DOI: 10.1109/...TODO...

Comparing Image Labeling Based on Art Historical Criteria to Automatic Clustering

Johannes Schuiki † Miriam Landkammer * Isabella Nicka * Andreas Uhl †

Paris Lodron University of Salzburg
† Department of Artificial Intelligence and Human Interfaces
* Institute for Medieval and Early Modern Material Culture

Abstract

Motivated by discrepancies of image ground truth labels and results from a science-to-public event, investigations were carried out trying to quantify the agreement between ground truth labels and automatic clustering. Experimental results indicate that labels on depicted materials in medieval images assigned according to art historical criteria and cluster labels generated using a variety of approaches do not necessarily overlap. The findings imply divergences between domain-specific art historical classifications and computational clustering methods, revealing the complexity of automated art categorization. Further, the demand to establish ground truth based on an inter-annotator agreement to mitigate subjective biases is suggested.

Contents

This work has been partially supported by the Salzburg State Government project "How Material Came into the Picture: Exploring Cultural Innovations Interdisciplinarily with Artificial Intelligence (KIKI)."

1 Introduction

Image classification is a cornerstone task in computer vision, involving the assignment of a specific label or category to an image based on its visual elements. The field has seen major advancements with the rise of deep learning techniques, which have dramatically improved the accuracy and efficiency of classification models. These improvements have enabled machines to interpret visual data with a level of precision approaching human capabilities. However, the success of these models heavily relies on the quality of the training data labels. These labels, often referred to as "ground truth", are crucial for supervised learning approaches. They provide the model with the correct answers to learn from, and inaccuracies in these labels lead to propagated errors in the model's verdict. Therefore, ensuring the integrity of these labeled data is paramount to achieving reliable and satisfactory results in image classification tasks.

Throughout various domains of science employing computer vision tools, the understanding of ground truth and also the process of obtaining slightly differs. Woodhouse [1] for example argues that the term had its origin in the domain of remote sensing, referring to the information actually measured on the ground – however advocating to abolish the term in the same work. In the domain of medicine and medical imaging, the term "gold standard" is used for the benchmark that is the available under reasonable conditions [2]. For data annotations, experts often rely on the results from gold standard methods. For example the gold standard for diagnosing celiac disease is considered performing an endoscopy together with a biopsy, i.e. remove tissue from the body for examination by a medical pathologist [3].

Another, more philosophical, perspective is that the term originates from the german word "Grundwahrheit" [4], whose definition can be translated to "fundamental, irrefutable statement or fact". Which is ironic because there is work suggesting that an inter-observer agreement between experts is not always the case [5], [6]. Also, another common approach for obtaining ground truth is to rely on "crowdsourcing", i.e. distributing the task for labeling images to multiple, possibly non-expert, workers [7]. Such an approach contrasts the definition that ground truth should be viewed as something indisputable.

A particularly intriguing application of image classification and ground truth generation lies in the field of art history, specifically within the study of depicted materials in paintings. Within medieval painting in the Latin West, in the 14th century, the depiction of material properties and surface qualities was paid increased attention for the first time since antiquity – a development that reached a new peak in the 15th century. Annotations are needed in order to be able to trace the course of this development in detail in the future and to be able to examine the rendering of specific materials in different contexts, e.g. iconographic subjects. One highly significant material attributed with complex ideas in the Middle Ages is the wood of Christ's cross. How it was represented in painting is therefore of particular interest (see Section 2 in [8]). Because working through large corpora of images manually can be quite cumbersome, the discipline of computer vision presents itself as a potential partner for cooperation. So far, very scarce literature can be found regarding automatic analysis of Christ's cross: In [8] it is tested, whether patches from natural images of wood could be used to retrieve patches of painted wood from Christ's cross. A research group from the University of Heidelberg developed an algorithm to search within 3620 crucifixion images from an online image archive [1], however their research focused on content based image recognition rather than analysis of depicted material properties. While there are openly available databases [9], [10] that include painted wood, none of them focus on wooden crosses. It can be derived that automatic analysis of painted wood and especially the wood of Christ's cross is still a current topic. Thus, the data used within the experiments in this work consist of samples from the wood of Christ's cross, annotated using an annotation guide established by an expert in the field of art history. Details given in Section 3.1.

