical to those in the original Transformer model developed for language applications.

A diagram of the ViT is shown in Figure 4·1. Assuming image X of width W and height H with C color components (channels) is the input to ViT, below we summarize the key steps performed by ViT.

- **Preprocessing**: ViT transforms image $X \in \mathbb{R}^{H\times W\times C}$ into a sequence of "flattened" 2D patches, resulting in a structure $X_p \in \mathbb{R}^{N\times(P\times P\times C)}$, where $P \times P$ is the spatial size of each patch and N is the number of patches.

- **Patch embedding**: The Transformer Encoder expects a two-dimensional matrix as input: $X_{input} \in \mathbb{R}^{N\times D}$, where N is the length of the sequence, and D (usually 256) is the dimension of each vector in the sequence. In order to map $X_p \in \mathbb{R}^{N\times(P\times P\times C)}$ to $X_{input} \in \mathbb{R}^{N\times D}$, usually a linear transformation (e.g., fully-connected layer) is applied to each vector, to reduce vector length $(P \times P \times C)$

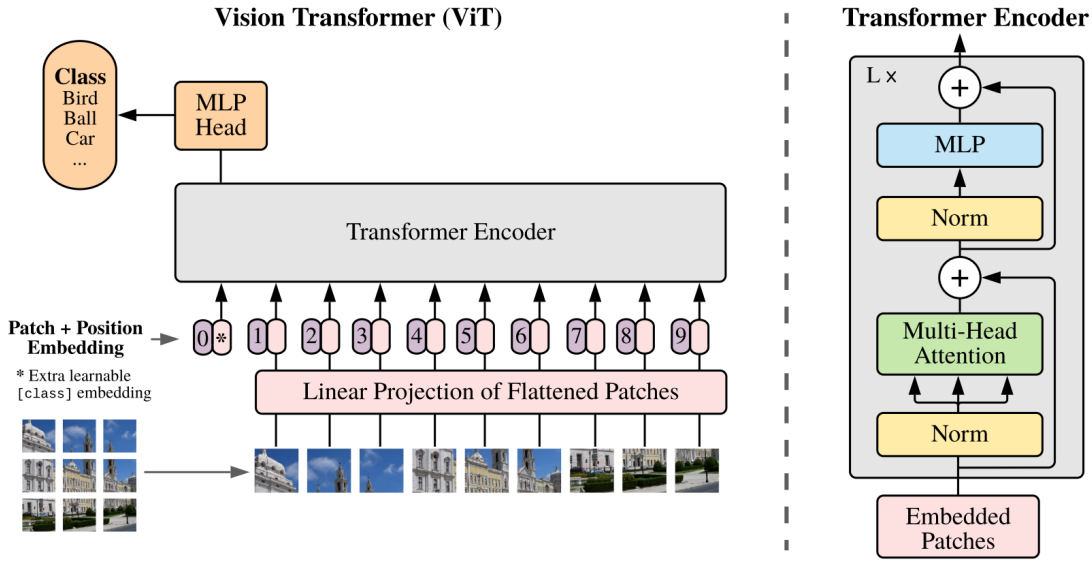

Figure 4·1: ViT model overview. Each image is first divided into blocks, that are reshaped into vectors and position information of each block is added. Then, these reshaped blocks are fed into an encoder. To allow classification, an additional trainable "classification token" is added to the sequence. Diagram from (Dosovitskiy et al., 2020).

to D. This is referred to as patch embedding.

Since the Transformer Encoder does not explicitly allow for image-class label, this is performed implicitly as shown in Figure 4·1. An additional vector (with label 0 in the diagram) is artificially added before applying positional embeddings and feeding into the transformer. This added vector is a learnable embedding, that represents the comprehensive information of the entire sequence (in this context, the whole image), thus providing a basis for classification.

- **Positional encoding**: For a transformer architecture to be effective in classifying images, all blocks must be jointly considered along with information about their location (e.g., blue sky at the top of the image). The Transformer Encoder uses a self-attention mechanism to analyze the relationships between individual

blocks, but it lacks capacity to leverage the order of blocks in the sequence. This affects the understanding of image structure. By adding positional encoding to each block, this model can capture the placement of blocks within the image, improving its understanding of image content and structure.

- **MLP head**: After a series of self-attention and feed-forward network layers, the model generates a set of high-dimensional feature vectors. These vectors contain information that the model has learned from the input image. The vectors first pass through one or more fully-connected layers, which constitute the Multi-Layer Perceptron (MLP). If there are multiple layers, activation functions (such as ReLU) are typically used to introduce non-linearity, helping the model to better capture more complex features. The output of the final fully-connected layer is mapped to a dimension that matches the number of categories (9 in CIR classification). Each element of the output vector represents the probability that the image belongs to the corresponding class.

4.2 Implementation of ViT-based classification

Since ViT is a large model (ViT used in our experiments has parameter size of 330 MB), training it from scratch with a dataset of only 1,323 images is counterproductive. To address this challenge, we employ a transfer-learning approach. We use `vit_base_patch16_224` model from the `timm` library (Wightman, 2022) initialized with weights pre-trained on the ImageNet dataset (Deng et al., 2009). We adapt this pre-trained model to our task by setting the number of output classes to 9, and fine-tuning the MLP head using our clutter-image dataset while keeping the transformer unchanged.

To optimize the ViT performance for our dataset, we performed grid search to find optimal training parameters. We explored various combinations of the learning rate

(0.0001, 0.001, 0.01), and of its decay period (5, 7, 9 epochs). We used the stochastic gradient descent (SGD) for training and found out that the learning rate of 0.001 that drops by half after every 5 epochs, performs best. For consistency with Tezcan at al.'s experiments, we adopted a momentum of 0.9, mini-batch size of 32 and their single-/multi-label loss function with our class-weight modifications.

4.3 Experimental results

Similarly to Section 3.3, we performed an ablation study using all 1,323 images available in the original dataset (Table 3.1). Table 4.1 shows ViT performance against ResNet-18 for various combinations of data augmentation and class-weighted loss function. The bottom part of the table repeats ResNet-18 results from Table 3.2 for ease of comparison.

The ablation study of ViT reveals similar patterns as those observed for ResNet-18. The baseline augmentation minimally improves the Average Max CCR-1 but slightly reduces CCR compared to no augmentation at all. The inclusion of class-weighted loss along with baseline augmentation slightly improves both metrics. The enhanced data augmentation improves the performance further. Finally, the combination of enhanced data augmentation and class-weighted loss gives the best performance, outperforming the method without data augmentation and no class-weighted loss by 0.0514 (over 5% points) in CCR and by 0.0256 (over 2.5% points) in Average Max CCR-1.

We observe that our ViT model, employing the same combination of data augmentation and class-weighted loss as ResNet-18, results in significant improvements in both CCR and CCR-1 metrics. For example, the corresponding improvements in CCR range from 0.0170 (1.7% points) to 0.0554 (5.54% points). Similarly, the corresponding improvements in the Average Max CCR-1 range from 0.0391 (3.91%

Table 4.1: Comparison of ViT and ResNet-18 performance for various combinations of data augmentation and class-weighted loss function.

ViT

Class-Weighted Loss	Data Augmentation	CCR Average	CCR-1 Average Max [Max/Min]
No	No	0.4546	0.8558 [0.8862/0.8293]
No	Baseline	0.4532	0.8567 [0.8832/0.8234]
Yes	Baseline	0.5000	0.8575 [0.8892/0.8234]
No	Enhanced	0.5058	0.8793 [0.8982/0.8563]
Yes	**Enhanced**	**0.5060**	**0.8814 [0.9012/0.8623]**

ResNet-18

Class-Weighted Loss	Data Augmentation	CCR Average	CCR-1 Average Max [Max/Min]
No	No	0.4376	0.8121 [0.8347/0.7994]
No	Baseline	0.4341	0.8176 [0.8293/0.7867]
Yes	Baseline	0.4446	0.8183 [0.8383/0.7964]
No	Enhanced	0.4551	0.8216 [0.8383/0.7754]
Yes	**Enhanced**	**0.4723**	**0.8299 [0.8503/0.8144]**

points) to 0.0577 (5.77% points). Most importantly, however, the ViT model with enhanced data augmentation and class-weighted loss achieves 0.8814 in Average Max CCR-1 compared to 0.8176 for the baseline method of Tezcan et al. (Tezcan et al., 2018), a very significant gain of 0.0638 (6.38% points). The gain in CCR is even more impressive, from 0.4341 to 0.506, a difference of 0.0719 (7.19% points).

Clearly, the replacement of ResNet-18 by ViT significantly improves the CIR classification performance, and its combination with enhanced data augmentation and class-weighted loss brings CCR-1 fairly close to 90%, a very desirable accuracy.

Chapter 5

AI-Driven Data Augmentation

Thus far, we showed that data augmentation, by creating more diverse data samples, helps a model learn more generalized feature representations enabling it to perform better on unseen images.

In this chapter, in addition to traditional data augmentation, we introduce AI-driven data augmentation. This method leverages Generative AI tools to produce new, synthesized images that are similar to natural images in our current dataset, but that are visually distinct. This AI-driven data augmentation essentially acts as an expansion of our current dataset.

5.1 Generation of room-clutter images using DALL·E

As the Generative AI tool of choice, we opted for DALL·E (Ramesh et al., 2021). DALL·E is an advanced deep-learning model developed by OpenAI, specifically designed for image generation. It builds on GPT-3, a large-scale language-processing model, enabling the creation of images from natural-language descriptions. It can generate highly-relevant and diverse images from specific, complex text prompts. Considering the difficulty in acquiring real-life images of hoarding-related room clutter with a CIR rating above 7, using DALL·E to generate images with equivalent ratings could potentially prove beneficial for addressing class-imbalance in our dataset.

