# I Spy With My Little Eye
# A Minimum Cost Multicut Investigation of Dataset Frames

Katharina Prasse[1], Isaac Bravo[2], Stefanie Walter[2], Margret Keuper[1,3]
[1]University of Mannheim, Mannheim, Germany, katharina.prasse@uni-mannheim.de
[2]Technical University Munich, Munich, Germany, {isaac.bravo, stefanie.walter}@tum.de
[3]Max-Planck-Institute for Informatics, Saarland Informatics Campus, keuper@uni-mannheim.de

## Abstract

*Visual framing analysis is a key method in social sciences for determining common themes and concepts in a given discourse. To reduce manual effort, image clustering can significantly speed up the annotation process. In this work, we phrase the clustering task as a Minimum Cost Multicut Problem [MP]. Solutions to the MP have been shown to provide clusterings that maximize the posterior probability, solely from provided local, pairwise probabilities of two images belonging to the same cluster. We discuss the efficacy of numerous embedding spaces to detect visual frames and show its superiority over other clustering methods. To this end, we employ the climate change dataset ClimateTV which contains images commonly used for visual frame analysis. For broad visual frames, DINOv2 is a suitable embedding space, while ConvNeXt V2 returns a larger number of clusters which contain fine-grain differences,* i.e. *speech and protest. Our insights into embedding space differences in combination with the optimal clustering - by definition - advances automated visual frame detection. Our code can be found at https://github.com/KathPra/MP4VisualFrameDetection.*

## 1. Introduction

Frame analysis [11] plays a key role in social science research. This method extracts the main concepts from a dataset, which is generally collected for a single analysis and not shared with the community. While frame analysis was originally text-centric, visual frame analysis has gained traction in the field, as images have become ever more present in communication. Visual frames can be formal/stylistic or content-oriented [44]. The detection of formal/stylistic frames can be easily automated with the help of style detection algorithms. Content-oriented frame detection is a much harder task to automate. Authors can either use zero-shot classification by selecting suitable frames from previous works, rely on clustering approaches, or a combination thereof. Automated frame detection is not well-researched and datasets are often manually annotated. The detection of abstract and diverse concepts remains a challenging task [53] and an active field of research.

We argue that clustering is the most suitable approach as it does not rely on pre-defined frames and thus imposes less bias on the outcome of the investigation. The emergence of novel frames, as investigated among others by O'Neill [40], can only be detected through this approach - or manually. Mooseder *et al*. employ clustering in their work to reduce the number of manual annotations. Similarly, Oquab *et al*. [39] use clustering to extend their dataset, while Zhou *et al*. employ clustering to evaluate their online tokenizer. Zhang [53] employs clustering to clean their dataset, by removing outliers. To provide clusters within a probabilistically meaningful framework, we propose to phrase the clustering problem as a Minimum Cost Multicut Problem [MP] [8], in which the images are the nodes of a graph with weighted edges indicating the images' probability of being in the same cluster. Without any hyperparameters, one can obtain a clustering with the maximum posterior probability.

The MP takes image similarities as an input which we generate using several vision or vision and language models. The annotated datasets ImageNette [19] and ImageWoof [20] are our independent validation and test set for finding the optimal calibration. On the dataset ClimateTV [41], we show and discuss the effectiveness of our approach for social science research. Our work further investigates embedding space differences both in the embedding and the resulting clusterings.

Our contributions are: (1) We advance automated visual frame detection by extensive clustering analysis. (2) To this aim, we formulate semantic embedding-based clustering as a Minimum Cost Multicut Problem that maximizes the posterior probability of the clustering. (3) We analyse the efficacy of powerful vision foundation models for this novel application and provide concrete recommendations on which embedding spaces are most suitable for this task.

## 2. Related Work

Visual frame analysis investigates frames in communication. Alone in the context of climate change, numerous works have used this method [5,7,12,32,36,38,40,43,44,49, 50]. Advances in automating the annotation process have recently started. Mooseder *et al*. [36] employ the *k-means* algorithm to cluster their images' VGG16 features and then manually annotate 100 random images per cluster. They set k= 5,000, which results in the manual annotation of 50,000 images. In total, they found 21 distinct frames with 8% of images excluded based on the clustering results. Given an optimal clustering for their data, their manual workload would have been drastically lowered, with a lower bound of 2,100 images to manually annotate.

Phrasing a problem as an MP has many applications in computer vision. The most prominent use case is multi-person tracking [16, 17, 37, 46, 47], where it is used to link and cluster person hypotheses over time. Furthermore, Andres *et al*. employ the MP to generate a probabilistic image segmentation [3], followed by [10, 21–23]. Keuper *et al*. build upon this work and use the MP for efficient image and mesh graph decomposition [28] and motion segmentation [27] with several follow-up works [24,26,29,30]. Ho *et al*. have employed the MP for image clustering [14, 18]. In contrast to their work, we do not train a deep neural network when computing the inputs for the MP. While MP clustering has been done in the past [15], we are the first to use MP clustering in conjunction with foundation model embedding spaces. Given the recent advances in computer vision, we employ models with highly expressive embedding spaces to create image features and use their pair-wise cosine similarities as inputs. This approach to obtaining edge costs has been proven successful by Swoboda *et al*. [1], who use ResNet-50 features [13]. The strength of cosine similarities has been ever present and its efficacy for foundation models was highlighted by Radford *et al*. [42].

In this work, we leverage vision models' and VLMs' embedding spaces for feature generation. While clustering using traditional vision models is aptly, foundation models clustering is just gaining traction within the community [4,48,52]. We want to highlight the concurrent work of Wagner *et al*. who show the efficacy of DINO features for data exploration [48]. In our work, we compare the expressiveness of performant vision model's embedding spaces w.r.t. their capabilities to represent abstract visual concepts. Visual frame analysis is precisely interested in such concepts, as its goal is to understand the common themes in a given dataset. The discussion of embedding space differences both adds to our understanding and advances their applicability in data exploration. We further strengthen our point by evaluating on OmniBenchmark [53], a computer vision dataset which is designed to test how universal vision features are.

