IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 5 January 2023, accepted 2 February 2023, date of publication 6 February 2023, date of current version 10 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242982

RESEARCH ARTICLE

Visual Probing: Cognitive Framework for Explaining Self-Supervised Image Representations

WITOLD OLESZKIEWICZ [1], DOMINIKA BASAJ [1,5], IGOR SIERADZKI [2], MICHAŁ GÓRSZCZAK [2], BARBARA RYCHALSKA [1,4], KORYNA LEWANDOWSKA [3], TOMASZ TRZCINSKI [1,2,5], (Senior Member, IEEE), AND BARTOSZ ZIELIŃSKI [2,6]

[1] Warsaw University of Technology, 00-661 Warszawa, Poland
[2] Faculty of Mathematics and Computer Science, Jagiellonian University, 31-007 Kraków, Poland
[3] Cognitive Neuroscience and Neuroergonomics, Institute of Applied Psychology, 30-060 Kraków, Poland
[4] Synerise, 30-383 Kraków, Poland
[5] Tooploox, 53-601 Wrocław, Poland
[6] Ardigen, 30-394 Kraków, Poland

Corresponding author: Witold Oleszkiewicz (witold.oleszkiewicz@pw.edu.pl)

This work was supported in part by the Foundation for Polish Science (carried out within the Team-Net Program co-financed by the European Union under the European Regional Development Fund) under Grant POIR.04.04.00-00-14DE/18-00; and in part by the National Science Centre, Poland, under Grant 2020/39/B/ST6/01511. The work of Dominika Basaj and Barbara Rychalska was supported by the National Science Centre, Poland, under Grant 2018/31/N/ST6/02273.

ABSTRACT Recently introduced self-supervised methods for image representation learning provide on par or superior results to their fully supervised competitors, yet the corresponding efforts to explain the self-supervised approaches lag behind. Motivated by this observation, we introduce a novel visual probing framework for explaining the self-supervised models by leveraging probing tasks employed previously in natural language processing. The probing tasks require knowledge about semantic relationships between image parts. Hence, we propose a systematic approach to obtain analogs of natural language in vision, such as visual words, context, and taxonomy. Our proposal is grounded in Marr's computational theory of vision and concerns features like textures, shapes, and lines. We show the effectiveness and applicability of those analogs in the context of explaining self-supervised representations. Our key findings emphasize that relations between language and vision can serve as an effective yet intuitive tool for discovering how machine learning models work, independently of data modality. Our work opens a plethora of research pathways towards more explainable and transparent AI.

INDEX TERMS Computer vision, explainability, probing tasks self-supervised representation.

I. INTRODUCTION

Visual representations are cornerstones of a multitude of contemporary computer vision and machine learning applications, ranging from visual search [8] to image classification [9] and Visual Question Answering, VQA [10]. However, learning representations from data typically requires tedious annotation. Therefore, recently introduced self-supervised representation learning methods concentrate on decreasing the need for data labeling without reducing their performance [1], [20], [21]. Because of the fundamental role representations play in real-life applications, much research focuses on explaining these embeddings [6], [15], [17]. Nevertheless, most of them concentrate on fully supervised embeddings [11] rather than on their self-supervised counterparts. Moreover, the majority of the proposed approaches rely on pixel-wise image analysis [13], [14], while general semantic concepts present in the images are often ignored.

Here, we attempt to overcome these shortcomings and draw inspiration from a simple yet often overlooked

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

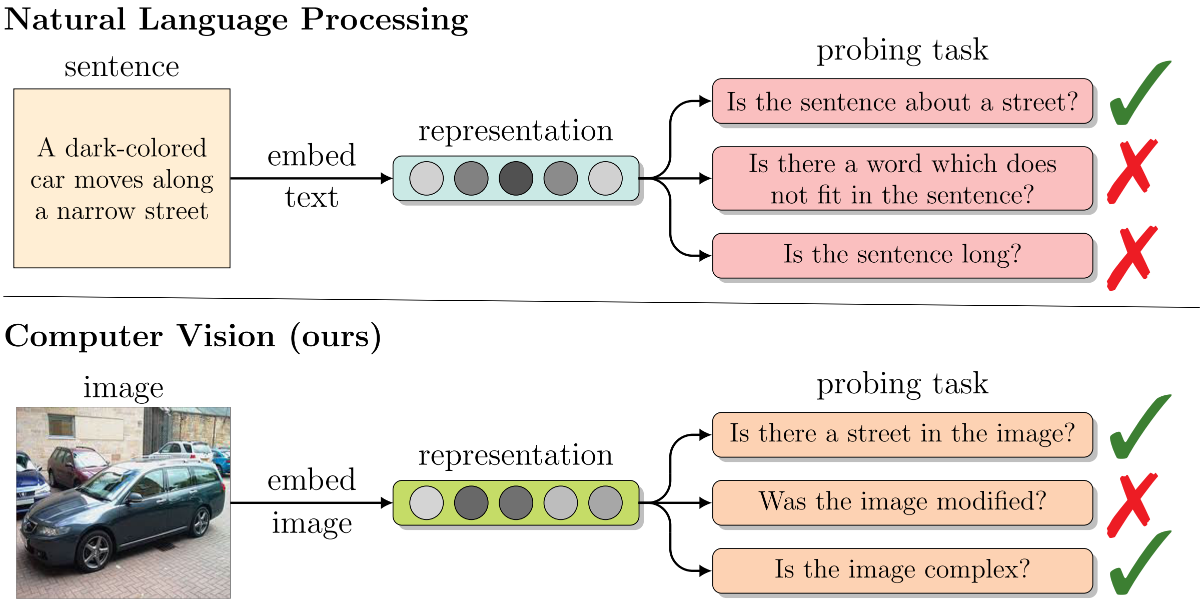

Natural Language Processing
sentence
A dark-colored car moves along a narrow street
embed text
representation
probing task
Is the sentence about a street? ✓
Is there a word which does not fit in the sentence? ✗
Is the sentence long? ✗
Computer Vision (ours)
image
embed image
representation
probing task
Is there a street in the image? ✓
Was the image modified? ✗
Is the image complex? ✓

FIGURE 1. ***Probing tasks*, widely used in natural language processing, validate if a *representation* implicitly encodes a given property, e.g., a sentence topic or its length. We introduce a visual taxonomy along with the corresponding probing framework that allows building analogous *visual probing tasks* and explain the self-supervised image representations. As a result, we e.g., discover that the information stored by self-supervised representations is more biased towards lines and forms than textures.**

observation that humans use language as a natural tool to explain what they learn about the world through their eyes [16]. Therefore, considering that the very same machine learning algorithms can be successfully applied to solve both vision and Natural Language Processing (NLP) tasks [23], [24], we postulate that the methods used to analyze text representation can also be employed to investigate visual inputs.

Very popular tools for explaining textual embeddings are *probing tasks* [18]. As shown in the upper part of Figure 1, a probing task in NLP is a simple classifier that asks if a given textual representation encodes a particular property, such as a sentence length or its semantic consistency, even though this property was not a direct training objective. For instance, we can create a textual probing task by substituting a word in a sentence and checking if a simple classifier that takes the representation of the original and altered sentence can detect this change. Furthermore, by analyzing the accuracy of a probing task, one can verify if the investigated representation contains certain information and understand the rationale behind embedding creation. However, while probing tasks are straightforward, intuitive, and widely used tools in NLP, their computer vision application is limited [12], mainly due to the lack of appropriate analogs between textual and visual modalities.

In this paper, we address this limitation by introducing an intuitive mapping between vision and language that enables applying the NLP probing tools in the computer vision (CV) domain. For this purpose, in Section III, we propose a taxonomy of visual units that includes *visual sentences*, *words*, and *characters*. We describe them using visual features presented in Marr's computational theory of vision [36], such as texture, shapes, and lines. Finally, we employ them as building blocks for a more general visual probing framework that contains a variety of NLP-inspired probing tasks, such as *Word Content*, *Sentence Length*, *Character Bin*, and *Semantic Odd Man Out* [17], [18]. The results we obtain provide us with unprecedented insights into semantic knowledge, complexity, and consistency of self-supervised image representations, e.g. we discover that semantics of the image only partially contribute to target task accuracy. One of our key findings is that the information stored by self-supervised representations is much more influenced by lines and forms than textures. What confirms the design choices behind hand-crafted visual representations such as SIFT [52] or BRIEF [53]. Our framework also allows us to compare the existing self-supervised representations from a novel perspective, as shown in Section VI.

Our contributions can be therefore summarized as follows:

- We propose an intuitive mapping between visual and textual modalities that constructs a visual taxonomy.
- We introduce novel visual probing tasks for comparing self-supervised image representations inspired by similar methods used in NLP.
- We show that leveraging the relationship between language and vision serves as an effective yet intuitive tool for discovering how self-supervised models work.

II. RELATED WORKS

The visual probing framework aims to explain image representations obtained from self-supervised methods. Moreover, it is inspired by probing tasks used in NLP. Therefore, in this section, we consider related works from three research areas: self-supervised computer vision models, probing tasks in natural language processing, and explainability methods in computer vision.

A. SELF-SUPERVISED COMPUTER VISION MODELS

Earliest self-supervised methods were based on a pretext task, for example, image colorization [45], or rotation prediction [46] using cross-entropy loss. However, recently published state-of-the-art methods usually base on contrastive loss [30], which measures the similarities of patches in representation space and aims to discriminate between positive and negative pairs. The positive pair contains modified versions of the same image, while the negative pairs correspond to two images in the same dataset. One of the methods, called MoCo v1 [27] trains a slowly progressing encoder, driven by a momentum update. This encoder plays the role of a large memory bank of past representations and delivers information about negative examples. Another method, called SimCLR v2 [1], proposes a different way of generating negative pairs, using a large batch size of up to 4096 examples. Other important improvements proposed by SimCLR v2 are the projection head and carefully tuned data augmentation. The projection head maps representations into space where contrastive loss is applied to prevent the loss of information. On the other hand, BYOL [20] also uses the projection head, but unlike MoCo v1 and SimCLR v2, it achieves a state-of-the-art performance without the explicitly defined contrastive loss function, so it does not need negative examples. Finally, SwAV [21] first obtains "codes" by assigning features to prototype vectors and then solves a "swapped" prediction problem wherein the codes obtained from one data augmented view are predicted using the other view. Our paper provides a framework for analyzing the representations generated by those methods regarding the semantic knowledge they encode.

