# LANGUAGE-GUIDED IMAGE CLUSTERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Image clustering methods have rapidly improved their ability to discover object categories. However, unsupervised clustering methods struggle on other image attributes, e.g. age or activity. The reason is that most recent clustering methods learn deep features that are designed to be sensitive to object category, but less so to other image attributes. We propose to overcome this limitation by introducing the new setting of language-guided image clustering. In this setting, the model is provided with an exhaustive list of phrases describing all the possible values of a specific attribute, together with a shared image-language embedding (e.g. CLIP). Our method then computes the subset of $K$ attribute phrases that form the best clustering of the images. Differently from standard clustering methods, our method can cluster according to image attributes other than the object category. We evaluate our method on attribute clustering tasks and demonstrate that our method significantly outperforms methods that do not use language-guidance.

## 1 INTRODUCTION

One of the core tasks of computer vision is to classify images into semantically similar classes. In the most common case of supervised classification, we explicitly indicate which images we deem as semantically similar using class labels. When such labels are unavailable, unsupervised image clustering aims to group the images to semantically similar groups, guided by the clustering methods' inductive bias and the statistics of the data (Shiran & Weinshall, 2021). One major limitation of current clustering methods is their assumption that a particular grouping of the data is inherently preferable to others. In fact, multiple semantic groupings of the data can be defined by many different image attributes including: object identity, age, activity, position or pose. As an illustrative example, we can consider the "drinking" and "brushing teeth" images seen in Fig.1 taken from the "Stanford Activity" dataset - note that two distinct groupings are possible. To resolve this ambiguity, language guidance is needed.

Current clustering methods typically wish to cluster the images according to the image category. Yet, grouping by other attributes such as activity or age is also perfectly valid. To partially resolve this ambiguity, clustering methods make design choices, such as the data augmentation (Shiran & Weinshall, 2021), that often guide the representation towards suppression of some properties (such as color or object position). The objective is that by suppression ("denying") of these properties, the component in the representation related to the object category will be dominant allowing the algorithm to cluster by category. However, this "deny-list" approach is problematic for two reasons: i) manually creating a deny-list for excluding all attributes but the ones we wish to group by is laborious and non-trivial (if possible). ii) Even if such a complete deny-list is given, there may not be a simple data augmentation for removing every undesired attribute.

The main idea in this work is to replace this "deny-listing" approach with "allow-listing", specifying which image attributes we deem as acceptable for grouping. In our approach, we provide an extensive list of words, or short sentences, describing the concepts which we "allow" as legitimate clusters for the data at hand. While this "allow-list" does assume some prior knowledge about the relevant attributes, it provides a significant advantage: giving guidance for the attributes that we wish to cluster by. Unlike "zero-shot classification", which assumes full knowledge on the dataset-specific labels, we only assume a list describing the generally allowed attribute values (and not the specific clusters names for this dataset).


Figure 1: While unsupervised clustering methods infer classes by visual features only, this may not be sufficient to identify ground truth classes. Our method guides the desired grouping by language.

Our key technical challenge is selecting the $K$ words (or phrases) out of a long list of English language phrases that best cluster the images into semantically consistent groups. We show that this task can be formulated as an unconstrained $K$ facility location problem, a commonly studied NP-hard problem in algorithmic theory. Although advanced methods exist for the solution with approximation guarantees, the most popular methods do not scale to our task. Instead, we suggest to solve the optimization task using a more scalable approach, that can be seen as a discretized version of K-means. Using an extensive dictionary of words or phrases also presents another issue: some uninformative words such as "entity" have an embedding similar to a large proportion of images in the dataset. We propose an unsupervised criterion for selecting a sublist of performant, informative words.

We evaluate our method on a range of attribute-clustering tasks, grouping according to different attributes including: activity or age. We show that language-guidance outperforms top methods such as SCAN. We take care to verify that this is not only merely due to our use of strong pre-trained features. Our method also outperforms zero-shot classification with a naive use of the given dictionary.

Our main contribution are:

1. Introducing the setting of language-guided image clustering and demonstrating its effectiveness for a variety of clustering tasks.
2. Reducing language-guided image clustering to the well-studied facility location problem.
3. Suggesting a scalable and empirically effective solution for solving the optimization task.
4. Proposing an unsupervised criterion for removing uninformative nuisance words from the word list.