We compare CLIP's general-purpose features which have been trained on a web-scale image-text dataset and have a high zero-shot classification accuracy [42] to DINOv2 features, which are trained using optimized training data collection with the goal of increasing features' robustness [39]. With ConvNeXt V2 we include another model which achieves a high classification accuracy [51]. This architecture contains fully convolutional masked autoencoders and a global response normalization. Moreover, we employ ResNet-50, VGG19-BN, and ViT features to assess the expressiveness differences between smaller and larger models. Liang *et al*. observe the CLIP images' cosine similarities resulting in a narrow cone, and conclude that the image embeddings occupy a small part of the embedding space [31]. Based on data distribution comparisons and image cluster analysis, we provide a guideline for the choice of embedding model.

## 3. Methods

The Minimum Cost Multicut problem is a graph problem which involves finding the cutting of the graph into distinct clusters such that the cost, the sum of the cut edges' weights, is minimal. As we phrase the clustering problem as a Minimum Cost Multicut Problem [MP] [8], we map the images to a graph structure. To this end, we embed all images and construct a fully connected graph which edge weights are the images' cosine similarities. We investigate embedding differences with respect to their expressiveness and effectiveness of finding visual frames. We use six models, differing in terms of parameters and classification performance. ResNet-50 [13], VGG19-BN [45], Vision Transformer B/32 [9], ConvNeXt V2 [51], and DINOv2 [39] are pure vision models, while CLIP ViT-B/32 [42] is trained in a multi-modal setting using image-text pairs. An overview of the employed models and their characteristics can be found in Appendix A.

### 3.1. Image Clustering

We phrase the image clustering task as a MP, also referred to as weighted Correlation Clustering.

**Definition 1** *A finite, undirected graph G = (V,E) with cost* $w : E \to \mathbb{R}$ *associated with the edges is separated into detached components such that the cost is minimal*

$$\min_{\boldsymbol{y}\in\{0,1\}^{|E|}} \quad c(\boldsymbol{y}) = \boldsymbol{y}^T \boldsymbol{w} = \sum_{e\in E} w_e y_e, \tag{1}$$

*where y is the binary edge label indicating whether the edge should be cut. This is subject to the linear constraint*

$$\forall C \in \mathit{cycles}(G), \forall e \in C : (1 - y_e) \leq \sum_{e'\in C\setminus\{e\}} (1 - y_{e'}). \tag{2}$$

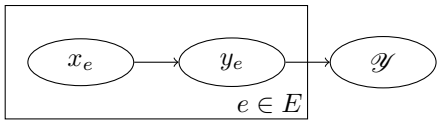



Figure 1. The Multicut Problem can be understood as a Bayesian Network which aims to predict the optimal partitioning $\mathscr{Y}$.

In line with previous work [3,28], the MP can be understood as a Bayesian Network, where the optimal partitioning $\mathscr{Y}$ depends on the individual edge decisions $y_e \in \{0, 1\}$. They are dependent on the image-pair features $x_e \in \mathbb{R}^n$ for all $e \in E$ of the graph G, as shown in Fig. 1.

Given appropriately set edge costs $w_e = \log\left(\frac{1-p(y_e|x_e)}{p(y_e|x_e)}\right)$, solving the MP is equivalent to maximizing the posterior probability $p_{y|x,\mathscr{Y}}$ with

$$p(y \mid x, \mathscr{Y}) \propto p(\mathscr{Y} \mid y) \cdot p(x, y) \tag{3}$$

which can be rewritten as

$$p_(y \mid x, \mathscr{Y}) \propto p(\mathscr{Y} \mid y) \cdot p(x \mid y) \cdot p(y). \tag{4}$$

This proportionality holds under the assumption that $x$ and $\mathscr{Y}$ are conditionally independent. The right-hand side of Eq. (4) contains three parts, the likelihood of a clustering $p(\mathscr{Y} \mid y)$ which is set to zero, if $y$ differs from the optimal clustering, and to a constant otherwise, as we do not have any prior knowledge about the clustering. The second part is the likelihood of the image similarity feature $p(x \mid y)$ and the third part is the bias term $p(y)$. By choosing strong embedding models, we assume all $p(x \mid y)$ are high for their respective $\mathscr{Y}$ and that different embedding spaces allow to detect different frames. We compare the visual frames detected using MP clustering to centroid-based k-means, density-based DBSCAN, and hierarchical agglomerative clustering using WARD linking.

### 3.2. Image-graph mapping

To formulate our clustering task as a MP, we map the images to nodes in a fully connected graph as Ho *et al*. suggests [14]. In the graph, the edge weights represent the cosine similarity between image embeddings. While the cosine similarity is defined for the range $-1 \leq c_s \leq 1$, our analysis in Sec. 4.2 shows that its distribution differs greatly between embedding models. First, min-max scaling is used to confine the weights to the range $[0, 1]$. Then, the weights are transformed such that the decision boundary for cutting an edge is at zero, with positive weights corresponding to the likelihood of images belonging to the same cluster and negative weights to different ones. Naïvely, the decision boundary is at the transformed, normalized cosine similarity of 0.5, using

$$w_{ab} = \log \frac{1 - p(y_{ab} \mid x_{ab})}{p(y_{ab} \mid x_{ab})} \propto \log \frac{s_c(a, b)}{1 - s_c(a, b)} \tag{5}$$

where $a$,$b$ are nodes in the graph, corresponding to images, and $s_c(\cdot, \cdot)$ is their cosine similarity. Depending on the embedding space, the inherent decision boundary can be located at different positions. To assign appropriate pseudo-probabilities and account for different calibration of the cosine similarity w.r.t. probabilities, we ablate a calibration term $cal$ for each embedding space and set

$$w_{ab} = \log \frac{s_c(a, b)}{1 - s_c(a, b)} + \log \frac{1 - cal}{cal}. \tag{6}$$

We ablate for $0.1 \leq cal \leq 0.9$ in Sec. 3.6 on independent validation and test sets. To this end, we use two annotated datasets and compare the clusterings to the dataset's classes.