B. PROBING TASKS IN NLP

NLP probing tasks aim to probe word or sentence representations for interesting linguistic features to discover whether they contain linguistic knowledge [48]. Probing is usually achieved with a binary or multi-class classifier, which takes one or two-word embeddings as input, and predicts the existence or absence of a chosen linguistic phenomenon in the input representation(s) [18]. The qualities of a good probing classifier are the subject of debate, as too expressive probes could learn important features on their own, even if the information is not present in the representations [5]. Thus, probing is usually achieved with simple classifiers.

Classic probing literature considers various linguistic aspects, from the simplest to very complex ones. In [18], the probed linguistic features are, for example, the depth of the sentence parse tree or whether the sentence contains a specific word. Other works propose to focus on lexical knowledge concerning the qualities of individual words more than the whole sentences [6], [17], probing token embeddings for qualities such as gender, case, and tense, or differentiation between real words and pseudowords [17]. Other approaches focus on certain kinds of words, e.g., function words, such as *wh*-words and propositions [4]. We consider all these objectives in our approach, i.e., we study probing tasks on individual concepts and their compositions. Moreover, while most works on probing tasks focus on one selected language, the others [17] are designed with multilingual settings in mind. It has been shown that it is possible to create NLP probing tasks that are transferable across languages, even if the languages vary considerably in their structure, which means that probing tasks can touch upon more universal cognitive phenomena [2]. This paper also aims at the flexibility and universality of our probing tasks, as our approach can be applied to various image domains.

C. EXPLAINABILITY METHODS IN CV REPRESENTATION LEARNING

eXplainable Artificial Intelligence (XAI) gains popularity fuelled by the black-box character of today's deep neural networks [19], [39]. Popular explainability approaches for model explanations are saliency or attention maps, which provide the importance of weights to pixels based on the first order derivatives [13], [14], [40], [41] but do not fully explain the reasoning behind the actual decision [54] and do not describe the concrete semantic concepts. Moreover, some of the methods are even agnostic of the model itself [13] and thus are not able to explain it. Another common local approach is perturbation-based interpretability, which applies changes to either data [55] or features [56] and observes the influence on the output.

Some methods verify the relevance of network hidden layers. For example, [12] uses linear classifiers trained on representations from these layers to measure how suitable they are for the classification. Subsequent efforts focused on understanding the function of hidden layers led to the introduction of network dissection [42], [43], which enables quantifying the interpretability of latent representations by evaluating the alignment between their hidden units and a set of visual semantic concepts obtained from human annotators.

More recent methods are inspired by the human brain and how it explains its visual judgments by pointing to prototypical features that an object possesses [57]. I.e., a certain object is a car because it has tires, a roof, headlights, and a horn. For example, prototypical part network [44] applies this paradigm by focusing on parts of an image and comparing them with prototypical parts of a given class. At the same time, the extension proposed in [58] uses data-dependent merge-pruning of the prototypes to allow sharing them among the classes. Another promising approach is Concept Activation Vector (CAV), defined in the feature space to quantify the degree to which a predefined concept is vital for a prediction [26]. This approach has recently been extended to automatically discovered concepts [19] and to interactive techniques used by pathologists to indicate what characteristics are essential when searching for similar images [59].

We propose to continue and extend this line of research by introducing visual word probing, which systematically explains the self-supervised representations. Our work presents a framework that focuses on model analysis. It interprets the internal representation of the deep learning model using visual probing tasks, e.g., it shows which semantic concepts are included in the representation and to what extent.

III. VISUAL PROBING

This section introduces a novel visual probing framework that analyzes the information stored in self-supervised image representations. For this purpose, in Section III-A, we propose a mapping between visual and textual modalities that constructs a visual taxonomy. As a result, the image becomes a "visual sentence" constructed from "visual words" and can be analyzed with visual probing tasks inspired by similar methods used in NLP (see Section III-C). Moreover, for in-depth analysis of the concepts trained by self-supervised methods, in Section III-B, we provide a cognitive visual systematic that identifies a visual word with structural features from Marr's computational theory [36].

A. MAPPING BETWEEN VISION AND NLP

After defining the images as analogous to sentences within our framework, the question remains which parts of an image should be considered equivalent to individual words and characters? There are multiple possible answers to this question. One of the intuitive ones is to divide an image into non-overlapping superpixels that group pixels into perceptually meaningful atomic regions, e.g., using SLIC algorithm [28]. As a result, we obtain an image built from superpixels, an analogy of a sentence built from words. The superpixels, similarly to words, have their order and meaning (see Section III-B). Moreover, each superpixel contains a specific number of pixels, like the number of characters in a word. As a consequence, we obtain an intuitive mapping between visual and textual domains.

However, superpixels differ conceptually from their linguistic counterparts in one important aspect: they do not repeat between different images, while in text, the words often repeat between sentences. Therefore, we propose to define visual words as the clusters of all training superpixels in representation space and assign each superpixel to the closest centroid from such a dictionary. For this purpose, we could use the original definition of visual words from [25]. However, it does not take into consideration the importance of those words for a model's prediction. Therefore, we use TCAV methodology [19], [26] that generates high-level concepts, which are important for prediction and easily understandable by humans. Such an approach requires a supervisory training network but generates visual words independent of the analyzed self-supervised techniques, which is crucial for a fair comparison. To summarize, the process of dividing an image into visual words consists of three steps: segmentation into superpixels, their encoding, and assignment to visual words (see Figure 2).

B. COGNITIVE VISUAL SYSTEMATIC

In contrast to words in NLP, visual words do not have a well-defined meaning required for in-depth analysis of self-supervised representations. Hence, in this section, we introduce cognitive visual systematic, considering that generating visual words is similar to the process of concept formation. This process, described in psychology and cognitive science, is traditionally understood as an internal cognitive representation of a set of similar objects, i.e., "an idea that includes all that is characteristically associated with it" [37]. In other words, concepts are created in relation to features that constitute similarity amongst included objects.

What features could then be the basis for the formation of visual words? Reference to Marr's computational theory of vision [35], [36] seems to be an appropriate aid in answering this question. Marr assumed that perception is achieved by detecting an object's specific structural features, which are then organized in a series of visual representations. Among those, three constitute the major representations: the "primal sketch", the "2.5D sketch" and the 3D model representation" [36]. The primal sketch is a two-dimensional image that uses information on light intensity changes, featuring blobs, edges, lines, boundaries, bars, and terminations. Colors and textures are also thought to be detected on this level [34], [38]. The 2.5D sketch represents mostly two-dimensional shapes and their orientation towards a viewer-centered location (the sense of image depth is achieved in this stage [35]). Finally, the 3D model is a representation suitable for object recognition. In this stage, the observer can imagine the object from different views. This includes surfaces that are currently invisible to the observer [35], [36].

To simplify visual word description in terms of Marr's theory, we decided to use concepts of light intensity (brightness), color, texture, and lines in relation to primal sketch, shape in relation to 2.5D sketch, and form in relation to 3D model (examples are depicted in Figure 3). Our initial analysis of individual visual words shows that these six categories from Marr's theory describe very well the particular types of our visual words. In the process of creating visual words (see Figure 2), similar superpixels cluster together. As a result, we generate different visual words consisting of similar lines, similar shapes, similar colors, etc. To confirm that our observations are not accidental, we conducted user studies that confirmed our assumptions and categorized visual words into specific categories from Marr's theory, such as brightness, color, texture, lines, forms, and shapes. This user study helps us to establish the meaning of the visual words we use.

C. VISUAL PROBING TASKS

After dividing an image into visual words, its representation can be analyzed by the visual probing framework that can adapt most NLP probing tasks. Here, we describe adaptations

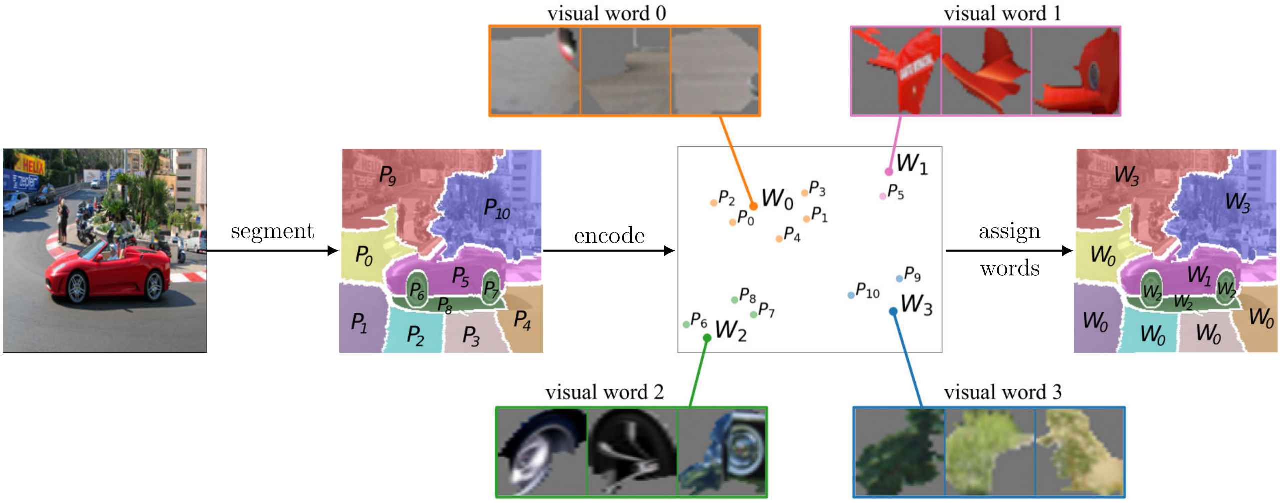


FIGURE 2. The process of dividing an image into visual words. First, an image is segmented into multiple superpixels: $P_0, P_1, \ldots, P_{10}$. Then, each superpixel is embedded in the latent space previously used to generate the dictionary of visual words: W_0, W_1, W_2, W_3. Finally, each superpixel is assigned to the closest word in the visual word dictionary. This results in mapping between vision and language and enables using the visual probing framework that includes a variety of NLP-inspired probing tasks.

of five of them, including those well known in the NLP community [17], [18] together with their original NLP definitions to make the paper self-contained.