Currently, the two commonly-used versions of DALL·E are: DALL·E-2 which can generate images from image prompts and DALL·E-3 which can generate images from

text prompts. Below is a brief introduction to these two models.

- **DALL·E-2**: Unlike DALL·E-3, which only supports text-based image generation, DALL·E-2 includes a variation function. This function enables the direct creation of variations from an existing image, thereby avoiding information loss that can occur during the conversion between image and text. Figure 5·1 shows an original image from our dataset and its two variations generated by DALL·E-2 using the variation function.

- **DALL·E-3**: Compared to DALL·E-2, DALL·E-3 significantly improves the process of generating images from text captions, greatly improving the realism and visual appeal of the synthesized images. However, currently DALL·E-3 can only handle text inputs; it does not support direct image inputs. Therefore, we must first use `GPT-4-vision-preview` API (Cai et al., 2023) to understand an image and generate a caption, before using the DALL·E-3 to create a new image from the obtained text. Effectively, to generate a new, synthesized image from a natural image, we follow an *Image-to-Text-to-Image* sequence of steps. This can lead to a considerable misrepresentation or loss of image content in the transition from image to text, particularly problematic if we want to generate an image with CIR value identical to that of the original image. Figure 5·2 shows an original image from our dataset and two images generated by GPT-4 and DALL·E-3 using such steps. For more examples of DALL·E-3-generated images, see Appendix A.

As is evident from Table 3.1, classes 1, 3, 4, 8, and 9 have significantly fewer image samples than other classes. To mitigate effects of this imbalance, we propose to use DALL·E to generate enough images to ensure that each class has approximately 160 samples. In order to test the impact of text-based and image-based image generation on CIR-classification performance, we study both approaches, DALL·E-2 and

Original image from the dataset.

Variation 1 generated by DALL·E-2.

Variation 2 generated by DALL·E-2.

Figure 5·1: Comparison of an original (natural) image from the dataset with its two variations generated by DALL·E-2.

Original image from the dataset.

Image 1 generated by DALL·E-3.

Image 2 generated by DALL·E-3.

Figure 5·2: Comparison of an original (natural) image from the dataset with two images generated by DALL·E-3 by means of intermediate caption generation by GPT-4.

Table 5.1: Number of images per CIR class in the new, expanded dataset. The original dataset of training and testing images is complemented by images generated by DALL·E to improve class balance.

CIR	1	2	3	4	5	6	7	8	9	All
Training	118	153	117	97	146	181	215	119	87	1,233
Testing	10	10	10	10	10	10	10	10	10	90
Generated	**32**	**0**	**34**	**50**	**0**	**0**	**0**	**34**	**60**	**210**
Total	160	163	161	157	156	191	225	163	157	1,533

DALL·E-3, and generate an equal number of images by each method. These generated images are combined with the original dataset to form a new, expanded dataset. Table 5.1 lists class distribution of the original dataset and the number of new images that we generate by both approaches for CIR classes 1, 3, 4, 8, 9.

5.2 Challenges in AI-driven data augmentation

To address the class-imbalance problem in the original dataset, we use two AI-driven data-augmentation methods described in Section 5.1. We generate new images only for the under-represented CIR classes 1, 3, 4, 8, and 9. The number of images generated in each class is shown in Table 5.1. The same number of images are generated by each method, DALL·E-2 and DALL·E-3.

5.2.1 Impact of unverified DALL·E-generated images

We aim to explore whether the proposed AI-driven data augmentation leads to improved performance for the underrepresented classes. In this test, for simplicity, rather than performing 4-fold cross-validation (applied in Sections 3.3 and 4.3), we perform training on the union of 1,233 original images and 210 DALL·E-generated images (1,443 images in total), and perform testing on the original test set of 90 natural images (Table 5.1).

We conducted experiments by using three distinct training sets:

- Original: 1,233 images,
- Original + DALL·E-2: 1,443 images,
- Original + DALL·E-3: 1,443 images,

but always testing using the same 90 test images. In all cases, we used the best-performing ViT model from Table 4.1 (with enhanced data augmentation and class-weighted loss), the same learning rate, momentum and mini-batch size, and repeated the training process 10 times, running the model for 50 epochs each time.

Table 5.2 reports the overall results across all CIR classes using the same metrics as used in Sections 3.3 and 4.3, namely Average CCR, Average Max CCR-1 and Maximum/Minimum CIR-1 across 10 last epochs of 10 runs. The ViT model's overall performance reported is similar to that reported in Table 4.1 although slightly lower (Average CCR of 0.4811 compared to 0.5060 and Average Max CCR-1 of 0.8711 compared to 0.8814). The difference is due to testing on 90 images instead of 4-fold cross-validation as explained earlier. Training on the "Original + DALL·E-2" dataset slightly improves performance; average CCR increased to 0.4878 and Average Max CCR-1 improved to 0.8767. The best performance, however, is accomplished by training on the "Original + DALL·E-3" dataset, increasing the metrics to 0.5156 and 0.8911, respectively.

Table 5.3 reports performance for each CIR class by using slightly different metrics. For each class, we recorded the maximum and minimum CCR-1 value in the last 10 epochs of all 10 runs and calculated the average of these maximum and minimum CCR-1 values, shown as "[Average Max/Average Min]" pair of numbers in Table 5.3. Note that we are not showing the maximum and minimum values recorded over the last 10 epochs in all 10 runs, that were shown as "[Max/Min]" in Tables 3.2 and

Table 5.2: Impact of AI-driven data augmentation (unverified images) on the overall performance of the ViT model tested on the original test set of 90 natural images.

Training Set	Training Set Size	CCR Average	CCR-1 Average Max [Max/Min]
Original	1,233	0.4811	0.8711 [0.9111/0.8556]
Original + DALL·E-2	1,433	0.4878	0.8767 [0.9111/0.8556]
Original + DALL·E-3	1,433	0.5156	0.8911 [0.9111/0.8511]

4.1, to simplify the table. We note that performance for individual classes is not as consistent as the overall performance across all classes. For instance, by training on the "Original + DALL·E-3" dataset, the performance for CIR classes 4 and 8 improved significantly, remained roughly unchanged for class 9, but slightly decreased for class 3. One possible reason for the improvement not extending to all underrepresented classes is that the CIR value of some DALL·E-generated images may be inconsistent with that of their source images, and therefore may create bias during training. We address this issue in the next section.

Table 5.3: Impact of AI-driven data augmentation (unverified images) on the per-class performance of the ViT model tested on the original test set of 90 natural images. The reported pairs of numbers are "[Average Max/Average Min]" CCR-1 values explained in detail in the text.

Training Set	CIR = 1	CIR = 2	CIR = 3	CIR = 4	CIR = 5
Original	[1.00/1.00]	[1.00/1.00]	[1.00/0.96]	[0.91/0.84]	[0.81/0.67]
Original + DALL·E-2	[1.00/1.00]	[1.00/1.00]	[1.00/1.00]	[0.96/0.92]	[0.82/0.75]
Original + DALL·E-3	[1.00/1.00]	[1.00/1.00]	[0.98/0.96]	[1.00/0.98]	[0.73/0.62]

Training Set	CIR = 6	CIR = 7	CIR = 8	CIR = 9
Original	[0.75/0.65]	[1.00/0.98]	[0.80/0.63]	[0.63/0.55]
Original + DALL·E-2	[0.80/0.68]	[1.00/0.96]	[0.78/0.78]	[0.59/0.51]
Original + DALL·E-3	[0.83/0.72]	[1.00/0.99]	[0.91/0.75]	[0.61/0.48]

5.2.2 Impact of verified DALL·E-generated images

In order to make sure that DALL·E-generated images are associated with CIR ratings consistent with their respective source images, we have asked trained professionals specializing in hoarding disorder to review the synthesized images in terms of their CIR rating and visual realism. Their task was three-fold:

- **confirm** that the CIR value inherited by a DALL·E-generated image from its source image is correct, or
- **change** the CIR value inherited by a DALL·E-generated image from its source image to a new value, or
- **reject** a DALL·E-generated image because of its poor visual quality, "cartoon-ish" look, lack of realism, etc., thereby removing it from the dataset.

Since in Section 5.2.1 we found that data augmentation using DALL·E-3 outperforms that of DALL·E-2, from now on we focus only on the DALL·E-3-enhanced dataset. As shown in Table 5.4, out of 210 images generated by DALL·E-3 (Table 5.1) only 128 images were retained after review. Clearly, many images were removed due to sub-par quality and/or lack of realism. Also, many of them were assigned a new CIR value. The most affected was CIR class 9 (extreme clutter) where only 4 DALL·E-3-generated images were considered of acceptable quality and realism, and whose depiction of extreme clutter was accurate enough to be retained at level 9. Although this effort has largely reduced class imbalance, CIR class 9 continues to be highly underrepresented. This is consistent with observations made by Tooke et al. (Tooke et al., 2016) and by Tezcan et al. (Tezcan et al., 2018) regarding huge difficulty with on-line collection of natural images with extreme clutter.

We performed a similar evaluation to that reported in Section 5.2.1, but instead we used 1,361 training images by combining the original natural training images (1,233)

Table 5.4: Number of images per CIR class in the new, expanded and verified dataset. The original dataset of training and testing images is complemented by images generated by DALL·E-3 and verified for CIR accuracy by trained professionals.

CIR	1	2	3	4	5	6	7	8	9	All
Training	118	153	117	97	146	181	215	119	87	1,233
Testing	10	10	10	10	10	10	10	10	10	90
Generated/verified	**23**	**0**	**31**	**49**	**0**	**0**	**0**	**21**	**4**	**128**
Total	151	163	158	156	156	191	225	150	101	1,451

with 128 images generated by DALL·E-3 and verified by professionals, and performed testing on the 90 natural test images as before.