The present study aims to quantify the agreement between labels based on art historical criteria and labels assigned through automatic clustering using a variety of image representation techniques and clustering algorithms. To evaluate this agreement, image representations are first grouped into clusters using a number of clustering algorithms. Subsequently, the cluster labels obtained are compared to the ground truth labels using a clustering comparison metric. Section 2 gives an incentive why such an investigation is of relevance. In Section 3, the employed dataset is introduced, the used methodology is explained and also details on implementation are given. Results are reported within Section 4, followed by a discussion. Finally, Section 5 concludes the present study.

[1] https://hci.iwr.uni-heidelberg.de/prototype_image_search_crucifixion

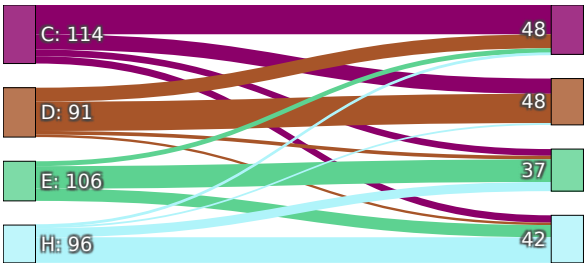

Figure 1: Migration plot depicting how participants labeled the wooden parts. Left: ground truth class; Right: assigned class.

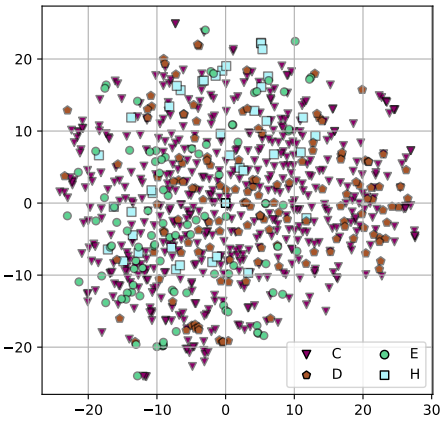

Figure 2: Scatter plot with t-SNE coordinates of the wooden parts. Colors indicate the ground truth texture classes.

2 Motivation

In the course of an interdisciplinary project between art history and computer science, images depicting the wood of Christ's cross were manually annotated using polygonal lines such that the wooden parts are isolated. Four examples of such wooden parts can be seen in Fig. 3. Further, every wooden part received a label (considered as ground truth) describing the structural pattern (i.e. texture) used to depict the material *wood*. Texture categories were chosen based on criteria that seemed interesting for art historical analyses.

During a science-to-public event, participants were tasked to assign wooden parts to a texture class based on some given reference samples. The migration plot in Fig. 1 shows the ground truth texture classes on the left and the texture classes assigned by the participants on the right. Alphabetic identifiers correspond to the classes in Fig. 4. The numbers on the left indicate the total occurrence of the texture classes throughout the experiment, and the numbers on the right display the number of samples where the assigned class and the ground truth class coincide. Samples were drawn randomly from a uniform distribution of the classes. The diagram provides a visual representation of how well the participants matched the samples with the ground truth labels. About half of the samples were assigned according to the expert annotation.

In an attempt to further investigate the discrepancy between participant assignments and expert annotations, a pre-trained ResNet-50 [11] was used to embed the wooden parts into a 2048-dimensional feature representation. Note that this can be done because modern implementations have an "adaptive pooling" layer at the end of the feature extraction part of the model, ensuring a fixed-dimensional feature representation at the penultimate layer regardless of input resolution. Next, t-SNE [12] was used to reduce these feature representations into two-dimensional space and the resulting 2D vectors were used to depict the wooden parts as points in a scatter plot as can be seen in Fig. 2. While t-SNE effectively reduces dimensionality, its results should be interpreted cautiously due to the potential distortion of global structure and distances. However, the visualization generated by t-SNE reveals an intermingling of data points across different classes, suggesting a high degree of overlap and complicating clear distinctions between classes. This motivated further analysis using methods as described in the upcoming sections.

3 Experimental Setup

3.1 Data

For dataset generation, a total of 287 images of 14th and 15th century paintings were selected from the REALonline image database [2], all of which depict a scene that includes at least one cross where wood grain is depicted. To isolate parts containing a depiction of wooden structures within the paintings, areas were annotated manually using the computer vision annotation tool [3]. Distinctions are made between 13 different types of texture within the wooden structures. The texture classes were chosen according to the needs of the art historical research questions. Every wooden piece with homogeneous texture type was surrounded with its own polygon. Doing so allows for clear separation of the texture types.