1) WORD CONTENT (WC)

The *Word Content* probing task aims to identify which visual words are present in an image (see Figure 4). The *input* of this probing task is a self-supervised representation of the image. The *target labels* represent the presence of a particular visual word. As we describe in Section IV, all visual words are clustered into 50 clusters. Hence, there are 50 binary *target labels*. Figure 2 illustrates the process of determining which visual words are present in the image. This is similar to the bag of words representation.

The NLP inspiration of this task probes for surface information, i.e., the type of information that does not require any linguistic knowledge [18]. In contrast, its adaptation requires *semantic knowledge* to understand which concept is represented by a superpixel.

2) SENTENCE LENGTH (SL)

The aim of the *Sentence Length* probing task is to distinguish between simple and complex images, as presented in Figure 5. The *input* of this probing task is a self-supervised representation of the image. The *target label* is the number of unique visual words in the image, which can be determined based on the WC labels. The original NLP probing task predicts the number of words (or tokens) and retains only surface information [18]. In CV, it serves as a proxy for *semantic complexity*, requiring the semantic understanding of the image.

3) CHARACTER BIN (CB)

The aim of the *Character Bin* probing task is to check whether the representation stores information about the complexity of the visual word represented by a superpixel. The *input* of this probing task is a self-supervised representation of the image's superpixel, and we define two *target labels* that are commonly used in CV literature to describe superpixels. The first target label is the compactness (CO) [49] of the superpixel S defined as the area of the superpixel $A(S)$ divided by the area $A(C)$ of a circle C with the same perimeter as S:

$$CO(S) = \frac{A(S)}{A(C)}.$$

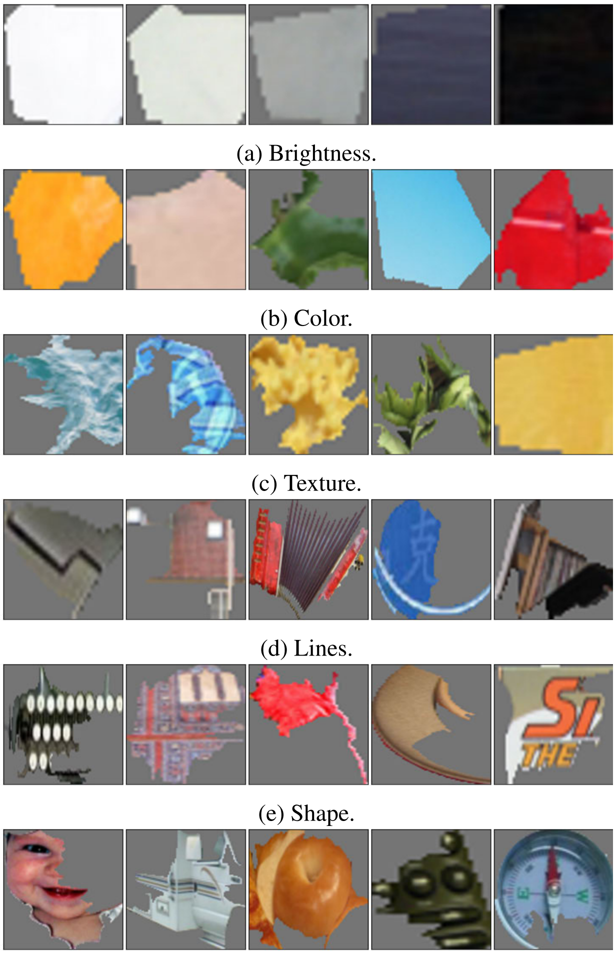
Sample superpixels with various ranges of CO are presented in Figure 6a. The second target label is Intra-Cluster Variation (ICV) [50] defined as the average standard deviation $\sigma_c(S)$ of channels $c \in C$ for superpixel S:

$$ICV(S) = \frac{1}{|C|}\sum_{c \epsilon C} \sigma_c(S).$$

Sample superpixels with various ranges of ICV are in Figure 6b. The original NLP probing task is defined as a classifier of the number of characters in a single word [17]. From this perspective, the *Character Bin* retains only surface information in both domains.

4) SEMANTIC ODD MAN OUT (SOMO)

The objective of the SOMO probing task is to predict whether the image was modified. We replace a center-biased superpixel in the image with a similarly shaped superpixel from another image that corresponds to different visual words. We pick a superpixel using a two-dimensional Gaussian distribution center in the middle of the image. Regarding replacement, we consider two setups, SOMO close and far, depending on how often two visual words co-occur in the training images. In SOMO close, we replace a center-biased superpixel with visual words that often co-occur with the replaced visual word. In SOMO far, we replace superpixel


(a) Brightness.

(b) Color.

(c) Texture.

(d) Lines.

(e) Shape.

(f) Form.

FIGURE 3. Sample superpixels illustrate six visual concepts from the Marr's computational theory of vision.

with the rarely co-occurring visual word (see Figure 7). In both cases, the *input* of the probing task is a self-supervised representation of the image. The *target label* is binary, i.e., whether the image was modified or not. The original NLP task predicts if replacing a random noun or verb alters the sentence [18]. In both domains, it requires the ability to detect alterations in *semantic consistency*.

5) MUTUAL WORD CONTENT (MWC)

The *Mutual Word Content* (MWC) probing task aims to discover which visual words bring two self-supervised representations close to each other and which ones push them farther away (see Figure 8). The *input* of this probing task is a pair of self-supervised representations of two images. The *target labels* represent the presence of a particular visual word in both images. The probing task classifier is validated on equally-sized subsets $\{S^i_{val}\}$ corresponding to the increasing cosine distance between pairs of representations. Discrepancies in the classifier accuracy show the impact of visual words on the representations' distance. More precisely, if the probing task performance drops with increasing representations' distance, the visual word information in both representations brings them closer. To quantify this relationship, we introduce the attraction coefficient. To calculate this coefficient, we use the Linear Regression fit to the points (i, AUC_i), where i is the index of the subset and AUC_i is the MWC probing task performance on this subset. Thus, the attraction coefficient is the first derivative of the fitted model.

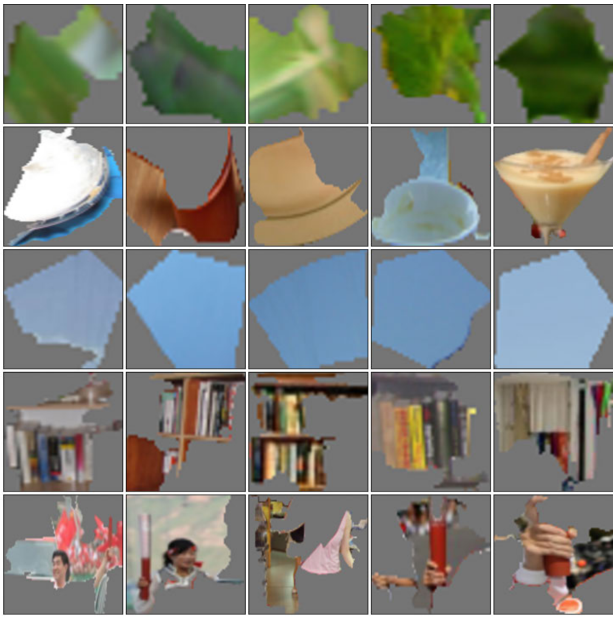


FIGURE 4. Sample visual words, each represented by one row of five superpixels.

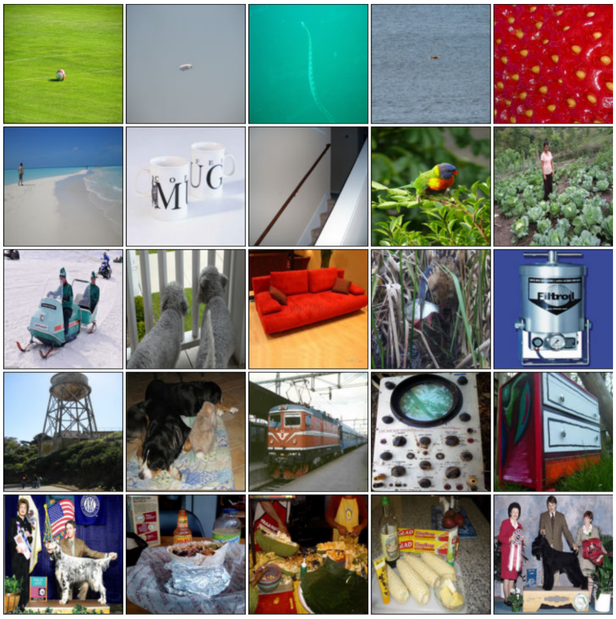
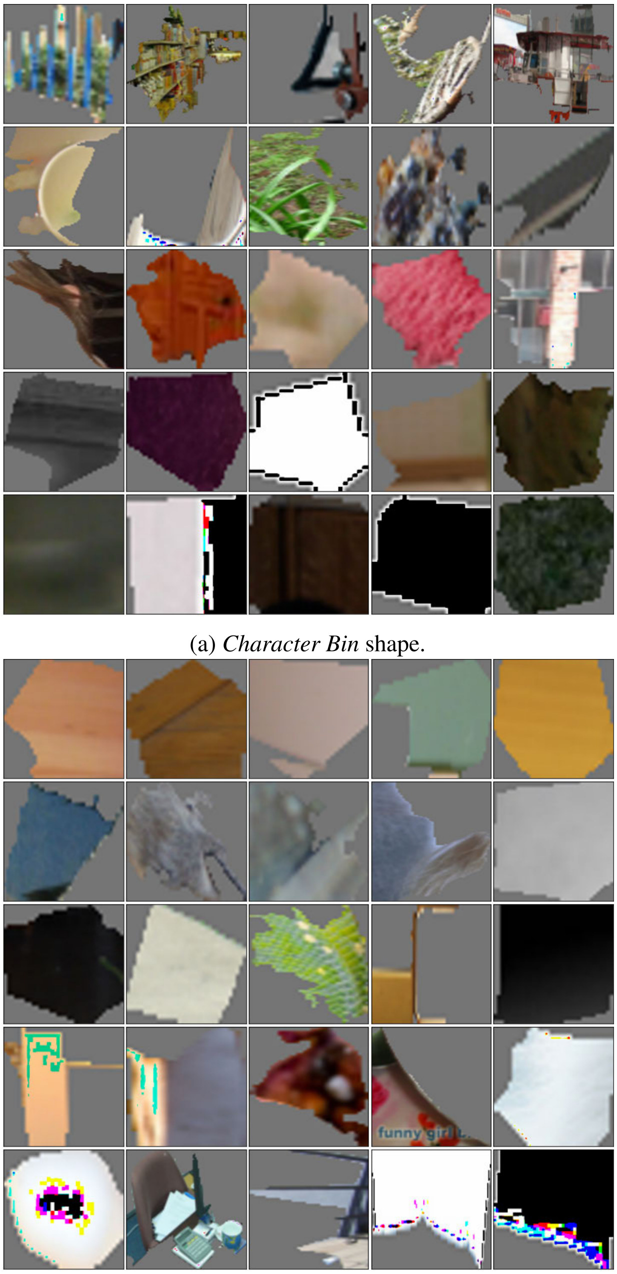


FIGURE 5. Sample images grouped into rows with increasing value of *Sentence Length* from top to bottom. One can observe that SL correlates with the semantic complexity of the image.



