Classification of Indoor Clutter from Images: Application to Hoarding Assessment

Zhenghao Sun
ECE Department
Boston University
Boston, MA, USA
szh1007@bu.edu

Jordana Muroff
School of Social Work
Boston University
Boston, MA, USA
jmuroff@bu.edu

Janusz Konrad
ECE Department
Boston University
Boston, MA, USA
jkonrad@bu.edu

***Abstract*—Hoarding is a mental-health problem manifested by excessively saving items irrespective of their value. One factor considered in the assessment of hoarding severity is the amount of clutter in a dwelling, usually quantified through a visual scale called the "Clutter Image Rating" (CIR). This requires a visit to an individual's home and rating clutter on the CIR scale, a time-consuming, subjective, and often non-repeatable process. To date, several methods were proposed for automatic rating of clutter from images but were evaluated on relatively narrow and unbalanced datasets. In this paper, we introduce a new 1,800-image, balanced dataset of clutter images that has been CIR-rated by health professionals. We also propose a new method for rating clutter that is based on the Vision Transformer. We evaluate the proposed method against a state-of-the-art clutter-rating method on two datasets via 4-fold cross-validation. We also perform two ablation studies (loss-function parametrization and data augmentation). In quantitative comparisons, we measure accuracy and accuracy within ± 1 since even health and human-service professionals admit to challenges in assigning exact CIR values. The proposed method is shown to outperform the best method to-date by 4.50-7.12% points in exact CIR matching and by 5.80-6.53% points in matching with a slack of ± 1. Even more importantly, the new method achieves accuracy of over 93% with a slack of ± 1 suggesting it can be a reasonable proxy for ratings by health professionals and a valuable tool in the assessment and treatment of hoarding disorder.**

***Index Terms*—Hoarding, Room clutter, Image clutter, Vision Transformer, ResNet, Deep learning**

I. Introduction

Hoarding disorder (HD) is a complex and impairing mental-health and public-health problem characterized by persistent difficulty and distress associated with discarding ordinary items regardless of their value and resulting in clutter in the living space [1]. In severe cases, hoarding poses health risks, including fires, falls, and poor sanitation [2]. In general, the quality of life of a person with HD is markedly, adversely affected [3], and family relationships are often strained [4]. In the United States, the prevalence of HD is about 5% of adult population [5] and is a serious social issue [4].

HD is identified through a detailed psychological assessment with the individual involved, preferably carried out in their home to properly evaluate the clutter and how it affects their life [6]. In 2008, a novel method, called the "Clutter Image Rating" (CIR), was introduced [7]. It proposes a set of 9 reference images with varying levels of clutter in a living room, bedroom and kitchen (Fig. 1) for assessing the severity of hoarding. The CIR method allows individuals with hoarding challenges, their family members, trained experts, or independent evaluators to measure the clutter in an individual's living space by visually matching the space to one of reference images (CIR = 1 corresponds to an uncluttered space, whereas CIR = 9 corresponds to a fully-cluttered space). However, this approach is time-consuming, subjective, and can lack consistency in its repeatability.

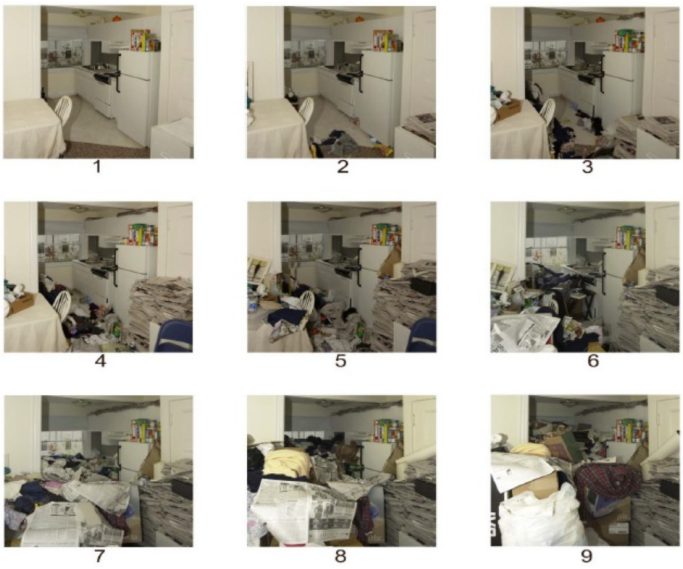

Fig. 1. Reference kitchen images proposed by Frost et al. [7] for image-based assessment of hoarding clutter according to CIR scale. Numbers shown below images are the assigned CIR values.