### 3.3. Solvers

Efficient heuristics [28] are used to solve the MP *i.e*. finding image clusters. We employ the Greedy Additive Edge Contraction [GAEC] and the Kerninghan-Lin [KL] algorithm [25] to efficiently cut the graph into distinct components in the implementation of Keuper *et al*. [28] in [2]. More algorithm details can be found in the documentation.

**Algorithm 1** Greedy Additive Edge Contraction

**Require:** $G = (V, E)$, Edge Weights $w_e \; \forall e \in E$
**Ensure:** Final set of clusters $C$ and total cost
  Initialize $C$ with each vertex in its own cluster
  Initialize total cost = 0
  Create a priority queue $Q$ to store edges $(u, v)$ sorted by their weights in descending order
  **while** $Q$ is not empty **and** highest edge weight $\geq 0$ **do**
    Extract edge $(u, v)$ with the highest weight from $Q$
    Merge the clusters containing $u$ and $v$
    Update the cluster set $C$ accordingly
    **for** each edge $(x, y)$ adjacent to $u$ or $v$ **do**
      Update the weight of the edge $(x, y)$ if needed
      Reinsert updated $(x, y)$ into $Q$
    **end for**
    Update total cost += weight of merged edge $(u, v)$
  **end while**
  **return** the final set of clusters $C$ and total cost

First, we use GAEC, described in Algorithm 1, to compute a preliminary clustering. This algorithm starts with each image in a separate cluster and iteratively merges the two clusters connected by the largest, positive edge weight. The graph size is continuously reduced during the execution and the algorithm stops when merging any two additional clusters would have a negative weight. The resulting, preliminary clustering is then optimized using the KL algorithm described in Algorithm 2, where three types of changes are possible, (1) exchange nodes of two neighbouring clusters, (2) move nodes to an new cluster, and (3) join

two neighbouring clusters. The clustering refinement using the KL algorithm improves the clustering results, as shown in C. Conceptually, the KL algorithm tries to improve an initial clustering by iteratively making small adjustments to the clustering and tracking their effect in terms of overall clustering cost. The cost of a clustering consists of the the sum of all cut edges' weights and the algorithm aims to minimize it.

**Algorithm 2** Kernighan-Lin Algorithm with Joins (KLj)

**Require:** Graph $G = (V, E)$, Edge Weights $w_e \forall e \in E$
**Ensure:** Partitioning $P$ with $|P| > 1$
  **procedure** EXTERNAL COST($a, P$)
    **return** $\sum_{v \in P \setminus \{P_a\}} w(a, v)$
  **end procedure**
  **procedure** INTERNAL COST($a, P$)
    **return** $\sum_{v \in P_a} w(a, v)$
  **end procedure**
  Compute initial D-values for all $v \in V$ where $D(v) =$ EXTERNALCOST$(v, P)$ − INTERNALCOST$(v, P)$
  **repeat**
    **for** each edge $e = (a, b) \in E$ **do**
      **if** node_changed($a$) **or** node_changed($b$) **then**
        Update $D(a)$ and $D(b)$
        Update partition $P'$
      **end if**
    **end for**
    **for** each node $a \in V$ **do**
      **if** node_changed($a$) **then**
        Update $D(a)$
        Update partition $P'$
      **end if**
    **end for**
  **until** no further changes in $P'$
  **return** the optimized partitioning $P'$

### 3.4. Metrics and Evaluation

To measure the differences between clusterings, we use the variation of information [VI] proposed by Meilă [35]

$$VI(C, C') = H(C \mid C') + H(C' \mid C), \tag{7}$$

where $VI$ is the sum of the two conditional entropies of the two clusterings. Each clustering result may consist of 1 to x clusters. The $VI = 0 \iff C = C'$ has the upper bounded $VI(C, C') \leq \log n$ where n is the number of nodes in the graph *i.e.* the number of images. It is a true metric and the triangle inequality holds. We ablate the calibration term by investigating the two conditional entropies, comp. Eq. (8) on an independent validation and test set.

$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y \mid x) \tag{8}$$

We use the implementation of [6], which iterates over all combinations of clusters in the two clusterings.

Additionally, we use standard cluster statistics to compare the clusterings in terms of cluster sizes and diversity. For ClimateTV, we manually investigate the largest clusters by randomly selecting 10 images per cluster. The differences between embedding models are further assessed in terms for common clusters and clustering differences. We extend this by investigating which clusters are entirely contained within another dataset's larger cluster.

### 3.5. Datasets

We employ two curated datasets, ImageNette [19] and ImageWoof [20], for determining the optimal $cal$ term for each embedding space. ImageNette is selected to assess the efficacy for the distinction between broad concepts and ImageWoof for fine-grained concepts. $Cal$ is selected based on the training set clustering and its effectiveness is validated using the validation set. In both cases we compare the MP clusterings to the dataset's classes. Moreover, we use OmniBenchmark to measure how many realms and concepts the assessed embedding spaces can distinguish between. It contains 21 realms which each consist of several concepts, in total 7,372 non-overlapping concepts are included and classification top-1 accuracy is currently below 50%. To this end, we randomly select 10k images from the train set and compare the clusters to the original labels, both on realm level and concept level. Finally, we employ the ClimateTV dataset [41] to exemplify the efficacy of our method in detecting visual frames. This classification dataset contains animal classes, social media visuals of political protest, conferences, climate change solutions such as wind energy, and several climate change consequences, *e.g.* floods, droughts, economic instability, and human rights infringements. The authors have collected all images that were shared on X (formerly Twitter) in the year 2019 in the context of climate change. We focus our evaluation on the images tweeted in January 2019. More details can be found in their work.