[2] https://realonline.imareal.sbg.ac.at/
[3] https://www.cvat.ai/

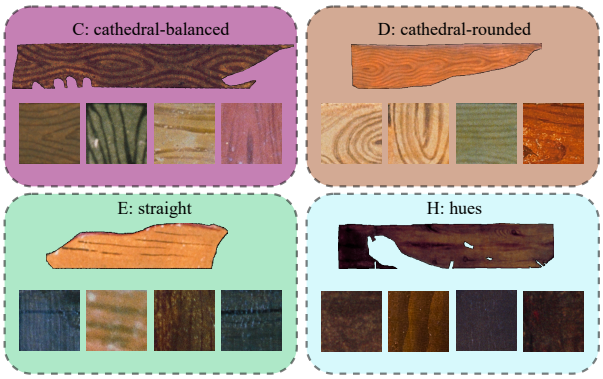

Figure 3: Examples of wooden parts cut directly from annotation polygons (one horizontally rotated example per class) and examples of wooden patches (four patches per class).

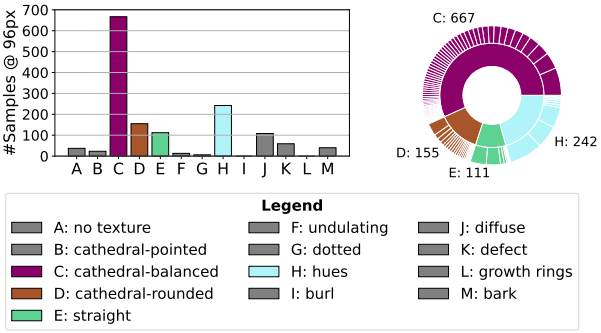

Figure 4: Dataset statistics

Wooden parts, as can be seen in Fig. 3 (one example per texture class), often contain large (transparent) background areas caused by exclusion of all non-wooden image contents such as fingers or parts of the hand. Hence, quadratic non-overlapping patches were cropped from annotated areas using a fixed patch size of 96 pixel side length for further experiments. Example patches for the used classes are seen in Fig. 3 (four patches per texture class). Dividing the data into smaller patches is a common strategy for texture classification because it artificially increases the dataset, while it is assumed that the defining structures are fine-grained and can thus be found on every patch. Dataset statistics are presented in Fig. 4. Due to the imbalance in the data in terms of samples per texture class and because for some classes very few patches are available, only four texture classes (indicated by colored bars; alphabetic identifiers C, D, E and H) are used for further analysis. The nested pie chart on the right visualizes the distribution of cropped patches based on their origin image, where each segment in the outer ring of a texture class represents the proportion of samples originating from the same source image. The number of patches per source image varies both because of differing resolutions of the images and differently sized wooden crosses within the paintings.

3.2 Image representation

In the experiments, various image representations are employed to evaluate the intrinsic correspondence between ground truth labels and image content. This approach is necessitated by the absence of a universally optimal image representation. By evaluating multiple representations, the aim is to comprehensively assess the intrinsic relationship between image content and ground truth labels across different levels of abstraction. The following image representations are utilized:

- Dimensionality reduction: Principal component analysis (PCA) is applied to reduce the dimensionality of the raw pixel data. This step helps to mitigate the curse of dimensionality and potentially reveals latent structures in the data.

- Classical texture features: SIFT [13] descriptors are extracted from grayscale patches on a dense grid with a step size of 8 pixels to capture local gradient-based texture information. The resulting descriptors are aggregated using Fisher vector encoding [14] followed by PCA, providing a compact representation per patch.

- Neural Networks: Feature embeddings extracted from pre-trained deep learning models are utilized. These representations excel at identifying relevant structures in images due to their ability to capture complex relations within the data. However, it is important to note that these networks, trained on ImageNet, introduce biases inherent to their training data. In pursuit of finding a "natural description" of the images and to avoid learning something unintended by force, fitting a model to the data is explicitly refrained from. Network model architectures are taken from the PyTorch Image Models [4] collection, which encompasses a large variety of state-of-the-art model architectures. In particular, six different architectures are employed: TinyNet [15], Vision Transformer (ViT) [16], ResNet-18 [11], Mobile-ViT [17], DinoV3 [18] ConvNext [19] & DinoV3 ViT. Within this work,

[4] https://github.com/rwightman/pytorch-image-models

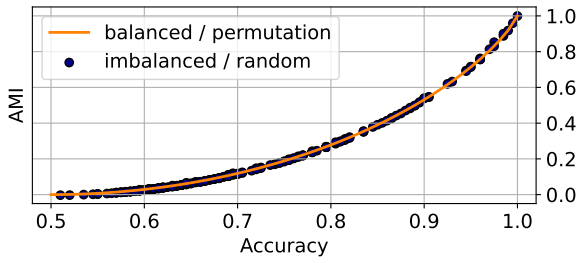

Figure 5: Simulated relationship between AMI and accuracy.

models are utilized such that images are embedded into a feature representation using a pretrained network where the final classification layer is removed.