In the last few years, automated CIR assessment methods have been developed with benefits of being instantaneous, objective (not dependent on assessor's mood, subjectivity, etc.) and repeatable (the same image always results in the same CIR value). Tooke *et al.* [8] introduced two methods combining Histogram of Oriented Gradients (HOG) [9] feature extractor with Support Vector Machine (SVM), used either as regressor or classifier, to assess CIR value from an image. On a set of 620 images, their SVM-based classifier outperformed SVM-based regressor, and achieved 72% accuracy in assessing CIR value within ± 1 off the ground truth using 4-fold cross-

validation. Subsequently, Tezcan *et al.* [10] proposed to use ResNet-18 deep-learning model and expanded the clutter-image dataset to 1,323 images. Pre-trained on ImageNet [11] and then fine-tuned and tested on the new dataset their approach achieved 81% accuracy within ± 1 off the ground-truth CIR in 4-fold cross-validation. This was a significant result since the HOG+SVM approach of Tooke *al.* [8] re-tested on the new dataset achieved only 60%.

In this paper, we expand the dataset developed by Tezcan *et al.* [10] to 1,800 images; we increase the variety of clutter scenarios and ensure balanced class memberships. We also develop a new image-clutter classification method based on the Vision Transformer (ViT) [12]. Since the collection and rating of hoarding-related images is very difficult and time-consuming, the new dataset is still relatively small. Therefore, we expand data augmentation introduced by Tezcan *et al.* [10] to support training. We compare our method with Tezcan *et al.*'s ResNet-18 model [10] on their dataset and on the new dataset using two accuracy metrics (with and without slack). We also perform two ablation studies, one regarding data augmentation and the other regarding loss-function parametrization to balance two accuracy objectives.

We make 3 contributions in this paper:

1) we introduce a new clutter-image dataset,
2) we propose a new clutter classification algorithm and enhanced data augmentation for its training,
3) we evaluate performance of the new algorithm against state of the art in rating clutter from images.

II. New Clutter-Image Dataset

Collecting and rating images of hoarding clutter is difficult and labor-intensive. First, finding such images is very challenging. Although quite a few videos can be found on-line, one has to select frames that are sufficiently different from one another, field of view is sufficiently wide, no people are recognizable (privacy), logos are not obtrusive, etc. Secondly, each image must be rated by a health professional to assign a CIR value. This can be problematic since even professionals have sometimes challenges with precise assignment of a CIR value (e.g., the rating may be between a 4 and a 5). This ambiguity impacts how we define the loss function and how we measure a method's performance.

We expanded the dataset developed by Tezcan *et al.* [10] from 1,323 to 1,800 images, a 36% increase, and made both publicly available as HINDER (Hoarding and INDoor cluttER) datasets. Unlike the previous dataset, the new dataset is balanced - all CIR classes contain 200 images, and includes a wider range of clutter scenarios. Table I lists the number of images for each CIR class in both datasets and provides download URLs. Fig. 2 shows one sample image from each class of the new dataset. Clearly, the value of CIR assigned to an image grows as the degree of clutter in a space increases. However, for higher degrees of clutter even professionals may have difficulty assigning an exact CIR value. As can be seen in the third row of Fig. 2, images rated as CIR = 7 and CIR = 8 have clutter reaching up to about one-half of room's height. However, in the CIR = 7 image the window is almost fully visible and the two door frames at the back of the room are filled up with items up to about half of their height. On the other hand, in the CIR = 8 image items reach up to about 80% of the door-frame height and are more evenly spread-out and up to a higher level, resulting in larger volume of clutter. This assessment is relative and subjective, but absent physical measurement of clutter volume it is the only option.