## 2 Related Work

**Self-supervised deep image clustering:** Deep features trained using self-supervised criteria are extensively used for image clustering. Early methods learned deep features using an auto-encoder with a reconstruction constraint (Xie et al., 2016; Yang et al., 2017). More recent approaches directly optimize clustering objectives during feature learning. Specifically, a common approach is to cluster images according to their learned features, and use this approximate clustering for further improvement of the features (this can be done iteratively or jointly) (Caron et al., 2018; Chang et al., 2017; Haeusser et al., 2018). An issue that remains is that such approaches are "free" to learn arbitrary sets of features, and therefore might clusters according to attributes not related to the ground truth labels. To overcome this issue, a promising line of approaches use carefully selected augmentations to remove the nuisance attributes and direct learning towards more semantic features (Wu et al.,

2019; Niu et al., 2020; Shiran & Weinshall, 2019). These ideas are often combined with contrastive learning (Tsai et al., 2021). The work by Van Gansbeke et al. (Van Gansbeke et al., 2020) suggested a two stage approach, where features are first learned using a self-supervised task, and then used as a prior for learning the features for clustering. In practice, we often look for clusters which would be balanced in size, at least approximately. Many works utilize an information theoretic criterion to impose such balancing (Hu et al., 2017; Ji et al., 2019; Darlow & Storkey, 2020). Many recent works boost the performance on clustering algorithms with self-labelling(Niu & Wang, 2021). This is a promising approach, which can usually be added as an extra-stage, on top of initial results from self-supervised or other clustering methods.

**Clustering using pretrained features**: Some research has also been done on image clustering using features pretrained on some auxiliary supervised data (Guérin et al., 2021). While pretrained features are not always applicable, they are often general enough to boost performance on datasets significantly different than the auxiliary data (Kornblith et al., 2019).

**Color name-based features:** Color quantization, divides all colors into a discrete number of color groups. Although simple K-means approaches are common, it has been argued that grouping according to a list of colors that have names in the English language provides superior results to simple clustering only based on the pixel color statistics (Van De Weijer et al., 2009; Yu et al., 2018; Mojsilovic, 2005). Color name-based identification was further applied to other tasks, such as image classification, visual tracking and action recognition (Van De Weijer & Khan, 2015). As one example, for the task of person re-identification in surveillance, color names were used as a prior in order to define better similarity metrics, which led to better performance (Yang et al., 2014), and scalability (Prates et al., 2016). Our approach can be seen as extending these ideas from pixel colors to whole images.

**Joint embedding for images and text:** Finding the joint embedding of images and text is a long-standing research task (Mori et al., 1999). A key motivation for looking into such joint embedding is reducing the requirement for image annotations, needed for supervised machine learning classifiers. This can instead be done by utilizing freely-available text captions from the web (Quattoni et al., 2007; Joulin et al., 2016; Sariyildiz et al., 2020). It was also suggested that such learned representations can be used for transfer learning (Mahajan et al., 2018; Desai & Johnson, 2020). (Radford et al., 2021) presented a new method, CLIP, that also maps images and sentences into a common space. CLIP was trained using a contrastive objective and provides encoders for images and text. It was shown that CLIP can be used for very accurate zero-shot classification of standard image datasets, by mapping all category names to embeddings and then for each image choosing the category name with the embedding nearest to it. Our method relies on the outstanding infrastructure provided by CLIP but tackles image clustering rather than zero-shot classification. The essential difference is that in CLIP, the set of image labels is provided whereas in clustering the set of categories is unknown.

**Uncapacitated facility location problem (UFLP)**: The UFLP problem is a long-studied task in economics, computer science, operations research and discrete optimization. It aims to open a set of facilities, so that they serve all clients at a minimal cost. Since its introduction in the 1960s (e.g. (Kuehn & Hamburger, 1963)), it attracted many theoretical and heuristic solutions. It has been shown by (Guha & Khuller, 1999) that the metric UFLP can be solved with a constant approximation guarantee bounded by $\rho > 1.463$. Different solutions methodologies have been applied to the task including: greedy methods (Arya et al., 2004), linear-programming with rounding (Shmoys et al., 1997) and linear-programming primal-dual methods (Jain & Vazirani, 2001). Here, we are concerned with the Uncapacitated K-Facility Location Problem (UKFLP) (Cornuéjols et al., 1983; Jain et al., 2002), which limits the number of facilities to $K$. We formulate our optimization objective as the UKFLP and use a fast, relaxed variant of the method of (Arya et al., 2004).