3.3 Consensus between ground truth and image content

To express the agreement between ground truth and the image descriptors outlined in Section 3.2 numerically, these descriptors are grouped using a variety of algorithmic approaches for data clustering. Four such clustering techniques are employed in this work: (i) k-means clustering, (ii) hierarchical (agglomerative) clustering, (iii) deep embedded clustering [20] and (iv) SubKmeans[21]. For vanilla k-means and hierarchical clustering, implementations from the scikit-learn [22] python library are used while for the latter two methods the ClustPY [23] library is utilized. Subsequently to clustering, images hold both a ground truth label and a newly generated cluster label. Similarity can thus be expressed resorting to clustering comparison metrics such as Adjusted Mutual Information (AMI) [24]. AMI is a version of mutual information that introduces a correction for chance by considering the expected similarity of all pair-wise comparisons. For calculation of an expected similarity value, a random model must be chosen based on how the data points can be clustered. According to [25], the right random model to choose for experiments within this work is the *one-sided* (always comparing against the same basis, i.e. ground truth) comparison with a *fixed number of clusters*. Hence, this work uses the reference implementation from Gates and Ahn [26] for calculation of the AMI. Note that employing a mutual information based metric to compare clustered images to its ground truth labels is common strategy for evaluating the performance of clustering algorithms [27].

4 Results & discussion

Results in Fig. 6 show the AMI values obtained using the procedure described in Sections 3.2 & 3.3. Two classes are always considered in isolation to evaluate the separability from another, resulting in six texture class constellations. The four texture classes used in the experiments correspond to the classes introduced in Section 3.1. Experiments including a random component were carried out ten times. Error bars indicate the standard deviation.

4.1 Relationship between AMI and accuracy

In an attempt to establish a relationship between the AMI and an estimation of the classification accuracy, two artificial clusterings were gradually alienated and the corresponding metrics calculated. Fig. 5 depicts the simulated relationship between AMI and accuracy for a two-class setup. Two cases for label randomization are considered: (i) permutation, where the starting point is an equal amount of elements per cluster and on every step of scrambling the labels, every cluster receives exactly one element from the other cluster. Doing so, the number of elements per cluster does not change, merely the labels get permuted. (ii) randomized, which also considers imbalanced clusters. Depending on the step, a number of elements are picked and assigned to the other cluster in a random fashion. The relationship can be understood such that an AMI value of y indicates that x% of samples are grouped (clustered) together with others from the same class. Note that 0.5 accuracy corresponds to random guessing in a two-class scenario.

4.2 Interpretation of the results

Since the AMI value is a measurement of cluster similarity, high values indicate high accordance of expert label and automatic label. Hence, higher AMI values suggest some separability in the feature space, which in turn indicates separability based on image-intrinsic properties. While PCA and the deep learning network embeddings mostly agree, the dense sift variant generally seems to yield lower AMI values. This can be explained by conversion to grayscale for this texture descriptor, thereby discarding relevant color information. The results, however, allow to deduce some tendencies. Because the various representations together with different clustering techniques only agree on most occasions, we should not focus on the precise numerical values itself but view