III. Problem Statement

We formulate CIR assessment as a supervised classification problem. Let $\{\mathbf{I}_k \in \mathbb{R}^{w\times h\times 3}\}_{k=1}^{N}$ be a set of N color images of width w and height h, and let $\xi_k \in \{1, 2 \ \ldots \ 9\}$ be the ground-truth CIR rating for image $\mathbf{I}_k$. Given N image-CIR pairs $(\mathbf{I}_1, \xi_1), (\mathbf{I}_2, \xi_2), \ \ldots \ , (\mathbf{I}_N, \xi_N)$, the goal is to find a mapping $\mathbf{I}_k \rightarrow \widehat{\xi}_k$ to predict the CIR rating $\widehat{\xi}_k$ of an unseen clutter image $\mathbf{I}_k$. Note, that in all experiments we use cross-validation so each image is considered unseen at some point.

To measure performance, we use Correct Classification Rate (CCR) for it captures multi-class classification accuracy (sum of diagonal entries over sum of all entries in the confusion matrix). However, as we already discussed, assigning ground-truth CIR values bears some uncertainty. Therefore, we follow earlier work and, in addition to CCR, we also use its variant proposed in [8] to measure performance:

$$CCR_1 = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}(|\xi_k - \widehat{\xi}_k| \ \leq \ 1) \tag{1}$$

where $\mathbf{1}(x)$ is an indicator function (1 if x is true, 0 if x is false). The use of CCR_1 in addition to CCR is motivated by the fact that professionals encounter challenges when assigning CIR values. While CCR measures the exact accuracy of CIR estimates, CCR_1 measures accuracy within ± 1.

IV. CIR Prediction Using the Vision Transformer

A. Architecture adaptation

We adapt the Vision Transformer architecture [12] to our clutter classification problem as follows.

- **Preprocessing**: We resize all images to 224×224 pixels, and divide each image into 16×16 patches (blocks). This results in a structure $\mathbf{S} \in \mathbb{R}^{K\times(16\times 16\times 3)}$, where $K = 196$ is the number of patches in each image.
- **Patch embedding and positional encoding**: We map each patch to a vector of length $D = 256$ by means of a fully-connected layer. This results in a 2-D matrix $\mathbf{X} \in \mathbb{R}^{K\times D}$ of patch embeddings that is passed to the Transformer Encoder along with positional encoding (learnable 1-D embedding) of each patch. Another learnable embedding is prepended to $\mathbf{X}$ to convey image information to transformer output and then to MLP head.
- **Transformer Encoder**: Subsequent encoding operations are identical to those in the original Transformer model developed for language applications [13].
- **MLP head**: The output of the Transformer Encoder is fed into a simple MLP (single fully-connected layer) with a 9-class output to allow CIR classification.

TABLE I
NUMBER OF IMAGES IN EACH CIR CLASS AND DOWNLOAD URLS FOR THE 2018 AND NEW DATASETS.

CIR	Download URL	1	2	3	4	5	6	7	8	9	Total
HINDER-2018 dataset [10]	`vip.bu.edu/hinder-2018`	128	163	127	107	156	191	225	129	97	1,323
HINDER-2025 dataset (new)	`vip.bu.edu/hinder-2025`	200	200	200	200	200	200	200	200	200	1,800

Fig. 2. Sample images from all CIR classes in the new HINDER-2025 dataset. Note that dataset images have varying dimensions and aspect ratios. Images presented above were selected to have similar aspect ratios for visualization purposes and were resized to the same horizontal dimension.

B. Implementation of ViT-based CIR classification

Since ViT is a large model (330MB in our adaptation), training it from scratch with a dataset of 1,800 images is counterproductive. Instead, we employ transfer learning; we use `vit_base_patch16_224` model from the `timm` library [14] initialized with weights pre-trained on ImageNet [15]. We adapt this model to CIR classification by setting the number of output classes to 9, and we fine-tune the MLP head using our 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 stochastic gradient descent (SGD) for training and found that the learning rate of 0.001 that drops by half after every 5 epochs, performs best. For consistency with Tezcan *et al.*'s [10] experiments, we adopted a momentum of 0.9 and mini-batch size of 32.