## 3 Image Clustering with the Single-Phrase Prior

Our goal is to cluster images according to a particular set of attributes for which we receive linguistic guidance. We are given $N_I$ images, which are mapped into feature vectors $\{v_1 \dots v_{N_I}\}, v_i \in \mathbb{R}^d$. We further assume a list of $N_W$ phrases, such that every phrase is mapped into a vector embedding $\{u_1 \dots u_{N_W}\}, u_i \in \mathbb{R}^d$. The list of all phrase embeddings is denoted as $\mathcal{W}$. The images and phrases are assumed to be embedded in the same feature space, in the sense that for each image, its nearest

phrase in feature space provides a good description of the content of the image. We obtain this joint embedding using CLIP (Radford et al., 2021), a recent state-of-the-art approach. We aim to divide the images into $K$ clusters $\{S_1..S_K\}$. Each cluster should consist of semantically similar images. We denote the cluster centers by a corresponding set of vectors $\{c_1..c_K\}, c_k \in \mathbb{R}^d$.

### 3.1 THE SINGLE-PHRASE PRIOR

Our main proposed idea is to further constrain the clustering task beyond merely the visual inter-cluster similarity requirement as posed by the feature embedding of our images. Unlike previous methods that use augmentations as a way of specifying the attributes we do not wish to cluster by ("deny-listing") - we provide a list of the possible values of the attributes that may be used for clustering ("allow-listing"). The allow-listing approach has an inherent advantage over the deny-listing approach, as the number of unwanted attributes is potentially infinite and augmentations that remove all those attributes may not be known.

Specifically, we define a "single-phrase" prior. We utilize a pre-trained network for mapping images and phrases into a common feature space. We require that embeddings of images in any given cluster $v \in S_k$ will be similar to the embedding of a single-phrase of our dictionary $w \in \mathcal{W}$ describing the relevant possible clusters. The set of plausible phrases is chosen from the (much longer) allowed list of phrases. We show that our method can guide the clustering process towards the desired attributes.

### 3.2 REMOVING NON-SPECIFIC PHRASES

Although we assume all plausible phrases are contained in the list $\mathcal{W}$, some phrases in the list may have a meaning that is too general, which may describe images taken out of more than one ground truth class, or even be related to all the images in the dataset. Examples for such phrases are: 'entity', 'abstraction', 'thing', 'object', 'whole'. We would like to filter out of our list those phrases that are ambiguous w.r.t. the ground truth classes of each dataset, in order to prevent "false" clusters. To this end, we score the "generality" of each phrases by calculating the average phrase embedding:

$$u_{avg} = \frac{1}{N_W} \sum_{i=1..N_W} u_i \tag{1}$$

We calculate the generality score $s$ for each phrase, as the inner product between its embedding $u_i$ and the average phrase embedding $u_{avg}$:

$$s(u_i) = u_i \cdot u_{avg} \tag{2}$$

We find that this score is indeed higher for the less specific phrases described earlier. We remove from the list all phrases that have a "generality score" $s$ higher than some quantile level $0 < q \leq 1$, and define the new sublist $\mathcal{W}_q \subseteq \mathcal{W}$ ($|\mathcal{W}_q| \approx q \cdot |\mathcal{W}|$, where $|.|$ denotes the length of a set).

To choose the quantile $q$ for each dataset using an unsupervised criterion we consider the *balance* of the resulting clusters. We try a set of values for $q$ (see implementation details 5.1), and run our algorithm with each of them. For each value of $q$, we obtain cluster assignments, and calculate the entropy. We choose to use the $q$ value for which our phrase list $W_q$ gives the most balanced clustering for each dataset, measured as the highest entropy cluster assignment. An ablation for this part of the method can be found in Sec.6.

### 3.3 CLUSTERING WITH THE SINGLE PHRASE PRIOR

We consider a cluster $S_k$ describable by a single phrase $c_k$ if the embeddings of its associated images are near the embedding of the phrase $c_k \in \mathcal{W}_q$. We formulate this objective, using the within-cluster sum of squares (WCSS) loss:

$$\begin{aligned} \min_{\{c_1..c_K\},\{S_1..S_K\}} \quad & \sum_{j=1}^{K} \sum_{v \in S_j} \|v - c_j\|^2 \\ \text{s.t.} \quad & c_j \in \mathcal{W}_q \end{aligned} \tag{3}$$

The objective is to find assignments $\{S_1..S_k\}$ and phrases $\{c_1..c_k\} \subseteq \mathcal{W}_q$, so that the sum of square distances for each cluster between the assigned images and the corresponding phrase is minimal.