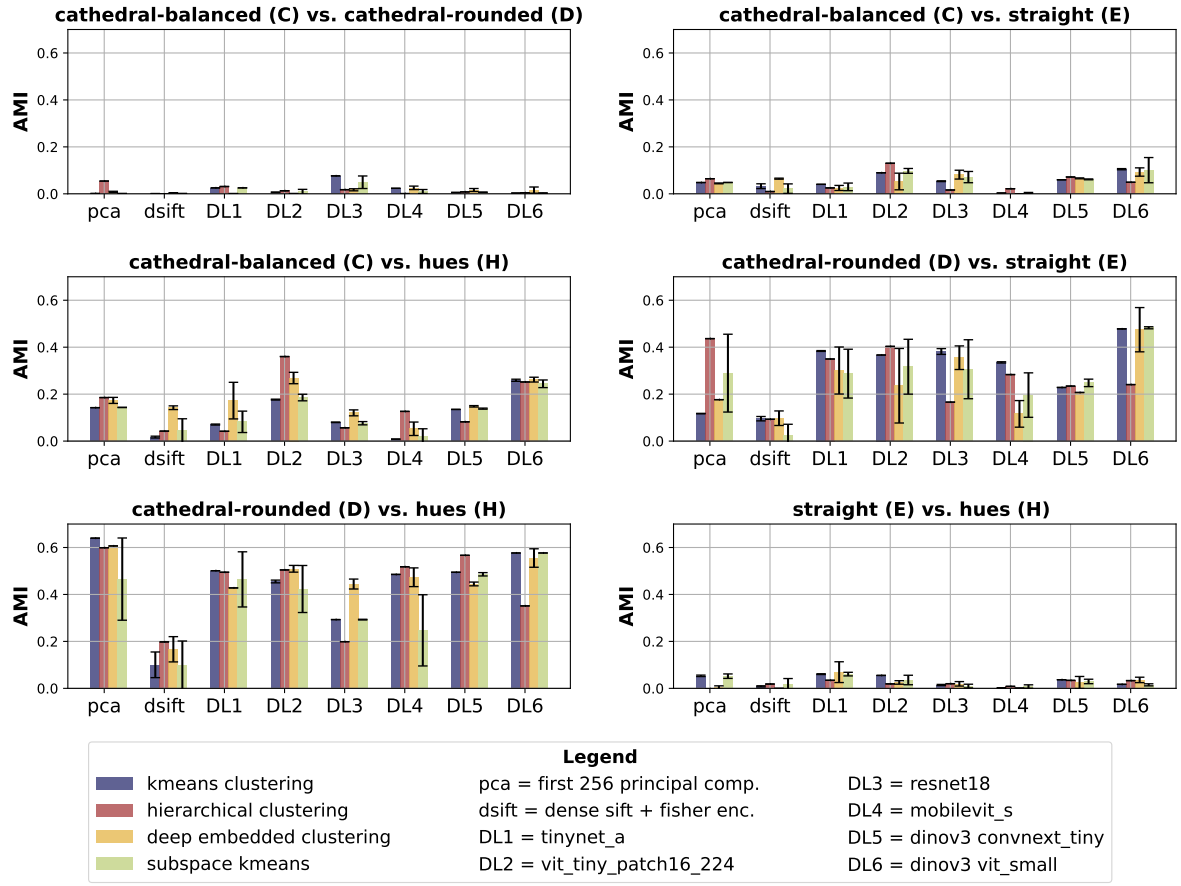

Figure 6: Experimental results. AMI values for every considered pair of classes.

the results as a general trend. This indicates that there exists some correspondence of ground truth and clustering labels for some classes but not for others. With the exception of C vs H, experiments including class C yield overall low results, indicating its role as a general-purpose class encompassing numerous texture structures. While class C seems related to both, class D & E, these classes appear to have some properties distinctive enough for AMI values up to 0.5. Note that this relation does not pose a contradiction since class C most likely spans a broader range of textures. The strongest evidence for intrinsic separability within the scope of this work pose the experiments D vs. H, reaching slightly above 0.6 AMI, which, according to Section 4.1, corresponds to around 92% classification accuracy. Low AMI values for experiments considering case E vs. H point again to a high overlap in feature space, i.e. appear visually similar.

5 Conclusion

This study explored the intrinsic correspondence between image labels assigned to painted materials based on art historical criteria and labels assigned through automatic clustering. Experiments included a variety of image representations and clustering techniques. The observations allow to derive the following conclusions: (i) Texture classes assigned based on art historical criteria do not necessary correspond to automatically generated classes based on their inherent visual properties. Hence, automatic separation according to the domain-specific labels can be challenging. A potential limitation of the used approach is the employment of patching, possibly destroying relevant structures in the process. However, it should be noted that data points in Fig. 2 correspond to the complete wooden parts, supporting the notion that the patching strategy has negligible effect on the overall findings. (ii) Although not addressed directly within the scope of this work, subjective bias could also be a factor. Non-experts (refer-

ring to the participants during the science-to-public event) evidently faced difficulties assigning the texture classes according to the expert label, hereby casting doubt on the reliability of the latter. In the domain of medical imaging, an inter annotator discrepancy exists and "collective-agreement" countermeasures are sometimes taken to mitigate subjective bias. The results in this study can be seen as an indication that Digital Humanities faces a similar problem as medical areas and multiple-annotator labels should be considered in future works.