Table 5.5 shows the overall performance in terms of CCR and CCR-1. The verification/revision of CIR value and pruning of unrealistic images resulted in a small drop of CCR to 0.5022 and a very small increase of Average Max CCR-1 to 0.8944. In terms of "[Min/Max]" values, while the maximum CCR-1 over the last 10 epochs of all runs remained unchanged at 0.9111 for all scenarios, the minimum CCR-1 increased considerably to 0.8778 when using verified images. Despite the reduction of available augmentation images from 210 to 128, the range of CCR-1 values when the algorithm is close to convergence became narrower.

Table 5.5: Impact of AI-driven data augmentation (verified images) on the overall performance of the ViT model tested on the original test set of 90 natural images.

Training Set	Training Set Size	CCR Average	CCR-1 Average Max [Max/Min]
Original	1,233	0.4811	0.8711 [0.9111/0.8556]
Original + DALL·E-3	1,433	0.5156	0.8911 [0.9111/0.8511]
Original + verified DALL·E-3	1,361	0.5022	0.8944 [0.9111/0.8778]

The per-class performance is shown in Table 5.6. In most of CIR classes, the

average maximum and minimum CCR-1 values shown in brackets have changed very little. However, there is a significant improvement for CIR class 9, likely due to the use of verified images, despite the fact that only 4 generated images were retained in this class after verification. Another significant improvement is for CIR class 5, although no AI-generated images were added to this class. Most likely, AI-generated images for class 4 have affected the learning of ViT model parameters and performance for class 4. On the flip side, performance markedly dropped for CIR class 8 in which 21 images were generated and verified. It seems that the decision boundary between classes 8 and 9 is quite fluid because of the similarity of the two clutter scenarios and even slight changes in model parameters may lead to a substantial change in performance for these two classes. Overall, however, per-class performance is excellent (CCR-1 close to 1.0) for classes 1, 2, 3, 4, 7, is good (CCR-1 0.7-0.8) for classes 5, 6 and 8, but only passable (CCR-1 of 0.6-0.7) for class 9.

Table 5.6: Impact of AI-driven data augmentation (verified images) on the per-class performance of the ViT model tested on the original test set of 90 natural images. The reported pairs of numbers are "[Average Max/Average Min]" CCR-1 values explained in detail in the text.

Training Set	CIR = 1	CIR = 2	CIR = 3	CIR = 4	CIR = 5
Original	[1.00/1.00]	[1.00/1.00]	[1.00/0.96]	[0.91/0.84]	[0.81/0.67]
Original + DALL·E-3	[1.00/1.00]	[1.00/1.00]	[0.98/0.96]	[1.00/0.98]	[0.73/0.62]
Original + verified DALL·E-3	[1.00/1.00]	[1.00/1.00]	[0.99/0.94]	[0.99/0.96]	[0.85/0.68]

Training Set	CIR = 6	CIR = 7	CIR = 8	CIR = 9
Original	[0.75/0.65]	[1.00/0.98]	[0.80/0.63]	[0.63/0.55]
Original + DALL·E-3	[0.83/0.72]	[1.00/0.99]	[0.91/0.75]	[0.61/0.48]
Original + verified DALL·E-3	[0.82/0.75]	[1.00/0.99]	[0.82/0.71]	[0.70/0.62]

5.3 Cross-validation using AI-enhanced datasets

In the previous section, we demonstrated that augmenting the original dataset with DALL·E-3-generated and professionally-verified images can lead to a small performance gain in CCR-1, which is more relevant than CCR due to the difficulty trained professionals have with the exact assignment of CIR value. However, that evaluation was performed only on the test set of 90 natural images. Here, we perform an evaluation using four-fold cross-validation using all available images in 3 scenarios described below. In order to facilitate a direct comparison with results reported by Tezcan et al. (Tezcan et al., 2018), we use 1,323 natural images from the original dataset's training and test sets (Table 5.4), and 128 DALL·E-3-generated and verified images.

1. **Four-fold cross-validation on the original dataset (scenario #1):** This is our baseline scenario involving four-fold cross-validation on the original dataset of 1,323 = 1,233 + 90 images. Results using this cross-validation scenario were reported in Tables 3.2 and 4.1.

2. **Four-fold cross-validation using AI-generated images only in training (scenario #2):** In this scenario, the original 1,323 training/test natural images are combined with 128 AI-generated and verified images for the total of 1,451 images. All these images are divided into 4 folds. While in each run 3 folds are used for training, during evaluation of the 4-th (test) fold all AI-generated images are excluded. This ensures that evaluation is performed solely on original (natural) images, thereby allowing direct comparison with baseline results. The AI-generated images are only used to enhance training, but are not evaluated.

3. **Four-fold cross-validation using AI-generated images in training and testing (scenario #3):** This scenario involves standard four-fold cross-validation on a mixed set of 1,451 images used in the second scenario above, but without

excluding AI-generated images from the test fold. The AI-generated images are used to enhance training and are also evaluated in the test fold.

In each of these cross-validation scenarios, we examine performance of both ViT and ResNet18 models in configurations that achieved the best performance in Table 4.1 (enhanced data augmentation and class-weighted loss function). We use the same performance metrics as in previous tables reporting the overall results (not class-specific), namely Average Max CCR-1 and the corresponding Average CCR, and Maximum/Minimum CCR-1 over the last 10 epochs in 10 runs.

Table 5.7: Impact of AI-driven data augmentation (verified images) on the performance of ViT and ResNet-18 models in 3 scenarios of four-fold cross-validation.

Cross-Validation Scenario	Model	CCR Average	CCR-1 Average Max [Max/Min]
Scenario #1: 4-fold cross-validation on the original dataset	ViT	0.5060	0.8814 [0.9012/0.8623]
	ResNet-18	0.4723	0.8299 [0.8503/0.8144]
Scenario #2: 4-fold cross-validation using AI-generated images only in *training*	ViT	0.4900	0.8839 [0.9129/0.8529]
	ResNet-18	0.4389	0.8313 [0.8559/0.8168]
Scenario #3: 4-fold cross-validation using AI-generated images in *training* and *testing*	ViT	0.5157	0.8887 [0.9044/0.8689]
	ResNet-18	0.4745	0.8397 [0.8716/0.8197]

Table 5.7 shows the performance of ViT and ResNet-18 for 3 cross-validation scenarios described earlier. Scenario #2 compared to scenario #1 (baseline) shows a slight increase in CCR-1 for both the ViT and ResNet-18 models. In scenario #2, ViT achieves Average Maximum CCR-1 of 0.8839 compared to 0.8814 in scenario #1, and similarly for ResNet-18 the improvement is from 0.8299 to 0.8313.

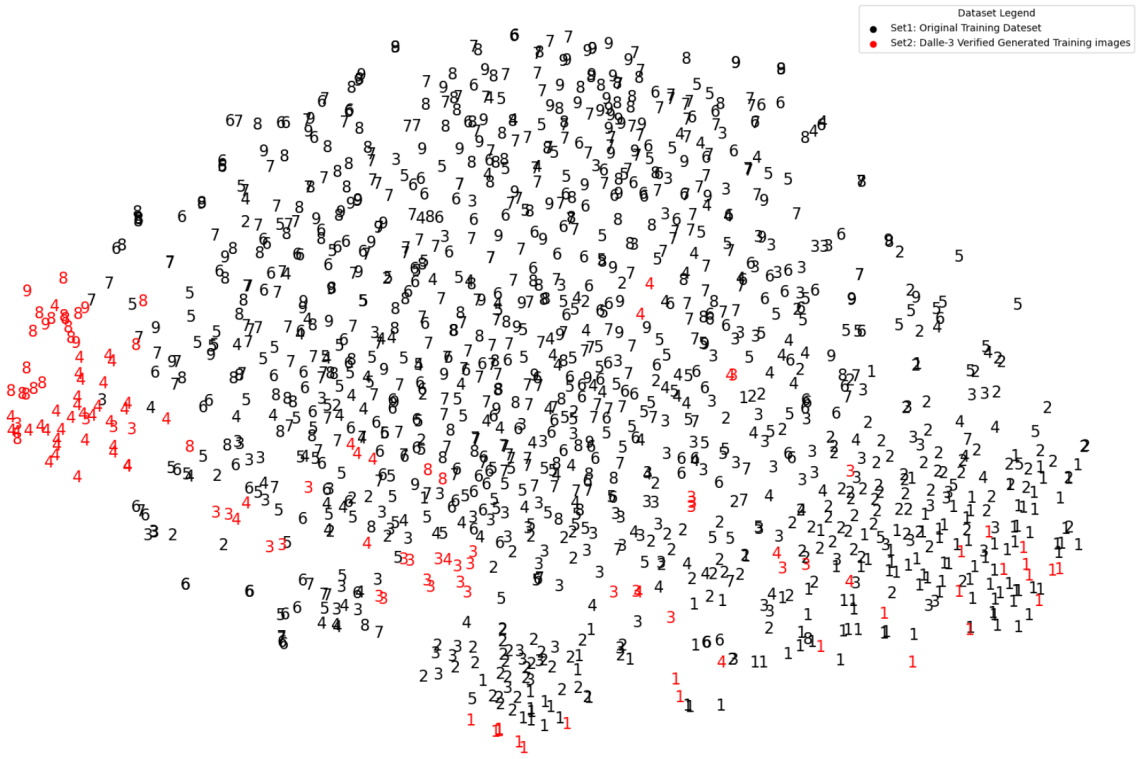

Figure 5·3: T-SNE visualization of image embeddings jointly for the original dataset of 1,233 natural images (black digits denote CIR classes) and 128 clutter images generated by DALL·E-3 and verified (red digits denote CIR classes).