(a) *Character Bin* shape.

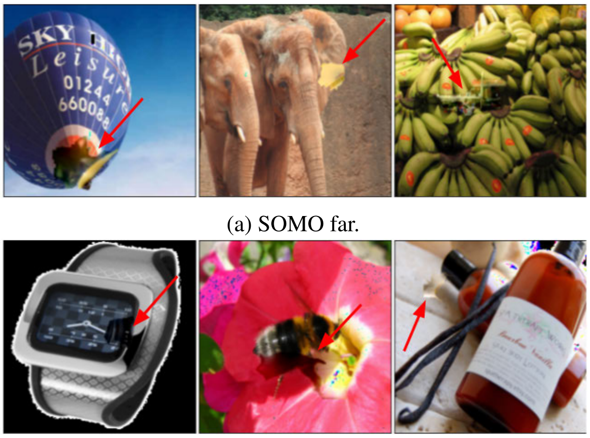
(b) *Character Bin* color.

FIGURE 6. Sample superpixels grouped into rows with increasing values of *CO* (a) and *ICV* (b) from top to bottom. One can observe that bottom rows contain superpixels of rounder shape (a) and higher contrast (b).

This probing task does not have a direct counterpart in the NLP domain.

IV. EXPERIMENTAL SETUP

This section describes how we generate visual words and self-supervised representations, assign visual words to images, and train the probing tasks.[1]



(a) SOMO far.

(b) SOMO close.

FIGURE 7. Sample images from SOMO far (a) and close (b) setup. Replacements in SOMO close come from a set of visual words that often co-occur with the replaced visual word. In SOMO far, we replace superpixel with rarely co-occurring visual words. One can observe that in the case of SOMO close, the differences are less visible. Notice that red arrows indicate alterations to the image.

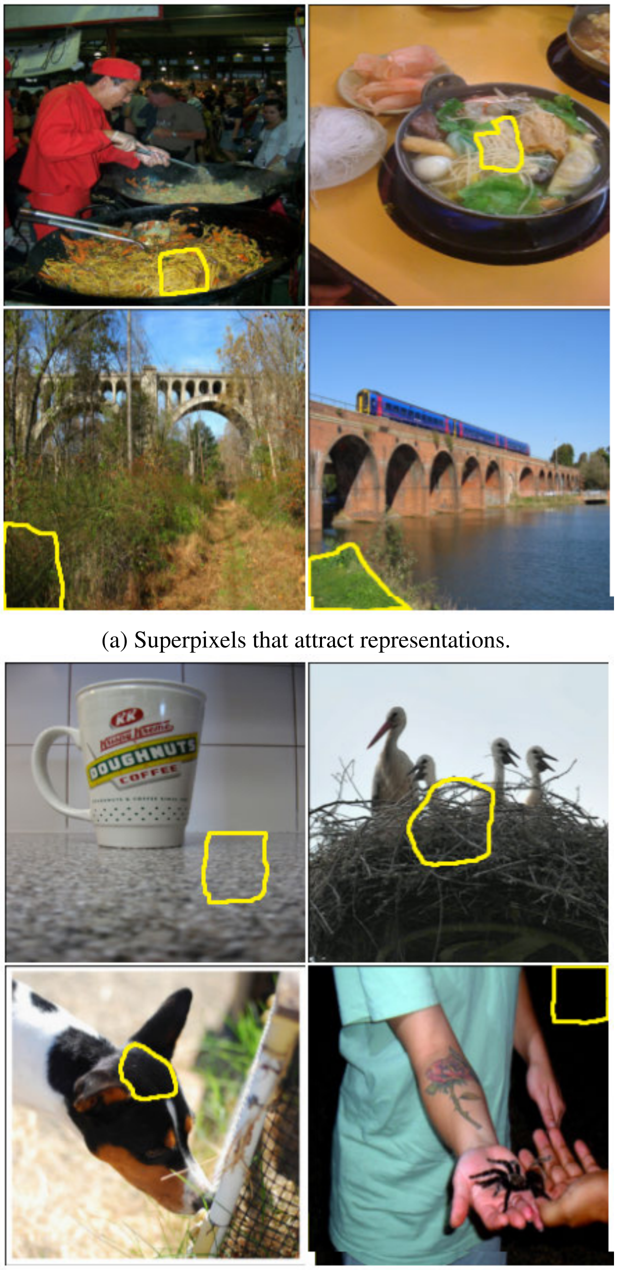
A. GENERATING VISUAL WORDS

This paper presents a general framework that can be used with various methods of generating visual words. However, choosing a high-quality method is crucial to draw meaningful conclusions from the probing tasks. That is why we use the established ACE algorithm [19]. It first divides images into superpixels using SLIC (Simple Linear Iterative Clustering) algorithm [28]. This algorithm clusters pixels in the combined five-dimensional color and position space to generate compact, nearly uniform superpixels. This approach has only one parameter that specifies the number of output superpixels. We use the SLIC algorithm with three resolutions of 15, 50, and 80 segments for each image. Next, it generates representations of these superpixels as an output of the *mixed4c* layer of GoogLeNet [7] trained on the ImageNet dataset. Then, for each class separately, corresponding representations are clustered using the k-means algorithm with $k = 25$ and filtered to remove infrequent and unpopular clusters (as described in [19]). This results in around 18 concepts per class and approximately $18,000$ concepts for the whole ImageNet dataset. They could be directly used as visual words. However, such words would be exclusive for particular classes, and some of them would be ambiguous due to the small TCAV score [26]. Hence, to obtain a reliable dictionary with visual words shared between classes, we filter out $12,000$ concepts with the smallest TCAV score and cluster the remaining $6,000$ concepts using the k-means algorithm into 50 clusters treated as visual words (see Figures 4 and 9). We do not treat the number of clusters as a tunable hyperparameter. Instead, we set a fixed number of clusters, ensuring various concepts and making user studies feasible.

B. GENERATING A SELF-SUPERVISED REPRESENTATION

We examine four self-supervised methods: MoCo v1 [27], SimCLR v2 [1], BYOL [20], and SwAV [21]. For all of them,

[1] The code available at github.com/BioNN-InfoTech/visual-probes



(a) Superpixels that attract representations.

(b) Superpixels that push representations away.

FIGURE 8. **Sample pairs of images with visual words (represented by marked superpixels) that attract (a) or push away (b) the representations. Visual words that attract representations include words with complicated forms and "green" words. On the other hand, visual words that push representations away contain fine textures.**

we use publicly available models trained on ImageNet.[2] Although they all use the penultimate layer of ResNet-50 to generate representations, their training hyperparameters differ, which is presented in Table 5.

TABLE 1. **Bins ranges corresponding to the classes in *Sentence Length* (SL) and *Character Bin* (CB) probing tasks.**

bin	SL	CB shape	CB color
0	< 18	< 0.153	< 0.063
1	$[18, 21)$	$[0.153, 0.207)$	$[0.063, 0.085)$
2	$[21, 23)$	$[0.207, 0.263)$	$[0.085, 0.104)$
3	$[23, 26)$	$[0.263, 0.336)$	$[0.104, 0.125)$
4	$[26, 28)$	$[0.336, 0.462)$	$[0.125, 0.155)$
5	$28 \leq$	$0.462 \leq$	$0.155 \leq$

C. ASSIGNING VISUAL WORDS

To assign a superpixel to a visual word, we first pass it through the GoogLeNet to generate a representation from the *mixed4c* layer (similarly to generating visual words). Since all concepts considered in Section IV-A are grouped into 50 clusters (visual words), we use a two-stage assignment. First, we find the closest concept and then assign the superpixel to the visual word containing this concept.

D. TRAINING PROBING TASKS

We use a logistic regression classifier with the LBFGS solver [61] to train all diagnostic classifiers. As input, we use representations generated by the self-supervised methods. The output depends on the probing task. In the case of *Word Content*, we train 50 classifiers corresponding to 50 visual words. Furthermore, we expect an image to be assigned to a particular visual word if at least one of its superpixels is assigned to it. Finally, we report the average AUC scores over 50 classifiers (see Table 2). To formulate a classification setup in the *Sentence Length* probing task, we group the possible output into six equally-sized bins (see Table 1), resulting in one-vs-one OVO AUC, which is resistant to class imbalance. A similar procedure is applied to the *Character Bin* probing tasks. SOMO is formulated as a binary classification task in which we predict whether the image was modified. We train two separate classifiers for two use cases, SOMO far and SOMO close, with balanced training and validation sets.

We conduct all of our experiments on the ImageNet dataset [29] with standard train/validation split. Moreover, we apply random over-sampling if needed to deal with the imbalanced classes.

V. USER STUDIES

While the cognitive visual systematic introduced in Section III-B presents the possible way of obtaining the meaning of visual words, it requires human observers to reliably decide which visual features should be assigned to particular visual words. Hence, in this section, we describe user studies conducted to establish this assignment.