V. LOSS FUNCTION

To achieve high accuracy (CCR), a commonly-used loss function is the cross-entropy. Applied in the context of one image-CIR pair number k, a *single-label* loss function can be written as follows:

$$\mathcal{L}_k^S = -\sum_{i=1}^{9} \boldsymbol{\xi}_k^1[i] \log \frac{\exp(\widehat{\boldsymbol{\xi}}_k[i])}{\sum_{j=1}^{9} \exp(\widehat{\boldsymbol{\xi}}_k[j])} \tag{2}$$

where $\boldsymbol{\xi}_k^1$ is a one-hot encoded vector of the ground-truth CIR value for image number k, and $\widehat{\boldsymbol{\xi}}_k$ is the output of the last layer of the MLP head (before *softmax*). The goal of this loss function is to achieve high accuracy without consideration for potential uncertainty in the ground-truth values. During prediction of CIR for image number k, we select the largest component of $\widehat{\boldsymbol{\xi}}_k$ (corresponding to the highest probability):

$$\widehat{\xi}_k = \arg\max_i(\widehat{\boldsymbol{\xi}}_k[i]). \tag{3}$$

As we pointed out, the CCR_1 metric (1) tolerates ± 1 errors. This requires a different problem definition - the ground truth is considered a *multi-label* value. The training data now include image-label pairs $(\mathbf{I}_k, \boldsymbol{\Xi}_k)$, where $\boldsymbol{\Xi}_k$ is a set of three consecutive CIR values, namely $\boldsymbol{\Xi}_k = \{\xi_k - 1,\ \xi_k,\ \xi_k + 1\}$. In this formulation, an image is associated with three different CIR labels (except for boundary cases of CIR = 1 and CIR = 9, when it has two labels only). Since we cannot assign three different labels during prediction, we need to find a function that maps an unseen image $\mathbf{I}_k$ to a CIR label $\widehat{\xi}_k \in \boldsymbol{\Xi}_k$.

To afford this type of classification, we use *multi-label*, binary cross-entropy between the *sigmoid* output of MLP head's last layer and a three-hot encoded ground truth. Applied to a single input-CIR pair number k, this loss function can be written as follows:

$$\mathcal{L}_k^M = -\sum_{i=1}^{9}\left(\boldsymbol{\xi}_k^3[i]\log\frac{1}{1+\exp(-\widehat{\boldsymbol{\xi}}_k[i])} + (1-\boldsymbol{\xi}_k^3[i])\log\frac{\exp(-\widehat{\boldsymbol{\xi}}_k[i])}{1+\exp(-\widehat{\boldsymbol{\xi}}_k[i])}\right) \tag{4}$$

where $\boldsymbol{\xi}_k^3[i]$ is a three-hot encoded vector of the ground truth, i.e., $\boldsymbol{\xi}_k^3[i]$ equals 1 for $i \in \boldsymbol{\Xi}_k$, and 0 otherwise. During prediction, we again choose $\widehat{\boldsymbol{\xi}}_k$'s largest component (3).

We follow Tezcan *et al.*'s [10] approach and linearly combine single- and multi-label loss functions for N images in a mini-batch:

$$\mathcal{L} = \sum_{k=1}^{N}(1-\lambda)\mathcal{L}_k^S + \lambda\mathcal{L}_k^M, \tag{5}$$

where parameter λ can be used to adjust the balance between CCR and CCR_1 performance.

VI. Data Augmentation

Due to a relatively small dataset size, both Tooke *et al.* [8] and Tezcan *et al.* [10] applied data augmentation by means of horizontal and vertical image shifts by 5, 10, or 15 pixels, and a horizontal "flip". However, the maximum shift of 15 pixels is very small even for 224×224 images, so very little visual information (clutter) is changed. To allow more significant visual "jitter", we increased the maximum range of shifts to ± 30 pixels while keeping 5-pixel increments. We also applied a horizontal "flip". Furthermore, since pictures of clutter are taken at a variety of angles (frequently not aligned with room features, e.g., door or window frames, room corners), we added an additional geometric augmentation by means of random image rotation up to ± 9 degrees in 1-degree increments. Finally, because of the diversity of cameras used as well as a 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. Examples of such augmentations can be found in [16].

VII. Experimental Results

In experiments below, we ran each scenario 10 times, each time over 50 epochs, and computed average CCR and CCR_1 from the highest respective values in the last 10 epochs.