Note that this is different from K-means as the cluster centers are constrained within the discrete set of phrases $\mathcal{W}$ whereas in K-means they are unconstrained.

# 4 OPTIMIZATION

## 4.1 THE UNCAPACITATED FACILITY LOCATION PROBLEM

We formalize our optimization problem, by restating it as an uncapacitated K-facility location problem (UKFLP). The UKFLP is a long studied discrete optimization task (see Sec. 2). In the UKFLP task we are asked to "open" $K$ "facilities" out of a larger set of sites $\mathcal{W}_q$, and assign each "client" to one of the $K$ facilities, such that the sum of distances between the "clients" and their assigned "facilities" is minimal. In our case, the clients are the image embeddings $v_1, v_2..v_{N_I}$, which are assigned to a set of $K$ phrase embeddings selected from the complete list $\mathcal{W}_q$. We look to optimize an assignment variable $x_{ij} \in \{0, 1\}$ indicating whether the "client" $v_i$ is assigned to the "facility", the phrase $u_j$. We also use a variable $y_j \in \{0, 1\}$ to determine if a facility was opened in site $j$ (if the phrase $u_j$ is the center of a cluster). The optimal assignment should minimize the sum squared distance between each image and its assigned phrase. The squared distance between image $v_i$ and phrase $u_j$ is denoted $d_{ij}$. We can now restate our loss as:

$$\begin{aligned} \min_{x_{ij}, y_j} \quad & \textstyle\sum_{i \in 1..N, j \in 1..N_W} d_{ij} x_{ij} \\ \text{s.t.} \quad & \forall i \in 1..N : \textstyle\sum_{j \in 1..N_W} x_{ij} = 1 \\ & x_{ij} \leq y_j \\ & \textstyle\sum_{j \in 1..N_W} y_j \leq K \end{aligned} \tag{4}$$

Where the bottom two constraints limit the number of phrases to be at most $K$.

Solving UKFLP is NP-hard, and the problem of approximation algorithms for UKFLP have been studied extensively both in terms of complexity and approximation ratio guarantees (see Sec.2). Yet, as the distance matrix $d_{ij}$ is very large, we could not run the existing solutions at the scale of many of many datasets (e.g. there may be as many as $82k$ phrases-"facilities" and a few hundred thousands images-"clients"). We therefore suggest a relaxed version of the popular Local Search algorithm.

## 4.2 LOCAL SEARCH ALGORITHM

The Local Search algorithm (Arya et al., 2004) is an effective, established method for solving facility location problems. Instead of looking for the optimal assignment at once, it looks for swaps between open and closed facilities that decrease the loss. It starts with "forward greedy" initialization: in the first $K$ steps, we open the new facility (choose a new phrase as center) that minimizes the loss the most, among all unopened sites (unselected phrases). After initialization, we iteratively perform the following procedure: In each step, we look to swap $p$ of our selected phrases by $p$ unselected phrases, such that the loss is decreased. If such phrases are found, the swap is applied. We repeat this step until better swaps cannot be found or the maximal number of iterations is, making it slow to run even for a small dataset.

## 4.3 LOCAL SEARCH LOCATION RELAXATION METHOD

As our task is very high-dimensional, running Local Search (or similar UKFLP algorithms) becomes too slow to be practical. Therefore, we suggest an alternative, a continuous relaxation approach which is much faster to compute (with complexity $O(N_I + N_W)$). Our method iterates the following steps until convergence. We initially assign each of our images $v_1..v_N$ to clusters $\{S_1..S_K\}$ according to the nearest cluster center ("Voronoi tessellation")

$$S_{k'} = \{v_i | \left\| v_i - c_{k'} \right\|^2 \leq \left\| v_i - c_k \right\|^2, \forall k\} \tag{5}$$

After assignment, the center locations $\{c_1..c_K\}$ are set to be the average feature in each cluster, which minimizes the WCSS (Eq.3) loss without the constraint. Precisely, we recompute each cluster center according to the image assignment $S_k$: $c_j = \frac{1}{|S_j|} \sum_{v \in S_j} v$. However, this is an infeasible

solution as cluster centers will generally not be in $\mathcal{W}_q$. We therefore replace each cluster center $c_j$ with its nearest neighbour phrase in $W_q$. The result of this step is a new set of $K$ phrases that form the cluster centers. Instead of using this new list of phrases as the new cluster centers, we keep $K-p$ phrases from the previous iteration and select $p$ phrases from the current set, such that the new combined set of centers decreases the loss in Eq. 3. This is similar to the swap in the Local Search algorithm, but differently from it, we limit our search space to the new set of $K$ phrases rather than the entire phrase list $\mathcal{W}_q$. If no loss decreasing swap is found, we terminate.