Overall, 40 volunteers participated in the study (30 males and 10 females aged 29 ± 10 years) recruited online. 62.5% of the participants were students/graduates of computer science and related fields, and the remaining attendees represented various backgrounds.

The description and questions of the study were in English and Polish. Participants ranged from 18 to 66 years of age.

[2] We use the following implementations of self-supervised methods: https://github.com/{google-research/simclr, yaox12/BYOL-PyTorch, facebookresearch/swav, facebookresearch/moco}. We use ResNet-50 (1x) variant for each self-supervised method.

The average age of the participant is 29, and 68% of the participants are between 20 and 38 years old. 75% of the participants declared themselves as male, 25% as female and 0% chose other options. Participants were recruited online. 62% of the participants were students or graduates of computer science and related fields, and the remaining attendees represented various backgrounds, e.g., medicine, law, and psychology. 35% of participants have at least a bachelor's degree.

Users completed an online questionnaire. Their task was to assess the similarity of superpixels representing a visual word and provide key features associated with this visual word. To this end, users were presented with 20 visual words consisting of 12 representative superpixels (close to the visual word center) each. Participants were instructed to use Likert scales with seven numerical responses with only endpoints labeled (1 and 7) for clarity. First, they were asked to evaluate the homogeneity of a given set (scale endpoints: great variety; great homogeneity; see Figure 14). Next, they evaluated to what extent a given feature was essential for visual word creation. In reference to Marr's computational theory of vision [36] (see Section III-B), six features were taken into consideration: light intensity (brightness), color, texture, lines, shape (Marr's 2.5D sketch) and form (Marr's 3D model representation). Scale endpoints were labeled as a not significant feature and a key feature (see Figure 14).

Before the main task, users obtained an instruction that included sample visual words with particular features (selected by a cognitivist). They also underwent two training trials to familiarize themselves with the task. There were no time constraints for trial or task completion. The order of visual words and on-screen localization of superpixels were semi-randomized for each participant.

Due to the high number of visual words, the assessment of all 50 visual words would be tedious for the users. That is why we decided to limit our user study to the twenty most reliable visual words. They were chosen based on the results of *Word Content* probing task by selecting best and worst-performing clusters, as well as the ones with the largest performance difference between considered self-supervised models.

Based on the results of the user studies, we select the most representative visual words for each of the six features: brightness, color, texture, lines, shape, and form. Those words are then used to obtain detailed results of the *Word Content* probing task presented in Table 3.

VI. RESULTS AND DISCUSSION

As we show in Table 2, all self-supervised representations retain information about semantic knowledge, complexity, and image consistency. However, SimCLR v2 surpasses other methods in all probing task except CB color. Moreover, the performance on probing tasks does not correlate with the accuracy of the target task. In the following, we analyze those aspects in greater detail.

A. SELF-SUPERVISED REPRESENTATIONS CONTAIN SEMANTIC KNOWLEDGE WHICH DOES NOT CORRELATE WITH THE TARGET TASK

As reported in Table 2, the AUC scores for *Word Content* probing task vary from 0.793 for MoCo v1 to 0.811 for SimCLR v2. This shows that considered self-supervised methods can predict which visual words are present in the image, i.e., they code the semantic knowledge in the generated representations.

Surprisingly, although the examined self-supervised methods have diverse target task accuracy, they all have a similar level of semantic knowledge. For instance, MoCo v1 obtains the worst target task accuracy (60.6%), but its results for the WC probing task are on par with more accurate self-supervised methods. Moreover, although SwAV has the highest accuracy on the target task, it does not provide the best performance in terms of semantic knowledge. This finding supports the conclusion from [31] that semantic knowledge only partially contributes to the target task accuracy.

B. CERTAIN TYPES OF VISUAL WORDS ARE REPRESENTED BETTER THAN THE OTHERS, DEPENDING ON THE METHOD

According to the results presented in Table 3 and Figure 9, self-supervised representations have more knowledge about visual words containing forms and lines than about those containing shapes and textures. This may indicate that the representations are lines- and form-biased, which sheds new light on this problem, considering that according to [31], self-supervised representations are texture-biased. Moreover, the information encoded by various self-supervised methods differs. It is especially visible for brightness and color, where the MoCo v1 works significantly worse than the remaining methods. We assume that it is caused by the lack of projection head in the former, which is important due to the loss of information induced by the contrastive loss [33].

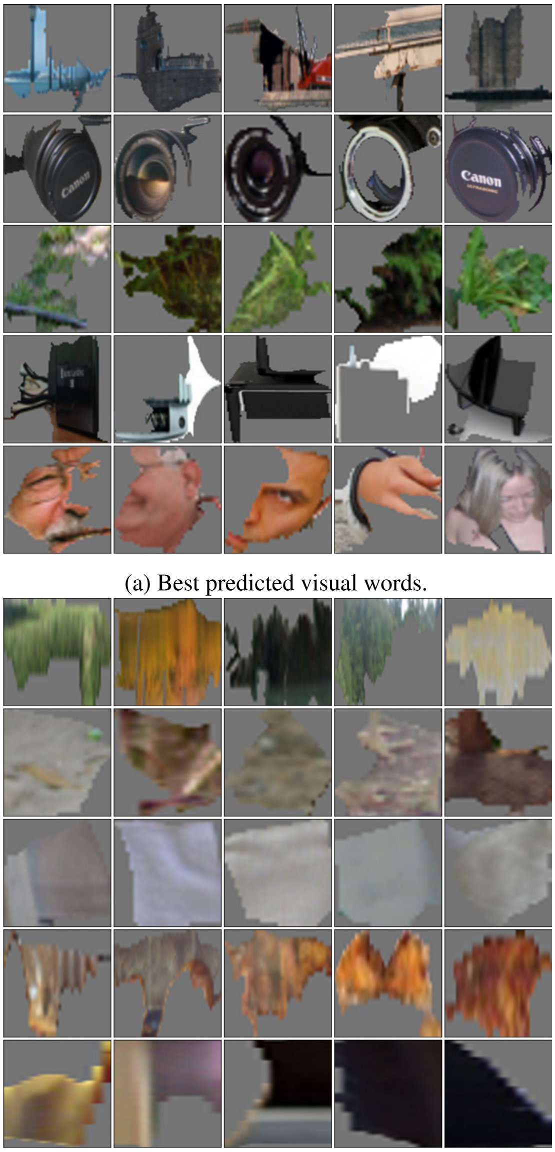
C. THE SAME VISUAL WORD IN A PAIR OF IMAGES USUALLY BRINGS THEIR REPRESENTATIONS CLOSER

The results of the MWC probing task presented in Table 4 show that the same visual word in a pair of images usually brings their representations closer. This is true for almost all visual words (45 out of 50), and especially for those presented in Figure 10a that contain complicated forms and lines or green areas. The remaining five visual words, usually corresponding to fine textures (see Figure 10c), are neutral or pushing representations away.

Interestingly, SwAV differs from the other methods in the case of lines and shapes, and BYOL differs in the case of shape. As in both cases, the presence of those features usually does not bring representations learned by those methods closer together. Differences in the training procedures of these methods might partially explain these results. SwAV and BYOL are trained without negative pairs, whereas

TABLE 2. AUC score for all our probing tasks (WC, MWC, SL, CB, and SOMO) and accuracy on the linear evaluation (Target) for the considered self-supervised methods.

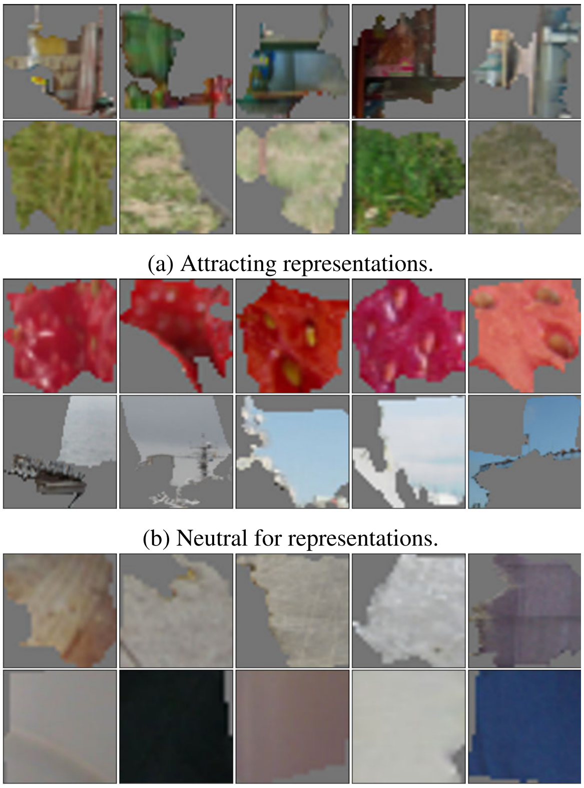
	Target	Probing tasks (ours)						
		WC	MWC	SL	CB shape	CB color	SOMO far	SOMO close
MoCo v1	0.606	0.793	0.763	0.771	0.797	0.872	0.850	0.830
SimCLR v2	0.717	**0.811**	**0.777**	**0.775**	**0.850**	0.876	**0.878**	**0.857**
BYOL	0.723	0.803	0.775	0.770	0.844	**0.893**	0.845	0.817
SwAV	**0.753**	0.802	0.776	0.769	0.842	0.879	0.856	0.839


(a) Best predicted visual words.

(b) Worst predicted visual words.

FIGURE 9. Visualization of the best (a) and the worst (b) predicted visual words according to the results of the WC probing task. It supports the results from Table 3 that self-supervised representations contain more information about lines and forms than textures.

SimCLR v2 and MoCo v1 use both positive and negative pairs during the training. Therefore, we hypothesize that this may cause differences in the MWC probing task result.


(a) Attracting representations.