However, both models exhibit a decline in CCR performance, with ViT dropping by 0.0160 (from 0.5060 to 0.4900) and ResNet-18 dropping more significantly by 0.0343 (from 0.4723 to 0.4389). To gain insight into these unexpected results, we conducted a joint t-SNE distribution analysis of the original (natural) dataset and the AI-generated and verified dataset. The t-SNE visualization in Figure 5·3 reveals that the majority of the newly-generated images in categories 4, 8, and 9 do not overlap the original dataset (see Appendix B for more t-SNE visualizations of the generated images against dataset images in each CIR class separately). This indicates that AI-generated images bring new information to the training step. However, because these generated images are excluded from the test set, the newly-introduced information

may actually interfere with the model's ability to make correct predictions, thus causing a decline in CCR and only a very small gain in CCR-1.

In contrast to scenario #2, where we observed a small increase in CCR-1 but an appreciable decrease in CCR compared to scenario #1, scenario #3 shows a different pattern. When the verified AI-generated images are also included in the test folds in scenario #3, both CCR and CCR-1 for ViT and ResNet-18 improve relative to scenario #1 (baseline). Specifically, ViT shows an increase of 0.0073 in Average Maximum CCR-1 value and ResNet-18 shows an increase of 0.0098. In terms of CCR, ViT shows an increase of 0.0097 and ResNet-18 shows an increase of 0.0022. The increase in CCR value in this scenario can be attributed to the inclusion of AI-generated images in the training phase, which enabled the models to more easily achieve correct CIR predictions on AI-generated images, now present in the test folds, thereby improving the overall CCR performance.

Chapter 6

Conclusions and Future Work

6.1 Thesis Summary and Conclusions

Building on earlier research by Tooke et al. (Tooke et al., 2016) and Tezcan et al. (Tezcan et al., 2018), in this thesis we have proposed several advanced strategies to enhance the classification accuracy of clutter images using deep learning. We briefly summarize the main contributions of this thesis below.

6.1.1 Enhanced Data Augmentation and Class-Weighted Loss Function

In Chapter 3, to address the issue of insufficient size of clutter-image datasets, we proposed an enhanced data-augmentation method. This method doubles the maximum range of horizontal and vertical image shift, adds rotation, and incorporates color-jitter transformation, thus significantly diversifying the original dataset. Furthermore, we proposed a class-weighted loss function to mitigate the impact of class imbalance in the original dataset. By assigning weights that are inversely-proportional to the number of samples in each class, this approach ensures equitable learning across all classes. These enhancements led to an improvement in CCR by 3.82% points and in CCR-1 by 1.23% points compared to the baseline method of Tezcan et al. (Tezcan et al., 2018). Clearly, increasing dataset size by additional data augmentation and addressing class imbalance by modifying the loss function has brought substantial performance gains.

6.1.2 Replacement of ResNet-18 by Vision Transformer

Chapter 4 explored the implementation of the Vision Transformer model, as a replacement for the ResNet-18 model proposed by Tezcan et al. (Tezcan et al., 2018). Employing ViT with the aforementioned data augmentation and class-weighted loss function resulted in a 7.19% points improvement in CCR and a 6.38% points improvement in CCR-1 compared to ResNet-18 with baseline augmentation. These are very large performance gains resulting from using a completely different deep-learning architecture, originally developed for language applications. The main question is what other architectures can be considered to further boost the CIR classification performance.

6.1.3 AI-Driven Data Augmentation

In Chapter 5, we proposed AI-driven data augmentation to further expand our dataset. We evaluated the impact of image generation by DALL·E-2 and DALL·E-3 on classification of a 90-image dataset using ViT and ResNet-18 and concluded that DALL·E-3-generated images offer a substantial performance advantage over DALL·E-2-generated images. Interestingly, this improvement occurred despite some images being unrealistic and even corresponding to an incorrect CIR value (clutter appearance was changed during AI-based image generation). Therefore, with help from professionals specializing in hoarding disorder, we revised the AI-generated images by either accepting them as is, changing their CIR value or removing them. This reduced their number by about half. Unfortunately, the revised AI-generated dataset did not result in further improvement of performance - CCR dropped by over 1% point but CCR-1 increased by about 0.3% points. In a final test, we assessed the impact of DALL·E-3-generated images on classification performance using three 4-fold cross-validation scenarios. Unfortunately, the results were not what we hoped for.

Expanding only the training folds by AI-generated and verified images resulted in a slight increase in CCR-1 but appreciable drop in CCR. We believe that the new AI-generated images bring new characteristics into the dataset (visualized using t-SNE), that skew the learning process and lead to performance drop. Still, AI-based dataset augmentation seems to be a promising direction, but requires further research.

In conclusion, this thesis introduced enhanced data-augmentation techniques, a novel class-weighted loss function, and replaced a traditional ResNet-18 model with a more advanced ViT model. These modifications significantly improved performance metrics of clutter-image classification, elevating CCR-1 from a baseline of 0.8176 to 0.8814, and CCR from 0.4341 to 0.5060. An additional AI-driven data augmentation further increased CCR-1 to 0.8839, but dropped CCR to 0.4900, suggesting a further exploration of this direction of study.

6.2 Future Work

There exist several directions that may further improve the proposed methodologies:

- **Dataset Expansion**: While this work proposes to mitigate the impact of small and unbalanced datasets through data-augmentation methods, a significant expansion of the original dataset is necessary to fully address these issues.

- **Mobile Integration**: The ultimate goal of automatically rating CIR is deployment in the field with social workers and other health and human service professionals who would use a mobile device for real-time clutter-image classification. Selecting a new model that is faster, more accurate, and lighter than ViT, such as EfficientFormer (Li et al., 2022), is needed for practical adoption of the proposed methodology at the edge.

- **Enhancing Model Robustness with Regularization**: To prevent overfit-

ting when training larger models on limited datasets, future research should explore integrating L1 and L2 regularization techniques into the training process. Regularization helps reduce model complexity and ensures generalization by penalizing large weights, thereby enhancing the model's practical usability and reliability.

Appendix A

DALL·E-3 Image-Generation Examples

On the next 5 pages, we show sample images of room clutter generated by DALL·E-3 by means of *Image-to-Text-to-Image* sequence of steps for CIR classes 1, 3, 4, 8, 9. First, the original (source) image from our dataset is fed into GPT-4 to generate a caption. Then, this caption is fed into DALL·E-3 to produce an image. For each value of CIR we show two sample images produced by DALL·E-3.

Original image from the dataset: CIR = 1.

Caption: The image shows a tidy and comfortably arranged bedroom. The room's design suggests a modern aesthetic with a neutral color palette. Here are the key features of the room: A large bed with a wooden headboard is centered against the main wall. It is neatly made with crisp white linens and a decorative comforter that has a pattern of leaves or vines. There are multiple pillows arranged at the head of the bed: two large white pillows at the back, two medium-sized pillows, and two smaller black decorative pillows at the front. Above the bed, the wall inclines upward, revealing that the room is likely on the top floor with a sloped ceiling, a feature often seen in attic conversions or rooms with dormer windows. Skylights are installed in the sloped ceiling, letting in natural light and likely offering views of the sky. On the left side of the bed, there is a nightstand with a lamp and what appears to be a photo frame or small artwork.

Image generated by DALL·E-3.

Image generated by DALL·E-3.

Figure A·1: Original image, its generated caption and DALL·E-3-generated images for CIR = 1.

Original image from the dataset: CIR = 3.

Caption: This image depicts a residential kitchen with a cluttered countertop. It gives the impression of being in a state post-meal preparation or cooking, with a variety of utensils and ingredients left out. On the left, there's a white refrigerator adorned with various magnets, notes, and what looks like a collection of children's drawings. The fridge is situated next to wooden cabinetry that extends around the kitchen. The countertop is covered with an assortment of kitchen items: there are several stainless steel pots and lids, a colander, mixing bowls, and a measuring cup. Also visible are a pepper mill, salt shaker, some glass bottles (one being a clear soda bottle), a white coffee maker, and a small stack of white plates. Other items like a whisk, a spatula, kitchen towels, a box of plastic wrap, and numerous other containers and utensils are scattered across the countertop. In the back, open shelves store more kitchenware, including plates, cups, and bowls. Through the window above the counter, one can see a glimpse of the outdoor area, suggesting the kitchen has plenty of natural light during the day. The room appears lived-in and functional, with a warm and casual atmosphere, possibly reflecting everyday family life.

Image generated by DALL·E-3.

Image generated by DALL·E-3.

Figure A·2: Original image, its generated caption and DALL·E-3-generated images for CIR = 3.

Original image from the dataset: CIR = 4.

Caption: The image depicts a cluttered room, which appears to be a living space that is quite messy and disorganized. In the foreground, there is a bed with crumpled white bedding, and no one is visible on the bed. The floor is covered with various items such as clothing, bags, boxes, papers, and other miscellaneous objects, which give the impression of disarray. Towards the center of the room, there is a wooden TV stand with a CRT television on top and some home decor items, such as vases and picture frames. To the left of the TV stand is another wooden furniture piece, possible a cabinet or kitchenette area, with appliances on top, including what looks like a microwave and a toaster. The counter space seems to be occupied by various kitchen items, and a glimpse of the sink area suggests it is full as well. The room extends further back, with a partially visible shelf holding books or other items. On the right, you can see an open door that leads to another room, where some clothes and additional items are visible. The walls are adorned with posters, pictures, and other hanging items, adding to the personal touch of the living space.

Image generated by DALL·E-3.

Image generated by DALL·E-3.

Figure A·3: Original image, its generated caption and DALL·E-3-generated images for CIR = 4.