**Empty and excessively large clusters:** In some cases, the discrete nature of the single-phrase constraint results in excessively large clusters, or one of our $K$ centers being "empty" of samples. In the case of empty clusters, we replace the center location with that of a phrase which would "attract" most samples. Specifically, we choose that phrase that has the most samples as its nearest neighbours (among those not already in use). To address the problem of excessively large clusters, we split the samples in that cluster among the phrases in $\mathcal{W}_q$ (by distance), and replace the center of the largest cluster with the phrase that was chosen by the largest number of images. Images that only loosely fit the cluster are therefore likely to be reassigned to other clusters.

**Cluster initialization:** We initialize the cluster assignments using Ward's clustering on the image embeddings $v_1, v_2..v_{N_I}$.

## 5 EXPERIMENTS

In this section, we evaluate our method on several attribute clustering tasks. We demonstrate that: i) our method can identify phrases that are closely aligned with the ground truth ii) it can achieve high clustering accuracy.

**Datasets:** Aiming to evaluate our clustering method in cases where the language guidance is necessary, we selected evaluation datasets where the ground truth class attribute is often not the largest object in the image. We evaluate the following datasets:

*Stanford Activity (Yao et al., 2011)*: A dataset presenting people performing $40$ different activities ("fixing a bike", "fixing a car", "riding a horse", etc...). The images are of very high inter class variability, see Fig.1. As many of the class names are composed from a verb and object, we composed our phrase list by taking the given verbs i.e. the dataset's activity names and related verbs aggregated from *Thesaurus.com*. Phrases are composed by pairing each verb with each object category.

*BU Action (Ma et al., 2017)*: A dataset containing $101$ different activities. The class names in this dataset consist of a verb, a noun or a combination. Therefore, we evaluated it using the complete list of over $82K$ WordNet (Miller, 1995) nouns. While the list does not contain the "ground truth" phrases for class names such as "Rock Climbing", they can often be identified with a related noun such as "Mountain Climbing".

*All-Age-Faces Dataset (Cheng et al., 2019)*: A dataset containing over $82K$ images of human faces. The ground truth annotation of each person's age (between 2 and 80) is available. For evaluation, we split that dataset to five age groups in even intervals. We ran our method using the complete list of all ages in intervals of one year without filtering (due to lack of balance in the dataset).

*People Playing Musical Instrument (PPMI) (Yao & Fei-Fei, 2010)*: A dataset of people interacting with $12$ different musical instruments (we use the *PPMI+* version). While this dataset can be viewed as an object category classification dataset, the musical instrument itself is usually not the single most visually dominant object in the scene (see Fig.3, Right). We also evaluate this dataset using the $82K$ WordNet (Miller, 1995) noun list.

**Compared methods:** We evaluated our method against standard clustering methods using the same feature representation, and against other variants of zero-shot classification methods.

*ZS-Naive*: We apply zero-shot classification using CLIP for all the dataset images using the entire list of phrases. We use the classifier to assign each image to its most likely phrase. We then choose the $K$ phrases to which most images were assigned. The number $K$ will typically be much smaller than the length of the original list $W$. We finally perform zero-shot classification for each image using CLIP with the reduced list of $K$ phrases. We set this as the final cluster assignment.

Figure 2: Representative images of the different datasets, from left to right: Stanford Activity, BU-action, All-Age-Faces, PPMI

Table 1: Clustering of Attributes Classification Datasets (%)

| | Stanford Activity | | | BU-action | | | All-Age-Faces | | | PPMI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI |
| PT Only | 61.4 | 66.4 | 49.0 | 61.0 | 77.4 | 51.9 | 47.5 | 28.6 | 19.7 | 34.2 | 26.5 | 15.8 |
| PT+SCAN | 54.0 | 66.0 | 45.3 | 63.1 | 78.8 | 56.9 | 48.8 | 33.4 | 20.2 | 27.5 | 24.1 | 12.3 |
| ZS-Naive | 66.0 | 74.5 | 55.3 | 52.5 | 72.9 | 44.3 | 50.6 | **38.7** | 24.2 | 35.6 | 29.6 | 16.6 |
| Ours | **72.8** | **76.3** | **64.5** | **65.9** | **79.5** | **57.8** | **58.5** | 38.4 | **26.6** | **39.4** | **33.7** | **21.9** |
| ZS-GT | 82.8 | 80.1 | 72.0 | 77.0 | 83.6 | 67.2 | 60.7 | 40.8 | 30.3 | 54.5 | 44.5 | 33.5 |

*PT Only*: Classical Ward's clustering based on CLIP's visual features but without the language priors.