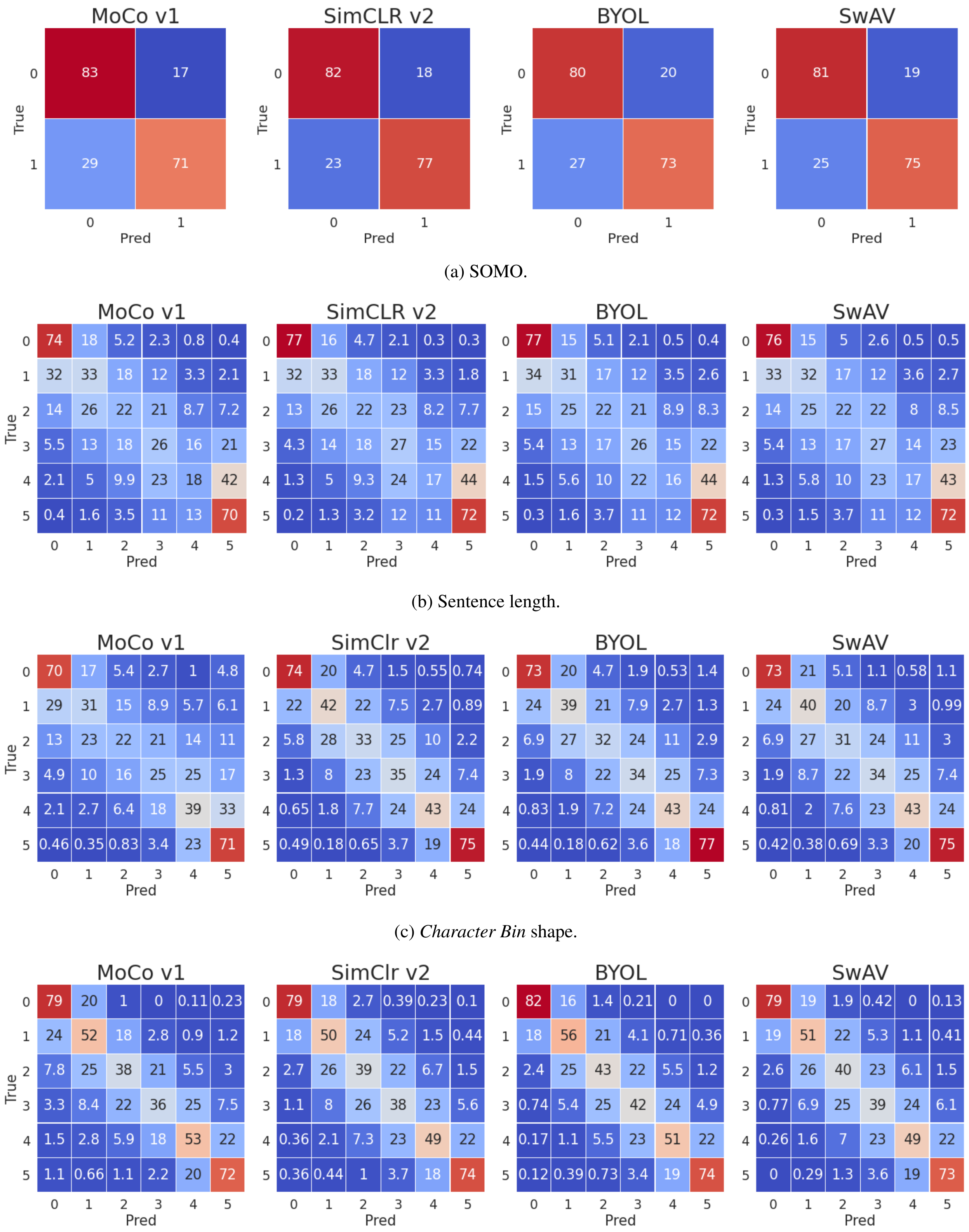
(b) Neutral for representations.

(c) Pushing representations away.

FIGURE 10. Sample visual words that attract (a), are neutral (b), or push away (c) the representations of two images. One can observe that the visual words that attract representations include words with complicated forms or green areas. On the other hand, visual words that push representations away contain fine textures.

D. SELF-SUPERVISED REPRESENTATIONS CONTAIN INFORMATION ABOUT SEMANTIC COMPLEXITY THAT DIFFERS BETWEEN METHODS

Based on the results in Table 2 and Figure 11, we observe that considered self-supervised methods code the level of semantic complexity, as they all obtain approximately 0.77 AUC for *Sentence Length*, and even higher AUCs are observed for CB shape and color (from 0.797 to 0.893 AUC). Moreover, when it comes to recognizing variance in superpixel color, BYOL works best, in contrast to all other probing tasks, where SimCLR v2 has the highest AUC. The potential reason for this behavior is the fact that a positive pair with similar color histograms provide more information in BYOL than in



(a) SOMO.

(b) Sentence length.

(c) *Character Bin* shape.

(d) *Character Bin* color.

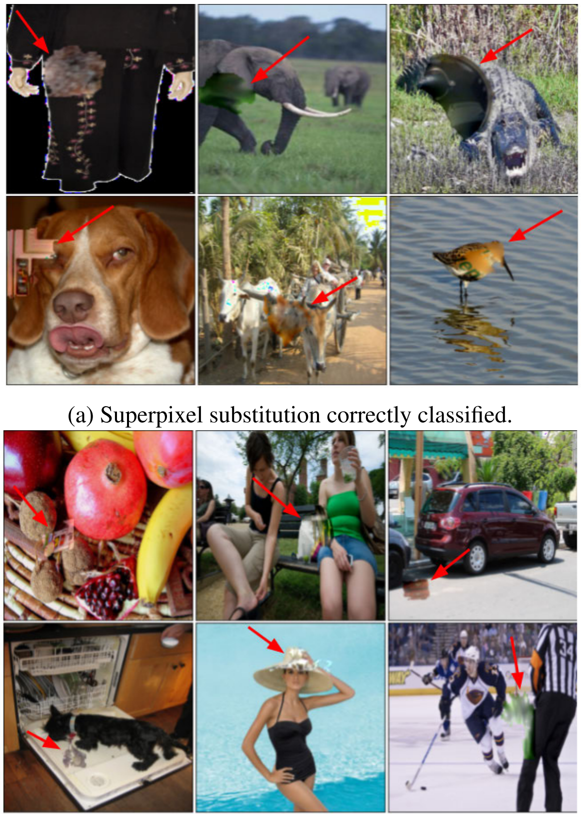
FIGURE 11. **Confusion matrices for *Sentence Length* and *Character Bin* probings (results in %). The results indicate that the ability of self-supervised representations to retain information about complexity differs depending on the level of image complexity. Moreover, even though the final AUC of SL and CB for self-supervised methods are similar, their confusion matrices differ.**

TABLE 3. Biases of the representations measured by WC probing tasks. The colors indicate a higher (orange) or lower (blue) AUC score compared to the overall performance for all visual words.

		Types of visual words					
	all visual words	**brightness**	**color**	**texture**	**lines**	**shape**	**form**
MoCo v1	0.793	0.777	0.769	0.785	0.847	0.784	0.836
SimCLR v2	0.811	0.831	0.832	0.804	0.852	0.810	0.854
BYOL	0.803	0.809	0.807	0.795	0.852	0.798	0.848
SwAV	0.802	0.819	0.820	0.796	0.851	0.802	0.849

TABLE 4. The attraction coefficient of MWC probing task. The attractions of representations containing the same visual word are marked in orange.

	Types of visual words					
	brightness	color	texture	lines	shape	form
MoCo v1	0.872	0.875	0.937	1.839	0.677	1.928
SimCLR v2	0.409	0.424	0.442	0.803	0.500	0.593
BYOL	0.396	0.379	0.502	0.336	−0.007	0.566
SwAV	0.382	0.445	0.150	−0.119	−0.118	0.248



(a) Superpixel substitution correctly classified.

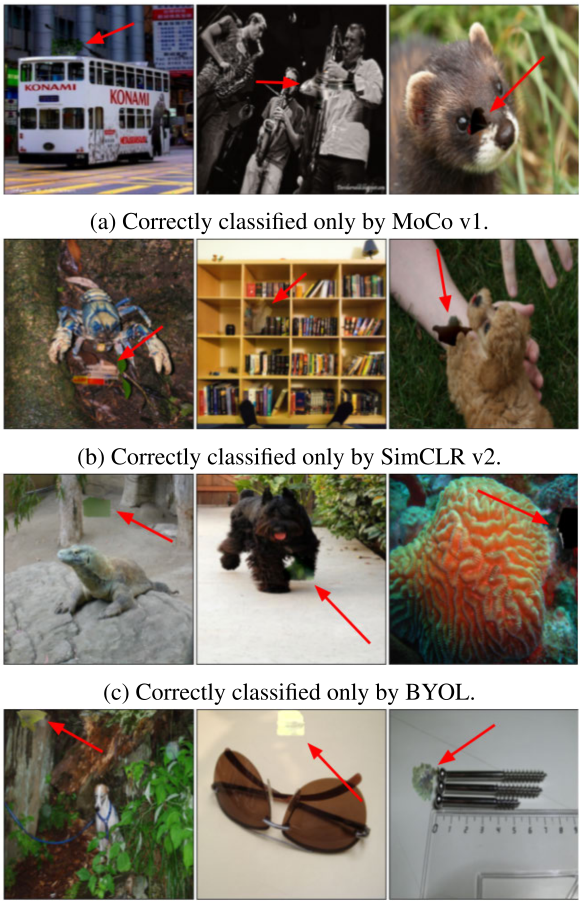
(b) Superpixel substitution incorrectly classified.

FIGURE 12. Sample images from the SOMO probing task. Images for which the superpixel substitution was correctly classified based on the representations by all methods (a) and by none of them (b). One can observe that probing struggles with more subtle changes, which are still visible to the human eye. Notice that red arrows indicate alterations to the image.

SimCLR, as presented in Section V of [20]. Therefore, BYOL puts more attention on the color characteristic.

E. SELF-SUPERVISED REPRESENTATIONS CONTAIN INFORMATION ABOUT SEMANTIC CONSISTENCY THAT DIFFERS BETWEEN METHODS

The results of the SOMO probing task in Table 2 and Figure 11 show that self-supervised representations reflect changes in the center of the image. However, as presented in Figure 12, the probing classifier struggles with more subtle changes, which are still visible to the human eye. Moreover, SimCLR v2 has the highest ability to recognize altered images, but surprisingly, BYOL has the lowest performance. However, as shown in Figure 12 and 13, there are no visible reasons for this result. Overall, our results are in line with [32], which claims that self-supervised methods improve out-of-distribution detection. However, they contradict our previous results [51], where the replaced superpixel is selected entirely randomly (without center bias). Nevertheless, we decided to change replacement to center-biased



(a) Correctly classified only by MoCo v1.