Original image from the dataset: CIR = 8.

Caption: This image shows an extremely cluttered room with a vast array of items haphazardly strewn about, indicating a severe case of hoarding. There are many cardboard boxes, some labeled with the word "DINEX," suggesting they may have originally contained dinnerware or kitchen items. The room appears to be filled to capacity with an assortment of objects, including books, bags, a wicker basket, clothing, and what looks like a chandelier on the floor among the chaos. In the middle, there's a brown chair piled with various items, making it unusable. A wide assortment of items can be seen, suggesting that the space serves multiple purposes, or that items have simply been placed without any organizational system. Among the visible items, there are also what appear to be yarn balls, possibly indicating that someone in the home may have knitting as a hobby. The lighting in the photo is dim, which might imply that the room lacks proper lighting or that the windows are obstructed, which is common in hoarding situations. Overall, the state of the room suggests a living space that has become dysfunctional due to excessive clutter.

Image generated by DALL·E-3.

Image generated by DALL·E-3.

Figure A·4: Original image, its generated caption and DALL·E-3-generated images for CIR = 8.

Original image from the dataset: CIR = 9.

Caption: The image depicts a cluttered and disorganized indoor space, seemingly a kitchen that is overwhelmed with various items, suggesting a severe case of hoarding. Here are the key elements observed: appliances and Furniture There is a white refrigerator on the left side with its door barely visible due to the clutter. On top of the refrigerator, there are more stacked items, mostly small boxes and loose pieces of paper. In front of the fridge, there is a gas stove, with its surface covered by various items, including what appears to be a blackened pot. The right side of the photo shows a part of a gray plastic storage container. The bulk of the image shows an enormous amount of clutter. Piles of cardboard boxes, both intact and broken down, are strewn about. There are also visible plastic containers, several bottles, and cans that appear to be food items or condiments. The mess covers the floor, making it difficult to distinguish a clear pathway. A variety of items can be seen amidst the chaos, such as kitchen utensils, food packaging, a striped piece of fabric.

Image generated by DALL·E-3.

Image generated by DALL·E-3.

Figure A·5: Original image, its generated caption and DALL·E-3-generated images for CIR = 9.

Appendix B

T-SNE Distribution Visualization for Individual CIR Classes

In Figure 5·3, we visualized a joint t-SNE distribution of image embeddings of the dataset of 1,233 natural images and of 128 images generated by DALL·E-3 and verified by professionals specializing in hoarding disorder. Since that visualization is quite dense, it is difficult to discern overlap patterns of symbols. Therefore, in Figures B·1-B·9 in this appendix, we show visualizations of t-SNE distributions of image embeddings of 128 DALL·E-3-generated and verified images (for CIR classes 1, 3, 4, 8, 9) jointly with the embeddings of natural images from a *single* CIR class from our dataset at a time. This allows easier interpretation of results.

It is clear from Figure B·1 that the generated images with CIR class 1 (red symbols "1") have substantial overlap with CIR class 1 images from the dataset (black symbols "1"); with few exceptions the red symbols "1" are located among the black symbols "1". However, there is minimal overlap between the generated images with CIR class 3 and natural images with the same class (Figure B·3), and almost no overlap for generated images with classes 4, 8 and 9 (Figures B·4, B·8 and B·9). Finally, as expected, the 128 generated images with CIR classes 1, 3, 4, 8, and 9 have effectively no overlap with natural images from the dataset for classes 2, 5, 6, 7. This observation aligns with our conjecture in Section 5.3, suggesting that the newly-generated images indeed introduce new content. Consequently, this leads to skewed results when the generated data are used only during the training process but not during testing.

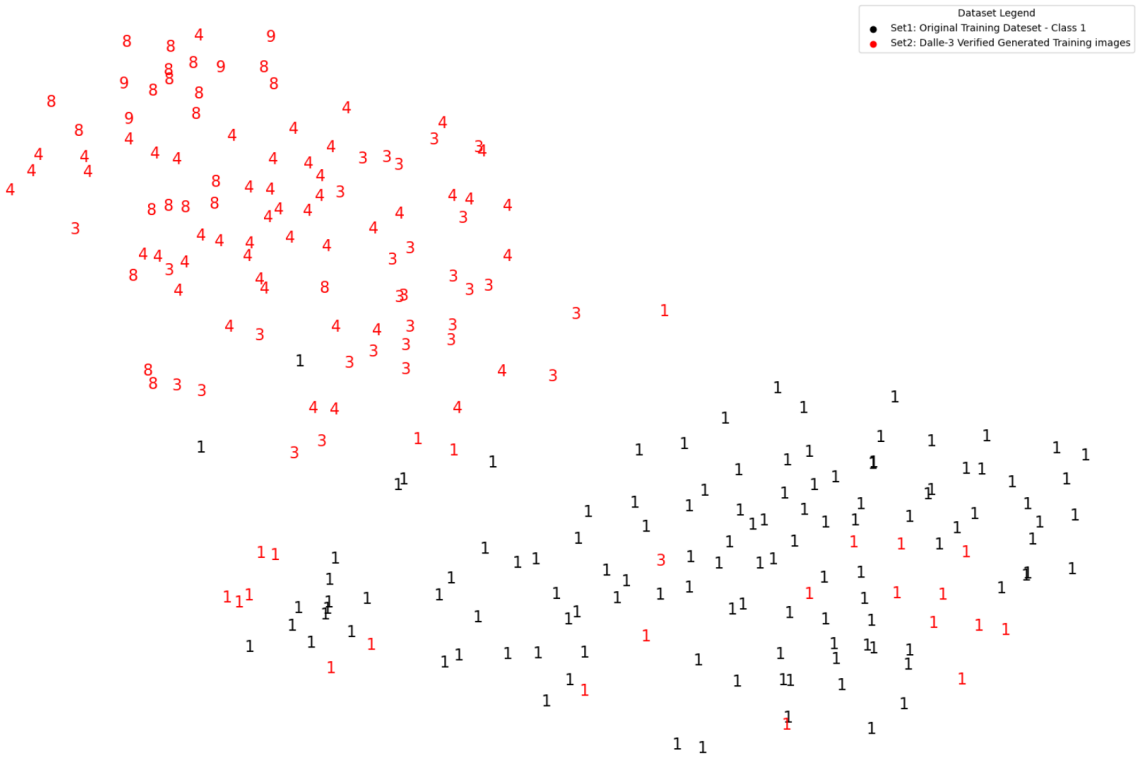

Figure B·1: Joint t-SNE visualization of embeddings for dataset images with CIR class 1 (118 natural images shown as black digits "1") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

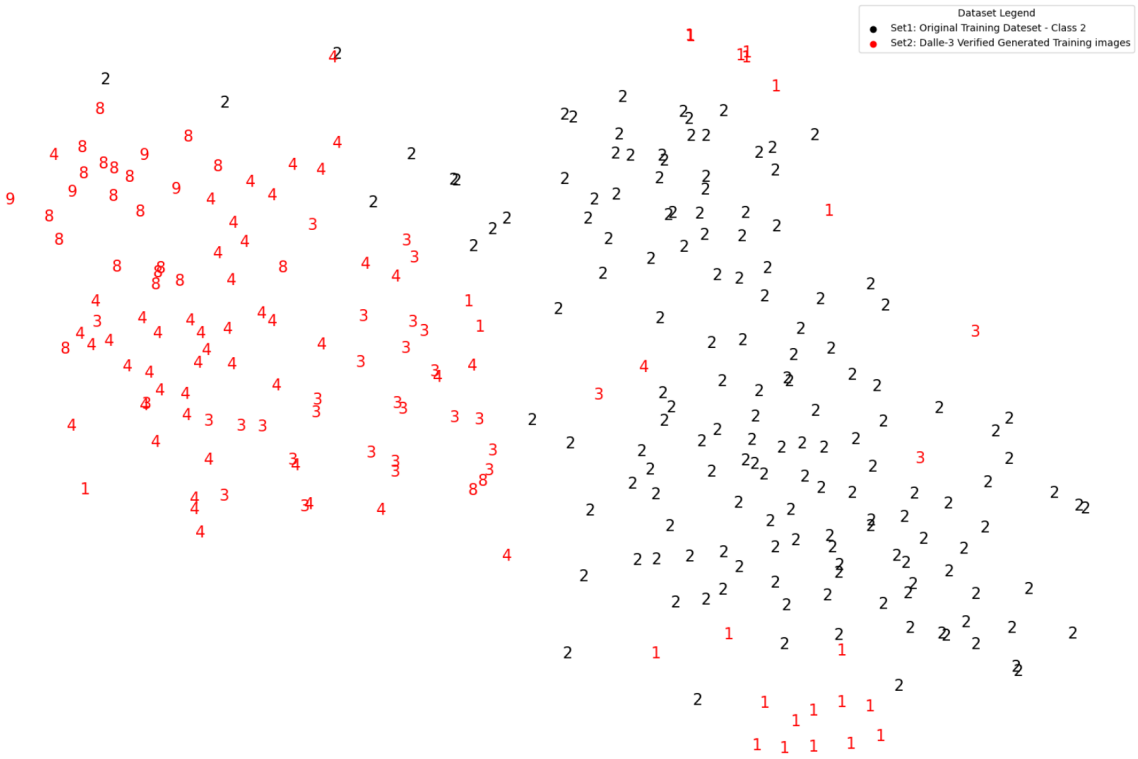

Figure B·2: Joint t-SNE visualization of embeddings for dataset images with CIR class 2 (153 natural images shown as black digits "2") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