### 3.6. Experimental Setup

We create image embeddings using ResNet50, VGG19-BN, ViT, ConvNeXt V2, CLIP ViT, and DINOv2 (details in Appendix A), and build complete image graphs with weighted edges based on the image embedding's cosine similarity. This graph is then divided into clusters using heuristic solvers for the MP. The experiments were run on Intel Xeon CPU E5 and the image embeddings are created using NVIDIA GeForce RTX 4090. The clusterings are compared with respect to their VI. Additionally, we report further cluster statistics, *e.g.* cluster sizes, distribution, and cleanliness. The comparison of quantitative and qualitative results concludes our investigation.

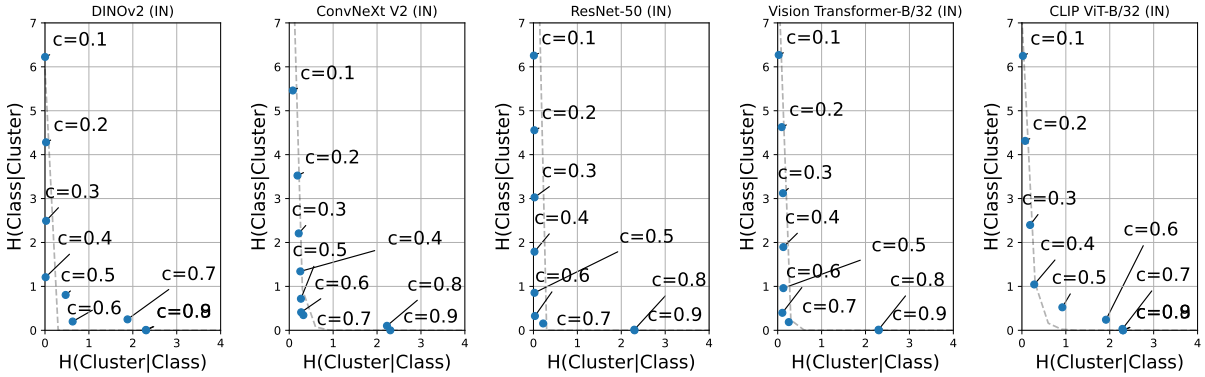



Figure 2. Calibration term [c] ablation across embedding spaces on ImageNette's train set shows that embedding spaces where the distance between different data points is increased during training require a smaller $cal$ compared to traditionally trained embedding spaces.

## 4. Results

We determine the calibration term $cal$ (defined in Sec. 3.2) for each embedding space using the annotated ImageNette [19] and ImageWoof datasets [20]. Moreover, we report on the characteristics of the embedding spaces such as their data distributions and the overlap between embedding spaces using UMAP [33] visualizations using the official implementation [34]. Finally, we discuss clusters for the ClimateTV [41] dataset to show how our method can support social science research. We compare the cluster statistics and results to other clustering approaches. This includes an excursion into multi-modality where we combine image and text input for the CLIP model.

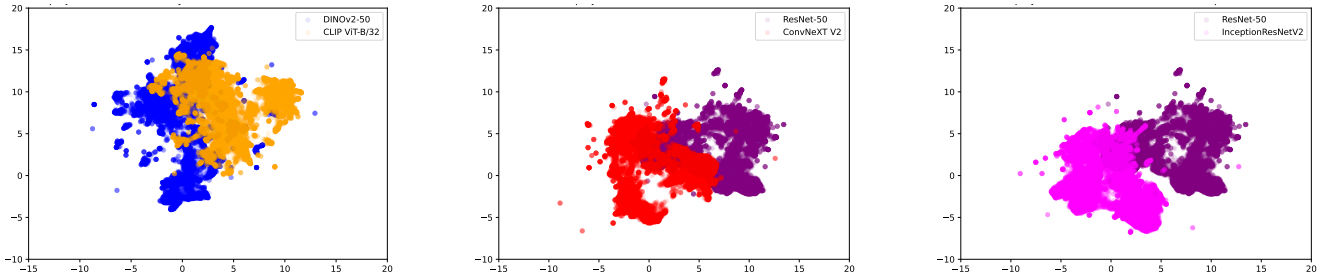
### 4.1. Calibaration Term Validation

We determine $cal$ by assessing the similarity of image classes and clusters. The train set is used for experiments using $0.1 \leq cal \leq 0.9$ with a step size of 0.1, whereof the best performing $cal$ is independently validated. This is done for ImageNette, which has highly diverse classes, and for ImageWoof, which has fine-grained differences between classes. Conceptually, the larger $cal$ is chosen, the more edge weights are larger than zero; reducing $cal$ has the opposite effect. Fig. 2 shows the most optimal $cal$ term in terms of two conditional entropies, $H(Class \mid Cluster)$ and $H(Cluster \mid Class)$. Given that the framework is probabilistic, we expect minor differences between runs. We select $cal$ such that $H(Class \mid Cluster)$ and $H(Cluster \mid Class)$ are balanced, as we aim for a high overlap between the clusters and the classes, *i.e.* a low VI. VI's differ slightly between the train and the validation set. Overall, CLIP ViT-B/32, DINOv2, and VGG19-BN have the worst clustering performance in terms of overlap with the original classes. All other models perform well, with ResNet-50 achieving the highest VI overall, as Tab. 1 shows. The same trends can be observed for fine-grained differences, as the experiments on ImageWoof show (Tab. 2). Again, CLIP ViT-B/32 achieves the worst VI and ResNet-50 the best. However, in this setting, all embedding models' performances, except CLIP's, are more alike. For the majority of embedding spaces, reducing $cal$ by a factor 0.1 improves the clustering similarity to the original classes. We suggest selecting $cal$ based on our ablation with the application in mind. If the data contains small differences, $cal$ can be reduced by 0.1.