*PT+SCAN:* We evaluate SCAN(Van Gansbeke et al., 2020) image clustering method using CLIP's pretrained visual features. The pretrained features are both used for selecting the neighbors in the first stage, and as the initialization of the second stage.

*CLIP-GT*: Zero-shot classification using CLIP with the ground truth class names as the K phrases. Note that this method uses more supervision than our method. In fact, recovery of the ground truth phrases is the aim of the core part of our method. We consider it as an approximate upper bound on our method.

We use the three most common clustering metrics: accuracy (ACC), normalized mutual information (NMI) and adjusted rand index (ARI).

**Results:** It is clear from Tab.1 that our method is able to utilize the language guidance for better clustering performance. We achieve strong results on the four evaluated datasets. Yet, our setting assumes the availability of two components not assumed by previous methods: pretrained visual features and a feature embedding of the phrase dictionary. We therefore compare our method to baseline methods enjoying similar supervision. We compare to *PT Only* and *PT+SCAN* which rely only on CLIP's visual features. While these methods show some image clustering capabilities, we see that the language guidance provides significant improvement on top of the "visual only" baselines. As our classes may differ in many visual properties the language guidance helps to limit the possible clusters to the ones we may look for.

The given phrase dictionary, although extensive, is another kind of supervision used by our method. Therefore, we compare our method also to the *ZS-Naive* baseline, utilizing the same supervision. Our method outperforms here as well. This stresses the contribution of the suggested method. A naive use of the language guidance such as *ZS-Naive* might find among its $K$ most common phrases ones that are not descriptive of a single class (or ones that describe well images spanning multiple classes). Further empirical results into the performance of our algorithm for object identification dataset features can be found in Sec.6.

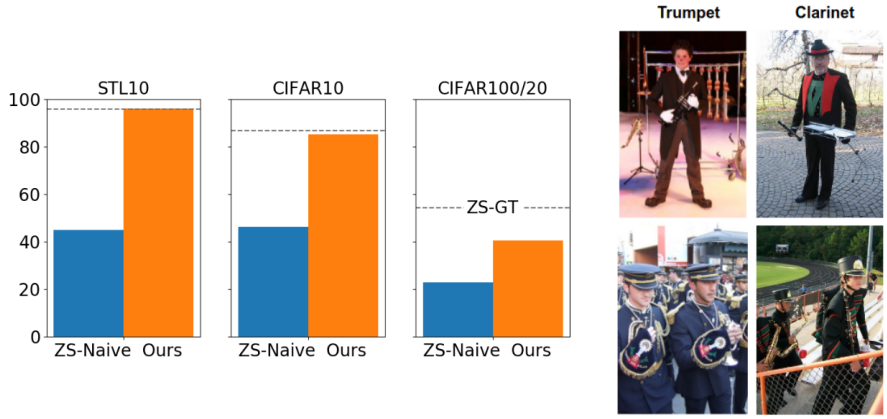


Figure 3: Left: Accuracy (%) of the naive zero-shot classification using the entire phrase list compared to our method. The dashed line notes the zero-shot accuracy using the "ground truth" phrases (the given class names). Right: While the ground truth labels in datasets such as "People Playing Musical Instrument" are based on identification of objects, these objects are often far from being the most salient item in the image.

| Accuracy (%) | STL10 | CIFAR10 | CIFAR20/100 |
|---|---|---|---|
| No filtering | 71.6 | 67.4 | 38.2 |
| With filtering | 72.8 | 77.0 | 39.4 |

Table 2: Our method with and without filtering non-specific phrases (see Sec.3.2)

### 5.1 Implementation details

**Optimization:** We run our algorithm with $p = \frac{\#classes}{2}$ swaps per iteration. For every experiment we run our algorithm for 30 iterations which was found to be enough for convergence for all datasets.