(b) Correctly classified only by SimCLR v2.

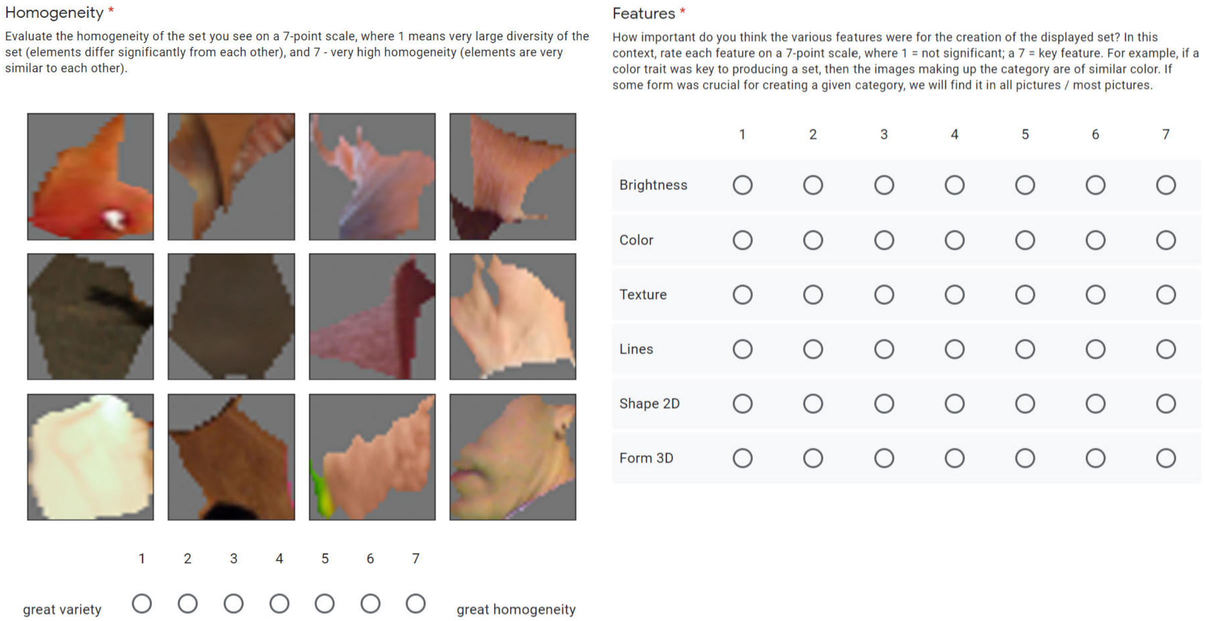
(c) Correctly classified only by BYOL.

(d) Correctly classified only by SwAV.

FIGURE 13. Sample images from SOMO probing task, correctly classified when embedded by MoCov1 (a), SimCLR v2 (b), BYOL (c), or SwAV (d) only. One can observe no clear differences between the types of inconsistencies classified correctly and incorrectly by the particular methods. Notice that red arrows indicate alterations to the image.

TABLE 5. Differences between architecture and training of the considered self-supervised methods.

		MoCo v1	SimCLR v2	BYOL	SwAV
Architecture	InfoNCE	yes	yes	no	no
	Positive pairs	yes	yes	yes	yes
	Negative pairs	yes, minibatches queue	yes, large batches	no	no
	Online to target network	copied with momentum	same	copied with momentum	same
	Size of patches	224x224	224x224	224x224	224x114 and 96x96
	Augmentations	resize, crop, color jittering, horizontal flip, grayscale conv.	crop, resize, horizontal flip, color distortion, grayscale conv., Gaussian blur, solarization	like in SimCLR v2	two types of crops, small and original, the rest like in SimCLR v2
	Projection	no	yes	yes	yes
Training	epochs	200	600	300	800
	Batch size	256	2048	4096	4096
	Time of training	53	170	not mentioned	49



Homogeneity *

Evaluate the homogeneity of the set you see on a 7-point scale, where 1 means very large diversity of the set (elements differ significantly from each other), and 7 - very high homogeneity (elements are very similar to each other).

	1	2	3	4	5	6	7	
great variety	○	○	○	○	○	○	○	great homogeneity

Features *

How important do you think the various features were for the creation of the displayed set? In this context, rate each feature on a 7-point scale, where 1 = not significant; a 7 = key feature. For example, if a color trait was key to producing a set, then the images making up the category are of similar color. If some form was crucial for creating a given category, we will find it in all pictures / most pictures.

	1	2	3	4	5	6	7
Brightness	○	○	○	○	○	○	○
Color	○	○	○	○	○	○	○
Texture	○	○	○	○	○	○	○
Lines	○	○	○	○	○	○	○
Shape 2D	○	○	○	○	○	○	○
Form 3D	○	○	○	○	○	○	○

FIGURE 14. Sample question from our user study that allows in-depth analysis of the Word Content probing task using Marr's computational theory of vision.

in this work because they better correspond to semantic inconsistency.

F. THE ABILITY TO DISTINGUISH ALTERED IMAGES DEPENDS ON HOW OFTEN THE REMOVED VISUAL WORD CO-OCCURS WITH THE REPLACEMENT

As presented in Table 2, SOMO far has higher performance than SOMO close. It is expected because recognizing alterations obtained by replacing a visual word with a non-fitting one is simpler. However, this difference in performance for all self-supervised representations leads us to believe that there is a family of alterations that might not be reflected well enough in a self-supervised representation. Hence, considering that even minor alterations might lead to a change in the prediction [60], this disability might pose a risk to the stability of the classification results.

VII. CONCLUSION

In this work, we introduce a novel visual probing framework that analyzes the information stored in self-supervised image representations. It is inspired by probing tasks employed in NLP and requires similar taxonomy. Hence, we propose a set of intuitive mappings between visual and textual modalities to construct visual sentences, words, and characters. Moreover, we provide a cognitive visual systematic that identifies a visual word with structural features from Marr's computational theory [36] and provide the meaning of the words.

Our cognitive framework reveals insights into high-level concepts that the model has learned. Such insights can be applied to promote the safer use of self-supervised learning, being aware of the biases encoded in the representations. The results of the provided experiments confirm the effectiveness and applicability of this framework in understanding self-supervised representations. We verify that the representations contain information about semantic knowledge, complexity, and consistency of the images. Moreover, a detailed analysis of each probing task reveals differences in the representations encoded by various methods, providing complementary knowledge to the accuracy of linear evaluation.

Our framework goes beyond per-sample explanations to identify higher-level human-understandable visual concepts that apply across the entire dataset. The existing work, the closest to ours, is a work [19] in which high-order concepts are automatically determined for images from each class. We build upon this work and propose a new approach using probing tasks. The advantage of our method is measurability, which enables us to compare to what extent individual self-supervised models encode information about concepts in their representations.

We note a couple of limitations of our framework. Our framework only applies to concepts in the form of groups of pixels. This assumption gives us plenty of insight into the model, but more complex and abstract concepts might be difficult to be noticed. In addition, the success of our approach depends on the quality of the generated labels for probing tasks. In our work, we presented five probing tasks inspired by NLP. However, in future work, one can consider a more generic approach to creating desirable probing tasks. One possible way to do this is to use a model that generates images based on a given text description, e.g., DALL-E 2 [62]. Thanks to this, we can create training pairs of image and text, generate probing labels based on text and use a probing classifier to image representation. In this case, creating new probing tasks would be equivalent to formulating appropriate queries for the image-generating model. These queries describe the human-understandable concept influencing the model's trait we want to investigate.

Potentially, our method may have a wider application, not only for self-supervised learning but for any learning methods for which individual layers of representation can be explained using our cognitive framework. The advantage of our framework is that it is generic. For example, using a different segmentation algorithm will lead to a different visual vocabulary. Therefore, conducting additional user studies to evaluate visual words generated using different methods and parameters would be worthwhile.

Finally, we show that the relations between language and vision can serve as an effective yet intuitive tool for explainable AI. Hence, we believe that our work will open new research directions in this domain.

ACKNOWLEDGMENT

The authors have applied a CC BY license to any Author Accepted Manuscript (AAM) version arising from this submission, in accordance with the grants' open access conditions.

REFERENCES

[1] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," 2020, *arXiv:2006.10029*.

[2] K. Krasnowska-Kieraś and A. Wróblewska, "Empirical linguistic study of sentence embeddings," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5729–5739.

[3] J. Hewitt and C. Manning, "A structural probe for finding syntax in word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Language Technol.*, vol. 1, Jun. 2019, pp. 4129–4138.

[4] N. Kim, R. Patel, A. Poliak, P. Xia, A. Wang, T. McCoy, I. Tenney, A. Ross, T. Linzen, B. Van Durme, S. R. Bowman, and E. Pavlick, "Probing what different NLP tasks teach machines about function word comprehension," in *Proc. 8th Joint Conf. Lexical Comput. Semantics (SEM)*, 2019, pp. 235–249.

[5] J. Hewitt and P. Liang, "Designing and interpreting probes with control tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1–11.

[6] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaś, and A. Korhonen, "Probing pretrained language models for lexical semantics," 2020, *arXiv:2010.05731*.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[8] J. Sivic and A. Zisserman, "Video Google: Efficient visual search of videos," in *Toward Category-Level Object Recognition*. Berlin, Germany: Springer, 2006, pp. 127–144.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[11] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018.

[12] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," 2016, *arXiv:1610.01644*.

[13] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," 2018, *arXiv:1810.03292*.

[14] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[15] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8662–8672.

[16] S. Kumar and P. Talukdar, "NILE: Natural language inference with faithful natural language explanations," 2020, *arXiv:2005.12116*.

[17] M. Eichler, G. G. Şahin, and I. Gurevych, "LINSPECTOR web: A multilingual probing suite for word representations," 2019, *arXiv:1907.11438*.

[18] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," 2018, *arXiv:1805.01070*.