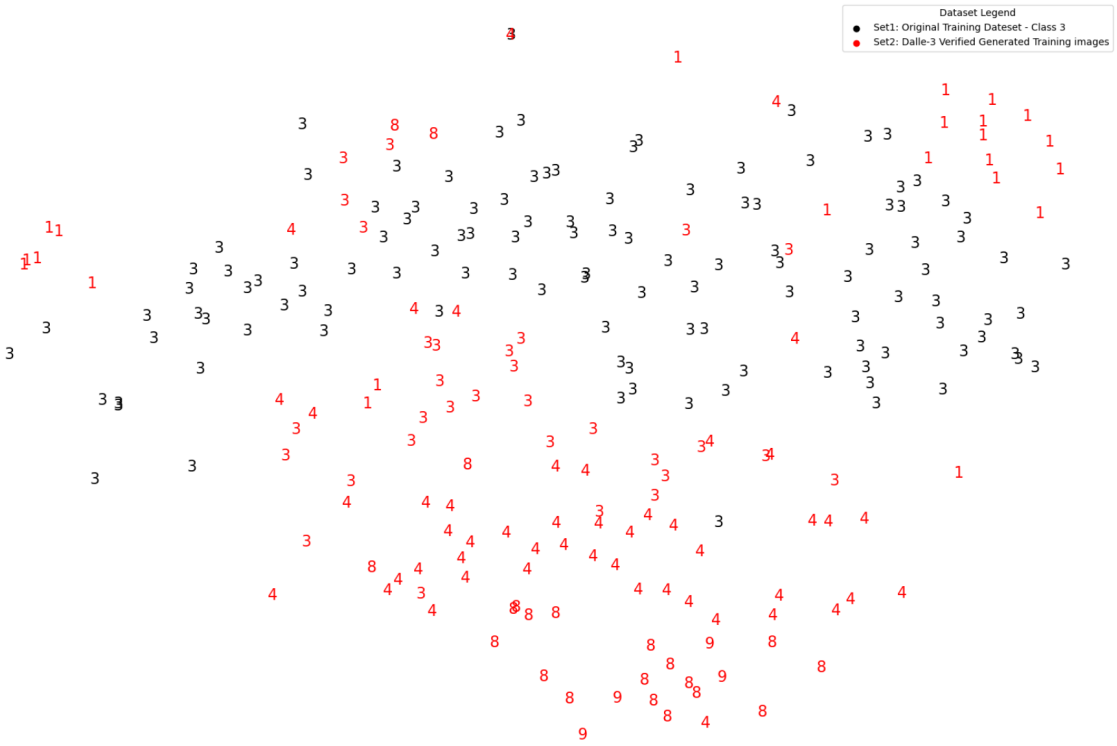

Figure B·3: Joint t-SNE visualization of embeddings for dataset images with CIR class 3 (117 natural images shown as black digits "3") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

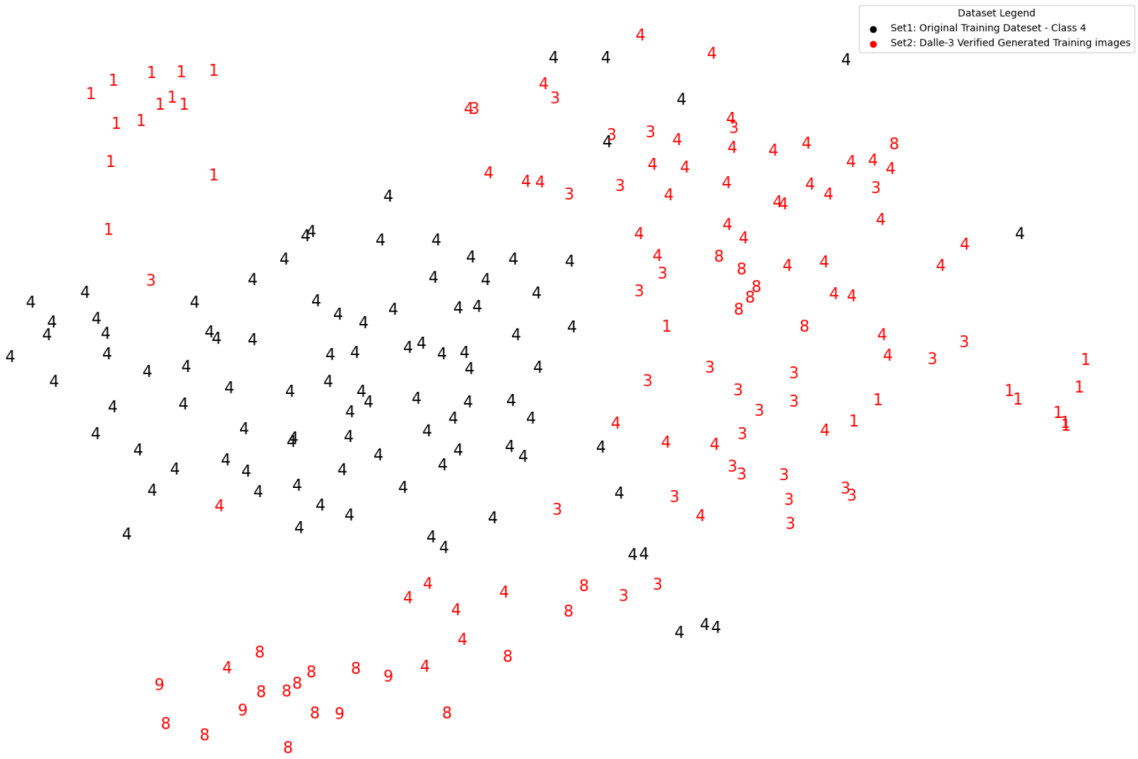

Figure B·4: Joint t-SNE visualization of embeddings for dataset images with CIR class 4 (97 natural images shown as black digits "4") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

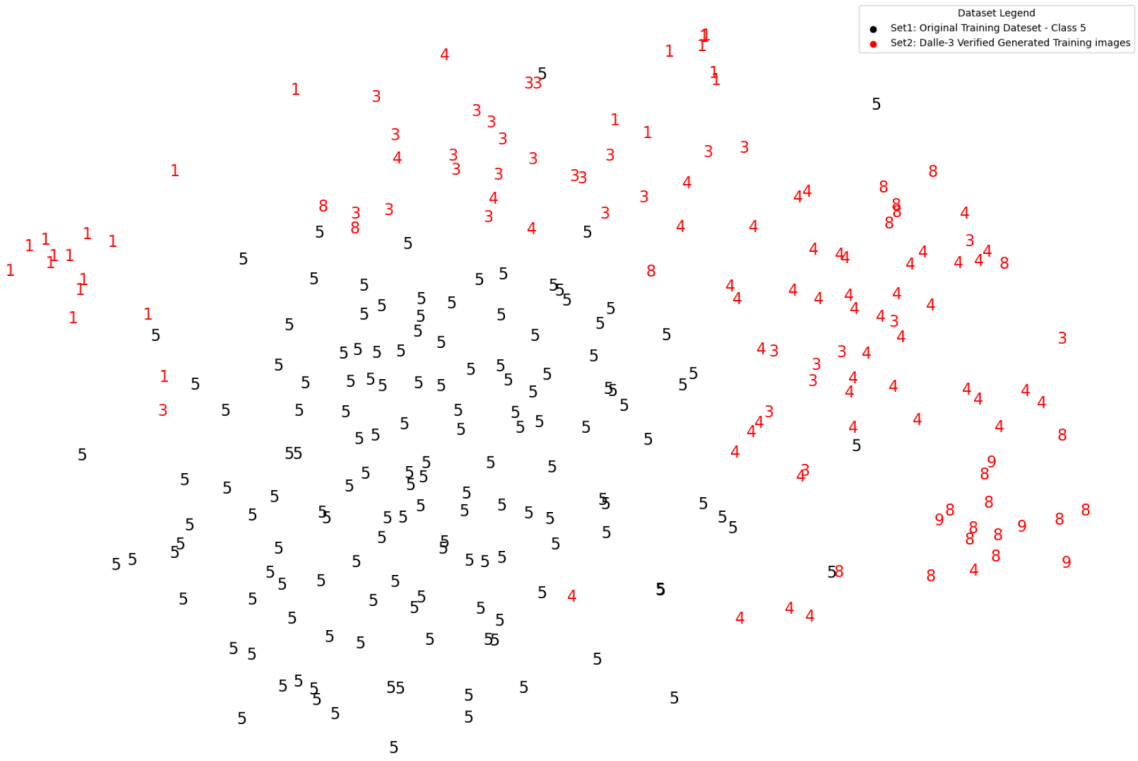

Figure B·5: Joint t-SNE visualization of embeddings for dataset images with CIR class 5 (146 natural images shown as black digits "5") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