**Dictionary:** For each dataset, we performed "generality" filtering by testing different quantile levels $q$. We used 20 $q$ values between 0.05 to 1 in 0.05 intervals. We selected $q$ using our unsupervised criterion (Sec.3.2).

**Dataset:** We used the standard train vs. test split for all datasets. We Trained all the methods on the training data and evaluated on the test data.

**Features:** We used the CLIP(Radford et al., 2021) pretrained model for our pretrained visual and text features. The visual features were extracted using the CLIP pretrained ViT-B-32 network. For the text features we used the CLIP pretrained transformer. Prompted with a "This is a photo of a ***" prompt, where *** is the single-phrase from our dictionary.

**SCAN comparisons:** When evaluating the *PT+SCAN* baselines we replaced SCAN's backbone with the visual ViT head of CLIP. We then ran the first stage of SCAN without feature training. We ran the second stage with the same backbone. We used the parameters for the most relevant dataset evaluated by SCAN authors. For the further zero-short classification comparisons using the object categorization benchmarks we also ran a variation using our language guidance. This evaluation was conducted in a similar manner to *PT+SCAN*, but with the nearest neighbours of SCAN's first stage limited to within the same language-guided cluster (as was calculated by our method). In both variants we tried to train the entire feature extractor or the head only, and took the better performing mode between the two. For statistics, we ran each evaluation of SCAN three times.

## 6 ANALYSIS

**Further zero-short classification comparison:** For a further investigation into our method we report its performance on the different datasets using the other zero-shot classification variants discussed in Sec.5. As the commonly used benchmarks are driven from object-classification datasets, we use an extensive list of all WordNet (Miller, 1995) nouns as our phrase lists for all these datasets. We find that the performance we achieve is close to that of CLIP's zero-shot classification. This demonstrated our algorithm's ability to recover phrases resembling the ground-truth concepts used for the construction of the dataset. Accordingly, our method significantly outperforms our *ZS-Naive* baseline and approaches *ZS-GT* performance noted by the dashed line.

On object-identification datasets, SCAN can significantly improve its performance using the CLIP pretrained visual features, achieving $98.33$, $88.3$ and $46.7$ on STL10 (Coates et al., 2011), and CIFAR10 (Krizhevsky et al., 2009) and CIFAR20/100(Krizhevsky et al., 2009) respectively. As SCAN's inductive bias already learns to detect a single salient object in each image, we can obtain only minor gains from the language guidance by incorporating our method into SCAN's first stage (achieving $98.4$, $88.5$ and $47.1$, all within one standard deviation from *PT+SCAN*). Both these versions of SCAN outperform *ZeroShot-gt*, that approximates the maximal performance we can expect using zero-shot classification approaches such as our without feature adaptation. We note that while the CLIP (Radford et al., 2021) paper reports better zero-shot classification results on these datasets, it uses extensive prompt engineering which is beyond the scope of this paper.

**Filtering our phrase list:** Before running the algorithm, we filter out phrases whose "generality" score is above some quantile $q$, as mentioned in Sec.3.2. We show the performance of our method with and without filtering in Tab.2.

**Ground truth phrase retrieval:** While the class names used by the creators of the datasets are only rarely recovered exactly, for many classes the cluster centers are close in meaning to the original phrase (e.g., "firing an arrow" for "shooting an arrow" or "marching the dog" for "walking the dog"). The *ZS-Naive* method often finds less accurate phrases (e.g. "testing an arrow").

**Facility location optimization methods:** As explained in Sec.4.3, our optimization method can be viewed as a relaxed version of the Local Search algorithm. The original Local Search algorithm is very slow. Yet, we were able to run it with a single swap in each step (also known as the Partitioning around Medoids algorithm or PAM) for the All-Age-Faces dataset which is both small and utilizes a relatively short list. As can be seen in Tab.3, PAM reaches comparable losses to our method. Both methods achieve loss values that are lower than the loss with the ground truth phrases as center ($L_{gt} = 1.455$, for All-Age-Faces Dataset). These metrics suggest that both methods can effectively optimize the objective. Conversely, PAM is much slower than our method. For large dataset and larger lists the time complexity of PAM is infeasible and significantly greater than that of our relaxed version.