| Emb. model | $cal$ | $\Delta_{tr} H_1, H_2$ | $VI_{train}$ | $VI_{val}$ |
|---|---|---|---|---|
| CLIP ViT-B-32 | 0.5 | 0.40 | 1.55 | 1.34 |
| DINOv2 | 0.6 | 0.43 | 0.89 | 1.19 |
| ConvNeXt V2 | 0.7 | 0.02 | 0.28 | 0.42 |
| ViT-B-32 | 0.7 | 0.07 | 0.46 | 0.26 |
| ResNet-50 | 0.7 | 0.07 | 0.44 | 0.54 |
| Inc.-ResNetv2 | 0.5 | 0.17 | 0.49 | 0.40 |
| VGG19-BN | 0.7 | 0.19 | 0.73 | 0.94 |

Table 1. Clustering ImageNette using ConvNeXt V2 closely fits the training set's classes. We use $\Delta H_1, H_2 = H(Class \mid Cluster) - H(Cluster \mid Class)$ as an indicator of clustering performance. ResNet-50 has the best validation VI.

We hypothesize that the slightly larger $cal$ aids the clustering of broad concepts, as it allows more images to be clustered together by GAEC. This algorithm does not join any negative edge, thus any two image representations connected by it, cannot initially be in the same cluster. The KL algorithm may alter the initial cluster assignment, however, the larger the negative cost, the less likely is this scenario. Likewise, in the setting with fine-grained differences between classes, it appears optimal to have a lower number of initially positive edge weights to avoid clustering different classes together. We expected that CLIP and DINOv2 would require lower $cal$ terms, as their training includes contrastive loss and the KoLeo regularizer respectively.




(a) Contrastively trained spaces (b) CNN spaces: ResNet vs. ConvNeXt V2 (c) ResNet variants

Figure 3. UMAP visualization for image embeddings on ClimateTV reveals the differences in embedding space occupancy between different embedding models. DINOv2 and CLIP ViT-B/32 have a similar $s_c$ distribution but only a small overlap. The embedding space comparison of CNNs pre-trained on ImageNet1k shows almost no overlap.

### 4.2. Embedding Space Analysis

| Emb. model | $cal$ | $\Delta_{tr}H_1, H_2$ | $VI_{train}$ | $VI_{val}$ |
|---|---|---|---|---|
| CLIP ViT-B-32 | 0.4 | 1.05 | 2.31 | 2.26 |
| DINOv2 | 0.5 | 0.40 | 1.45 | 1.39 |
| ConvNeXt V2 | 0.6 | 0.01 | 0.61 | 1.03 |
| ViT-B-32 | 0.6 | 0.04 | 1.04 | 1.10 |
| ResNet-50 | 0.7 | 0.03 | 0.86 | 0.73 |
| Inc.-ResNetv2 | 0.5 | 0.04 | 0.72 | 0.97 |
| VGG19-BN | 0.6 | 0.41 | 1.25 | 1.13 |

Table 2. Clustering ImageWoof using ResNet-50 closely fits the dataset's classes. $\Delta H_1, H_2 = H(Class \mid Cluster) - H(Cluster \mid Class)$ is an indicator of clustering performance.

We analyse all image embeddings' un-normalized cosine similarities to asses the dataset's native distribution, as visualized in Fig. 10. All cosine similarity distributions have a long-tail towards $c_s = 1$. This is anticipated, as each image is only similar to the other $10\%$ of images which have the same class and are dissimilar to the remaining $90\%$ of the images. The tail of the ViT and the ConvNeXt V2 model appear almost disjoint from the remaining distribution. While some cosine similarity distributions appear more bell-shaped (DINOv2, CLIP RN50, CLIP ViT, Inception-ResNet-v2), others appear more skewed (ConvNeXt V2, RN50, ViT, VGG19). It appears that the more PDF resembles the normal distribution, the lower $cal$ term is optimal. Models, which have less similar embeddings require a higher $cal$ to have a large enough number of positive edge weights. Our findings also show *narrow cone effect* [31], *i.e.* embeddings having an above average cosine similarity and thus only cover a narrow cone in the embedding space hypersphere. The narrow cone effect [31] can be observed in the CLIP (ViT & RN50) embedding space and in the DINOv2 embeddings space, as their $\mu$ values are far beyond zero. When the same architecture (ViT & RN50) is pre-trained without contrastive loss, the narrow cone is not formed. We have included the ResNet-50 CLIP encoder here, to show another instance of the narrow cone forming through multi-modal training. The narrow cone is apparent for some of the self-supervised models, but not for ConvNeXt V2. Further causal investigations are beyond the scope of this work.

We find that the choice of embedding model has a large effect on the clusterings' VI, as no clear relation between $\mu$ and VI can be observed. Fig. 3 shows how little different embedding spaces overlap. We can observe, that DINOv2 embeddings have the largest spread, while CNN embedding spaces appear more dense.

The models with a more Gaussian distribution are naturally less affected by an alteration of $cal$. More details on the $s_c$ distributions can be found in Appendix A.1.

Based on this ablation, we advocate for setting $cal$ according to the $s_c$ distribution of the embedding model. When the inherent differences between embeddings are large, $cal = 0.7$ is required to create an impactful clustering. Models whose embeddings are approximately normally distributed require no calibration, *i.e.* $cal = 0.5$. We find that the shape of the distribution is more influential than its mean value. The granularity of the embedding space has a large effect on the final clustering.