A. Loss function tuning

In order to identify the value of weight parameter λ (5) that best balances algorithm performance in terms of CCR and CCR_1, we ran experiments for $0 < \lambda < 1$ with a step of 0.1. Fig. 3 shows CCR and CCR_1 as a function of λ for ResNet-18 and ViT-based algorithms on both datasets. Plots for both datasets exhibit similar trends. Unsurprisingly, CCR_1 increases with a growing λ since more and more weight is given to $\mathcal{L}_k^M$ (4) which allows ± 1 CIR mismatch. As for CCR, as expected, it is higher for $\lambda = 0$ than for $\lambda = 1$ since at $\lambda = 1$ a predicted CIR value is allowed to be within ± 1 off the ground truth, so CCR is likely to decrease. However, for $0.1 \leq \lambda \leq 0.9$ CCR slowly grows, which is surprising since less and less weight is given to $\mathcal{L}_k^S$ (2) so one would expect lower and lower CCR values. Taking a deeper dive into these results we observed that for $0.1 \leq \lambda \leq 0.9$ exact matches (contributing to CCR) may occur not only for images exactly-predicted by algorithms trained with either $\lambda = 0$ or $\lambda = 1$, but also for other images. We believe this is due to non-linear behavior of both ResNet-18 and ViT.

In Fig. 3, CCR (solid lines) is highest for $\lambda = 0.9$, while CCR_1 (dashed lines) is highest at $\lambda = 0.9$ for ViT and the new dataset, and otherwise it is a very close second. Therefore, we use $\lambda = 0.9$ in the remaining experiments.

B. Impact of data augmentation

One contribution of this work is enhanced data augmentation compared to earlier approaches. Table II shows CCR and CCR_1 performance for the ResNet-18 [10] and VIT-based methods with no augmentation, with baseline augmentation [8], [10], and with the new, enhanced augmentation (Section VI). As can be seen, the baseline augmentation improves performance of both methods compared to no augmentation by 1.19-3.60% points in terms of CCR and by 1.12-2.73% points in terms of CCR_1, across datasets and methods. The proposed enhanced augmentation further improves performance by 1.55-3.90% and 0.53-1.92% points, respectively.

Overall, the proposed method with new data augmentation outperforms ResNet-18 by 4.50-7.12% points in CCR and by 5.80-6.53% points in CCR_1 across the two datasets. Importantly, the new algorithm achieves CCR_1 of over 93% with CCR exceeding 53% on the new dataset. This suggests that *algorithmic* clutter rating close in performance to ratings by health professionals is within reach.

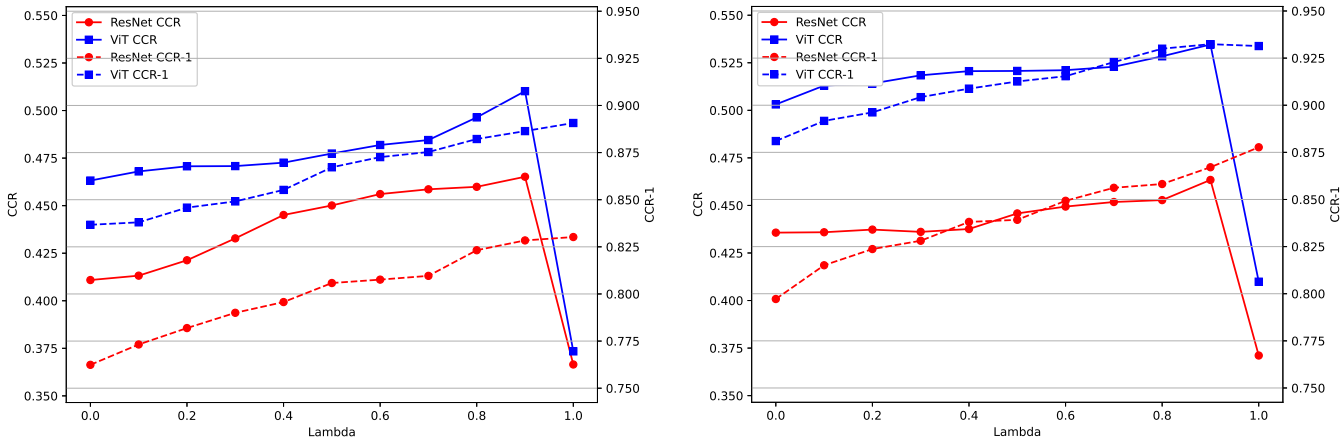

Fig. 3. Plots of CCR and CCR_1 for different values of λ (5) for ResNet-18 [10] and proposed ViT-based algorithms on both datasets.