Table 3: Comparison between PAM and our method (All-Age-Faces Dataset)

| | Ours | PAM |
|---|---|---|
| Train Accuracy (%) | 58.6 | 58.4 |
| Loss | 1.451 | 1.450 |

## 7 CONCLUSION

We proposed language-guided image clustering for attribute clustering. We reduce the task to a well-studied discrete optimization task, the uncapacitated K-facility location problem. To solve this task with acceptable runtime, we suggested an efficient optimization method for solving it. The phrases we find optimize the loss well. Therefore, a further improvement of our method is more likely to be achieved by improving our loss function or by extending our method's expressivity. A more expressive method may utilize a language model to produce more specific phrases, or combine our method with other clustering methods for finetuning the visual features based on the data.

## REFERENCES


Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on computing*, 33(3):544–562, 2004.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149, 2018.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pp. 5879–5887, 2017.

Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3):333–345, 2019.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Gérard Cornuéjols, George Nemhauser, and Laurence Wolsey. The uncapicitated facility location problem. Technical report, Cornell University Operations Research and Industrial Engineering, 1983.

Luke Nicholas Darlow and Amos Storkey. Dhog: Deep hierarchical object grouping. *arXiv preprint arXiv:2003.08821*, 2020.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.

Joris Guérin, Stephane Thiery, Eric Nyiri, Olivier Gibaru, and Byron Boots. Combining pretrained cnn feature extractors to enhance clustering of complex natural images. *Neurocomputing*, 423: 551–571, 2021.

Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. *Journal of algorithms*, 31(1):228–248, 1999.

Philip Haeusser, Johannes Plapp, Vladimir Golkov, Elie Aljalbout, and Daniel Cremers. Associative deep clustering: Training a classification network with no labels. In *German Conference on Pattern Recognition*, pp. 18–32. Springer, 2018.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International Conference on Machine Learning*, pp. 1558–1567. PMLR, 2017.

Kamal Jain and Vijay V Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. *Journal of the ACM (JACM)*, 48(2):274–296, 2001.

Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 731–740, 2002.

Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9865–9874, 2019.

Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pp. 67–84. Springer, 2016.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alfred A Kuehn and Michael J Hamburger. A heuristic program for locating warehouses. *Management science*, 9(4):643–666, 1963.

Shugao Ma, Sarah Adel Bargal, Jianming Zhang, Leonid Sigal, and Stan Sclaroff. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, 68:334–345, 2017.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Aleksandra Mojsilovic. A computational model for color naming and describing color composition of images. *IEEE Transactions on Image processing*, 14(5):690–699, 2005.

Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First international workshop on multimedia intelligent storage and retrieval management*, pp. 1–9. Citeseer, 1999.

Chuang Niu and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *arXiv preprint arXiv:2103.09382*, 2021.

Chuang Niu, Jun Zhang, Ge Wang, and Jimin Liang. Gatcluster: Self-supervised gaussian-attention network for image clustering. In *European Conference on Computer Vision*, pp. 735–751. Springer, 2020.

Raphael Prates, Cristianne RS Dutra, and William Robson Schwartz. Predominant color name indexing structure for person re-identification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 779–783. IEEE, 2016.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2, 2021.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. *arXiv preprint arXiv:2008.01392*, 2020.

Guy Shiran and Daphna Weinshall. Multi-modal deep clustering: Unsupervised partitioning of images. *arXiv preprint arXiv:1912.02678*, 2019.

Guy Shiran and Daphna Weinshall. Multi-modal deep clustering: Unsupervised partitioning of images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4728–4735. IEEE, 2021.

David B Shmoys, Éva Tardos, and Karen Aardal. Approximation algorithms for facility location problems. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pp. 265–274, 1997.

Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. Mi{ce}: Mixture of contrastive experts for unsupervised image clustering. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=gV3wdEOGy_V`.

Joost Van De Weijer and Fahad Shahbaz Khan. An overview of color name applications in computer vision. In *International Workshop on Computational Color Imaging*, pp. 16–22. Springer, 2015.

Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, 2009.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pp. 268–285. Springer, 2020.

Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8150–8159, 2019.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pp. 478–487. PMLR, 2016.

Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, pp. 3861–3870. PMLR, 2017.

Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *European conference on computer vision*, pp. 536–551. Springer, 2014.

Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9–16. IEEE, 2010.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*, pp. 1331–1338. IEEE, 2011.

Lu Yu, Lichao Zhang, Joost van de Weijer, Fahad Shahbaz Khan, Yongmei Cheng, and C Alejandro Parraga. Beyond eleven color names for image understanding. *Machine Vision and Applications*, 29(2):361–373, 2018.