### 4.3. Clustering using MP

We compare the clusterings' statistics to better understand model differences to further investigate on image level. For the annotated dataset, we compare the clustering to the classes. Here, the number of clusters was larger than the number of classes (10).This can be helpful in detecting outliers, *i.e.* unlikely representations of the class, as for all classes clusters of size one exist. Overall, the clusterings depend on the employed embedding model, as both statistics and image-level investigations show. The clusterings are also subject to the diversity of the dataset, as VIs are generally larger for ImageWoof than for ImageNette.

This section highlights selected findings, additional results can be found in Appendix D.

**ImageNette** clusters confirm our expectations that unusual representations are separated into single image clusters (first row), while images that contain visual elements common for other classes might be mis-clustered (Sec. 4.3). The image of the plain church, without any cross, which contains a beautiful sky is mixed with most of the parachute images, while the image with a parachute resembling a roof like structure is added into the large church cluster. This stands in contrast to the clustering obtained using the CLIP $s_c$, where 41% of clusters are mixed. They contain up to 9 classes with at times equal contributions.

Figure 4. The ImageNette church class is grouped into two single image clusters (r1), one fn as parachute and one fp parachute (r2).

**ImageWoof** clusterings contain fewer clusters for most models. Exclusively for DINOv2, more clusters are generated for ImageWoof as ImageNette. These clusters contain many images of a single class in combination with a few outliers. While this results in a high VI compared to the class labels, the clusters appear meaningful, as they are based on common features shown in Fig. 5

**ClimateTV image** clusterings are highly diverse, with $VI = 3.99$ between ConvNeXt V2 and DINOv2 clusterings. The number of clusters returned is more 2x higher for ConvNeXt V2 than for DINOv2. While the largest cluster's size is comparable between models (approx. 11,5k), the cluster size distributions differ. DINOv2 contains more large clusters in comparison, but also has a lower median value, due to more image clusters of size 1. The largest cluster's contents differ greatly, as the DINOv2 cluster contains *persons*, while the ConvNeXt V2 cluster contains *computer generated content e.g.* posters, visualizations, text. The next larger ConvNeXt V2 clusters contains speakers (5k), outdoor photographs (5k), protest (1k), portraits (427), satellite images/earth visualizations. The largest agreement (81%)

Figure 5. DINOv2 features appear to encode dog position and background, as r1 and r2 each show one cluster for ImageWoof.

between the two model's clusters is observed for frogs. Both clusters contain few additional images without frogs. Several ConvNeXt V2 clusters are contained in DINOv2 clusters, *i.e.* frequently in *persons*, but also in *animals*.

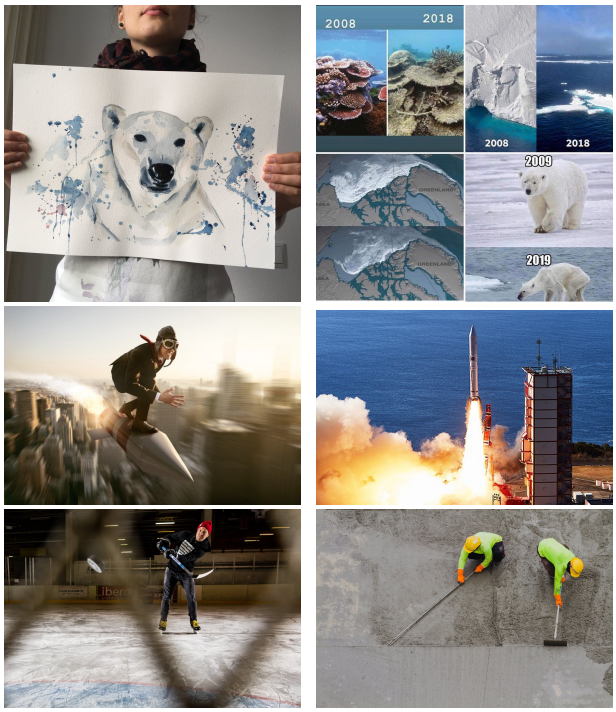



Figure 6. ConvNeXt V2 image features are clustered into content based topics, *e.g.* polar bears (r1) and rockets (r2). Image background and common object can lead to mis-clustering (r3) where a hockey player and road construction workers form a cluster.

In line with previous experiments, the ConvNeXt V2 features result in meaningful clusters, examples shown in Fig. 6, even for smaller cluster sizes. When comparing the polar bear cluster between models, its images appear in 12 DINOv2 clusters. All images containing persons (Fig. 6,

r1, left) are in *persons*. This includes both a person wearing a polar bear costume and a theater performance with stuffed polar bears, which we find remarkable. Polar bear photographs in nature are in both cases clustered together. However, for DINOv2, this cluster also contains images of other animals. The combination of the two clusterings further allows us to identify noise.

The small ConvNeXt V2 clusters of mixed images (Fig. 6, r3) can again be found in the same DINOv2 cluster, however, here in combination with almost 12k other images containing humans. Other large DINOv2 clusters contain event information or natural images, which indicates that DINOv2 clusters represent common patterns in the data. Fig. 7 shows that both embeddings have a shape bias which results in small clusters of uncommon objects with similar shapes. The DINOv2 cluster (Fig. 7, r1 & r2 left) is one example thereof, while ConvNeXt V2 clusters contain more variations of the curly shape (Fig. 7, c1).


Figure 7. While both models have a certain shape bias, DINOv2 strong tendency to clusters persons together offsets this. All images without humans form one DINOv2 cluster. The staircase images (r2, right & r1, right) by ConvNeXt V2.