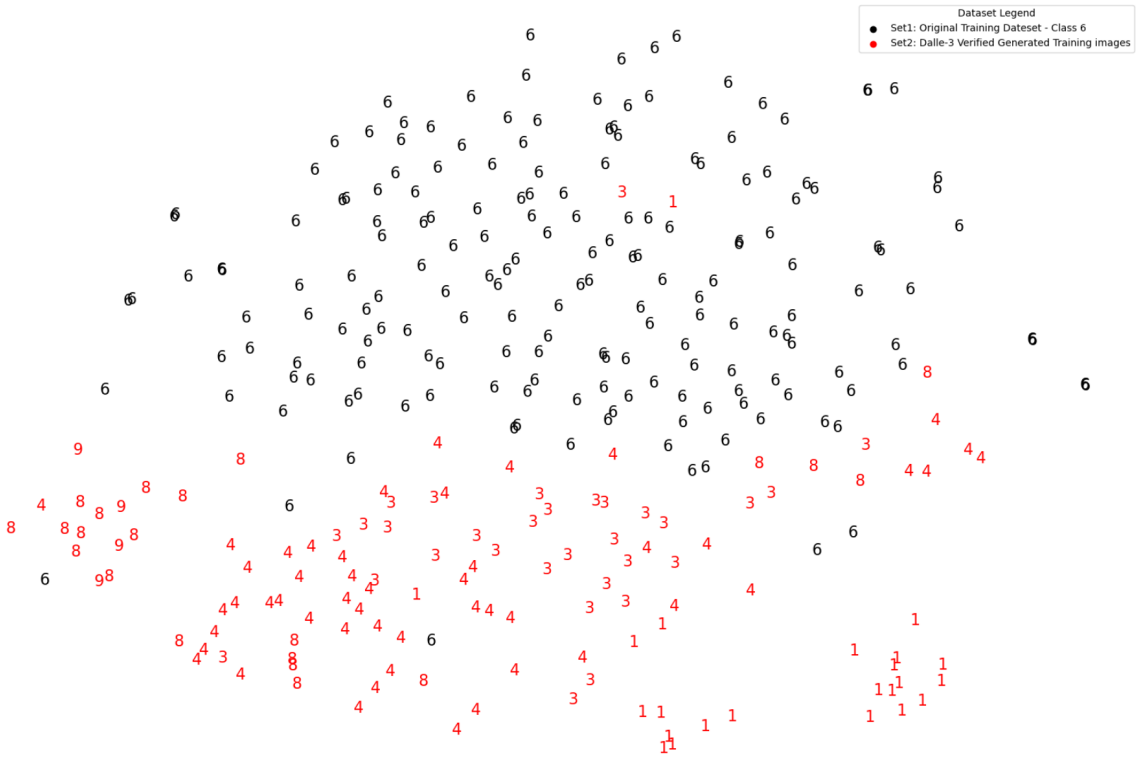

Figure B·6: Joint t-SNE visualization of embeddings for dataset images with CIR class 6 (181 natural images shown as black digits "6") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

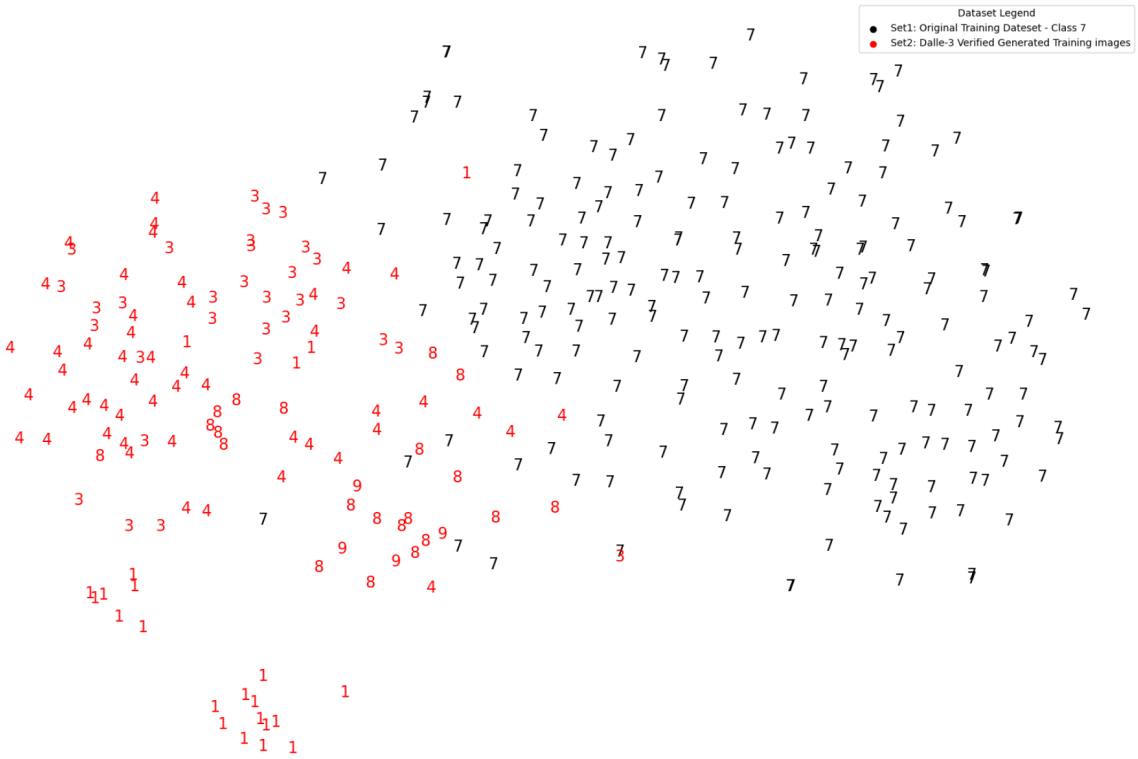

Figure B·7: Joint t-SNE visualization of embeddings for dataset images with CIR class 7 (215 natural images shown as black digits "7") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

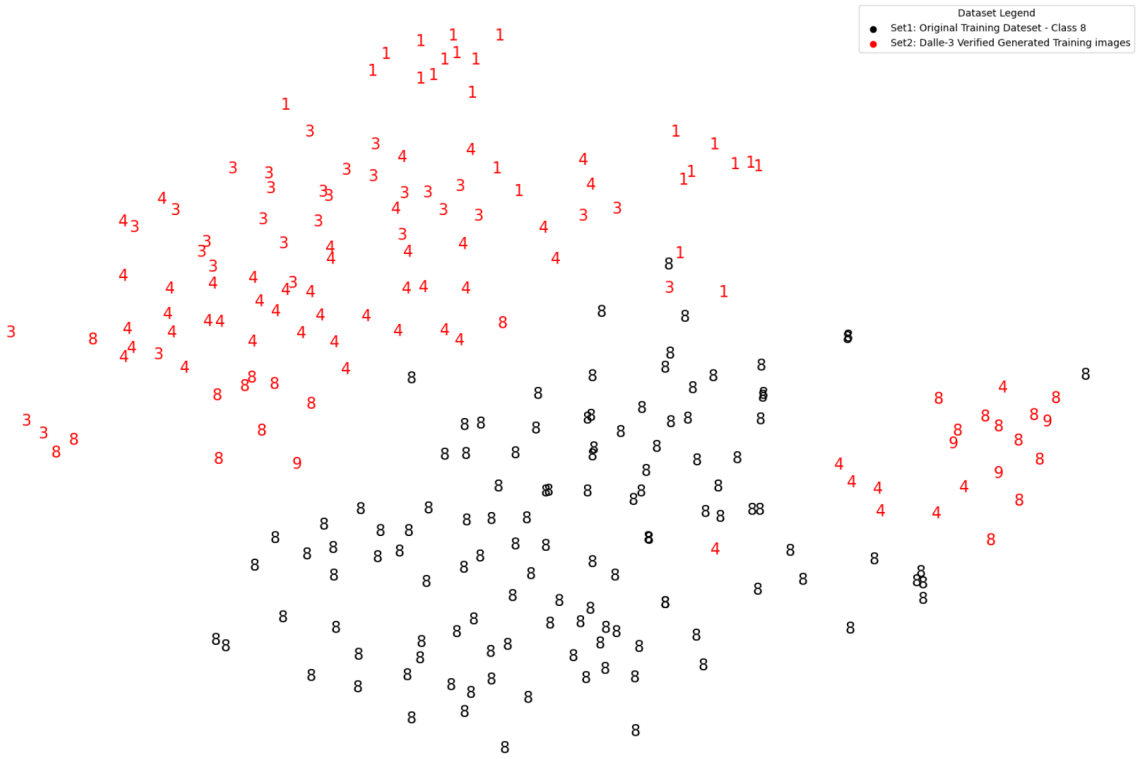

Figure B·8: Joint t-SNE visualization of embeddings for dataset images with CIR class 8 (119 natural images shown as black digits "8") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

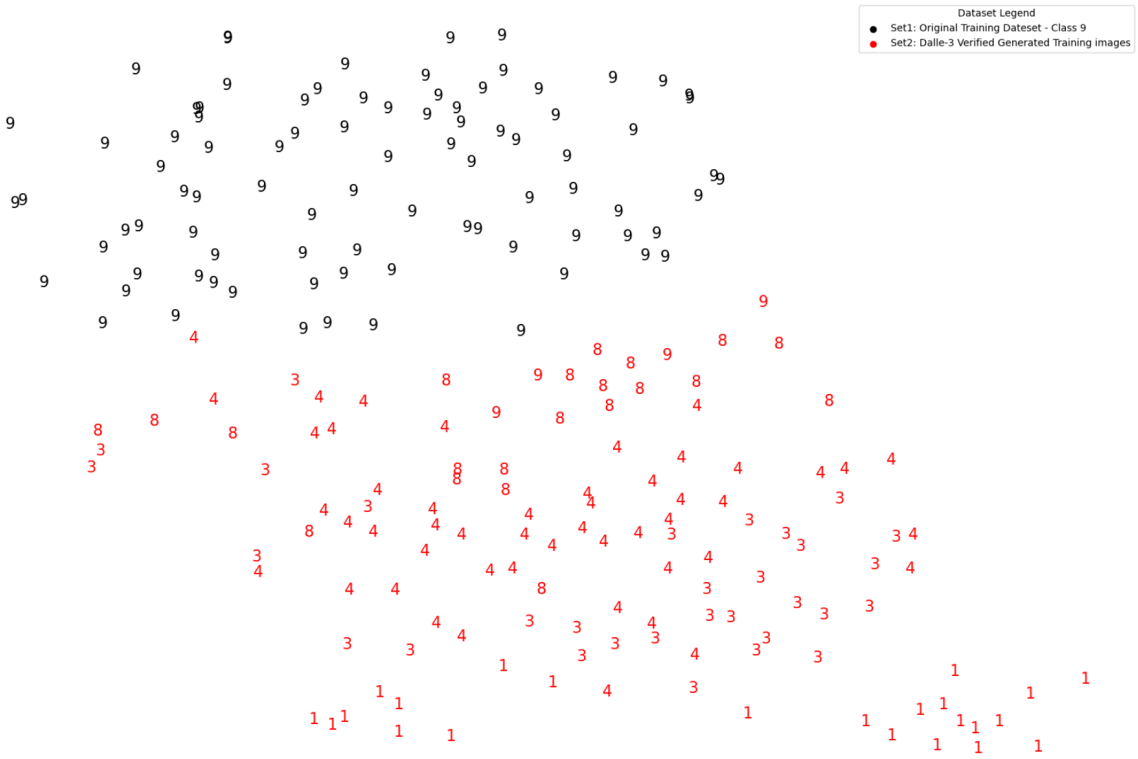

Figure B·9: Joint t-SNE visualization of embeddings for dataset images with CIR class 9 (87 natural images shown as black digits "9") and 128 DALL·E-3-generated and verified images (red digits denoting CIR classes).