TABLE II
IMPACT OF DATA AUGMENTATION ON PERFORMANCE OF RESNET-18 AND VIT-BASED CLUTTER-RATING METHODS ON TWO DATASETS ($\lambda = 0.9$).

Augmentation	ResNet-18 [10]		ViT (proposed)	
	CCR	CCR_1	CCR	CCR_1
	HINDER-2018 dataset [10]			
None	0.4243	0.8061	0.4476	0.8557
Baseline	0.4362	0.8173	0.4837	0.8811
Enhanced	0.4652	0.8284	0.5102	0.8864
	HINDER-2025 dataset (new)			
None	0.4064	0.8206	0.4897	0.9014
Baseline	0.4296	0.8479	0.5191	0.9187
Enhanced	0.4634	0.8671	0.5346	0.9324

VIII. CONCLUSIONS

We have developed a new method for rating living-space clutter from images which can be useful in the assessment of hoarding and other clutter-related health challenges. We also introduced a new dataset of clutter images, HINDER-2025, rated by health providers who work with people with hoarding, using the CIR. The new method, based on the Visual Transformer, outperforms the previous best clutter-rating method based on ResNet-18 by up to 7% points in terms of exact CIR matching and by up to 6.5% points in terms of matching within ± 1, when tested on two datasets. Importantly, the new method achieves accuracy of over 93% within ± 1 off the ground truth suggesting it can be a reasonable proxy for assessment by health and human-service professionals who sometimes have challenges with exact assignment of CIR value. Used by an individual with hoarding, family member or another trusted party *via* a smartphone/tablet app, our ViT-based clutter-rating method can be a valuable tool in the assessment and treatment of hoarding and other health challenges associated with clutter. We are currently developing such an app for field testing with housing authorities and other community partners.

REFERENCES

[1] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC: American Psychiatric Publishing, 2013.

[2] R. O. Frost, G. Steketee, and L. Williams, "Hoarding: A community health problem," Health & Social Care in the Community, vol. 8, no. 4, pp. 229–234, 2000.

[3] S. Saxena, C. R. Ayers, K. M. Maidment, T. Vapnik, J. L. Wetherell, and A. Bystritsky, "Quality of life and functional impairment in compulsive hoarding," J. of Psychiatric Research, vol. 45, no. 4, pp. 475–480, 2011.

[4] D. F. Tolin, R. O. Frost, G. Steketee, K. D. Gray, and K. E. Fitch, "The economic and social burden of compulsive hoarding," Psychiatry Research, vol. 160, pp. 200–211, 2008.

[5] A. C. Iervolino, N. Perroud, M. A. Fullana, M. Guipponi, L. Cherkas, D. A. Collier, and D. Mataix-Cols, "Prevalence and heritability of compulsive hoarding: A twin study," The American Journal of Psychiatry, vol. 166, no. 10, pp. 1156–1161, 2009.

[6] D. Mataix-Cols, "Hoarding Disorder," New England Journal of Medicine, vol. 370, no. 21, pp. 2023–2030, 2014.

[7] R. O. Frost, G. Steketee, D. F. Tolin, and S. Renaud, "Development and validation of the clutter image rating," Journal of Psychopathology and Behavioral Assessment, vol. 30, no. 3, pp. 193–203, 2008.

[8] A. Tooke, J. Konrad, and J. Muroff, "Towards automatic assessment of compulsive hoarding from images," in Proc. Int. Conf. on Image Processing (ICIP), 2016.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE/CVF Computer Vision and Pattern Recognition Conf. (CVPR), 2005.

[10] M. O. Tezcan, J. Konrad, and J. Muroff, "Automatic assessment of hoarding clutter from images using convolutional neural networks," in Proc. IEEE Southwest Symp. on Image Analysis and Interpret., 2018.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF Computer Vision and Pattern Recognition Conf. (CVPR), 2016.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. on Learning Representations (ICLR), 2021.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. of Advances in Neural Information Processing Systems (NeurIPS), 2017.

[14] R. Wightman, "Pytorch image models (timm)," https://timm.fast.ai/.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Proc. IEEE/CVF Computer Vision and Pattern Recognition Conf. (CVPR), 2009.

[16] Z. Sun, "Image-based classification of hoarding clutter using deep learning," Master's thesis, Boston University, May 2024. [Online]. Available: http://www.bu.edu/vip/files/pubs/theses/Sun24thesis.pdf