References

[1] I. H. Woodhouse, "On 'ground' truth and why we should abandon the term," *J. Appl. Remote Sens.*, vol. 15, no. 4, 2021. DOI: 10.1117/1.JRS.15.041501.

[2] J. Cardoso, L. Pereira, M. Iversen, *et al.*, "What is gold standard and what is ground truth?" *Dental Press J Orthod*, vol. 19, 2014. DOI: 10.1590/2176-9451.19.5.027-030.ebo.

[3] C. Li, B. Jing, L. Ke, *et al.*, "Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies," *Cancer Commun (Lond)*, vol. 38, no. 1, 2018. DOI: 10.1186/s40880-018-0325-9.

[4] M. Boenig, M. Federbusch, E. Herrmann, *et al.*, "Ground truth: Grundwahrheit oder ad-hoc-lösung? wo stehen die digital humanities?" In *Jahrestagung des Verbands Dig. Humanities im deutschsprachigen Raum*, 2018.

[5] A. Sethi, T. Doukides, D. V. Sejpal, *et al.*, "Inter-observer agreement for single operator choledochoscopy imaging: Can we do better?" *Diagnostic and Therapeutic Endoscopy*, vol. 2014, no. 1, 2014. DOI: 10.1155/2014/730731.

[6] J. Schuiki, M. Steiner, H. Hofbauer, *et al.*, "Quantifying inter-annotator agreement and generalist model limitations in imaging mass cytometry single cell segmentation," in *Medical Image Understanding and Analysis*, Springer Nature Switzerland, 2025. DOI: 10.1007/978-3-031-98694-9_11.

[7] S. N. Ørting, A. Doyle, A. van Hilten, *et al.*, "A survey of crowdsourcing in medical image analysis," *Hum. comput.*, vol. 7, no. 1, 2020. DOI: 10.15346/hc.v7i1.1.

[8] J. Schuiki, M. Landkammer, M. Linortner, *et al.*, "Towards using natural images of wood to retrieve painterly depictions of the wood of christ's cross," in *Image Analysis and Processing - ICIAP 2023 Workshops*, Springer Nature Switzerland, 2024. DOI: 10.1007/978-3-031-51026-7_31.

[9] M. J. P. van Zuijlen, H. Lin, K. Bala, *et al.*, "Materials in paintings (mip): An interdisciplinary dataset for perception, art history, and computer vision," *PLOS ONE*, vol. 16, no. 8, 2021. DOI: 10.1371/journal.pone.0255109.

[10] T. Mensink and J. van Gemert, "The rijksmuseum challenge: Museum-centered visual recognition," in *ACM International Conference on Multimedia Retrieval*, 2014. DOI: 10.1145/2578726.2578791.

[11] K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *2016 IEEE CVPR*, 2016. DOI: 10.1109/CVPR.2016.90.

[12] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, 2008.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.

[14] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE CVPR*, 2007. DOI: 10.1109/CVPR.2007.383266.

[15] K. Han, Y. Wang, Q. Zhang, *et al.*, "Model rubik's cube: Twisting resolution, depth and width for tinynets," ser. NIPS '20, Curran Associates Inc., 2020.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[17] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in *International Conference on Learning Representations*, 2022.

[18] O. Siméoni, H. V. Vo, M. Seitzer, *et al.*, *DINOv3*, 2025. arXiv: 2508.10104 [cs.CV].

[19] Z. Liu, H. Mao, C.-Y. Wu, *et al.*, "A convnet for the 2020s," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[20] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16, JMLR.org, 2016.

[21] D. Mautz, W. Ye, C. Plant, *et al.*, "Towards an optimal subspace for k-means," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17, Association for Computing Machinery, 2017. DOI: 10.1145/3097983.3097989.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, 2011.

[23] C. Leiber, L. Miklautz, C. Plant, *et al.*, "Benchmarking deep clustering algorithms with clustpy," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2023. DOI: 10.1109/ICDMW60847.2023.00087.

[24] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, 2010.

[25] A. J. Gates and Y.-Y. Ahn, "The impact of random models on clustering similarity," *J. Mach. Learn. Res.*, vol. 18, no. 1, 2017.

[26] ——, "Clusim: A python package for calculating clustering similarity," *J. Open Source Softw.*, vol. 4, no. 35, 2019. DOI: 10.21105/joss.01264.

[27] C. Song, F. Liu, Y. Huang, *et al.*, "Auto-encoder based data clustering," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Springer Berlin Heidelberg, 2013. DOI: 10.1007/978-3-642-41822-8_15.