**Clustering Comparison** When comparing our proposed MP clustering to KMEANS, DBSCAN, and agglomeration clustering, we can observe the same trends. The CLIP embeddings' clusterings are the least similar to the other clusterings, with DINOv2's embedding space being to closest to it. While the clusterings of KMEANS and agglomerative clustering are highly diverse, the results of DBSCAN all have inter-model VI's of less than 0.4. The main advantage of MP clustering over KMEANS is that it can detect clusters of varying length, *e.g.* DINOvs's people class would not be possible in this way. Additionally, KMEANS clustering is less suitable for outlier detection as it - by design - only has few small clusters.

**Multimodal clustering** We obtained the text corresponding to the ClimateTV images and used both as input to the MP. The CLIP ViT-B/32 embedding space is already aligned and we use $cal = 0.5$ as for the uni-modal model. Due to the increased computational cost of using twice the nodes in the graph, we selected a random subset of 1,000 images to investigate the performance of the multi-modal MP. We confirm prior findings of the modality gap [31] by having only image or text clusters.

## 5. Discussion

Image clusterings strongly depend on the image embedding model. We observe a *narrow cone* for both CLIP models, but also for the uni-modal DINOv2 and Inception-ResNet-v2 model. Given that the clusterings using CLIP embeddings were most distinct when compared to the class labels, our results of it reducing embedding expressiveness are in line with previous works [31]. The analysis of the ClimateTV dataset shows that DINOv2 features allow for a high-level understanding of the dataset by producing clusters such as *person*, *outdoor scene*, and *nature*. In spite of ResNet-50's strong performance on ImageNette and ImageWoof, its clustering on the abstract ClimateTV dataset contained a strong shape bias The investigation of ImageNette clustering appeared to still contain too specific classes, so this effect was only identified when manually analysing the clusterings. ConvNeXt V2 clusters however depend on fine-grained differences by *e.g.* differentiating between *speech* and *protest*, which contain shared visual elements. The analysis of ImageWoof indicates that DINOv2 embeds the background and position of elements in the image, which can be helpful both in its own clustering or the validation of clusterings returned by another model. The *polar bear* cluster was divided and cleaned using the DINOv2 clustering. In the context of climate change, we argue that the high-level DINOv2 and the fine-grained ConvNeXt V2 clusterings combined are a good starting point for frame analysis.

## 6. Conclusion

In this work, we propose a new method for social scientists to automate visual frame detection. Our probabilistic clustering is phrased as a MP and uses image similarities of strong vision (and language) foundation models as a proxy for clustering probabilities. In our experiments, we show the efficacy of MP clustering for detecting visual frames. We find that especially abstract frames such as *speech* can only be detected by foundation models. The intersection of clusterings can be used in order to reduce the number of noise in the clusters, given that the two clusterings are sufficiently distinct. Our analyses have shown the potential usefulness of inter and intra embedding model multi-stage clustering, which we plan to investigate in future work.

## Acknowledgements

This work is supported by the BMBF project 16DKWN027b Climate Visions. All experiments were run on University of Mannheim's server.

## References


[1] Ahmed Abbas and Paul Swoboda. ClusterFuG: Clustering fully connected graphs by multicut. In *International Conference on Machine Learning*. PMLR, 2023. 2

[2] Bjoern Andres, Duligur Ibeling, Giannis Kalofolias, Margret Keuper, Jan-Hendrik Lange, Evgeny Levinkov, Mark Matten, and Markus Rempfler. Graphs and graph algorithms in c++. http://www.andres.sc/graph.html, 2016. 3

[3] Bjoern Andres, Jörg H Kappes, Thorsten Beier, Ullrich Köthe, and Fred A Hamprecht. Probabilistic image segmentation with closedness constraints. In *International Conference on Computer Vision*. IEEE, 2011. 2, 3

[4] James Baker. Using multimodal foundation models and clustering for improved style ambiguity loss. *arXiv preprint arXiv:2407.12009*, 2024. 2

[5] Dorothea Born. Bearing witness? polar bears as icons for climate change communication in national geographic. *Environmental Communication*, 13(5), 2019. 2

[6] Jon Carr. Variation of information (vi). https://gist.github.com/jwcarr/626cbc80e0006b526688, 2017. 4

[7] Andreu Casas and Nora Webb Williams. Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*, 72(2), 2019. 2

[8] Sunil Chopra and Mendu R Rao. The partition problem. *Mathematical programming*, 59(1), 1993. 1, 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. OpenReview, 2021. 2

[10] Fabio Galasso, Margret Keuper, Thomas Brox, and Bernt Schiele. Spectral graph reduction for efficient image and streaming video segmentation. In *Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014. 2

[11] Erving Goffman. *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974. 1

[12] Sylvia Hayes and Saffron O'Neill. The greta effect: Visualising climate protest in uk media and the getty images collections. *Global Environmental Change*, 71, 2021. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016. 2

[14] Kalun Ho, Avraam Chatzimichailidis, Margret Keuper, and Janis Keuper. Msm: Multi-stage multicuts for scalable image clustering. In *International Conference on High Performance Computing*. Springer, 2021. 2, 3

[15] Kalun Ho, Avraam Chatzimichailidis, Franz-Josef Pfreundt, Janis Keuper, and Margret Keuper. Estimating the robustness of classification models by the structure of the learned feature-space. In *AAAI Workshop on Adversarial Machine Learning and Beyond*. OpenReview, 2022. 2

[16] Kalun Ho, Amirhossein Kardoost, Franz-Josef Pfreundt, Janis Keuper, and Margret Keuper. A two-stage minimum cost multicut approach to self-supervised multiple person tracking. In *Asian Conference on Computer Vision*. Springer, 2020. 2

[17] Kalun Ho, Janis Keuper, and Margret Keuper. Unsupervised multiple person tracking using autoencoder-based lifted multicuts. *arXiv preprint arXiv:2002.01192*, 2020. 2

[18] Kalun Ho, Janis Keuper, Franz-Josef Pfreundt, and Margret Keuper. Learning embeddings for image clustering: An empirical study of triplet loss approaches. In *International Conference of Pattern Recognition*. IEEE, 2020. 2