References

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Publishing, Washington, DC, 5th edition.

Cai, T., Chen, M., and Chu, C. (2023). Gpt-4. https://openai.com/gpt-4.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, I:886–893.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021 - 9th International Conference on Learning Representations.

Frost, R. O. and Gross, R. C. (1993). The hoarding of possessions. Behaviour research and therapy, 31(4):367–381.

Frost, R. O. and Hartl, T. L. (1996). A cognitive-behavioral model of compulsive hoarding. Behaviour research and therapy, 34(4):341–350.

Frost, R. O., Steketee, G., and Grisham, J. (2004). Measurement of compulsive hoarding: saving inventory-revised. Behaviour Research and Therapy, 42(10):1163–1182.

Frost, R. O., Steketee, G., Tolin, D. F., and Renaud, S. (2008). Development and validation of the clutter image rating. Journal of Psychopathology and Behavioral Assessment, 30(3):193–203.

Frost, R. O., Steketee, G., and Williams, L. (2000). Hoarding: a community health problem. Health & Social Care in the Community, 8(4):229–234.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem:770–778.

Iervolino, A. C., Perroud, N., Fullana, M. A., Guipponi, M., Cherkas, L., Collier, D. A., and Mataix-Cols, D. (2009). Prevalence and heritability of compulsive hoarding: A twin study. The American Journal of Psychiatry, 166(10):1156–1161.

Kellman-McFarlane, K., Stewart, B., Woody, S., Ayers, C., Dozier, M., Frost, R. O., Grisham, J., Isemann, S., Steketee, G., Tolin, D. F., and Welsted, A. (2019). Saving inventory – Revised: Psychometric performance across the lifespan. Journal of affective disorders, 252:358.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems, 25.

Li, Y., Yuan, G., Wen, Y., Hu, E., Evangelidis, G., Tulyakov, S., Wang, Y., and Ren, J. (2022). Efficientformer: Vision transformers at mobilenet speed. ArXiv, abs/2206.01191.

Mataix-Cols, D. (2014). Hoarding Disorder. New England Journal of Medicine, 370(21):2023–2030.

Ramesh, A., Gray, S., Goh, G., and Pavlov, M. (2021). Dall·E: Creating images from text. https://openai.com/research/dall-e.

Samuels, J. F., Bienvenu, O. J., Grados, M. A., Cullen, B., Riddle, M. A., Liang, K. Y., et al. (2008). Prevalence and correlates of hoarding behavior in a community-based sample. Behavioral Research and Therapy, 46:836–844.

Saxena, S., Ayers, C. R., Dozier, M. E., and Maidment, K. M. (2015). The UCLA Hoarding Severity Scale: Development and Validation. Journal of affective disorders, 175:488.

Saxena, S., Ayers, C. R., Maidment, K. M., Vapnik, T., Wetherell, J. L., and Bystritsky, A. (2011). Quality of life and functional impairment in compulsive hoarding. Journal of Psychiatric Research, 45(4):475–480.

Tezcan, M. O., Konrad, J., and Muroff, J. (2018). Automatic assessment of hoarding clutter from images using convolutional neural networks. Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, 2018-April:109–112.

Tolin, D. F., Frost, R. O., and Steketee, G. (2010). A brief interview for assessing compulsive hoarding: The Hoarding Rating Scale-Interview. Psychiatry Research, 178(1):147–152.

Tolin, D. F., Frost, R. O., Steketee, G., Gray, K. D., and Fitch, K. E. (2008). The economic and social burden of compulsive hoarding. Psychiatry Research, 160:200–211.

Tolin, D. F., Gilliam, C. M., Davis, E., Springer, K., Levy, H. C., Frost, R. O., Steketee, G., and Stevens, M. C. (2018). Psychometric properties of the hoarding rating scale-interview. Journal of Obsessive-Compulsive and Related Disorders, 16:76–80.

Tooke, A., Konrad, J., and Muroff, J. (2016). Towards automatic assessment of compulsive hoarding from images. Proceedings - International Conference on Image Processing, ICIP, 2016-Augus:1324–1328.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 2017-Decem:5999–6009.

Wightman, R. (2022). Pytorch image models (timm). https://timm.fast.ai/.

CURRICULUM VITAE

Zhenghao Sun

Boston, MA • szh1007@bu.edu

Education

- **M.S.** in Electrical and Computer Engineering
  Boston University, Boston, MA — Sept 2022 – May 2024
- **B.Eng.** in Telecommunication Engineering
  Huaqiao University, Xiamen, China — Sept 2018 – Jun 2022

Research Experience

- **Image-Based Classification of Hoarding Clutter Using Deep Learning**
  - Proposed a novel automatic clutter-assessment method from images using deep learning; the classification follows the Clutter Image Rating (CIR) scale and allows ± 1 tolerance from the ground-truth CIR value (similar to uncertainty exhibited by human raters).
  - Applied a combination of diverse data-augmentation techniques and cross-validation methods to mitigate overfitting and generalization issues due to relatively small dataset size for training large models.
  - Proposed a novel weighted-loss function based on class-size distribution and achieved a 8% accuracy increase by integrating data augmentation with the ViT model.
- **Image Classification with Directional Image Sensors**
  - Simulated the replacement of the first convolutional layer in LeNet-5 CNN with an Optical Transfer Function's (OTF) of a directional image sensor to reduce computational and power requirements.
  - Adjusted network architecture and parameters; analyzed Floating-Point Operations per Second (FLOPs) and accuracy to assess model complexity and performance.
  - Reduced the number of FLOPs of the combined OTF+LeNet optical-digital system by 16.8% with only a 0.1% accuracy drop on on MNIST-Digits dataset.

- **Image Tracking Technology Based on Particle Capture and Tracking**
  - Developed an enhanced nearest-neighbor algorithm for tracking, counting, and analyzing motion of ultrafine particles or water droplets under a microscope, including their average speed, acceleration, trajectories, and quantity.
  - Employed the Sobel gradient operator for edge detection in particle area and combined morphological processing with advanced threshold-based segmentation for analyzing each frame of particle videos under uneven lighting conditions.
  - Leveraged MATLAB's `regionprops` function for dynamic video frame detection and recognition, applying it to synthetic data to validate the reliability and accuracy of the algorithms in particle tracking scenarios.

Project Experience

- **Cervical Spine Fracture Detection**
  - Created a data loader for preprocessing training and test datasets from RSNA, including functions for segmentation-mask extraction, image cropping, intensity adjustment, and normalization.
  - Implemented 3D DenseNet121 multilabel classifier from MONAI library to evaluate each segmented vertebrae for fractures, as well as overall fracture estimation.
  - Post-processed data, achieving a localized accuracy of nearly 90% and fracture detection accuracy exceeding 95%.
- **Video Noise Reduction**
  - Implemented 2-D Non-Local Means (NLM) filtering algorithm for noise reduction in video and optimized its performance by incorporating Principal Component Analysis (PCA) intensity adjustment and normalization.
  - Extended NLM filtering to the temporal domain, resulting in a 3-D NLM algorithm that accounts for temporal relationships and enhances denoising performance.
  - Applied the Block-Matching and 3D Filtering (BM3D) algorithm, which leverages low-pass filtering in the transform domain, to eliminate high-frequency noise components in video.

- **Portrait Stylization**
  - Implemented the VGG-16 model, leveraging high-level semantic retention and low-level textural learning in CNNs, achieving the separation of style and content.
  - Used DualStyleGAN to address the limitations of VGG-16 in handling complex facial features in portraits, such as hair and wrinkles, while preserving facial characteristics.
  - Designed and integrated a user-centric Graphical User Interface (GUI) with QT, enabling efficient and precise visualization of stylized images, thus improving the overall user accessibility and experience.

Teaching Experience

- **EC503 Grader**
  - Assessed and graded assignments, providing detailed feedback to students to guide their understanding of key concepts in data science and machine learning.
  - Supported the course instructor by preparing assessment materials and contributing to the development of grading rubrics that align with course objectives.
  - Collaborated with the teaching team to ensure consistency in grading and to identify common areas where students required additional support or clarification.

Honors and Awards

- Outstanding Student Leaders Award, Huaqiao University 2020 & 2019
- First-class Scholarship, Huaqiao University 2021 & 2020
- Second-class Scholarship, Huaqiao University 2019

Other Skills

- **Programming Languages:** Java, Python, Matlab, JavaScript, C/C++, Kotlin
- **Web Development:** HTML, CSS, React, Node.js, Spring, REST, AWS, GCP
- **Others:** Linux, Git, PyTorch, TensorFlow