[19] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards automatic concept-based explanations," 2019, *arXiv:1902.03129*.

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.

[21] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.

[22] L. R. Sipe, "How picture books work: A semiotically framed theory of text-picture relationships," *Children's Literature in Educ.*, vol. 29, pp. 97–108, Jun. 1998.

[23] 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 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[25] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, pp. 29–44, Jun. 2001.

[26] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.

[27] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[28] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[29] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[30] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.

[31] R. Geirhos, K. Narayanappa, B. Mitzkus, M. Bethge, F. A. Wichmann, and W. Brendel, "On the surprising similarities between supervised and self-supervised models," 2020, *arXiv:2010.08377*.

[32] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," 2019, *arXiv:1906.12340*.

[33] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[34] C.-E. Guo, S.-C. Zhu, and Y. N. Wu, "Primal sketch: Integrating structure and texture," *Comput. Vis. Image Understand.*, vol. 106, pp. 5–19, Apr. 2007.

[35] P. Kitcher, "Marr's computational theory of vision," *Philosophy Sci.*, vol. 55, no. 1, pp. 1–24, Apr. 1988.

[36] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt, 1982.

[37] D. Medin, "Concepts and conceptual structure," *Amer. Psychologist*, vol. 44, pp. 81–1469, Dec. 1989.

[38] M. J. Morgan, "Features and the 'primal sketch,'" *Vis. Res.*, vol. 51, no. 7, pp. 738–753, Apr. 2011.

[39] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you': Explaining the predictions of any classifier," 2016, *arXiv:1602.04938*.

[40] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," 2014, *arXiv:1412.0035*.

[41] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," 2017, *arXiv:1710.11063*.

[42] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," 2017, *arXiv:1704.05796*.

[43] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," 2017, *arXiv:1711.05611*.

[44] C. Chen, O. Li, A. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," 2018, *arXiv:1806.10574*.

[45] R. Zhang, P. Isola, and A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.

[46] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.

[47] P. Goyal, Q. Duval, J. Reizenstein, M. Leavitt, M. Xu, B. Lefaudeux, M. Singh, V. Reis, M. Caron, P. Bojanowski, A. Joulin, and I. Misra. (2021). *VISSL*. [Online]. Available: https://github.com/facebookresearch/vissl

[48] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: A survey," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 49–72, Apr. 2019.

[49] A. Schick, M. Fischer, and R. Stiefelhagen, "Measuring and evaluating the compactness of superpixels," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, 2012, pp. 930–934.

[50] W. Benesova and M. Kottman, "Fast superpixel segmentation using morphological processing," in *Proc. Int. Conf. Mach. Vis. Mach. Learn.*, 2014, pp. 1–9.

[51] D. Basaj, W. Oleszkiewicz, I. Sieradzki, M. Górszczak, B. Rychalska, T. Trzcinski, and B. Zieliński, "Explaining self-supervised image representations with visual probing," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 592–598.

[52] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110 Nov. 2004.

[53] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.

[54] L. Sixt, M. Granz, and T. Landgraf, "When explanations lie: Why many modified BP attributions fail," 2019, *arXiv:1912.09818*.

[55] J. Bernard, M. Hutter, C. Ritter, M. Lehmann, M. Sedlmair, and M. Zeppelzauer, "Visual analysis of degree-of-interest functions to support selection strategies for instance labeling," in *Proc. EuroVA, Int. Workshop Vis. Anal.*, 2019.

[56] M. Ribeiro, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. SIGKDD*, 2016, pp. 97–101.

[57] R. Salakhutdinov, "One-shot learning with a hierarchical nonparametric Bayesian model," in *Proc. ICML UTL Workshop*, 2012, pp. 1–13.

[58] D. Rymarczyk, L. Struski, J. Tabor, and B. Zielinski, "ProtoPShare: Prototypical parts sharing for similarity discovery in interpretable image classification," in *Proc. SIGKDD*, 2021, pp. 1420–1430.

[59] C. J. Cai, E. Reif, N. Hegde, J. Hipp, and B. Kim, "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–14.

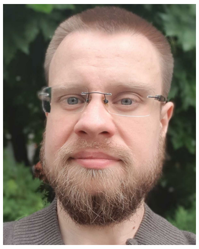
[60] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[61] H. Matthies and G. Strang, "The solution of nonlinear finite element equations," *Int. J. Numer. Methods Eng.*, vol. 14, no. 11, pp. 1613–1626, 1979.

[62] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

WITOLD OLESZKIEWICZ received the M.Sc. degree in computer science from the Institute of Computer Science, Warsaw University of Technology, in 2017, where he is currently pursuing the Ph.D. degree with the Division of Artificial Intelligence, Institute of Computer Science. He is currently an Assistant with the Division of Artificial Intelligence, Institute of Computer Science, Warsaw University of Technology. His professional appointments include work with Samsung, in 2013, and Braster, from 2015 to 2018, where he worked on the use of machine learning in breast cancer detection. He was a Visiting Scholar at Stanford University, in 2018, where he worked on privacy-preserving generative models, and New York University, in 2019, where he worked on understanding the robustness of deep learning for breast cancer screening.

DOMINIKA BASAJ received the master's degree in quantitative methods in economics and information systems from the Warsaw School of Economics, in 2016. She was developing machine learning models in financial institutions. In 2019, she was a Visiting Researcher at the Nanyang University of Technology, where she worked on discourse-aware neural machine translation, and at the University of California at Davis, where she worked on the prediction of protein structure. She is currently a Senior AI Engineer with Tooploox. Her research interests include the interpretability and robustness of neural networks.



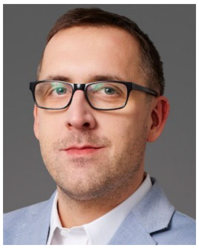
IGOR SIERADZKI received the M.Sc. degree in computer science on active learning in computer-aided drug design from Jagiellonian University, in 2016, where he is currently pursuing the Ph.D. degree with the Faculty of Mathematics and Computer Science, Institute of Computer Science and Computer Mathematics. He is currently an Assistant with the Faculty of Mathematics and Computer Science, Institute of Computer Science and Computer Mathematics, Jagiellonian University, Kraków, since 2019. Before the academic position, he worked with Applica.ai on the modern use of deep learning in natural language processing. His research internships include a stay at the University of Edinburgh, in 2015.

MICHAŁ GÓRSZCZAK recevied the B.Eng. degree in applied computer science from the University of Science and Technology, Kraków, in 2019. He is currently pursuing the master's degree with the Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków.

BARBARA RYCHALSKA received the master's degree in computer science from the Warsaw University of Technology in 2016, also studied applied linguistics at the Warsaw University, where she is currently pursuing the Ph.D. degree with the Faculty of Mathematics and Information Science. She is currently an AI Research Scientist with Synerise, where she works on topics ranging from natural language processing to recommender systems. Previously, she worked at Samsung Research and Development Research Institute Warsaw and Findwise AB, as an AI Researcher. She was a Visiting Scientist at the Nanyang Technological University, Singapore, in 2019.

KORYNA LEWANDOWSKA received the M.A. degree in psychology and the Ph.D. degree in psychology on the influence of decision bias on visual recognition memory from Jagiellonian University, Kraków, in 2011 and 2019, respectively. She is currently an Assistant with the Department of Cognitive Neuroscience and Neuroergonomics, Faculty of Management and Social Communication, Institute of Applied Psychology, Jagiellonian University. She is also a Lecturer with the College of Economics and Computer Science. Her research interests include the realization of projects concerning issues from the fields of cognitive psychology, cognitive neuroscience, chronopsychology, and consumer neuroscience. She is a member of the Polish Association for Cognitive and Behavioral Therapy.

TOMASZ TRZCINSKI (Senior Member, IEEE) received the M.Sc. degree in research on information and communication technologies from the Universitat Politècnica de Catalunya, the M.Sc. degree in electronics engineering from the Politecnico di Torino, in 2010, the Ph.D. degree in computer vision from the École Polytechnique Fédérale de Lausanne, in 2014, and the D.Sc. degree (Habilitation) from the Warsaw University of Technology, in 2020. He has been an Assistant Professor with the Division of Computer Graphics, Institute of Computer Science, Warsaw University of Technology, since 2015. His professional appointments include work with Google, in 2013; Qualcomm Corporate Research and Development, in 2012; and Telefónica Research and Development, in 2010. He was a Visiting Scholar at Stanford University, in 2017, and Nanyang Technological University, in 2019. He is a Co-Organizer of warsaw.ai a member of Computer Vision Foundation, an Expert of the National Science Centre and Foundation for Polish Science, as well as a member of the Scientific Board for PLinML and Data Science Summit conferences. He is a Chief Scientist and a Partner at Tooploox, where he leads a team of machine learning researchers and engineers. He has co-founded Comixify, a technology startup focused on using machine learning algorithms for editing videos. He is currently an Associate Editor of IEEE Access and frequently serves as a reviewer in major computer vision conferences (CVPR, ICCV, ECCV, ACCV, BMVC, ICML, MICCAI) and international journals (IEEE Transactions on Pattern Analysis and Machine Intelligence, *IJCV*, *CVIU*, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia).



BARTOSZ ZIELIŃSKI received the M.Sc. degree in computer science from Jagiellonian University, in 2007, and the Ph.D. degree in computer science with the Institute of Fundamental Technological Research, Polish Academy of Science, in 2012. He is currently an Assistant Professor with the Faculty of Mathematics and Computer Science, Institute of Computer Science and Computer Mathematics, Jagiellonian University, Kraków, since 2012. His professional appointments include work with Volantis Systems Ltd., in 2009, and Samsung, in 2018. He was a Visiting Scholar at the Vienna University of Technology, in 2015, and the Instituto Superior Técnico, Lisbon, in 2019. He is a Co-Organizer of the Cracow Cognitive Science Conference and Theoretical Foundations of Machine Learning. He is a Lead Data Scientist at Ardigen, where he leads a team of medical image analysis researchers and engineers. He frequently serves as a Reviewer in international journals on machine learning and medical image analysis (*AIR*, *CSBJ*, *CBM*, IEEE Transactions on Biomedical Engineering, *Trends in Microbiology*).