[19] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet. https://github.com/fastai/imagenette, 2019. 1, 4, 5

[20] Jeremy Howard. Imagewoof: a subset of 10 classes from imagenet that aren't so easy to classify. https://github.com/fastai/imagenette#imagewoof, 2019. 1, 4, 5

[21] Steffen Jung and Margret Keuper. Learning to solve minimum cost multicuts efficiently using edge-weighted graph convolutional neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022. 2

[22] Steffen Jung, Sebastian Ziegler, Amirhossein Kardoost, and Margret Keuper. Optimizing edge detection for image segmentation with multicut penalties. In *DAGM German Conference on Pattern Recognition*. Springer, 2022. 2

[23] Amirhossein Kardoost and Margret Keuper. Solving minimum cost lifted multicut problems by node agglomeration. In *Asian Conference on Computer Vision*. Springer, 2019. 2

[24] Amirhossein Kardoost and Margret Keuper. Uncertainty in minimum cost multicuts for image and motion segmentation. In Cassio de Campos and Marloes H. Maathuis, editors, *Conference on Uncertainty in Artificial Intelligence*, volume 161. PMLR, 2021. 2

[25] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal*, 49(2), 1970. 3

[26] Margret Keuper. Higher-order minimum cost lifted multicuts for motion segmentation. In *International Conference on Computer Vision*. IEEE, 2017. 2

[27] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *International Conference on Computer Vision*. IEEE, 2015. 2

[28] Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoué, Thomas Brox, and Bjorn Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *International Conference on Computer Vision*. IEEE, 2015. 2, 3

[29] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation and multiple object tracking by correlation co-clustering. *Transactions on Pattern Analysis and Machine Intelligence*, 42(1), 2018. 2

[30] Evgeny Levinkov, Amirhossein Kardoost, Bjoern Andres, and Margret Keuper. Higher-order multicuts for geometric mmdel fitting and motion segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 2022. 2

[31] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding

the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*. OpenReview, 2022. 2, 6, 8

[32] Aidan McGarry and Emiliano Treré. Fire as an aesthetic resource in climate change communication: Exploring the visual discourse of the california wildfires on twitter/x. *Visual Studies*, 2024. 2

[33] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, Feb. 2018. 5

[34] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29), 2018. 5

[35] Marina Meilă. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*. Springer, 2003. 4

[36] Angelina Mooseder, Cornelia Brantner, Rodrigo Zamith, and Jürgen Pfeffer. (social) media logics and visualizing climate change: 10 years of# climatechange images on twitter. *Social Media+ Society*, 9(1), 2023. 2

[37] Duy MH Nguyen, Roberto Henschel, Bodo Rosenhahn, Daniel Sonntag, and Paul Swoboda. Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2022. 2

[38] Saffron O'Neill, Sylvia Hayes, Nadine Strauß, Marie-Noëlle Doutreix, Katharine Steentjes, Joshua Ettinger, Ned Westwood, and James Painter. Visual portrayals of fun in the sun in european news outlets misrepresent heatwave risks. *The Geographical Journal*, 189(1), 2023. 2

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1, 2

[40] Saffron O'Neill. More than meets the eye: A longitudinal analysis of climate change imagery in the print media. *Climatic Change*, 163(1), 2020. 1, 2

[41] Katharina Prasse, Steffen Jung, Isaac B Bravo, Stefanie Walter, and Margret Keuper. Towards understanding climate change perceptions: A social media dataset. In *NeurIPS Workshop on Tackling Climate Change with Machine Learning*. Climate Change AI, 2023. 1, 4, 5

[42] 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. In *International Conference on Machine Learning*. PMLR, 2021. 2

[43] Stacy Rebich-Hespanha, Ronald E Rice, Daniel R Montello, Sean Retzloff, Sandrine Tien, and João P Hespanha. Image themes and frames in us print news stories about climate change. *Environmental Communication*, 9(4), 2015. 2

[44] Mike S. Schäfer and Saffron O'Neill. Frame analysis in climate change communication: Approaches for assessing journalists' minds, online communication and media portrayals. In Matthew Nisbet, Shirley Ho, Ezra Markowitz, Saffron O'Neill, Mike S. Schäfer, and Jagadish Thaker, editors, *Oxford Encyclopedia of Climate Change Communication*. Oxford University Press, New York, 2017. 1, 2

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. Computational and Biological Learning Society, 2015. 2

[46] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Multi-person tracking by multicut and deep matching. In *ECCV Workshop on Benchmarking Multi-Target Tracking: MOTChallenge*. Springer, 2016. 2

[47] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. 2

[48] Stefan Sylvius Wagner and Stefan Harmeling. Just cluster it: An approach for exploration in high-dimensions using clustering and pre-trained representations. In *International Conference on Machine Learning*. OpenReview, 2024. 2

[49] Susie Wang, Adam Corner, Daniel Chapman, and Ezra Markowitz. Public engagement with climate imagery in a changing digital landscape. *WIREs: Climate Change*, 9(2), 2018. 2

[50] Iain Weaver, Ned Westwood, Travis Coan, Saffron O'Neill, and Hywel TP Williams. Sponsored messaging about climate change on facebook: Actors, content, frames. *arXiv preprint arXiv:2211.13965*, 2022. 2

[51] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Conference on Computer Vision and Pattern Recognition*. IEEE, 2023. 2

[52] Han Yuan and Chuan Hong. Foundation model makes clustering a better initialization for cold-start active learning. *arXiv preprint arXiv:2402.02561*, 2024. 2

[53] Yuanhan Zhang, Zhenfei Yin, Jing Shao, and Ziwei Liu. Benchmarking omni-vision representation through the lens of visual realms. In *European Conference on Computer Vision*. Springer, 2022. 1, 2